



UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II



UNIVERSITÀ DEGLI STUDI DI
NAPOLI FEDERICO II
DIPARTIMENTO DI INGEGNERIA ELETTRICA
E TECNOLOGIE DELL'INFORMAZIONE



DIETI
DIPARTIMENTO
DI ECCELLENZA
2023 - 2027

Università degli Studi di Napoli Federico II

Ph.D program in

Computational and Quantitative Biology

XXXVII Cycle

Thesis for the Degree of Doctor of Philosophy

**Dual RNA-Sequencing assessment and
implementation to study transcriptome alteration
of *Solanum lycopersicum* during the interaction
with *Orobanchaceae* spp.**

by

CARMINE FRUGGIERO

Advisor: Nunzio D'Agostino

Co-advisor: Edoardo Pasolli



Scuola Politecnica e delle Scienze di Base

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

Dual RNA-Sequencing assessment and implementation to study transcriptome alteration of *Solanum lycopersicum* during the interaction with Orobanchaceae

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Computational and Quantitative Biology
by

CARMINE FRUGGIERO

December 2024



Approved as to style and content by

Prof. Nunzio D'Agostino, Advisor

Prof. Edoardo Pasolli, Co-advisor

Università degli Studi di Napoli Federico II

Ph.D. Program in Computational and Quantitative Biology

XXXVII Cycle - Chairman: Prof. Michele Ceccarelli



<http://itee.dieti.unina.it>

Candidate's declaration

I hereby declare that this thesis, submitted in fulfillment of the requirements for the degree of Philosophiæ Doctor (Ph.D.) in Computational and Quantitative Biology, is my own original work. I have used only the sources and resources explicitly referenced, and I have appropriately acknowledged all direct and indirect sources.

Portions of this dissertation have been published in international journals and/or conference proceedings (see the list of publications by the author at the end of the thesis).

Napoli, 03/02/2025

A handwritten signature in black ink, reading "Carmine Fruggiero". The signature is written in a cursive style with a large initial 'C'.

Carmine Fruggiero

Abstract

Transcriptomics aims to achieve a comprehensive understanding of each RNA molecule in a given cell (the transcriptome), under specific conditions. The primary method for transcriptome analysis is RNA-sequencing where, typically, each processed sample originates from a single organism. In the case of host-parasite interactions, the close biological interdependence complicates this approach, as the parasite relies on the host for nutrients. To study gene expression changes in both host and parasite simultaneously, dual RNA-seq has emerged as the method of choice. This technique enables the concurrent analysis of gene expression from both interacting organisms, even when their physical separation is difficult. A key challenge in dual RNA-seq lies in the accurate *in silico* discrimination of reads. The existence of shared homologous sequences between organisms can lead to cross-mapping, a phenomenon where reads from one organism map to the other or to both. Read discrimination becomes most reliable when the host and parasite belong to different kingdoms. In this study, host-parasite interactions were simulated to evaluate the feasibility of dual RNA-seq when both organisms belong to the *Plantae* kingdom. The simulations demonstrated that the method is reliable and highlighted the possibility of executing it using two approaches (**Chapter 2**). Dual RNA-seq execution requires programming skills, which may not be accessible to all researchers. In contrast, graphical user interfaces (GUIs) have been developed for the classical bulk RNA-seq, making the analysis process more accessible. To bridge this gap, inDAGO was developed as an open-access software, that enables users to conduct dual RNA-seq analysis without the need for programming. inDAGO provides the option to choose between two approaches and includes an extension for bulk RNA-seq analysis. The software was tested using data from the previous simulations (**Chapter 3**). The parasitic plant *Phelipanche ramosa*, a member of the *Orobanchaceae* family, and *Solanum* species provide an example of host-parasite interaction to apply dual RNA-seq pipeline using inDAGO. The two tomato species *Solanum pennellii* and *Solanum lycopersicum* differ in their response to *Phelipanche ramosa* infection, with the former showing tolerance and the latter being susceptible. An analysis was carried out to explore the genetic basis of *S. pennellii* tolerance mechanisms. Genes involved in the strigolactone pathway - known to stimulate *Orobanchaceae* germination - were identified in *S. lycopersicum*. Their orthologous in *S. pennellii* were inferred. Two of these genes exhibited polymorphisms that may explain the resistance mechanism related to parasite seed germination (**Chapter 4**). This thesis demonstrated three findings: (i) dual RNA-seq is reliable even for taxonomically close organisms; (ii) a GUI was developed to make the analysis accessible to biologists without programming skills; and (iii) the tolerance mechanisms of *S. pennellii* were preliminarily explored. These pave the way for future transcriptomic investigations of *S. pennellii* introgression lines during interactions with *P. ramosa* using the inDAGO software in a dual RNA-seq experiment.

Keywords: Transcriptomics, dual RNA-seq, plant host-parasitic plant interactions, cross-mapping, GUI (graphical user interface), Orobanchaceae, strigolactone pathway.

Sintesi in lingua italiana

Il metodo principale per analizzare il trascrittoma è il sequenziamento dell'RNA (RNA-seq), nel quale ogni campione analizzato appartiene ad uno specifico organismo. Tuttavia, nelle interazioni ospite-parassita, una stretta interconnessione tissutale si viene a formare tra due organismi affinché il parassita si nutra a spese dell'ospite. Per studiare le simultanee variazioni dell'espressione genica di entrambi gli organismi, il dual RNA-seq è emerso come metodo di riferimento. Questa tecnica consente di analizzare congiuntamente l'espressione genica dell'ospite e del parassita, anche quando risulta difficile separarli fisicamente. Una delle sfide principali nel dual RNA-seq riguarda la corretta assegnazione delle sequenze durante l'analisi. Infatti, la presenza di sequenze omologhe condivise tra gli organismi interagenti porta al fenomeno di allineamento incrociato, ossia sequenze appartenenti ad un organismo mappano correttamente sull'altro o su entrambi. In questo studio, sono state simulate due interazioni ospite-parassita per verificare la fattibilità del dual RNA-seq anche nel caso in cui entrambi gli organismi appartengano al regno *Plantae*. Le simulazioni hanno dimostrato che il metodo è affidabile, permettendo l'adozione di due approcci distinti (**Capitolo 2**). Sebbene il dual RNA-seq possa essere eseguito utilizzando due protocolli consolidati, entrambe richiedono competenze di base di programmazione, limitandone l'accessibilità a molti ricercatori. Al contrario, per facilitare l'analisi classica *bulk* RNA-seq e renderla più accessibile, sono stati sviluppati programmi ad interfaccia grafica. Per colmare questa lacuna, è stato sviluppato inDAGO, un programma *open source*, multiplatforma che consente di eseguire l'analisi dual RNA-seq senza necessità di conoscenze informatiche pregresse. inDAGO permette di scegliere tra due approcci analitici (sequenziale e combinato) e include anche un'estensione per l'analisi *bulk* RNA-seq. L'interfaccia è stata testata utilizzando i dati delle simulazioni precedenti confermandone l'efficacia (**Capitolo 3**). Un esempio di possibile applicazione del protocollo dual RNA-seq è rappresentato dall'interazione tra la pianta parassita *Phelipanche ramosa* e le specie del genere *Solanum*. In particolare, *Solanum pennellii* e *S. lycopersicum* mostrano risposte diverse all'infezione da *P. ramosa*: *S. pennellii* risulta tollerante, mentre *S. lycopersicum* è suscettibile. È stata condotta un'analisi preliminare per indagare i meccanismi di tolleranza di *S. pennellii*. Sono stati identificati, in *S. lycopersicum*, i geni coinvolti nella biosintesi degli strigolattoni (stimolanti della germinazione delle Orobanchaceae) e successivamente sono stati individuati i loro ortologhi in *S. pennellii*. Due geni presentano interessanti polimorfismi che potrebbero spiegare il meccanismo di resistenza al parassita (**Capitolo 4**). Questa tesi ha portato a tre principali risultati: (i) il dual RNA-seq è affidabile anche per organismi tassonomicamente affini; (ii) è stata sviluppata un'interfaccia per semplificare l'analisi anche a ricercatori privi di competenze di programmazione; (iii) sono stati preliminarmente esplorati i meccanismi di tolleranza di *S. pennellii*, aprendo la strada a future indagini trascrittomiche su linee di introgressione di *S. pennellii* nelle interazioni con *P. ramosa* utilizzando inDAGO in esperimenti di dual RNA-seq.

Parole chiave: Trascrittomica, dual RNA-seq, interazioni pianta ospite-pianta parassita, allineamento incrociato, programma ad interfaccia grafica, Orobanchaceae, strigolattoni.

Acknowledgements

I would like to acknowledge the Dago lab, the Bioinformatics laboratory at the Department of Agricultural Sciences, University of Naples Federico II, as the primary host institution. I am also grateful to "Sequentia Biotech SL, located at Carrer Dr. Trueta 179, 08005, Barcelona" for serving as the foreign host institution during my research period abroad.



UNIONE EUROPEA
Fondo Sociale Europeo



*Ministero dell'Università
e della Ricerca*



PON
RICERCA
E INNOVAZIONE
2014 - 2020

REACT EU

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020, risorse FSE REACT-EU
Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione”
e Azione IV.5 “Dottorati su tematiche Green”

Contents

Abstract

Keywords

Sintesi in lingua italiana

Acknowledgements

List of acronyms

List of figures

List of tables

List of symbols

List of supplementary tables

Chapter 1. Introduction

- 1.1. Capturing gene expression dynamics: a transcriptomic snapshot.
- 1.2 Investigating host-parasite interactions through dual RNA-seq transcriptome analysis
- 1.3 Bioinformatics made easy: a Graphical User Interface for beginners
- 1.4 Parasitism in the *Plantae* kingdom: mechanisms and evolutionary implications
- 1.5 The biology and impact of the parasitic plant *Phelipanche ramosa*
- 1.6 *Solanum lycopersicum*: a host for the parasitic plant *Phelipanche ramosa*
- 1.7 Aim of the thesis

Chapter 2. Refining dual RNA-seq mapping: sequential and combined approaches in host-parasite plant dynamics

- 2.1 Abstract
- 2.2 Introduction
- 2.3 Materials and methods
 - 2.3.1 Phylogenetic analysis of *Cuscuta spp.* and their host range

2.3.2 Retrieval of input data and their pre-processing

2.3.3 Merging of the transcriptome data

2.3.4 Dual RNA-Seq simulation via sequential and combined approach

2.3.5 Evaluation metrics

2.3.6 Investigating one-side cross-mapped reads in the sequential and combined approaches

2.3.7 Investigating two-side cross-mapped reads in the sequential and combined approaches

2.4 Results

2.4.1 Phylogenetic analysis of *Cuscuta spp.* and their host range

2.4.2 Dual RNA-seq simulation via sequential and combined approaches

2.4.2.1 Sequential approach

2.4.2.1.1 Mapping of *A. thaliana* and *C. campestris* reads

2.4.2.1.2 Mapping of *S. lycopersicum* and *C. campestris* reads

2.4.2.2 Combined approach

2.4.2.2.1 Mapping of *A. thaliana* and *C. campestris* reads

2.4.2.2.2 Mapping of *S. lycopersicum* and *C. campestris* reads

2.4.3. Investigating one-side cross-mapped reads

2.4.4 Investigating two-side cross-mapped reads

2.4.4.1 Sequential approach

2.4.4.1.1 *A. thaliana* and *C. campestris*

2.4.4.1.2 *S. lycopersicum* and *C. campestris*

2.4.4.2 Combined approach

2.4.4.2.1 *A. thaliana* and *C. campestris*

2.4.4.2.2 *S. lycopersicum* and *C. campestris*

2.5 Discussion

2.6 Conclusions

2.7 Data availability

Chapter 3. Introducing inDAGO: a graphical user interface for effortless dual RNA-seq.

- 3.1 Abstract
- 3.2 Introduction
- 3.3 Overview of the interface
- 3.4 Methods
 - 3.4.1 Quality control
 - 3.4.2 Filtering
 - 3.4.3 Genome indexing
 - 3.4.4 Mapping
 - 3.4.5 Summarization
 - 3.4.6 Exploratory data analysis
 - 3.4.7 DEGs analysis
 - 3.4.8 Graphical structure and dynamic documentation
 - 3.4.9 Dataset
- 3.5 Results
- 3.6 Conclusion
- 3.7 Code and data availability

Chapter 4. Investigating molecular mechanisms of pre-attachment defence against *Phelipanche ramosa*: identifying polymorphisms in strigolactone pathway genes between *Solanum lycopersicum* and its wild relative *Solanum pennellii*

- 4.1 Abstract
- 4.2 Introduction
- 4.3 Materials and methods
 - 4.3.1 Strigolactone pathway genes
 - 4.3.2 Identifying orthologous in *Solanum lycopersicum*
 - 4.3.3 Identifying orthologous in *Solanum pennellii*

4.3.4 Comparison of SL pathway genes between *S. lycopersicum* and *S. pennellii*

4.3.5 Primer design

4.3.6 Nucleic acid isolation, PCR amplification, and Sanger sequencing

4.4 Results

4.4.1 SL orthologous genes in *S. lycopersicum* and *S. pennellii*

4.4.2 Searching for polymorphisms between *S. lycopersicum* and *S. pennellii*

4.4.3 In vivo validation of detected polymorphisms

4.5 Discussions

4.6 Conclusions

Chapter 5. Concluding remarks

Supplementary tables

Bibliography

Author's publications

List of Acronyms

The following acronyms are used throughout the thesis.

BAM	Binary Alignment/Map
CCD7	Carotenoid Cleavage Dioxygenase 7
CCD8	Carotenoid Cleavage Dioxygenase 8
cDNA	complementary DNA
CLI	Command Line Interface
CRAM	Compressed Reference-oriented Alignment Map
CSS	Cascade Style Sheet
CSV	Comma-Separated Values
D27	Dwarf27
DEA	Data Exploration Analysis
DEGs	Differentially Expressed Genes
DGE	Differential Gene Expression
DNA	DeoxyriboNucleic Acid
DPI	Dots Per Inch
ENA	European Nucleotide Archive
EST	Expressed Sequence Tag
FCS	Functional Class Scoring
FDR	False Discovery Rate
GO	Gene Ontology
GSA	Gene Set Analysis
GUI	Graphical User Interface
HGT	Horizontal Gene Transfer
HTML	HyperText Markup Language
IL	Introgression Line

indels	insertions/deletions
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCM	Laser Capture Microdissection
log2 FC	log2 Fold-Changes
LRT	Likelihood Ratio Tests
MDS	Multi-Dimensional Scaling
mRNA	messenger RNA
NCBI	National Center of Biotechnology Information
ncRNA	non-coding RNA
NGS	Next Generation Sequencing
ORA	Over-Representation Analysis
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PRRs	Pattern Recognition Receptors
PT-based	Pathway Topology-based
QC	Quality Control
QLFTest	Quasi-Likelihood F-Tests
RNA	RiboNucleic Acid
RNA-seq	RNA sequencing
ROS	Reactive Oxygen Species
rRNA	ribosomal RNA
SAM	Sequence Alignment/Map
SNP	Single Nucleotide Polymorphisms
TC	Time Course
tRNA	transfer RNA

List of Figures

- 1.1 Dual RNA-seq workflow for investigating host plant-parasitic plant interactions at two temporal points.
- 1.2 A brief history of a bioinformatic analysis
- 1.3 Lifecycle of a broomrape
- 2.1 Workflow employed for analyzing dual RNA-seq data in this case study
- 2.2 Categorization of reads based on the employed mapping procedure
- 2.3 Identification of host-parasite two-side cross-mapped reads in the sequential approach
- 2.4 Phylogenetic tree of *Cuscuta spp.* and their host range
- 2.5 Workflow summary results
- 3.1 inDAGO pipeline for performing dual RNA-seq using either the sequential or combined approach
- 3.2 Graphical interface of inDAGO
- 3.3 Quality control module results
- 3.4 Results of the mapping module
- 3.5 Results of the data exploration module
- 3.6 Results of the differential gene expression analysis module
- 4.1 Strigolactones biosynthetic pathway
- 4.2 Dot plot between *S.lycopersicum* D27 gene (Soly09g065750) and *S. pennellii* D27 gene (XM_015232090.2).
- 4.3 Protein sequence of the *S. pennellii* CCD7 gene (Sopen01g036390.1)
- 4.4 Multiple alignment of D27 gene sequences

List of tables

- 2.1** The input RNA-seq data of the three selected species
- 2.2.1** Statistics on the number of mapped reads by considering a sequential approach (*A. thaliana*-*C. campestris*)
- 2.2.2** Sequential approach evaluation metrics (*A. thaliana*-*C. campestris*)
- 2.3.1** Statistics on the number of mapped reads by considering a sequential approach (*S. lycopersicum*-*C. campestris*)
- 2.3.2** Sequential approach evaluation metrics (*S. lycopersicum*-*C. campestris*)
- 2.4.1** Statistics on the number of mapped reads by considering combined approach
- 2.4.2** Combined approach evaluation metrics
- 4.1** Strigolactone pathway genes
- 4.2** PCR primer
- 4.3** Accession number regarding orthologous sequences of strigolactones genes.

List of supplementary tables

- 2.1 Protein sequences accession numbers involved in the phylogenetic analysis
- 2.2 Sequential approach cross-mapping
- 2.3 *A. thaliana* and *C. campestris* cross-mapping in sequential approach
- 2.4 *S. lycopersicum* and *C. campestris* cross-mapping in sequential approach
- 2.5 Cross-mapping in sequential approach
- 2.6 Cross-mapping in combined approach
- 2.7 One-side cross-mapping in sequential approach
- 2.8 One-side cross-mapping in combined approach
- 2.9 *A. thaliana* and *C. campestris* two-side cross-mapping in sequential approach
- 2.10 *S. lycopersicum* and *C. campestris* two-side cross-mapping in sequential approach
- 2.11 *A. thaliana* and *C. campestris* two-side cross-mapping in sequential approach
- 2.12 *S. lycopersicum* and *C. campestris* two-side cross-mapping in combined approach
- 4.1 PredictSNP results related to NCBI annotation of *S. lycopersicum* and *S. pennellii*
- 4.2 PredictSNP results related to SolGenomic annotation of *S. lycopersicum* and *S. pennellii*

Chapter 1. Introduction

1.1. Capturing gene expression dynamics: a transcriptomic snapshot.

The cell is the simplest organization of molecules that exhibits the properties of life. It is the basic unit capable of carrying out life functions. The information required to perform these functions is encoded in a macromolecule called “DNA” (deoxyribonucleic acid). The complete set of DNA molecules stored within a cell, is known as the genome. To better understand how genomic information contributes to the development of cellular functions, it is crucial to focus on transcription, the first key regulatory step of gene expression. Transcription involves copying the information stored in DNA into RNA molecules, commonly referred to as transcripts (Krebs, Goldstein and Kilpatrick, 2017). Once synthesized, these transcripts can undergo various modifications, forming what are known as “mature transcripts”. Typically, in Eukaryotes, newly synthesized transcripts consist of regions called exons, separated by non-coding regions, known as introns. The process of splicing removes introns and covalently joins the exons together. When splicing involves different splice sites (alternative splicing), it increases the coding potential of genes (Marasco and Kornblihtt, 2023). Additionally, RNA editing, which involves altering single bases in the RNA after transcription, introduces further complexity and diversity in gene expression (Farajollahi and Maas, 2010).

The complete set of RNA species in a cell or tissue is referred to as the transcriptome (Assis *et al.*, 2022). More specifically, the transcriptome includes messenger RNAs (mRNA), transfer RNAs (tRNA), ribosomal RNAs (rRNA), various non-coding RNAs (ncRNA) (Dong and Chen, 2013). The transcriptome reflects the RNA present in a specific cell or tissue at a particular developmental stage and/or under specific physiological conditions. Analyzing the transcriptome provides valuable insights into gene structure, regulation of gene expression, gene product functions, genome dynamics, and the regulation networks underlying biological processes (Jiang *et al.*, 2015). Transcriptome analysis captures both coding and non-coding RNAs, enabling quantification of their expression levels. The development of Next Generation Sequencing (NGS) technologies (Hu *et al.*, 2021) has made it possible for researchers to sequence and analyze the entire transcriptome of organisms efficiently.

One of the earliest methods for sequencing the transcriptome, developed in the early 1990s, was Expressed Sequence Tag (EST) sequencing. This technique generated fragments of mRNA sequences through single sequencing reactions on randomly selected complementary DNA (cDNA) libraries (Adams *et al.*, 1992). EST sequencing was based on Sanger sequencing technology. In contrast, RNA sequencing (RNA-seq), the widely used modern approach to transcriptome analysis, utilizes NGS technologies. RNA-seq, also known as whole-transcriptome shotgun sequencing, allows for the detection of transcript

expression level, RNA editing events and alternative transcript isoforms. It has become the preferred method for studying gene expression and identify novel RNA species.

RNA-seq has been widely used in various contexts, leading to numerous adaptations of the method to meet different needs. To gain insight into alternative splicing isoforms and chimeric RNAs (Li, Qin and Li, 2018), it is essential to sequence each RNA molecule from end to end. This requires a hybrid approach that combines long-read sequencing with short-read sequencing (Au *et al.*, 2012). However, analyzing lowly expressed genes across the entire transcriptome can be challenging. To overcome this, targeted RNA-seq analysis can be employed, focusing on specific genes using gene-specific probes (Mercer *et al.*, 2014) or primer pairs (Mamedov *et al.*, 2013).

Bulk RNA-seq is the most commonly used RNA-seq approach, analyzing RNAs from a pool of cells as a single sample (Wilhelm and Landry, 2009). However, bulk RNA-seq lacks the resolution needed to detect gene expression at the individual cell level, which is essential for understanding the intricate regulatory processes within cells. To address this limitation, single-cell RNA-seq was developed (Slovin *et al.*, 2021), allowing for reliable analysis even with low RNA concentrations compared to bulk RNA-seq. In addition to these approaches, transcriptome analysis can target specific RNA species, such as small RNAs (Benesova, Kubista and Valihrach, 2021) or circular RNAs (Li, Yang and Chen, 2018). It can also focus on different stages of RNA life cycle, including nascent RNA (Core, Waterfall and Lis, 2008), RNA-protein interactions (Zambelli and Pavesi, 2015) and RNA-RNA interactions (Engreitz *et al.*, 2014). For each of these specialized aspects, tailored RNA-seq methods have been developed to optimize data collection and analysis. The latest frontier in RNA-seq technology is spatial transcriptomics, which allows researchers to map single cell gene expression patterns within a tissue or organ, assigning each transcript to its precise spatial location. This method offers a powerful tool for gaining deeper insights into cellular heterogeneity, cell-to-cell communication, and the molecular mechanisms driving tissue development and differentiation (Valihrach *et al.*, 2024). By providing spatial context to gene expression data, spatial transcriptomics significantly enhances our ability to understand the complex architecture and function of biological systems.

1.2. Investigating host-parasite interactions through dual RNA-seq transcriptome analysis

Over the past decade, a specialized variant of RNA-seq has been refined to study complex interactions between biological entities from different species that are physically interconnected and difficult to separate.

Organisms rarely exist in isolation; they constantly interact with one other in varying degrees of intensity and duration. These interactions can range from brief and weak to long-lasting and tightly linked. In some cases, organisms live in close association, forming relationships such as symbiosis. One type of symbiotic relationship is parasitism, where one organism, the parasite, relies on another organism, the host, for nutrients and survival (Solomon *et al.*, 2015). Host-parasite interactions often involve a highly intimate connection, especially within the plant kingdom. Parasitic plants, for instance, must directly invade and extract nutrients from their host plants through specialized structures like

modified roots, shoots, or leaves known as haustoria. These structures facilitate a direct link with host tissues (Nickrent, 2020).

Studying the transcriptomics of host-parasite interactions, particularly at the contact zone between the two organisms, presents significant challenges. The main difficulty lies in physically separating the host and parasite tissues to ensure that RNA samples are derived from one organism alone. Jhu *et al.* (2022) successfully addressed this challenge by utilizing the Laser Capture Microdissection (LCM) technique. This approach allowed them to precisely isolate tissue from the host *Solanum lycopersicum* and the parasite *Cuscuta campestris*, enabling a more accurate bulk RNA-seq analysis of the interaction.

However, LCM presents several drawbacks, including high costs, technical complexity, and time-consuming processes. A promising alternative to overcome these challenges is dual RNA-seq, first introduced by Westermann *et al.*, (2012) to investigate gene expression changes in eukaryotic host cells infected by bacterial pathogens.

Unlike bulk RNA-seq, where each sample ideally represents the transcriptome of a single species, dual RNA-seq enables the simultaneous analysis of both host and parasite transcriptomes from a mixed sample of RNA derived from the interconnected region of the two organisms.

In a dual RNA-seq experiment, the total RNA extracted contains transcripts from both species, providing a comprehensive view of their interaction at the molecular level. As with other RNA-seq protocols, careful planning of the experimental design is essential to ensure that the data generated can answer relevant biological questions. Several critical factors must be considered, such as whether to use strand-specific protocols which preserve the directionality of each transcript. Strand-specific RNA-seq allows for more precise quantification of expression levels, particularly for antisense transcripts and genes with overlapping loci that are transcribed from opposite DNA strands (Zhao *et al.*, 2015). Additionally, sequencing depth - defined as the total number of reads sequenced per sample - is crucial for the downstream statistical analysis aiming at the detection of differentially expressed genes (DEGs). Higher sequencing depth enhances the power and robustness of downstream analyses, increasing the likelihood of detecting subtle changes in gene expression (Cai *et al.*, 2012).

High sequencing depth is especially crucial when one of the interacting organisms lacks an assembled genome or transcriptome. In such cases, *de novo* transcriptome assembly must be performed, and the quality and completeness of the assembly depend heavily on sequencing depth (Patterson *et al.*, 2019). Without sufficient depth, the assembly may fail to capture the full range of transcripts, resulting in incomplete data and potentially skewed interpretations.

Equally important is the inclusion of an adequate number of biological replicates to ensure robust and biologically meaningful results. Replication is key to improving the accuracy and reliability of detecting DEGs. According to Schurch *et al.*, (2016), a minimum of six biological replicates is recommended, with up to twelve replicates necessary when aiming to identify DEGs across all fold changes (i.e., the ratio of gene average expression between experimental groups). Without proper biological replication, no valid statistical inference can be made, rendering *p-value* calculations meaningless and compromising the reliability

of the results. An insightful perspective connecting sequencing depth and the number of biological replicates was provided by Liu et al., (2014). They emphasized that sequencing depth remains useful only up to the saturation point, where increasing the number of transcripts sequenced no longer significantly increases the number of detected genes. Liu et al. (2014) proposed guidelines for finding an optimal balance between sequencing depth and biological replicates to maximize the power of detecting DEGs. Before conducting dual RNA-seq, an important step to minimize biases in the analysis is the inclusion of exogenous reference transcripts, commonly known as spike-ins. RNA spike-ins are artificial transcripts of known sequence and quantity that are mixed with biological samples during RNA extraction. These spike-ins serve as internal controls, enabling more accurate library-size normalization by providing a reference for comparison with the transcripts detected in the sample (Jiang *et al.*, 2011; Lovén *et al.*, 2012).

Once RNA samples are collected and extracted, the transcripts must be converted into cDNA. The method used to build the cDNA library depends on the sequencing platform chosen. Most sequencing platforms generate raw data in FASTQ format (Cock *et al.*, 2010), which serves as the starting point for bioinformatics analysis. The first step of the analysis is data pre-processing, which includes performing quality control (QC) to guide subsequent data cleaning steps such as adapter removal, base trimming, quality filtering, and contamination removal. Quality control scans the FASTQ files to assess base quality scores, calculate GC content, detect the presence of adapters, identify over-represented k-mers, and assess the level of duplicate reads. More specifically:

- Adapters, oligonucleotides added during library preparation, need to be removed as they can interfere with accurate alignment to the reference sequences.
- Trimming reads is essential to eliminate sequencing errors at both 5' and 3' ends of reads. If a read has an overall low-quality score, it may be better to filter it out entirely to avoid ambiguous alignments in downstream analysis.
- Detecting and filtering out contamination, such as reads from other species or sources, is crucial to ensuring clean, species-specific results (Gondane and Itkonen, 2023).

By incorporating these steps, researchers can improve data accuracy and reduce potential sources of error in dual RNA-seq experiments, resulting in more reliable and interpretable findings.

Once reads are pre-processed, the next step is aligning them to a reference to identify each transcript. Ideally, each read should closely match the reference sequence. However, the chosen alignment algorithm must be flexible enough to account for natural variations like single nucleotide polymorphisms (SNPs), insertion/deletion (indels), and sequencing errors that may be present in the data. When a genome assembly is not available, reads can be mapped to the transcriptome. Aligning reads to the reference genome yields different results compared to aligning them to the reference transcriptome. Mapping to the genome typically results in a higher percentage of mapped reads because unannotated transcripts - those not yet included in the transcriptome - can still be captured. In contrast, mapping to the transcriptome may result in more reads mapping equally well to multiple transcripts, often due to shared exons between different isoforms, repetitive sequences, or common

domains found in paralogous genes (Conesa *et al.*, 2016). Additionally, aligning to the transcriptome limits the ability to discover new, unannotated transcripts, whereas genome mapping allows for the identification of previously unknown regions of transcription.

The outcome of the alignment process is usually stored in the Sequence Alignment/Map (SAM) format, or its more compact binary version (BAM: Binary Alignment/Map), which reduces storage space. Recently, the Compressed Reference-oriented Alignment Map (CRAM) format has been introduced as a more space-efficient alternative, and it is continually being optimized with the potential to eventually replace the BAM format (Liu *et al.*, 2023).

Read alignment is a crucial step in dual RNA-seq analysis because it enables the *in silico* separation of mixed reads based on the sequence divergence between the interacting species (Westermann, Gorski and Vogel, 2012). In dual RNA-seq, there are two primary approaches to mapping these mixed reads: the sequential approach and the combined approach, each involving different strategies for aligning reads to reference sequences.

The sequential approach involves mapping the reads in two stages. First the mixed reads are aligned to the reference genome or transcriptome of one of the interacting species. Any unmapped reads from this initial alignment are then mapped to the reference sequence of the second species. This stepwise method ensures that each species' reads are gradually separated based on their unique sequence characteristics.

In contrast, the combined approach maps the mixed reads in a single step. A concatenated reference sequence is created by merging the genomes or transcriptomes of both interacting species, and the reads are mapped against this combined reference all at once (Wolf *et al.*, 2018). After mapping, the single alignment output must be carefully processed to distinguish which transcripts belong to each species, ensuring accurate separation of the two transcriptomes.

Both approaches have their advantages: the sequential approach may offer more control over the mapping process, while the combined approach streamlines the workflow by reducing the need for multiple mapping steps. However, the combined approach requires careful handling of the merged alignment file to accurately allocate reads to the appropriate species.

In cases where neither a reference genome nor a transcriptome is available for one of two interacting species, the sequential approach becomes essential. In this scenario, the unmapped reads, coming from the initial alignment must be used to perform a *de novo* transcriptome assembly. The unmapped reads are assembled into longer contigs, effectively representing the expressed transcriptome. Subsequently, these contigs are used to remap the original unmapped reads (O'Keeffe and Jones, 2019). Both mapping approaches are susceptible to the issue of cross-mapping, where reads from one organism align well to the reference of another organism or even to both.

This phenomenon can significantly confound the analysis, as it introduces ambiguity in determining the true source of the reads. Therefore, it is advisable to filter out these cross-mapped reads to improve the reliability of the results.

This removal can be achieved in the sequential approach through an additive mapping step. In contrast, it is more feasible to address cross-mapping in the combined approach, with a careful alignment files handling.

The estimation of expression levels for specific biological entities (e.g., exons, transcripts or genes) is referred to as summarization. Mapped reads must be summarized in a "table of counts", which reports the number of sequences assigned to each selected entity. A portion of the reads often aligns within unannotated genomic regions, providing a potential starting point for enriching the current genome annotation of the species under investigation. In the case of annotated regions, four main scenarios may arise: (1) reads mapping to exons; (2) reads mapping to exon-exon junctions; (3) reads mapping to unannotated exon (which are annotated as introns); (4) reads mapping to unannotated exon-exon junctions. The last two scenarios also contribute to further enhancing the existing genome annotation (Young *et al.*, 2012). At this stage of the analysis, it is advisable to obtain a preliminary overview of data to characterize variation among replicates and across different experimental conditions. Common approaches for visualizing this variation include dimensionality reduction techniques, such as Multi-Dimensional Scaling (MDS) and Principal Component Analysis (PCA), as well as the calculation of correlation coefficients between samples. Dimensionality reduction techniques help visualize clustering among biological replicates allowing for the identification of potential outliers. Similarly, correlation analysis can be employed; a higher correlation coefficient between two samples indicated greater similarity in their expression profiles (Koch *et al.*, 2018). Another preliminary step, before conducting differential gene expression analysis is examining the distribution of expressed genes to establish a threshold for filtering out lowly expressed genes, which often introduce noise into the data. Removing these low expressed genes can enhance the sensitivity of detecting DEGs. Various methods exist for determining an appropriate filtering threshold, and selecting the optimal one can be challenging, as it may vary depending on the RNA-seq pipeline used. Ying Sha *et al.* (2015) evaluated several methods for setting filtering thresholds across different RNA-seq pipelines, concluding that optimal threshold is one that maximizes the number of detected DEGs.

The resulting "table of counts" may contain various sources of variation, often unwanted which can impact subsequent calling of differentially expressed genes. These variations can arise from differences in library sizes, leading to "between-sample" discrepancies, or from variations in gene lengths and GC content, causing "within-sample" differences. Therefore, data normalization is essential to ensure that expression estimates are more comparable across summarized biological entities and between samples (Zyprych-Walczak *et al.*, 2015). Finally, differential gene expression analysis identifies genes that are expressed at significantly different levels across one or more conditions by applying the appropriate statistical test.

For each gene, the null hypothesis assumes that the expression distributions of two groups are the same (i.e., genes are not differentially expressed). If this hypothesis is rejected, the alternative hypothesis is accepted, indicating that the two groups have significantly different expression distributions, and the genes are differentially expressed (Li and Li, 2018). In simpler terms, differences between conditions are considered significant when

the statistical distributions used to model gene count data differ between groups. Various methods have been developed to detect DEGs; each choosing a probability distribution that best fits the data. Since RNA-seq analysis yields discrete data in the form of quantified transcripts, they must be modeled using an appropriate discrete probability distribution.

Conesa *et al.* (2016) suggest that after normalization, RNA-seq data may lose its discrete nature, making a continuous distribution more appropriate for analysis. Several software packages have been developed for DEG calling, with some applying discrete distributions, such as the negative binomial distribution (Di *et al.*, 2011), while others utilize non-parametric approaches (Tarazona *et al.*, 2011; Li and Tibshirani, 2013). As a general guideline, it is recommended to perform differential gene expression (DGE) analysis using multiple software tools and then either choose the more suitable result or reach a consensus between them. The output of DGE analysis is typically a table listing DEGs, where genes with adjusted *p-values* < 0.05 are considered significant. This table includes *p-values*, adjusted *p-values* (accounting for multiple hypothesis testing), and log₂ Fold-Changes (log₂ FC) for each gene.

In the context of host-parasite interactions, these processes are dynamic and can be captured over time through multiple time points, conducting a so-called Time Course (TC) experiment (Figure 1.1), which aim to model the entire biological process. By collecting samples at multiple time points, TC experiments provide insights into the full set of activated genes involved in the interaction, allowing researchers to infer changes in gene expression rate, their temporal order, causal relationships, regulation, and coordination (Spies and Ciaudo, 2015). Several software packages have been developed specifically for detecting DEGs in TC experiments. Spies *et al.* (2015) conducted a benchmark analysis of these tools, including traditional methods such as edgeR and DESeq2, which are typically used for pairwise comparisons. Their results indicated a few tools performed best when a large number of time points were analyzed. However, classical DGE analysis tools like edgeR and DESeq2 also demonstrated similarly strong performance when compared to the top TC-specific tools, making them reliable options even in time course studies. With the decreasing costs of nucleic acid sequencing, experiments now typically involve multiple treatment groups or large numbers of samples.

This increase in data volume makes clustering algorithms highly valuable, as they can be applied to group genes based on their similarity in an unsupervised manner, helping researchers uncover hidden patterns and gain insights from the data. Clustering analysis aids in identifying clusters of genes that share similar expression patterns, which can be used for further downstream analysis (Koch *et al.*, 2018). Clustering can be applied to either genes or to samples. When applied to genes, it identifies co-expressed genes that are likely functionally related, facilitating gene network analysis and improving gene annotation. Conversely, clustering applied to sample groups based on gene expression levels, revealing relationships between the samples. The most commonly used clustering algorithms include K-means clustering, hierarchical clustering, model-based clustering, and hybrid hierarchical clustering (Liu and Si, 2014).

The list of DEGs generated from RNA-seq analysis is crucial for performing pathway enrichment analysis, also referred to as Gene Set Analysis (GSA), which helps identify the

biological processes affected by the experimental conditions (Maleki *et al.*, 2020). GSA methods aim to pinpoint the metabolic pathways significantly impacted by treatments by comparing RNA-seq results to annotated biological databases through statistical tests.

These tests rank the pathways or ontology terms that are significantly involved in the response (Xie, Jauhari and Mora, 2021). Currently, a variety of GSA methods exist, which can be broadly categorized into three classes: Over-Representation Analysis (ORA), Functional Class Scoring (FCS), and Pathway Topology (PT) methods (Khatri, Sirota and Butte, 2012; Bayerlová *et al.*, 2015).

ORA identifies pathways where more genes than expected are differentially expressed, FCS evaluates gene sets based on expression levels across the whole dataset, and PT methods incorporate pathway structure and gene interactions to provide more detailed insights.

Conducting enrichment analysis requires selecting an appropriate database, and over the past decade, the most commonly used databases for this purpose include Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and Gene Ontology (GO) (Mubeen *et al.*, 2022). These resources allow researchers to contextualize DEGs in biological pathways and processes, providing a deeper understanding of the experimental results.

Finally, by leveraging the correlation of gene expression profiles across multiple samples, gene co-expression networks can be constructed. In these networks, nodes represent genes, while edges signify connection between them. A connection is established when, there is significant co-expression between the two connected nodes, typically measured using the Pearson correlation coefficient (Zhang and Horvath, 2005). Notably, the co-expression of genes with known function alongside novel or poorly characterized genes can provide valuable insights into the potential functions of many genes, for which functional information is currently lacking (Liu and Si, 2014). At the conclusion of the analysis, the list of DEGs and the accompanying biological insights derived from downstream analysis serve to guide scientists in uncovering new genes involved in the host-parasite interaction. These insights highlight key genes and provide a comprehensive overview of the dynamic interplay between the host and the parasite throughout the infection process at the transcriptomic level.

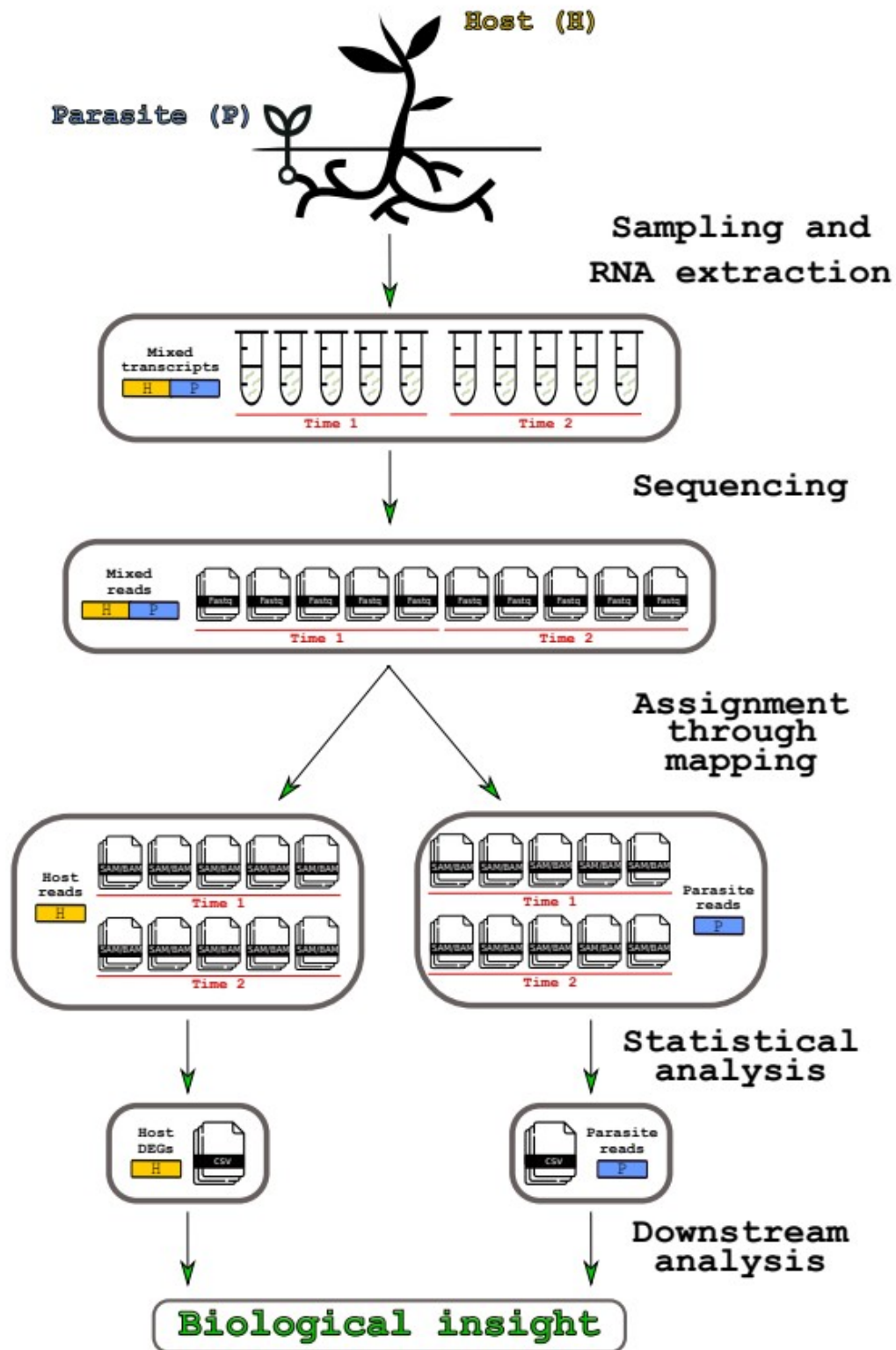


Figure 1.1. Dual RNA-seq workflow for investigating host plant-parasitic plant interactions at two temporal points. The host and parasitic plants are subjected to dual RNA-seq to study the transcriptome in their contact region. Mixed tissue samples are collected at two time points, from which RNA molecules are extracted and sequenced. The resulting sequences are stored in FASTQ files. To assign transcripts to each of the interacting species, the sequences are mapped to their respective reference sequences. Once the reads are assigned, a series of statistical analyses are conducted to identify differentially expressed genes (DEGs). These DEGs undergo several downstream analyses, providing valuable biological insights into the dynamics of the infection.

1.3. Bioinformatics made easy: a Graphical User Interface for beginners

Dual RNA-seq analysis is currently inaccessible to researchers without basic programming skills. However, since the inception of bioinformatics, there has been a trend of developing command line interface (CLI) software for emerging analyses. As these methods become more established, there is a growing tendency to make them accessible to a broader audience (Smith, 2013), as illustrated in Figure 1.2.

Bioinformatics emerged in the early 1960s when Margaret Dayhoff developed COMPROTEIN, the first bioinformatics software, for the IBM 7090 machine, entirely coded in FORTRAN (Dayhoff and Ledley, 1962). COMPROTEIN was also the first *de novo* assembler for protein sequences, utilizing Edman peptide sequencing data (Edman and Begg, 1967).

Early bioinformatics software was written in low-level programming languages like C and Fortran. This remained the standard until the late 1980s, when high-level programming languages such as Perl and Python were introduced. Unlike low-level languages, these interpreted languages do not require compilation, making development more flexible. Over time, access to bioinformatics tools became increasingly user-friendly. Until the early 1990s, computational biology analyses were exclusively performed through CLI, requiring knowledge of UNIX-like operating systems and programming skills.

However, developers soon began creating graphical user interfaces (GUI) to make basic analyses accessible to non-bioinformaticians. At present, user-friendly bioinformatic GUIs, such as Galaxy (Afgan *et al.*, 2018) and the BLAST website (Boratyn *et al.*, 2013), are continually expanding in availability and usability (Gauthier *et al.*, 2019). GUIs offer researchers a simplified way to communicate and interact with computers without needing to understand programming language. They provide various components, such as icons, menus, toolbars, ribbons, buttons, boxes, which enable users to interact with the system, typically through a pointing device. User actions trigger events that guide the GUI to perform specific tasks. GUIs come in various form, including operating systems, data analysis packages, spreadsheet, notebook interfaces, and, with the growth of bioinformatics, web-based platforms. To ensure a GUI meets user needs, developers avoid overly complex designs and provide clear, comprehensive documentation (Martinez, 2011). When developing a GUI, it is essential for developers to adopt a declarative programming style, where code or specific segments are executed based on user actions, allowing for more dynamic and user-driven interactions (Ari and Mamatnazarova, 2014).

The most commonly used programming languages for building GUIs are Python and Java (Khoirom *et al.*, 2020). These are often supplemented by three additional languages: HyperText Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript, which are crucial for designing and enhancing web-based interfaces (Kienle, 2010).

However, one of the most widely used programming language in bioinformatics is R (R Core Team, 2024), whose functionality is continuously enhanced by the development of open-source packages from the R community. Among these, the Shiny package enables the creation of interactive GUI applications within the R environment. Shiny provides a framework that allows the design of graphical components of the client side while performing data analysis on the server side, with specific protocols to seamlessly connect the two (Jia *et al.*, 2022). Given this, the development of a GUI for dual RNA-seq analysis can

be seen as a valuable responsibility for the bioinformatics community, ensuring broader accessibility to complex analyses.

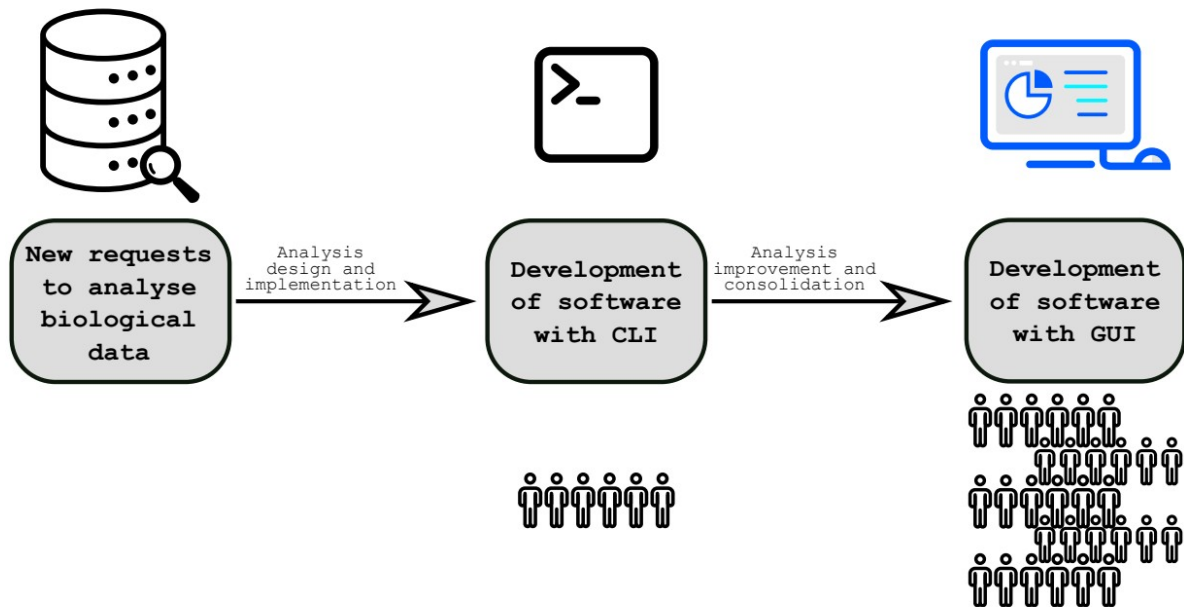


Figure 1.2 Evolution of bioinformatic analyses: from command line to user-friendly graphical interfaces. The emergence and probable evolution of bioinformatic analyses are closely tied to the development of new technologies for generating biological data or expanding existing datasets within databases. Such advancements often necessitate the design of innovative bioinformatic analyses. Typically, the implementation of a new analysis begins with the development of software utilizing a Command Line Interface (CLI), which requires basic programming skills that not all interested researchers possess. As the analysis matures and becomes more established, a Graphical User Interface (GUI) is usually created, facilitating access for a broader range of researchers and enhancing usability.

1.4. Parasitism in the *Plantae* kingdom: mechanisms and evolutionary implications

A specific interaction within the *Plantae* kingdom, that requires dual RNA-seq analysis, involves parasitic plants that use other plants as hosts.

Parasitism in the plant kingdom represents a fascinating example of the evolution of novel traits, where parasitic plants have developed the ability to establish direct connections with their hosts, invading them through specialized structures and extracting nutrients. Parasitic plants can be classified based on their degree of dependence on the host and their photosynthetic capabilities. In terms of host dependence, parasitic plants are categorized as either facultative or obligate. Facultative parasites can live autotrophically, parasitizing a host opportunistically, while obligate parasites rely entirely on the host for survival. Based on their photosynthetic status, parasitic plants are further divided hemiparasites—plants that retain photosynthetic capabilities—and holoparasites, which lack any photosynthetic activity and are fully dependent on their host for nutrients (Yang *et al.*, 2015). Parasitic plants represent about 1,2% of flowering plant species and have contrasting effects on agriculture and ecosystems. In agriculture, parasitic plants act like "Dracula", draining vitality from crops and significantly reducing their productivity. However, from an ecological perspective, they steal nutrients from dominant plant species, contrasting their overexpansion within the plant community, and, as a consequence, promoting biodiversity (Twyford, 2018).

Only a few parasitic plants cause severe damage to crops, with witchweed and broomrapes being the most destructive. *Striga* species (witchweed), obligate parasite of the *Orobanchaceae* family, are particularly harmful to cereal and legume cultivation in sub-Saharan Africa, Middle East and Asia, leading to billions of dollars in crop losses annually (Dafaallah, 2019). Similarly, *Orobanche* and *Phelipanche spp.* attack a variety of crops across southern and eastern Europe, the Middle East and North Africa, posing a major threat to agricultural productivity in these regions.

In contrast, grassland hemiparasites (mainly *Rhinantus spp.*) play an important ecological role by controlling the overgrowth of competitively dominant plants. This regulation helps maintain high biodiversity within ecosystems, benefiting various trophic levels (Chaudron *et al.*, 2021). Another positive aspect of parasitic plants is their use in horticulture. For instance, *Rafflesia arnoldii*, known for producing the largest flowers in the world, is admired for its ornamental value (Susatya, 2020).

Among parasitic plants, the genus *Cuscuta* (dodders) is one of the most thoroughly studied, with research dating back to the early 20th century. This widespread distribution of dodders and their significant negative impact on agricultural yields have made them a key focus of scientific inquiry (Hartenstein, Albert and Krause, 2023). As a result, studies on *Cuscuta* have provided valuable insights into the mechanisms and dynamics of plant parasitism. The evolution of parasitic functions in plants occurred primarily through genome duplication events, followed by the acquisition of novel gene functions (neofunctionalization). Comparative transcriptomic studies have suggested that many parasitism-related genes originated from genes associated with root and floral tissues (Yang *et al.*, 2015). In holoparasites, the loss of autotrophic functions has been extensively investigated through the molecular evolution of plastid genomes, where gene loss has been a major driving force in their adaptation to parasitism (Wicke and Naumann, 2018).

Despite significant progress in understanding parasitism-related genes, many questions remain unanswered. These include how much and when these genes are expressed, the interactions between genes during haustorium development, and the molecular details of host-parasite interaction dynamics (Twyford, 2018). Dual RNA-seq analysis, applied to the precise regions of host-parasite interaction, holds great potential for addressing these open questions and providing deeper insights into the molecular mechanisms of parasitism.

1.5. The biology and impact of the parasitic plant *Phelipanche ramosa*

Members of the *Orobanchaceae* family (*Striga*, *Orobanche* and *Phelipanche*) are among the most agriculturally destructive parasitic plants (Westwood *et al.*, 2010), particularly in the Mediterranean basin (Casadesús and Munné-Bosch, 2021).

Within this family, the root holoparasitic weed *Phelipanche ramosa* poses a significant threat to important crops, such as oilseed rape, flax, eggplant, potato, tobacco and tomato (Mutuku *et al.*, 2021). This species is widely distributed across the Mediterranean basin and central Europe. One notable characteristic of *P. ramosa* is its reproductive strategy: it produces thousands of tiny seeds that can remain viable in the soil for several years. Consequently, its broad host range, combined with challenges associated with controlling this persistent parasite, significantly elevates the threat posed by *P. ramosa* (Pérez-De-Luque *et al.*, 2009). The life cycle of *P. ramosa* is divided into two distinct phases: an underground phase and an aboveground phase.

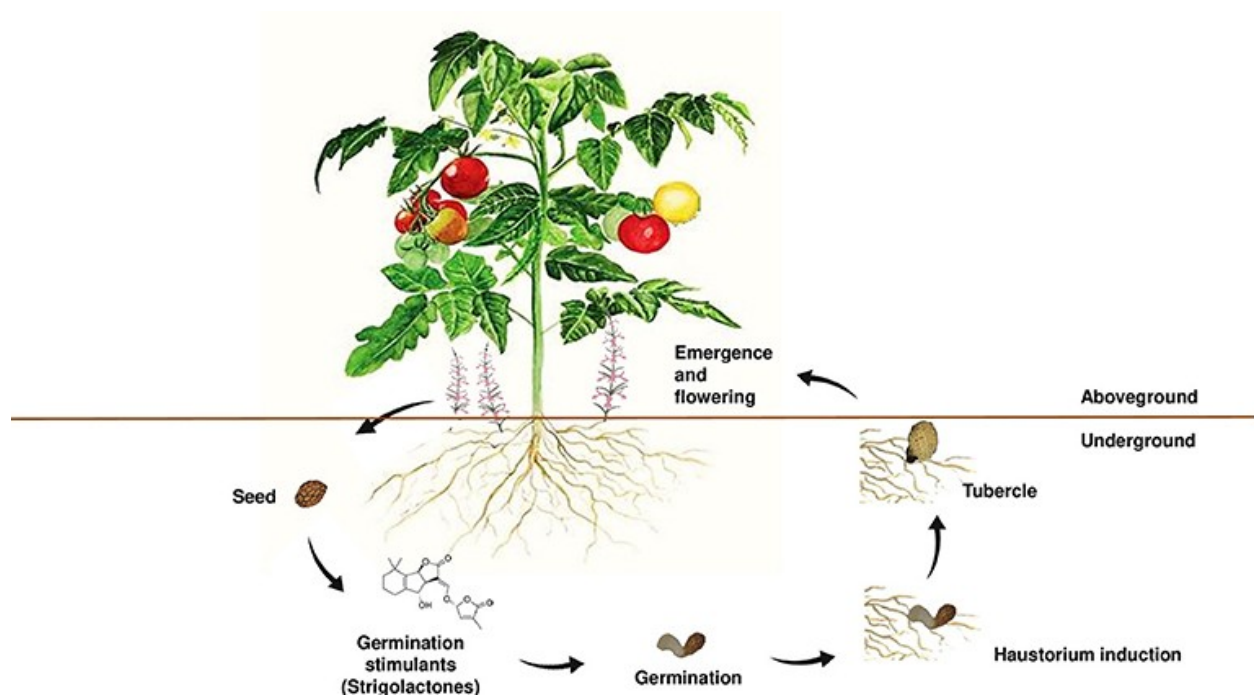


Figure 1.3 Lifecycle of a broomrape. Image coming from 10.3389/frym.2023.1207431. During underground phase, broomrape seeds germinate after SLs stimulation. After attachment to host roots, broomrapes develop haustorium organ first and tubercle structure then. Aboveground phase begins with the emergence of broomrapes from soil, ending with reproduction and seed dissemination.

The underground phase begins with seed conditioning, a process in which seeds break dormancy when exposed to specific moisture and temperature conditions. Once dormancy is broken, the seeds become capable of detecting germination stimulant molecules produced by the roots of a host plant. These stimulants primarily belong to the strigolactone family, a group of plant hormones that are recognized as the main triggers for parasitic seed germination (Huizinga and Bouwmeester, 2023). Strigolactone are released by host roots, especially under condition of phosphorous deficiency, as a way to promote symbiosis with arbuscular mycorrhizal fungi (Mashiguchi, Seto and Yamaguchi, 2021).

Upon stimulation, germination begins with the development of a germ tube, known as a procaulome, which grows towards the host roots via positive chemotropism. Once the procaulome reaches the host, it forms a specialized structure called the appressorium. This structure expands radially and penetrates the host roots using both mechanical pressure and enzymes that degrade the root's surface. After the initial penetration, the haustorium, a specialized organ responsible for nutrient absorption, starts to develop. The haustorium creates a physical and physiological connection between the parasite and the host's vascular system, particularly the phloem, allowing the parasite to extract nutrients. As the haustorium matures, it transforms into a tubercle, a tuberous mass with small adventitious roots and shoot apical meristems. The tubercle serves as a temporary storage site for resources. In response to specific signals from the host, the tubercle then produces stems that lack chlorophyll, signaling the end of the underground phase. The emergence of the parasite plant above the soil marks the beginning of the aboveground phase, which continues with flowering, self-fertilization, and seed production (Delavault, 2015).

The molecules exchanged between host and parasite are not limited to photosynthates and reduced nitrogen compounds. Nucleic acids, particularly RNA molecules, are transferred bidirectionally between the two organisms. This exchange occurs through the tight connection established during the parasitic infection process. RNA molecules can move not only between cells of the same organism but also across species boundaries, as observed between host plants and parasitic plants (LeBlanc, Kim and Westwood, 2012).

These mobile RNA molecules have the potential to influence the growth and development of the recipient cells. Beyond transient effects on gene expression after transcription, these transferred RNAs may also exert long-term impacts by modifying gene expression patterns acting on DNA methylation patterns (Westwood and Kim, 2017; Clarke *et al.*, 2019). This suggests that RNA exchange plays a significant role in the parasitic relationship, affecting both immediate cellular processes and more stable epigenetic changes.

Interestingly, the phenomenon of Horizontal Gene Transfer (HGT, i.e., the movement and the genomic integration of DNA across connections between two different species) has also been observed in interactions between host and parasite plants (Kado and Innan, 2018). A more recent study (Yang *et al.*, 2019) proposed that HGT could potentially occur via the exchange of RNA molecules. This RNA could then be reverse-transcribed, leading to the genomic integration of retro-processed sequences. A key signature supporting this hypothesis is the absence of introns in the identified HGT sequences, which is characteristic of retro-transcription events.

Considering these factors, *P. ramosa* emerges as an ideal model for advancing our understanding on plant parasitism dynamics. Its significant impact on the yield of numerous economically important crops makes it a crucial species for studying both the molecular mechanisms of parasitism and the potential role of HGT in plant evolution.

1.6. *Solanum lycopersicum*: a host for the parasitic plant *Phelipanche ramosa*

One of the most economically relevant hosts of *P. ramosa* is the cultivated tomato, *Solanum lycopersicum*. This species belongs to the *Solanaceae* family and is primarily recognized as a global food crop, while also serving as an ornamental plant in some regions. Among the *Solanaceae* spp. utilized for agricultural purposes, *S. lycopersicum* is the second most cultivated after *Solanum tuberosum* (potato). It originated in South America. However, its introduction as a food source can be traced back to Mexico, and following the Spanish colonization of the Americas, *S. lycopersicum* was disseminated worldwide. Cultivated *S. lycopersicum* has 12 wild relatives, which are characterized by substantial genetic diversity. This diversity plays a crucial role in the genetic improvement of modern cultivars.

Additionally, from a scientific perspective, *S. lycopersicum* is a diploid model species ($2n=2x=24$) that is widely used for genetic and genomic research (Gatahi, 2020), alongside *Arabidopsis thaliana* (Hays, 2002). *P. ramosa* poses a challenge to *S. lycopersicum* production, particularly in Mediterranean climates, as it contributes to reduced yields (Mauromicale, Monaco and Longo, 2008).

Unfortunately, parasitic weeds such as *P. ramosa* are challenging to control, leading to significant production losses. Over the years, various control strategies of different type have been implemented. Cultural methods include targeted fertilization to reduce the production of germination stimulants from host roots, intercropping to reduce host density, and crop- rotation with trap plants (which do not belong to the parasite's host range) that stimulate seed germination. In addition to cultural methods, physical strategies such as (solarization), chemical approaches involving artificial germination stimulants, germination inhibitors, fumigants, and herbicides, as well as biological control methods that utilize fungi and bacteria capable of infecting seeds and the application of phytophagous insects, have been explored. However, for various reasons detailed in the literature (Fernández-Aparicio, Reboud and Gibot-Leclerc, 2016; Fernández-Aparicio, Delavault and Timko, 2020), all of these methods have demonstrated limited applicability and success. The optimal strategy to combat the threat posed by *P. ramosa* may involve identifying genetic sources of resistance to this parasitic weed, ultimately leading to the development of a breeding program aimed at producing tolerant lines with desirable agronomic traits. Sources of resistance can be explored within the wild relatives of *S. lycopersicum*. To date, the most resistant wild relatives to *Phelipanche* spp. infections include *Solanum pennellii*, *Solanum hirsutum*, *Solanum pimpinellifolium* and *Solanum chilense*; all of which exhibit either no or significantly reduced numbers of emerged parasitic shoots (Cuccurullo, Nicolai and Cardi, 2022).

Among these, *S. pennellii* exhibited the most promising results. El-Halmouch et al, (2006) demonstrated that root exudates from *S. pennellii* inhibited the germination of *Phelipanche aegyptiaca* (a *Phelipanche* spp. close to *P. ramosa*) seeds. Furthermore, after infection, the development of tubercles was significantly restricted. Additionally, (Bai et al., 2020) reported similar levels of resistance to *P. aegyptiaca*, in both in *S. pennellii* and two introgression lines (ILs) of *S. pennellii* (Alseekh et al., 2013). Introgression lines (ILs) are plant lines that contain homozygous chromosomal regions from a wild donor species (in this case, *S. pennellii*) incorporated into the genetic background of an elite cultivar (the recipient), specifically *S. lycopersicum* cv. M82. These lines are created through selective breeding and backcrossing, allowing researchers to study the effects of specific wild alleles

on desirable traits while maintaining the overall agronomic characteristics of the cultivated variety. Resistance to *P. ramosa* can generally be classified into three main categories: pre-attachment, during-attachment and post-attachment, each focused on the parasite's attempt to establish a connection with the host's vascular bundles.

- Pre-attachment defense: This involves the release of low levels of germination stimulants to drastically reduce seed germination (Sisou *et al.*, 2021).
- During-attachment defense: This defense mechanism is characterized by the accumulation of phenolic compounds, which enhance the hardness of the root surface, making it more difficult for the haustorium to penetrate the host's vasculature (Mutuku *et al.*, 2019).
- Post-attachment defense: In this stage, the host blocks the translocation of nutrients to the parasite by accumulating mucilage-like substances in its vascular cells, gradually diminishing the vitality of the parasite (Yoder and Scholes, 2010).

These three categories highlight the various strategies that host plants employ to defend themselves against parasitic invasion. Given its economic significance, cultivated *S. lycopersicum* is an excellent candidate for investigating the dynamics between host plants and *P. ramosa* with the aim to develop strategies for managing parasitic plants, while minimizing disruptions to ecosystem balance. Understanding these interactions will facilitate the formulation of effective containment measures against parasitic invasions, promoting sustainable agricultural practices. To achieve these goals, an analysis of genomic polymorphisms in wild tomato species compared to cultivated tomato genotypes, coupled with a dual-RNA-seq analysis at the plant- parasite interaction site, could prove to be an effective solution. This approach involves two interconnected omics layers, offering a comprehensive perspective.

1.7. Aim of the thesis

The present PhD thesis has three primary objectives:

- Assess the feasibility of dual RNA-seq: evaluate the application of dual RNA-seq to two interacting organisms within the same taxonomic kingdom.
- Develop a user-friendly GUI: create a graphical user interface (GUI) designed to provide an optimized tool for conducting dual RNA-seq analysis, making it accessible to users without programming skills.
- Identify genetic polymorphisms: investigate genetic polymorphisms between *S. lycopersicum* and its wild relative *S. pennellii*, serving as an intriguing starting point for dissecting the genetic basis of tolerance mechanisms.

Chapter 2. Refining dual RNA-seq mapping: sequential and combined approaches in host-parasitic plant dynamics

2.1 Abstract

Transcriptional profiles of host plant-parasitic plant interactions are challenging to obtain due to the close physical connection at their interface and the high percentage of homologous sequences shared between the two organisms. Biases arising from the physical separation of these interacting species can be mitigated using dual RNA-seq, a method that allows for the *in silico* separation of mixed transcripts collected from the interface region. However, dual RNA-seq must contend with multiple mapping reads within a single reference genome as well as cross-mapping between the host and parasite genomes. Furthermore, the incidence of cross-mapping events tends to increase as the evolutionary divergence between organisms decreases. Discarding multi-mapped and cross-mapped reads typically results in underestimating gene/transcript abundance.

This chapter evaluates the feasibility of dual RNA-seq by applying it to two simulated interactions involving the parasitic plant *Cuscuta campestris* and two host plants: *Arabidopsis thaliana* (as a model organism) and *Solanum lycopersicum* (as a crop phylogenetically closer to *C. campestris*).

The closeness among the three plant species was confirmed through a phylogenetic analysis, followed by the merging of the transcriptomes of each host plant with that of the parasitic plant. Read assignment was performed using by two different alignment approaches: mapping reads sequentially to two separate reference genomes (sequential approach) and mapping reads to a single concatenated genome (combined approach). In both approaches, the mapping rate was approximately 90%, while only about 1% of reads being discarded due to cross-mapping, indicating the method effectiveness in separating mixed transcripts *in silico*. Both the sequential and combined approaches proved reliable, with only minor differences between them. The combined approach yielded slightly more accurate alignments, reducing the number of incorrectly assigned reads. Additionally, it was less computationally demanding than the sequential approach. Interestingly, the taxonomic closeness of *S. lycopersicum* and *C. campestris* did not result in a higher rate of incorrect assignment, despite a higher percentage of homologous sequences compared to the assignment results for *A. thaliana* and *C. campestris*.

This study demonstrated the reliability of dual RNA-Seq as an effective method for analyzing interactions between hosts and parasites within the same taxonomic kingdom. This approach paves the way for new experiments aimed at expanding our understanding of the key genes involved in plant parasitism.

2.2 Introduction

Parasitism in angiosperms involves a parasitic plant deriving nutrients from host plants. This complex ecological strategy has evolved independently approximately a dozen times, resulting in more than 290 genera and 4,700 species of parasitic plants (Westwood *et al.*, 2010; Nickrent, 2020). Notably, certain parasitic species exhibit generalist behaviors, enabling them to parasitize multiple host species. At least 25 genera are recognized crop pathogens, including *Striga* (witchweeds), *Orobanche*, *Phelipanche* (broomrapes), and *Cuscuta* (dodder), posing significant threats to agriculture (Nickrent and Musselman, 2004). Quantifying yield losses can be challenging, yet the impact of parasitic weeds on international agriculture is undeniably on the rise (Hegenauer, Körner and Albert, 2017).

The evolutionary transition from non-parasitic ancestors to parasitic plants marked a shift from autotrophy (self-sustained nutrition through photosynthesis) to varying degrees of heterotrophy (reliance on external sources for sustenance) (Westwood *et al.*, 2010). One way for classifying parasitic plants is based on their photosynthetic capacity. Hemiparasites can photosynthesize but primarily rely on hosts for water and mineral nutrients, while holoparasites lack photosynthesis and depend entirely on hosts for nutrition. Another classification is based on their dependency from hosts to complete their lifecycle: obligate parasites require a host, whereas facultative parasites can reproduce independently. While holoparasites are necessarily obligate due to their lack of photosynthesis, also some hemiparasites still need hosts for greatly enhanced the reproductive success thanks to the increased intake of mineral elements (Klaren and Janssen, 1978; Lambers and Oliveira, 2019). These interactions between host and parasite can have a significant impact on growth, reproduction, physiology, and ecosystem dynamics of the host (Hegenauer, Körner and Albert, 2017).

The most renowned and widespread stem holoparasitic genus is *Cuscuta*, comprising of about 170-200 species (Park *et al.*, 2019). These plants have degenerated roots and leaves and stems spiral counterclockwise around their host plants.

The trophic connection between *Cuscuta* and its host relies on the development of a specialized structure called haustorium. The haustorium develops in stages, with haustorial cells forming hyphae that penetrate the vascular tissues of the host, mimicking xylem or phloem conduits. This intimate connection allows the parasitic plant to acquire water and nutrients and facilitates the horizontal transfer of macromolecules, including messenger RNA, small and long non-coding RNA, proteins, and even pathogens like viruses, phytoplasmas and viroids (Hosford, 1967; van Dorst and Peters, 1974; Kamińska and Korbin, 1999; Haupt *et al.*, 2001; Birschwilks *et al.*, 2006; Kim *et al.*, 2014; Johnson and Axtell, 2019; Subhankar *et al.*, 2021; Jhu and Sinha, 2022; Y. Wu *et al.*, 2022).

The exchange of RNA between hosts and *Cuscuta* can occur in both directions. For example, a study on *Cuscuta pentagona* parasitizing *Arabidopsis thaliana* found that 45% of the genes expressed in *Arabidopsis* were detected in *Cuscuta*, and conversely 24% of the genes expressed in *Cuscuta* were found in the *Arabidopsis* stem (Kim *et al.*, 2014). The exact role of these mobile RNAs is not fully understood, although different studies have suggested that some of them may be translated into proteins that affect plants' physiology or act as modulators of gene expression in response to abiotic and biotic stress (Westwood and Kim, 2017; Shahid *et al.*, 2018; Hudzik *et al.*, 2020; Maizel *et al.*, 2020; Park *et al.*, 2022).

The interface region acts as a battleground, where gene expression changes involve both parasite and host to disrupt various physiological processes of the other. These include recognition via pattern recognition receptors (PRRs), production of cytotoxic compounds, establishment of physical barriers, release of reactive oxygen species (ROS), and initiation of plant cell death (Albert, Axtell and Timko, 2020). Understanding this complex conflict requires a deep understanding of the gene expression changes occurring in both host and parasite plant as they engage in interaction.

Since its introduction, RNA-sequencing (RNA-seq; known as bulk RNA-seq) has emerged as the favored technology for this purpose. Traditionally, RNA-seq reveals mRNA and/or non-coding RNA to provide a snapshot of gene expression in a sample (Mortazavi *et al.*, 2008). Estimation of the average gene expression levels across a population of sampled cells provides insights into tissue-specific molecular mechanisms.

RNA-seq analysis comprises various steps: experimental design and sample acquisition; pre-processing of raw data for quality control; read mapping; gene/transcript level quantification; identification of differently expressed genes; and functional analysis (Conesa *et al.*, 2016; Galise, Esposito and D'Agostino, 2021).

In experiments designed to capture transcriptional profiles of interacting organisms at the interface region, precise tissue sampling and RNA isolation are crucial steps. Traditionally, the isolation of total transcripts has relied on preparing plant tissue sections for laser capture microdissection (LCM) followed by RNA isolation and high-throughput sequencing (Honaas *et al.*, 2013; Jhu and Sinha, 2022). LCM combines microscopy with laser beams to isolate tissue types from host-parasite combined samples. Although this method has undergone refinement to yield efficient outcomes within a feasible timeframe, the intricate nature of the interface tissue (comprising host plant, haustorial and hyphae tissue) poses challenges that require trained personnel and specialized equipment (Park *et al.*, 2022). An alternative to address biases in separation techniques has been represented by the dual RNA-seq approach. This technique relies in sampling the entire “host plant-parasitic plant” interface, and expression profiles are discerned computationally by mapping reads to the respective reference sequences (genome/transcriptome) (Westermann, Gorski and Vogel, 2012; Naidoo *et al.*, 2018; Wolf *et al.*, 2018). The sampling from multiple tissues simultaneously introduces computational challenges in managing reads coming from different organisms. As example, mapping becomes non-trivial as it involves managing multiple mapping events within a single reference sequence (i.e., when reads map equally well on multiple loci within single organism), in addition to cross-mapping events occurring between the two reference sequences.

In this study, we categorized cross-mapping events into two types: (1) one-side cross-mapping, when reads from one organism are exclusively assigned to the other organism (often due to missing mapping to the first organism genome, presumably reflecting genome incompleteness); (2) two-side cross-mapping, when reads from one organism are assigned to both organisms. Within-genome multiple mapping results from gene duplication events, while cross-mapping is mainly due to insufficient divergence in the gene sequences between interacting organisms. For this reason, several reads can ambiguously map within and between organisms, an issue that is further emphasized when using short reads technologies (Westermann, Barquist and Vogel, 2017; Deschamps-Francoeur, Simoneau and Scott, 2020).

The percentage of cross-mapped reads is influenced by various factors, particularly the evolutionary divergence between interacting organisms. Dual RNA-seq approaches have been broadly employed to study various “host plant-parasitic non-plant” interactions (Liao *et al.*, 2019; Du *et al.*, 2022; Walker *et al.*, 2024). In these cases, the interacting organisms are phylogenetically distant, resulting in substantial sequence divergence that reduces the probability of cross-mapping. Nevertheless, the entity of cross-mapping involving phylogenetically close organisms as in “host plant-parasitic plant” interaction, has not been well explored. A naïve approach to handle ambiguously assigned reads is to discard them outright. However, this could potentially underestimate gene/transcript abundance levels. Therefore, it is crucial to develop and implement more advanced techniques capable of accurately aligning reads to the reference genome in dual RNA-seq applications. The accuracy of this procedure depends on various factors, including the choice of alignment algorithm, the quality of the reference sequence, and the configuration of the algorithm parameters (Srivastava *et al.*, 2020).

In this study, we assessed the feasibility of using dual RNA-seq to investigate interactions between phylogenetically close parasite and host species, both belong to the Plantae kingdom, with focus on challenges related to multiple mapping and cross-mapping. More specifically, we simulated two *in silico* interactions involving the parasitic plant *Cuscuta campestris* with two different hosts: *A. thaliana* (as a model organism) and *Solanum lycopersicum* (as a crop phylogenetically closer to *C. campestris*). The mapping was performed using the available reference genomes of *C. campestris*, *A. thaliana* and *S. lycopersicum*. We did not consider de novo transcriptome assemblies due to the uncertainty about horizontal transfer between host and parasitic plants, which could result in misattributed transcripts.

We compared two approaches to differentiate mixed reads, as summarized in Figure 1: mapping reads sequentially to the genomes of both species (the sequential approach), and mapping reads to a single combined genome (the combined approach).

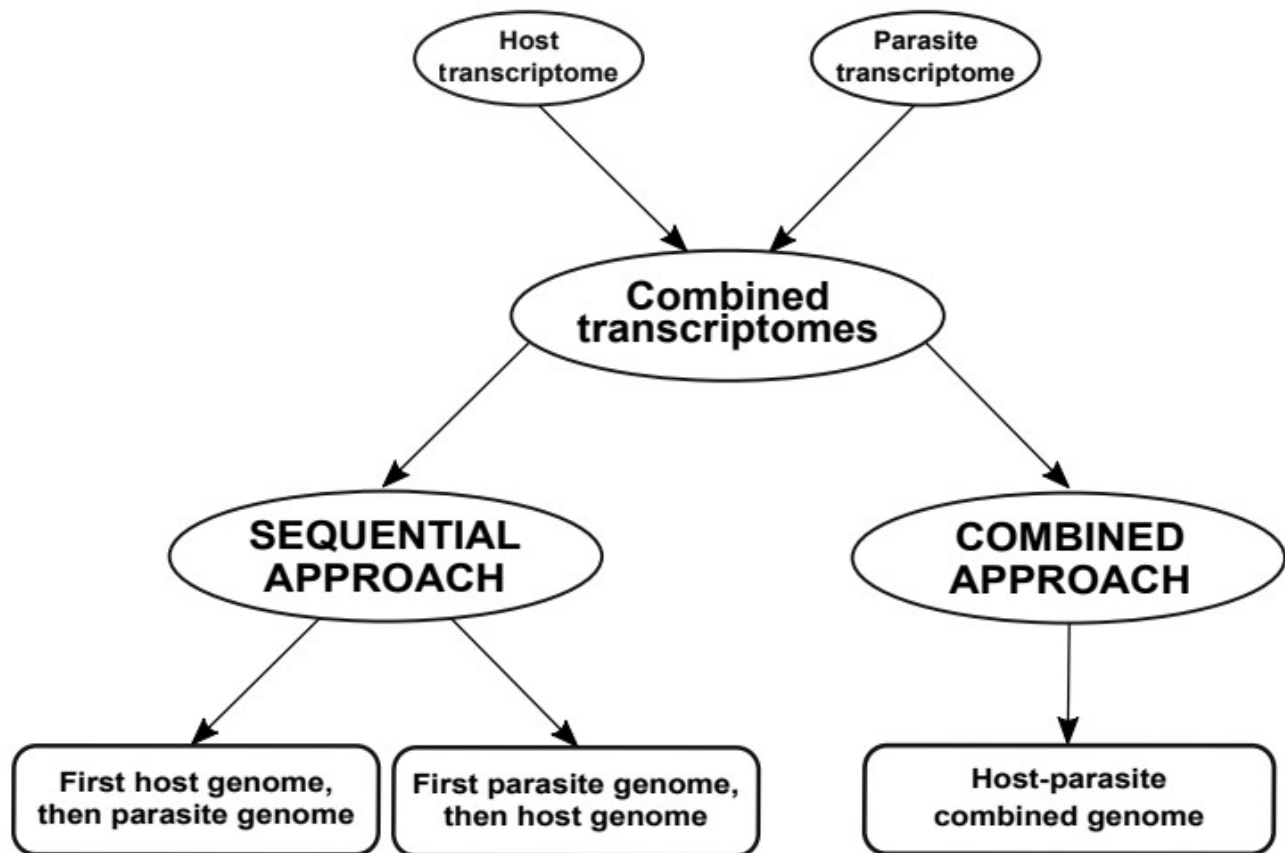


Figure 2.1 Workflow employed for analyzing dual RNA-seq data in this case study. A dual RNA-seq experiment was simulated by merging transcriptomes from two species involved in a parasitic relationship. Subsequently, merged transcriptomes were analyzed using two approaches: sequential and combined. In the sequential approach, the merged transcriptomes are aligned first to the host plant and then to the parasitic plant (or vice versa). In the combined approach, the merged transcriptomes are aligned to a single genome that combines both host and parasite genomes.

2.3 Materials and methods

To manage the data, scripts in R (version 4.3.3) and GNU bash (version 5.0.17(1)) were developed in house (GNU, 2007; R Core Team, 2024). The Sankey plot was generated using SankeyMATIC (<https://sankeymatic.com/>) and further modified with Inkscape version 1.3 (<https://inkscape.org/>).

2.3.1 Phylogenetic analysis of *Cuscuta* spp. and their host range

The large subunit of the ribulose-bisphosphate carboxylase (*rbcL*) gene (Manhart, 1994) was considered to perform a phylogenetic analysis with the aim of comparing *Cuscuta* species and their host plants. The *rbcL* protein sequences were used to build a multiple alignment using MAFFT v7.520 with the iterative refinement method (Katoh, 2002). IQ-TREE version 2.2.2.6 was used to construct a maximum-likelihood phylogenetic tree with 1,000 bootstrap replicates (Minh *et al.*, 2020). The LG+I+G4 substitution model was identified as the best-fitting model for the analysis. Finally, the resulting tree was visualized via the R packages Treeio v1.26.0 and ggtree v3.10.1 (Wang *et al.*, 2020; Yu, 2020). Detailed identifiers of the sequences used in tree construction can be found in Supplementary Table 2.1.

2.3.2 Retrieval of input data and their pre-processing

C. campestris as parasitic plant and *A. thaliana* and *S. lycopersicum* as host plants were considered. The reference genomes of these species were retrieved from the European Nucleotide Archive (ENA) repository: *A. thaliana* (GCF_000001735.4), *S. lycopersicum* (GCF_000188115.5) and *C. campestris* (GCA_900332095.2). For each species, RNA-seq data from independent studies were downloaded, ensuring that three biological replicates were included for each (Table 2.1): *A. thaliana* data from the stem tissue (ENA acc. no.: SRR22559142, SRR22559143, SRR22559144) of the Columbia ecotype (Col-0) sampled during the vegetative stage (~20.1 M of reads on average); *S. lycopersicum* data from the stem tissue (ENA acc. no.: SRR25558913, SRR25558914, SRR25558915) of the Heinz 1706 cultivar (~43.2 M of reads on average); *C. campestris* data from developing haustoria (ENA acc. no.: SRR12763776, SRR12763787, SRR12763788), without host contact (~14.7 M of reads on average) (Bawin *et al.*, 2022). All replicates were selected based on the use of NovaSeq 6000 sequencing technology and to ensure similar average read lengths (150 bases for the hosts and 100 bases for the parasite). This approach was chosen to minimize mapping performance biases related to sequencing technology platform. These RNA-Seq paired-end libraries were filtered using Trimmomatic version 0.39 with parameters: LEADING=20; TRAILING=20; SLIDING WINDOW=4 (Bolger, Lohse and Usadel, 2014). Only reads ≥75 nucleotides were retained.

Plant species	ENA study accession number	ENA run accession number	Number of raw reads	Number of pre-processed reads
<i>A. thaliana</i>	PRJNA899009	SRR22559142, SRR22559143, SRR22559144	20248052, 20784415, 22169416	19336553, 19864492, 21188340
<i>S. lycopersicum</i>	PRJNA1003223	SRR25558913, SRR25558914, SRR25558915	43307357, 41684242, 49831874	41602968, 40278306, 47801733
<i>C. campestris</i>	PRJNA666991	SRR12763776, SRR12763787, SRR12763788	15454773, 15497676, 15864650	14613644, 14578626, 14978212

Table 2.1. The input RNA-seq data of the three selected species. Specifically, the ENA bioproject and run accession numbers are reported along with the number of reads before and after the pre-processing step.

2.3.3 Merging of the transcriptome data

A dual RNA-seq experiment was simulated where the acquired reads included both host and parasitic plants. Specifically, the first replicate of one species was merged with the first replicate of the other one, and similarly for the other replicates. Consequently, the *C. campestris* transcriptome was merged with that of *A. thaliana* to create three merged transcriptome replicates (acc. no.: SRR22559142 + SRR12763776; SRR22559143 + SRR12763787; SRR22559144 + SRR12763788), which were used for downstream analysis. The same procedure was followed for *C. campestris* and *S. lycopersicum* (acc. no.: SRR25558913 + SRR12763776; SRR25558914 + SRR12763787; SRR25558915 + SRR12763788).

2.3.4 Dual RNA-Seq simulation via sequential and combined approach

In the sequential approach, host plant and parasitic plant genomes were individually indexed using STAR version 2.5.2b (Dobin *et al.*, 2013). The merged pre-processed reads were mapped using STAR with parameters `--outFilterMultimapNmax 10` and `--outFilterMismatchNmax 5`, to minimize multiple mapping and cross-mapping issues. The reads were aligned to the host genome, and the resulting unmapped reads were then aligned to the parasite genome. The same procedure was repeated by swapping the order of mapping. In the combined approach, host plant and parasitic plant genomes were first combined and then a single index was created. At this point, the merged pre-processed reads were mapped with STAR as previously described.

2.3.5 Evaluation metrics

Various performance metrics were computed for aligned reads. If one mate of a pair of reads mapped entirely while the other did not map at all, both mates were discarded and labelled as unmapped. As for the sequential, when considering reads initially mapped to the genome of the host plant, correct assignment of host plant reads to the host plant genome was defined as $S\text{-}TP_{h1}$ and $S\text{-}TN_{p1}$ (true positives for the host plant and true negatives for the parasitic plant). Conversely, incorrect assignment to the genome of the parasitic plant was labelled as $S\text{-}FN_{h2}$ and $S\text{-}FP_{p2}$ (false negatives for the host plant and false positives for the parasitic plant). Similarly, correct assignment of parasitic plant reads to the parasitic plant genome was defined as $S\text{-}TN_{h2}$ and $S\text{-}TP_{p2}$ (true negatives for the host plant and true positives for the parasitic plant), while incorrect assignment to the plant host genome was designated as $S\text{-}FP_{h1}$ and $S\text{-}FN_{p1}$ (false positives for the host plant and false negative for the parasitic plant) (Figure 2.2).

When considering reads initially mapped to the genome of the parasitic plant, correct assignment of parasitic plant reads to the parasitic plant genome was defined as $S\text{-}TP_{p1}$ and $S\text{-}TN_{h1}$ (true positive for the parasitic plant and true negatives for the host plant). Conversely, incorrect assignment to the genome of the host plant was labelled as $S\text{-}FN_{p2}$ and $S\text{-}FP_{h2}$ (false negatives for the parasitic plant and false positives for the host plant). Correspondingly, correct assignment of host plant reads to the genome of the host plant was defined as $S\text{-}TN_{p2}$ and $S\text{-}TP_{h2}$ (true negatives for the parasitic plant and true positives for the host plant), while incorrect assignment to the genome of the parasitic plant was designated as $S\text{-}FP_{p1}$ and $S\text{-}FN_{h1}$ (false positives for the parasitic plant and false negative for the host plant). As for the

combined approach, correct assignment of plant host reads to the genome of host plant was defined as C-TP_h and C-TN_p (true positives for the host plant and true negatives for the parasitic plant). Conversely, incorrect assignment to the genome of the parasitic plant was labelled as C-FN_h and C-FP_p (false negatives for the host plant and false positives for the parasitic plant). Correspondingly, correct assignment of parasitic plant reads to the genome was defined as C-TN_h and C-TP_p (true negatives for the host plant and true positives for the parasitic plant), while incorrect assignment to the genome of the host plant was designated as C-FP_h and C-FN_p (false positives for the host plant and false negatives for the parasitic plant). BAM files were processed using Samtools version 1.14 (Danecek *et al.*, 2021). Precision, sensitivity, specificity and accuracy metrics were calculated as follows:

$$precision = \frac{truepositive}{truepositive + falsepositive},$$

$$sensitivity = \frac{truepositive}{truepositive + falsenegative},$$

$$specificity = \frac{truenegative}{truenegative + falsepositive},$$

$$accuracy = \frac{truepositive + truenegative}{truepositive + truenegative + falsepositive + falsenegative}.$$

		First mapping step		Second mapping step	
SEQ	First Host	H > H S-TP _{h1} [S-TN _{p1}]	P > H S-FP _{h1} [S-FN _{p1}]	P > P S-TN _{h2} [S-TP _{p2}]	H > P S-FN _{h2} [S-FP _{p2}]
	First Parasite	P > P S-TN _{h1} [S-TP _{p1}]	H > P S-FN _{h1} [S-FP _{p1}]	H > H S-TP _{h2} [S-TN _{p2}]	P > H S-FP _{h2} [S-FN _{p2}]
Single Mapping step					
COM		H > H C-TP _h [C-TN _p]	P > H C-FP _h [C-FN _p]	P > P C-TN _h [C-TP _p]	H > P C-FN _h [C-FP _p]

Figure 2.2. Categorization of reads based on the employed mapping procedure. The "SEQ" row illustrates the sequential approach, detailing mappings for both host and parasitic plants, separately. In contrast, the "COM" row represents the combined approach, where a single mapping encompasses both host and parasitic plants. The mapping of reads originated from the parasite/host to their respective genomes is indicated by the symbol ">", with the assigned labels resulting from the mapping displayed below. Labels without brackets refer to the host plant, while those within brackets refer to the parasitic plant. For example, to indicate reads belonging to parasitic plant but mapping to the genome of the host plant in the first mapping step "P > H", the label "S-FP_{h1}" (Sequential - False Positive host plant first mapping step) is used for the host plant reference, and "S-FN_{p1}" (Sequential - False Negative parasitic plant first mapping step) is used for the parasitic plant reference.

2.3.6 Investigating one-side cross-mapped reads in the sequential and combined approaches

Exploration about one-side cross-mapped reads was performed considering reads labelled as S-FN_{h2} for host and S-FN_{p2} for parasite in sequential approach. Regarding combined approach, reads extractable through the complement C-FN_h (total cross-mapped reads) respect to C-TP_h (host reads correct assigned) were considered for host, while reads extractable through the complement C-FN_p (total cross-mapped reads) respect to C-TP_p (host reads correct assigned) were considered for parasite. Resulting reads were remapped to their respective genomes using STAR with less stringent parameters. Specifically, the parameters --outFilterMultimapNmax 10 and --outFilterMismatchNmax 10 were set. The statistical results were averaged between replicates. Summarization of these reads was achieved using the htseq-count function embedded in STAR version 2.5.2b and relative genome GFF3 annotation files.

2.3.7 Investigating two-side cross-mapped reads in the sequential and combined approaches

To investigate host loci where two-side cross-mapped reads align, S-TP_{h1} and S-FN_{h1} were intersected in sequential approach. Similarly, S-TP_{p1} and S-FN_{p1} were intersected to identify two-side cross-mapped reads for parasitic loci (Figure 2.3).

These reads were summarized using the htseq-count function embedded in STAR version 2.5.2b and relative genome GFF3 annotation files. From these, the relative functional description attributes (the “product” tag) were extracted.

In the combined approach, two-side cross-mapped reads were identified by intersecting C-TP_h with C-FN_h for host loci, and C-TP_p and C-FN_p for parasite loci. Next, the functional description attributes (the “product” tag) of each identified transcript were then extracted from the corresponding GFF3 annotation files.

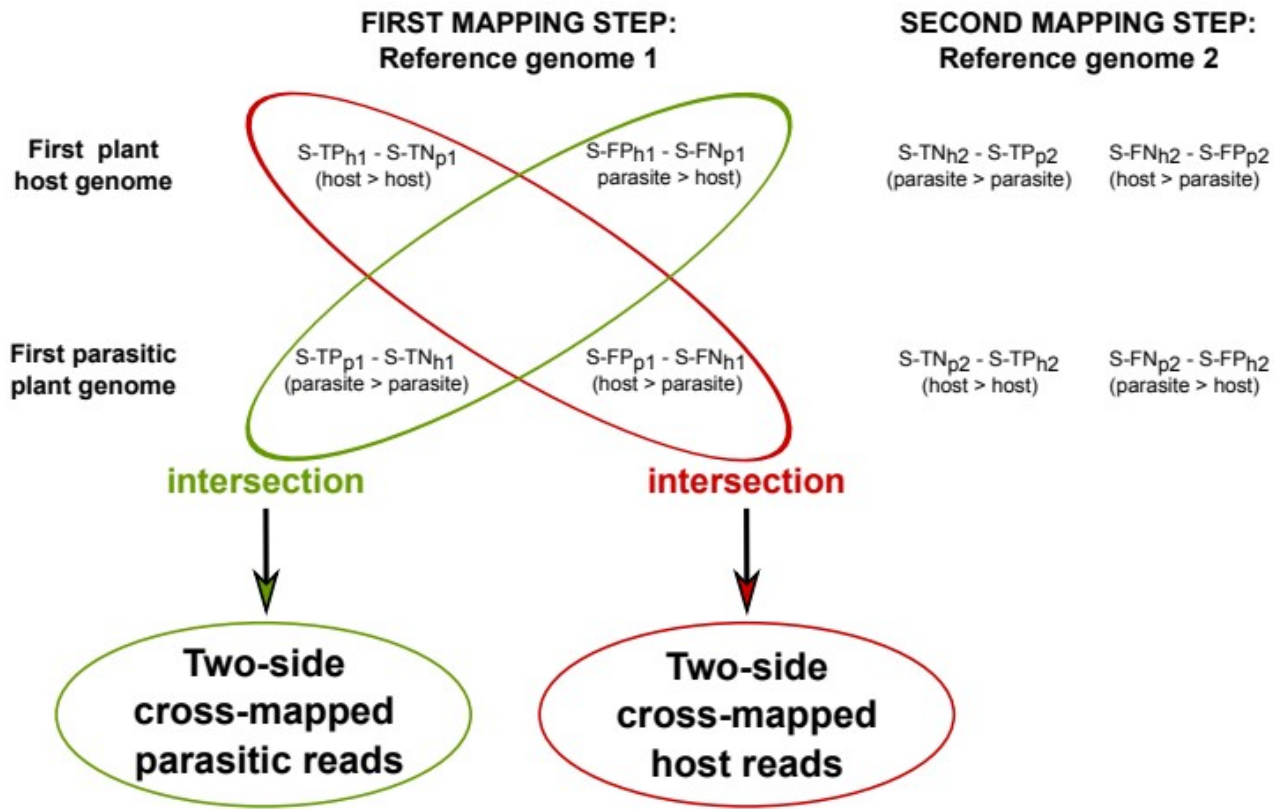


Figure 2.3. Identification of host-parasite two-side cross-mapped reads in the sequential approach. Two-side cross-mapped reads were identified through intersection, solely based on the outcome of the first mapping steps in the sequential approach. In particular, reads labeled as $S\text{-}TP_{h1}$ (also labeled $S\text{-}TN_{p1}$) and $S\text{-}FN_{h1}$ (also labeled $S\text{-}FP_{p1}$) were intersected to identify reads originated from host. Similarly, reads labeled as $S\text{-}TN_{h1}$ (also labeled $S\text{-}TP_{p1}$) and $S\text{-}FP_{h1}$ (also labeled $S\text{-}FN_{p1}$) were intersected to identify reads originated from parasite. In these labels, the symbol “>” denotes the mapping of reads originating from the parasite/host to their respective genomes.

2.4 Results

2.4.1 Phylogenetic analysis of *Cuscuta* spp. and their host range

A phylogenetic analysis was performed based on the *rbcL* protein sequences from six *Cuscuta* species and 28 host species. In the resulting tree (Figure 2.4), a large clade was identified supported by a high bootstrap value (0.93) that comprises the parasitic plant *C. campestris*, other *Cuscuta* species and six host species. Five of these species belonged to the *Solanales* order, including the crop *S. lycopersicum*, along with one species from the *Convolvulus* order. Other host species of interest, such as *A. thaliana* from the *Brassicales* order, were clearly phylogenetically distant from the clade that includes *Cuscuta* species. Based on this, *S. lycopersicum* was considered as host species phylogenetically closer to *C. campestris*, whereas *A. thaliana* represents a host with a more divergent evolutionary relationship.

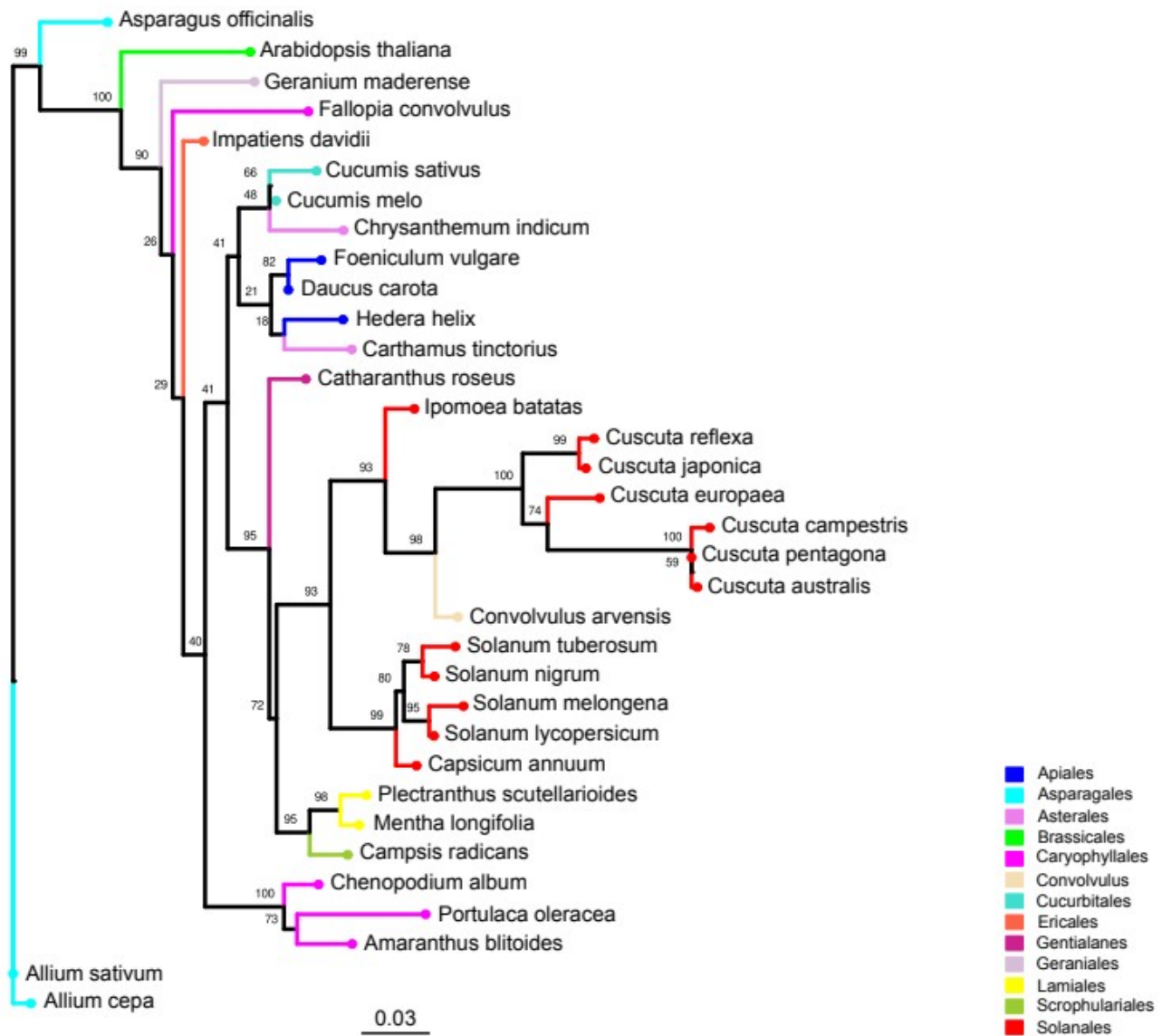


Figure 2.4. Phylogenetic tree of *Cuscuta* spp. and their host range. The phylogenetic tree was constructed using *rbcl* protein sequences derived from six *Cuscuta* species and with 28 host species spanning 12 orders. Leaves representing the same order are depicted with consistent colors throughout the tree. Bootstrap values are annotated above corresponding branches.

2.4.2 Dual RNA-seq simulation via sequential and combined approaches

After quality filtering and library merging, the three resulting merged libraries involving *A. thaliana* and *C. campestris* included an average of ~34.85 M reads, whereas the three resulting merged libraries involving *S. lycopersicum* and *C. campestris* included an average of ~57.95 M reads.

2.4.2.1 Sequential approach

2.4.2.1.1 Mapping of *A. thaliana* and *C. campestris* reads

Mapping first to host and then to parasite genome, approximately 19.81 M reads were assigned to *A. thaliana* and 14.22 M reads to *C. campestris*. On average, uniquely mapped reads assigned to host were ~18.99 M (~54.49%) while ~13.06 M (~37.46%) were assigned to parasite.

Multiple mapped reads assigned to host were ~0.82 M (~2.35%) and ~1.17 M (~3.35%) were assigned to parasite (Table 2.2.1, Figure 2.5).

Among the mapped reads, the total cross-mapped reads (both one-side and two-side), originating from *A. thaliana* but assigned to *C. campestris* were ~0.004 M (~0.002%). Conversely, reads originating from *C. campestris* and assigned to *A. thaliana* were ~0.13 M (~0.64%) (Supplementary Table 2.2, Supplementary Table 2.3). Notably, the evaluation metrics (i.e., precision, sensitivity, accuracy, and specificity) were all close to one (Table 2.2.2).

The analysis was repeated by swapping the order of mapping; in this scenario, reads were initially mapped to the *C. campestris* genome and then to the *A. thaliana* genome.

Approximately 19.24 M reads were assigned to *A. thaliana* and 14.79 M reads to *C. campestris*. On average, the uniquely mapped reads assigned to the host totaled ~18.78 M (~53.87%), while ~13.32 M (~38.22%) were assigned to the parasite. The multiple mapped reads assigned to the host accounted for ~0.47 M (~1.34%), whereas those assigned to parasite were ~1.47 M (~4.22%) (Table 2.2.1, Figure 2.5). Among mapped reads, the total cross-mapped (both one-side and two-side), originating from *A. thaliana* and assigned to *C. campestris* amounted to ~0.49 M (~3.33%). Conversely, reads originating from *C. campestris* and assigned to *A. thaliana* were ~0.05 M (~0.28%) (Supplementary Table 2.2, Supplementary Table 2.3). Also in this case, the evaluation metrics were close to one (Table 2.2.2).

Libraries (<i>A. thaliana</i> + <i>C. campestris</i>)		Order mapping	Processed reads	Uniquely mapped to host genome	Multiple mapped to host genome	Uniquely mapped to parasite genome	Multiple mapped to parasite genome	Unmapped
SRR22559142 + SRR12763776	Replicate 1	Host; parasite	33950197	18603604 (55%)	410240 (1%)	12948080 (38%)	1165669 (3%)	822604 (2%)
SRR22559143 + SRR12763787	Replicate 2	Host; parasite	34443118	18102724 (53%)	1618200 (5%)	12953498 (38%)	1126992 (3%)	641704 (2%)
SRR22559144 + SRR12763788	Replicate 3	Host; parasite	36166552	20268051 (56%)	426377 (1%)	13270916 (37%)	1208650 (3%)	992558 (3%)
SRR22559142 + SRR12763776	Replicate 1	Parasite; host	33950197	18564026 (55%)	327509 (1%)	12992108 (38%)	1243959 (4%)	822595 (2%)
SRR22559143 + SRR12763787	Replicate 2	Parasite; host	34443118	17553088 (51%)	752222 (2%)	13639241 (40%)	1857011 (5%)	641556 (2%)
SRR22559144 + SRR12763788	Replicate 3	Parasite; host	36166552	20210535 (56%)	324911 (1%)	13331684 (37%)	1306968 (4%)	992454 (3%)

Table 2.2.1. Statistics on the number of mapped reads by considering a sequential approach (*A. thaliana*-*C. campestris*). Reads were first mapped to the *A. thaliana* (host) genome and then to the *C. campestris* (parasite) genome (Order mapping: “Host; parasite”) or vice versa.

Libraries												
(<i>A. thaliana</i> + <i>C. campestris</i>)	Replicate	Order mapping	Processed reads	Mapped	Precision (host)	Sensitivity (host)	Accuracy (host)	Specificity (host)	Precision (parasite)	Sensitivity (parasite)	Accuracy (parasite)	Specificity (parasite)
SRR22559142 + SRR12763776	Replicate 1	Host; parasite	33950197	33127593 (97.58%)	0.9940452	0.9999979	0.9965810	0.9920416	0.9999972	0.9920416	0.9965810	0.9999979
SRR22559143 + SRR12763787	Replicate 2	Host; parasite	34443118	33801414 (98.14%)	0.9929953	0.9999623	0.9958913	0.9902841	0.9999475	0.9902841	0.9958913	0.9999623
SRR22559144 + SRR12763788	Replicate 3	Host; parasite	36166552	35173994 (97.26%)	0.9936527	0.9999860	0.9962574	0.9910098	0.9999801	0.9910098	0.9962574	0.9999860
SRR22559142 + SRR12763776	Replicate 1	Parasite; host	33950197	33127602 (97.58%)	0.9974339	0.9969504	0.9967967	0.9965926	0.9959512	0.9965926	0.9967967	0.9969504
SRR22559143 + SRR12763787	Replicate 2	Parasite; host	34443118	33801562 (98.14%)	0.9968245	0.9317532	0.9587400	0.9959116	0.9137517	0.9959116	0.9587400	0.9317532
SRR22559144 + SRR12763788	Replicate 3	Parasite; host	36166552	35174098 (97.26%)	0.9972441	0.9958850	0.9959853	0.9961265	0.9942195	0.9961265	0.9959853	0.9958850

Table 2.2.2. Sequential approach evaluation metrics (*A. thaliana*-*C. campestris*). Precision, sensitivity, accuracy, and specificity were used to assess read mapping based on the sequential approach applied to *A. thaliana* (host) and *C. campestris* (parasite) interaction. Each metric was calculated and reported separately for both the host and the parasite.

2.4.2.1.2 Mapping of *S. lycopersicum* and *C. campestris* reads

Mapping first to host and then to parasite genome, approximately 37.52 M of reads were assigned to *S. lycopersicum* and 14.38 M reads to *C. campestris*. In the scenario where reads were initially mapped to the *S. lycopersicum* genome, on average, the uniquely mapped reads assigned to the host were ~36.97 M (~63.79%), while ~13.1 M (~89.0%) were assigned to the parasite. The multiple mapped reads assigned to the host accounted for ~0.6 M (~1.3%), while those assigned to the parasite were ~1.3 M (~8.7%) were (Table 2.3.1, Figure 2.5). Among mapped reads, the total cross-mapped (both one-side and two-side), originating from *S. lycopersicum* and assigned to *C. campestris*, were ~0.13 M reads (~0.92%). Conversely, reads originating from *C. campestris* and assigned to *S. lycopersicum* were ~0.09 M reads (~0.25%) (Supplementary Table 2.2, Supplementary Table 2.4). Ultimately, all statistical metrics approached unity (Table 2.3.2).

Reversing the order of use of the reference genomes, approximately 37.33 M reads were assigned to *S. lycopersicum*, while 14.58 M reads to *C. campestris*. Among these, the uniquely mapped reads assigned to the host totaled ~36.8 M (~85.2%), whereas those assigned to the parasite were ~13.2 M (~89.4%). The multiple mapped reads assigned to the host accounted for ~0.5 M (~1.1%), while those assigned to the parasite were ~1.4 M (~9.6%) (Table 2.3.1, Figure 2.5).

The total cross-mapped reads (both one-side and two-side), originating from *S. lycopersicum* and assigned to *C. campestris* amounted to ~0.28 M (~1.93%). Conversely, reads originating from *C. campestris* and assigned to *S. lycopersicum* were ~0.05 M (~0.13%) (Supplementary Table 2.2, Supplementary Table 4.2). Consistent with previous mapping results, all statistical metrics approached unity (Table 2.3.2).

Libraries		Order mapping	Processed reads	Uniquely mapped to the host genome	Multiple mapped to the host genome	Uniquely mapped to the parasite genome	Multiple mapped to the parasite genome	Unmapped
(<i>S. lycopersicum</i> + <i>C. campestris</i>)	Replicate							
SRR25558913 + SRR12763776	Replicate 1	Host; parasite	56216612	35405921 (63%)	495390 (1%)	13001831 (23%)	1292197 (2%)	6021273 (11%)
SRR25558914 + SRR12763787	Replicate 2	Host; parasite	54856932	35128009 (64%)	538023 (1%)	12993036 (24%)	1226568 (2%)	4971296 (9%)
SRR25558915 + SRR12763788	Replicate 3	Host; parasite	62779945	40372884 (64%)	625278 (1%)	13314939 (21%)	1321988 (2%)	7144856 (11%)
SRR25558913 + SRR12763776	Replicate 1	Parasite; host	56216612	35253494 (63%)	441975 (1%)	13066350 (23%)	1433890 (3%)	6020903 (11%)
SRR25558914 + SRR12763787	Replicate 2	Parasite; host	54856932	35008504 (64%)	485523 (1%)	13052054 (24%)	1339693 (2%)	4971158 (9%)
SRR25558915 + SRR12763788	Replicate 3	Parasite; host	62779945	40238892 (64%)	562252 (1%)	13378826 (21%)	1454850 (2%)	7145125 (11%)

Table 2.3.1. Statistics on the number of mapped reads by considering a sequential approach (*S. lycopersicum*-*C. campestris*). Reads were first mapped to the *S. lycopersicum* (host) genome and then to the *C. campestris* genome (Order mapping: “Host; parasite) or vice versa.

Libraries (<i>S.lycopersicum</i> + <i>C. campestris</i>)												
	Replicate	Order mapping	Processed reads	Mapped	Precision (host)	Sensitivity (host)	Accuracy (host)	Specificity (host)	Precision (parasite)	Sensitivity (parasite)	Accuracy (parasite)	Specificity (parasite)
SRR25558913 + SRR12763776	Replicate 1	Host; parasite	56216612	50195339 (89.29%)	0.9976894	0.9956543	0.9952329	0.9941667	0.9976894	0.9956543	0.9952329	0.9941667
SRR25558914 + SRR12763787	Replicate 2	Host; parasite	54856932	49885636 (90.94%)	0.9971581	0.9969043	0.9957543	0.9928673	0.9971581	0.9969043	0.9957543	0.9928673
SRR25558915 + SRR12763788	Replicate 3	Host; parasite	62779945	55635089 (88.62%)	0.9976331	0.9968478	0.9959311	0.9933556	0.9976331	0.9968478	0.9959311	0.9933556
SRR25558913 + SRR12763776	Replicate 1	Parasite; host	56216612	50195709 (89.29%)	0.9988087	0.9910511	0.9927392	0.9970097	0.9988087	0.9910511	0.9927392	0.9970097
SRR25558914 + SRR12763787	Replicate 2	Parasite; host	54856932	49885774 (90.94%)	0.9985610	0.9934882	0.9943193	0.9964058	0.9985610	0.9934882	0.9943193	0.9964058
SRR25558915 + SRR12763788	Replicate 3	Parasite; host	62779945	55634820 (88.62%)	0.9987852	0.9931971	0.9940920	0.9966060	0.9987852	0.9931971	0.9940920	0.9966060

Table 2.3.2. Sequential approach evaluation metrics (*S. lycopersicum*-*C. campestris*). Precision, sensitivity, accuracy, specificity were used to assess read mapping based on the sequential approach applied to *S. lycopersicum* (host) and *C. campestris* (parasite) interaction. Each metric was calculated and reported separately for both the host and the parasite.

2.4.2.2 Combined approach

2.4.2.2.1 Mapping of *A. thaliana* and *C. campestris* reads

When aligning reads to the combined genome of *A. thaliana* and *C. campestris*, approximately 19.68 M reads were assigned to *A. thaliana*, while 14.30 M reads to *C. campestris*. Among these, the uniquely mapped reads assigned to host totaled ~18.95 M (~54.37%) while for the parasite they were ~13.08 M (~37.53%). The multiple mapped reads assigned to host accounted for ~0.73 M (~2.10%), whereas those assigned to the parasite were ~1.22M (~3.49%) (Table 2.4.1, Figure 2.5). The total cross-mapped reads (both one-side and two-side), originating from *A. thaliana* and assigned to *C. campestris* amounted to ~0.01 M (~0.03%). Conversely, reads originating from *C. campestris* and assigned to *A. thaliana* were ~0.01 M (~0.04%) (Supplementary Table 2.5, Supplementary Table 2.6). Ultimately, all evaluation metrics were close to unity (Table 2.4.2).

Interaction	Libraries (host + parasite)	Replicate	Processed reads	Uniquely mapped to the host genome	Multiple mapped to the host genome	Uniquely mapped to the parasite genome	Multiple mapped to the parasite genome	Two-side cross-mapped	Unmapped
<i>Arabidopsis-Cuscuta</i>	SRR22559142 + SRR12763776	Replicate 1	33950197	18572122 (54.7%)	330452 (0.97%)	12965830 (38.19%)	1211131 (3.57%)	1709 (0.01%)	868953 (2.56%)
<i>Arabidopsis-Cuscuta</i>	SRR22559143 + SRR12763787	Replicate 2	34443118	18046908 (52.4%)	1526290 (4.43%)	12979800 (37.68%)	1178915 (3.42%)	13583 (0.04%)	697622 (2.03%)
<i>Arabidopsis-Cuscuta</i>	SRR22559144 + SRR12763788	Replicate 3	36166552	20227851 (55.93%)	337908 (0.93%)	13292023 (36.75%)	1260436 (3.49%)	2831 (0.01%)	1045503 (2.89%)
<i>Solanum-Cuscuta</i>	SRR25558913 + SRR12763776	Replicate 1	56216612	35261137 (62.72%)	452535 (0.8%)	13019757 (23.16%)	1396199 (2.48%)	8192 (0.01%)	6078792 (10.81%)
<i>Solanum-Cuscuta</i>	SRR25558914 + SRR12763787	Replicate 2	54856932	35013923 (63.83%)	490718 (0.89%)	13016735 (23.73%)	1307742 (2.38%)	7000 (0.01%)	5020814 (9.15%)
<i>Solanum-Cuscuta</i>	SRR25558915 + SRR12763788	Replicate 3	62779945	40262060 (64.13%)	575318 (0.92%)	13335341 (21.24%)	1415630 (2.25%)	8307 (0.01%)	7183289 (11.44%)

Table 2.4.1. Statistics on the number of mapped reads by considering combined approach. Reads were mapped to the host plant-parasitic plant combined genome. The reported two-side cross-mapped reads represent the cumulative count between the two organisms.

Interaction	Libraries	Replicate	Processed		Precision	Sensitivit	Accuracy	Specificity	Precision	Sensitivity	Accuracy	Specificity	
	(host + parasite)		reads	Mapped	(host)	y (host)	(host)	(host)	(parasite)	(parasite)	(parasite)	(parasite)	
Arabidopsis-Cuscuta	SRR22559142 + SRR12763776	Replicate 1	33950197	33079535 (97.44%)	0.9998379	0.9999910	0.9999932	2	0.9997838	0.9999035	0.9999035	0.9997838	0.9999932
Arabidopsis-Cuscuta	SRR22559143 + SRR12763787	Replicate 2	34443118	33731913 (97.94%)	0.9997350	0.9998929	0.9999222	5	0.9996337	0.9998013	0.9998013	0.9996337	0.9999225
Arabidopsis-Cuscuta	SRR22559144 + SRR12763788	Replicate 3	36166552	35118218 (97.1%)	0.9997497	0.9999470	0.999962	5	0.9996464	0.9998315	0.9998315	0.9996464	0.9999625
Solanum-Cuscuta	SRR25558913 + SRR12763776	Replicate 1	56216612	50129628 (89.17%)	0.9999063	0.9832704	0.993291	7	0.9997641	0.9951223	0.9951223	0.9997641	0.9932917
Solanum-Cuscuta	SRR25558914 + SRR12763787	Replicate 2	54856932	49829118 (90.83%)	0.9998413	0.9881445	0.995238	9	0.9996021	0.9964788	0.9964788	0.9996021	0.9952389
Solanum-Cuscuta	SRR25558915 + SRR12763788	Replicate 3	62779945	55588349 (88.54%)	0.9998616	0.9863561	0.995095	1	0.9996118	0.9962778	0.9962778	0.9996118	0.9950951

Table 2.4.2. Combined approach evaluation metrics. Precision, sensitivity, accuracy, specificity were used to assess read mapping based on the combined approach. Each metric was calculated and reported separately for both the host and the parasite.

2.4.2.2.2 Mapping of *S. lycopersicum* and *C. campestris* reads

When aligning reads to the combined genome of *S. lycopersicum* and *C. campestris*, approximately 37.35 M of reads were assigned to *S. lycopersicum*, while 14.50 M reads to *C. campestris*. Among these, the uniquely mapped reads assigned to host accounted for ~36.85 M (~63.58%) while for the parasite they were ~13.12 M (~22.65%). The multiple mapped reads assigned to the host totaled ~0.51 M (~0.87%), whereas those assigned to the parasite were ~1.37 M (~2.37%) (Table 2.4.1, Figure 2.5). The total cross-mapped (both one-side and two-side), originating from *S. lycopersicum* and assigned to *C. campestris* amounted to ~0.21 M (~0.56%). Conversely, reads originating from *C. campestris* and assigned to *S. lycopersicum* were ~0.01 M (~0.05%) (Supplementary Table 2.5, Supplementary Table 2.6). Finally, all evaluation metrics were close to unity (Table 2.4.2).

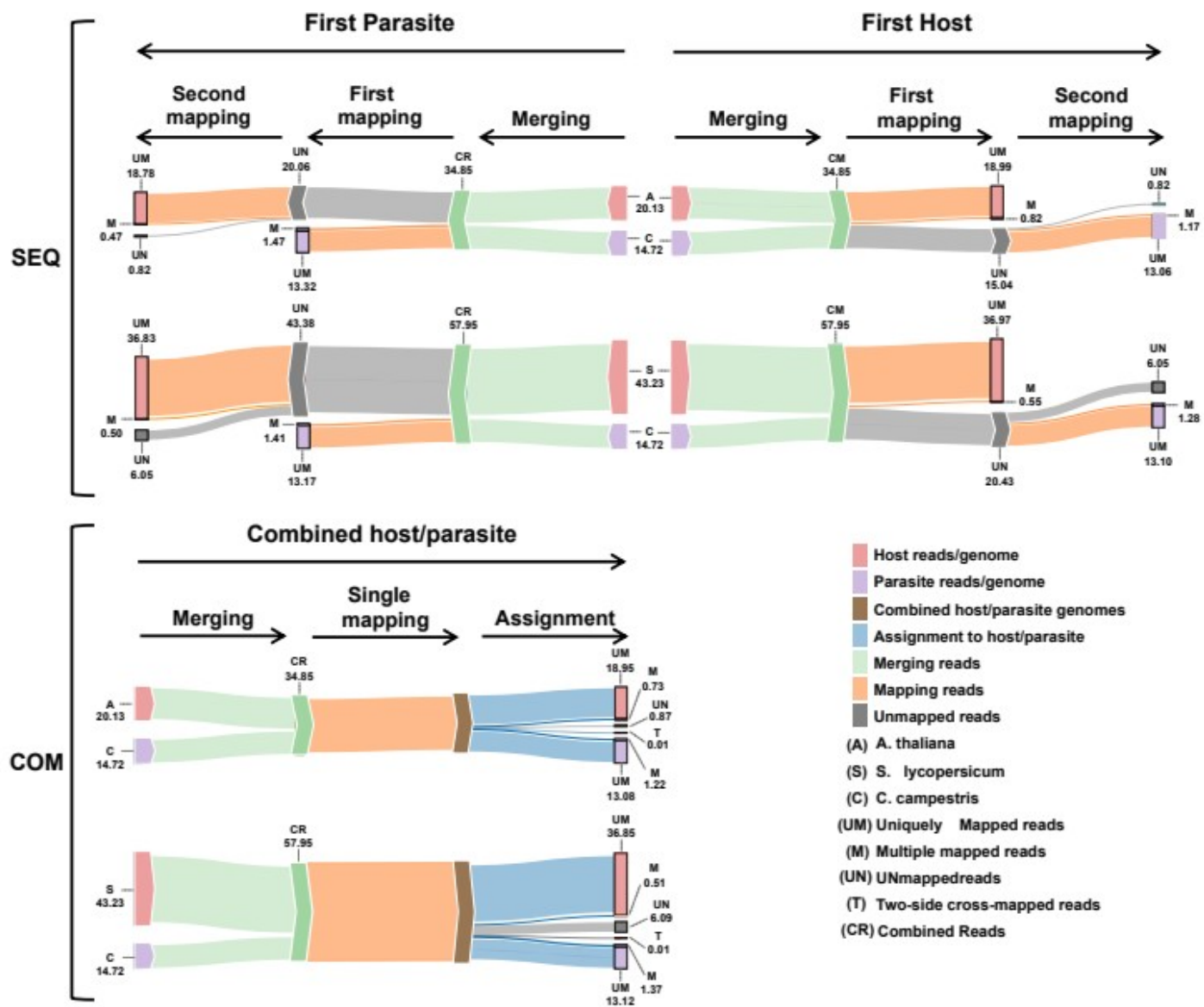


Figure 2.5. Workflow summary results. Sankey plot illustrating the workflow of both sequential (SEQ, top) and combined (COM, bottom) approaches used in the dual RNA-seq study. Each step displays the average number of reads (in millions). The legend at the bottom right explains the colors and abbreviations used in the plot.

2.4.3. Investigating one-side cross-mapped reads

On average, in the sequential approach involving *A. thaliana* and *C. campestris*, the number of reads labelled as S-FN_{h2} were not significant, while S-FN_{p2} accounted for ~0.05 M reads; of these only ~1.83% successfully remapped to the *C. campestris* genome. Similarly, when considering *S. lycopersicum* and *C. campestris*, the number of reads labelled as S-FN_{h2} were ~0.13 M, while S-FN_{p2} accounted for ~0.05 M; of these only ~59.82% and ~1.92% successfully remapped to the *S. lycopersicum* and *C. campestris* genomes, respectively (Supplementary Table 2.7).

In the combined approach involving *A. thaliana* and *C. campestris*, the number of reads labelled as C-FN_h and C-FN_p were both not significant. For *S. lycopersicum* and *C. campestris*, the number of reads labelled as C-FN_h were ~0.20 M, while the number of reads labelled as C-FN_p were not significant. The remapping rate of C-FN_h reads to *S. lycopersicum* were ~76.53%. Moreover, reads did not map to annotated loci (Supplementary Table 2.8).

2.4.4 Investigating two-side cross-mapped reads

2.4.4.1 Sequential approach

2.4.4.1.1 *A. thaliana* and *C. campestris*

The intersection between *A. thaliana* S-TP_{h1} reads and *C. campestris* S-FN_{h1} reads yielded an average of 0.49 M reads. Similarly, the intersection between *C. campestris* S-TP_{p1} reads and *A. thaliana* S-FN_{p1} reads resulted in an average of 0.07 M reads.

The resulting loci span across all five *A. thaliana* chromosomes (including organelle genomes) were reported in Supplementary Table 2.9. None of the loci were annotated in *C. campestris*.

2.4.4.1.2 *S. lycopersicum* and *C. campestris*

The intersection between *S. lycopersicum* S-TP_{h1} reads and *C. campestris* S-FN_{h1} reads resulted in an average of 0.17 M reads. Similarly, the intersection between *C. campestris* S-TP_{p1} reads and *S. lycopersicum* S-FN_{p1} reads yielded an average of 0.04 M reads. The resulting loci span across all twelve *S. lycopersicum* chromosomes (including organelle genomes) were reported in Supplementary Table 2.10. None of the loci were annotated in *C. campestris*.

2.4.4.2 Combined approach

2.4.4.2.1 *A. thaliana* vs *C. campestris*

On average of 0.005 M reads from *A. thaliana* and 0.001 M reads from *C. campestris* were tagged as two-side cross-mapped reads, by intersecting C-TP_h with C-FN_h for the host and C-TP_p with C-FN_p for the parasite. Annotated loci resulting from this process were identified on chromosome 2, chromosome 3, and the mitochondrial and plastidial genomes of *A. thaliana* (Supplementary Table 2.11). None of the loci were annotated in *C. campestris*.

2.4.4.2.2 *S. lycopersicum* vs *C. campestris* interaction

On average 0.01 M reads from *S. lycopersicum* and 0.002 M reads from *C. campestris* were tagged as two-side cross-mapped reads, by intersecting C-TP_h with C-FN_h for the host and C-

TP_p with C-FN_p for the parasite. Annotated loci resulting from this process were identified on chromosomes 3 and 11, and the mitochondrial genome of *S. lycopersicum* (Supplementary Table 2.12). None of the loci were annotated in *C. campestris*.

2.5 Discussion

Transcriptome investigations on “host plant-parasitic plant” interaction have typically involved separate analysis of the two organisms. For this purpose, the use of LCM to isolate the host from the parasite has represented the standard in the field (Jhu and Sinha, 2022). While valuable, LCM is costly, time-consuming, and requires specialized expertise limiting its accessibility. For this reason, *in silico* separation of transcripts via dual-RNA-seq offers a promising alternative.

Dual RNA-seq allows for the identification of core genes involved in the interaction by sampling and analyzing infected tissues. Examining gene expression changes in the parasite alongside the host response, uncovers critical mechanisms in the host-parasite interplay. This method is economical and practical since the tissues do not need to be physically separated. It requires that reads must assigned through read mapping onto their respective reference genomes. If one species lacks an assembled genome, its assembled transcriptome can be used. Anyway, the choice of genome as reference, instead of assembled transcriptome, is required in order to avoid the potential presence of transferred transcript. Two approaches are used: the sequential method, where reads are mapped sequentially to both species reference genomes, and the combined method, where reads are mapped to a single concatenated genome. The sequential method can be used in the presence of almost one reference genome (parasite or host), while the combined approach is exclusively employed when both host and parasitic reference genomes are accessible. Dual RNA-seq complexity arises from handling sequences from two organisms simultaneously, with a key challenge being cross-mapping reads; these latter could represent misleading or loss information. Here, cross-mapping events were classified into two main categories, namely “one-side cross-mapping” and “two-side cross-mapping”. The first indicates reads from one organism but exclusively assigned to the other, while the second refers to reads ambiguously assigned to both organisms.

Although distinguishing reads originating from two eukaryotes poses additional challenges, dual RNA-seq has been successfully applied to study plant-pathogen interactions across various plant species and eukaryotic pathogens and parasites, including fungi, oomycetes, and nematodes (Petitot *et al.*, 2016; Naidoo *et al.*, 2018). Despite demonstrated utility, performing this analysis between two plants is still relatively uncharted territory, which may seem daunting at first glance. In fact, the key to separating reads from bacterial and eukaryotic cells lies in their divergence and distinct content of their RNA molecules (Marsh *et al.*, 2017; Westermann, Barquist and Vogel, 2017). A previous study by Ikeue *et al.* (2015) attempted to separate *in silico* reads from two distinct plant species, *Impatiens balsamina* and *Cuscuta japonica*. They exclusively used sequential approach without prior knowledge of sequence origin.

This chapter aims to evaluate dual RNA-seq feasibility between host and parasite within the same taxonomic kingdom. Using assembled genomes as references, both sequential and combined approaches were employed. Artificial datasets were generated to replicate interaction of two host-parasite systems: *A. thaliana*-*C. campestris* and *S. lycopersicum*-*C. campestris*. *A. thaliana*, a *Brassicaceae* family member, and *S. lycopersicum*, a *Solanaceae* family member, were chosen to investigate whether a host plant, phylogenetically further from *C. campestris* (member of the *Convolvulaceae* family), could improve read assignment accuracy due to sequence divergence. A phylogenetic analysis using the sequences of the *rbcL* protein (Manhart, 1994), commonly used in plant phylogenetics, determined the evolutionary

distance between host and parasite. This analysis serves as preliminary step in comparing the extent of cross-mapping in two host plants with varying evolutionary distances from the parasitic plant. As anticipated, *S. lycopersicum* is phylogenetically closer to *C. campestris* than *A. thaliana*. These findings align with previous studies supporting the monophyletic nature of *Convolvulaceae* family as sister group of *Solanaceae* family, placing *Convolvulaceae* and *Brassicaceae* into distinct clades (Soltis *et al.*, 2011).

Accurately assigning reads to their respective genomes could be challenging, especially when interacting organisms belong to the same taxonomic kingdom. This challenge stems from the presence of homologous sequences, which are often highly similar. To verify this hypothesis, the RNA-seq libraries were chosen to utilize the maximum read length available in ENA repository, in order to enhance the accuracy of sequence assignment to their respective reference genomes. The libraries come from tissues that could be attacked during infection process, namely the stem for the two host plants and the dodder haustoria, closely mimicking genuine parasitic conditions. Furthermore, to prevent read contaminations between host into the parasitic and vice versa, due to RNA transfer phenomenon during the infection, we selected samples collected when they did not interact, ensuring confident attribution of reads to their sources. Host and parasitic plant transcriptomes were combined to simulate two distinct interactions, allowing accurate assessment of mapped reads, multiple mapped reads, cross-mapped reads, and computation of evaluation metrics: precision, sensitivity, specificity and accuracy. Both interactions were thoroughly analyzed through sequential and combined approaches using reference genome of interested species. Typically, RNA-seq analysis yields mapping percentages ranging 70%-90% (Conesa *et al.*, 2016). In this chapter, alignment of simulated mixed RNA-seq resulted in high mapping rates (around 90% in both approaches), demonstrating the method's effectiveness. Our findings suggest that, in sequential approach, when merged reads were firstly mapped to the host genome, this latter tends to retain a little percentage of reads (cross-mapping near 1%) belonging to the parasite and vice versa; however, in the combined approach, this trend was less pronounced (cross-mapping less than 0.2%). This variation can be attributed to the alignment step being a single operation in the combined approach. Thus, the mapping tool selects the genome to which each read aligns best, leveraging homologous sequence polymorphisms between the two species (Wolf *et al.*, 2018; Espindula *et al.*, 2019; O'Keeffe and Jones, 2019). The single operation of the combined approach is less time-consuming. Specifically, a halving of the processing time was observed in the combined approach compared to the sequential approach.

Moreover, combined approach offers a further advantage. It is possible to isolate the two-side cross-mapped reads after a single mapping step. Differently, the sequential approach requires a more intricate strategy, requiring up to three mapping operations and reordering the reference genomes. The sequential and combined approach exhibit minimal differences in the number of multiple mapped reads. Despite these differences, all evaluation metrics indicate near parity, affirming the equal reliability of both methods. No significant differences were observed in cross-mapping percentages between reads of *S. lycopersicum* (phylogenetically closer to *C. campestris* than *A. thaliana*) and *A. thaliana* when combined with those of *C. campestris* thanks to the stringent parameters adopted. By utilizing these parameters, we effectively addressed the struggle highlighted by O'Keeffe *et al.*, (2019), namely the direct relationship between read discrimination among host-pathogen species and their taxonomic divergence. Another intriguing observation, from both the sequential and combined approach pertains to the failure of a few reads labelled S-FN_{h2}, S-FN_{p2}, C-FN_h and C-FN_p (one-side cross-mapped reads) to align with their respective genomes. This could stem from either incomplete genome assemblies or excessively stringent mapping criteria. Our results indicate that relaxing these criteria would result in reallocating over 50% of previously unaligned reads to the respective genome. Notably, during the remapping process for *C. campestris* a significantly lower percentage of reads were reallocated, suggesting that genome

incompleteness may be the primary factor in this instance. Loci and the relative annotation were examined, to evaluate the extent of misleading or loss information, when similar reads map equally well in both genomes (two-side cross-mapping). In the sequential approach, a few hundred thousand two-side cross-mapped reads were identified, whereas in the combined approach, we observed a reduction of over 90%. These loci were annotated only for the hosts due to potential incompleteness in *C. campestris* annotation. They span all chromosomes, including the organelle genomes, and are involved in various cellular processes, such as protein synthesis, respiration, and some are annotated as non-coding RNA. All genes are part of the plant basal metabolism, suggesting they are unlikely to be crucial in the interaction. Moreover, given their low abundance, this loss of information may not be significant. This chapter underscores the reliability of dual RNA-seq as an effective choice, particularly when the host and parasite belong to the same kingdom. This insight paves the way to new experiments expanding our understanding of the key genes involved in plant parasitism.

2.6 Conclusions

Dual RNA-seq can be applied to “host plant – parasitic plant” interaction, providing a reliable in silico separation of mixed reads. Cross-mapped reads, due to homologous sequences between the two genomes, did not represent a significant misleading information. This chapter establishes that sequential and combined approach are both equally trustworthy. However, combined approach results to be less time-consuming and slightly mitigates the cross-mapping phenomenon.

2.7 Data availability

All the datasets we used in our analysis are publicly available, while the code used for pre-processing and mapping of reads are available in Zenodo open repository at <https://doi.org/10.5281/zenodo.11488003>.

Chapter 3. Introducing inDAGO: a graphical user interface for effortless dual RNA-seq.

3.1 Abstract

Transcriptomics studies of host-parasite interactions present a significant challenge, particularly when the contact region between organisms is difficult to manage. Dual RNA-seq offers a powerful solution to this issue, by enabling the *in silico* separation of mixed transcripts, providing a comprehensive view of transcriptome reprogramming in both host and parasite during infection. However, most current dual RNA-seq pipelines are not easily accessible to researchers without a background in programming. In this chapter, I introduce inDAGO, an open-source, cross-platform GUI designed specifically for biologists to perform dual RNA-seq analysis without the need for programming skills. inDAGO supports two operational modes: sequential and combined analysis. Additionally, it includes a module for performing classical bulk RNA-seq analysis. To assess the reliability of the interface, I tested inDAGO using data from *A. thaliana*-*C. campestris* interaction, as described in Chapter 2. From the pre-processing to the mapping stage, the analysis results were consistent with those obtained in Chapter 2 using both the sequential and combined approaches. Subsequent steps, including data summarization, exploration and differential gene expression analysis, also yielded reliable results, comparable to those produced by command-line-based R packages. Ultimately, inDAGO is a friendly interface able to automate the analysis process and provide a step-by-step guidance for end users, allowing users to easily navigate between different analysis steps and download intermediate results and publication ready plots. inDAGO is freely available on <https://github.com/CarmineFruggiero/inDAGO>.

3.2 Introduction

To date, the dual RNA-seq approach has been used to study a wide range of interspecies interactions, from mutualism (Mateus *et al.*, 2019) to host-pathogen dynamics (Westermann, Barquist and Vogel, 2017; Naidoo *et al.*, 2018)), – covering almost all taxonomic kingdoms, including bacteria, fungi, plants and animals. There are essentially two distinct methodologies for conducting the analysis: the sequential and combined approaches. Both share the core steps of traditional bulk RNA-seq analysis (quality control, pre-processing, reference genome indexing, mapping, summarization, exploratory data analysis and identification of differentially expressed genes), but with some key differences. In the sequential approach, the genomes of the interacting species must be indexed separately, whereas in the combined approach, all genomes in the combined approach are concatenated before indexing as a single entity. The combined approach produces a unified sequence alignment file (SAM/BAM/CRAM format) after mapping, unlike the sequential approach, which generates separate alignment. Consequently, for the combined approach, the resulting alignment file must be carefully partitioned to extract species-specific data (Espindula *et al.*, 2019).

In summary, dual RNA-seq shares many similarities with classical bulk RNA-seq, with the main differences occurring in the genome indexing step and the handling of alignment files before summarization.

Bulk RNA-seq analysis can be carried out via the command line by combining various software tools, allowing users to build pipelines in different programming languages. Additionally, there are graphical user interfaces (GUIs) specifically designed to facilitate this process, such as the Galaxy platform (Afgan *et al.*, 2018), RNAdetector (La Ferlita *et al.*, 2021), RNASeqGUI (Russo and Angelini, 2014), and RNALysis (Teichman *et al.*, 2023).

Despite the wide range of options available for bulk RNA-seq analysis, no dedicated tools or platforms have yet been developed for dual RNA-seq analysis. To perform dual RNA-seq analysis, it is currently necessary to develop customized pipelines, which requires proficiency in programming and scripting. This makes, dual RNA-seq analysis challenging for scientists who lack programming skills but wish to use a user-friendly interface that can guide them through the process step by step. To address this gap, this chapter introduces inDAGO, a standalone software with a GUI designed to automate dual RNA-seq analysis and guide users through each stage of the process. No programming expertise is required, allowing users to seamlessly navigate through the different steps of the analysis and download intermediate results as needed. Additionally, inDAGO offers supplementary modules to perform classical bulk RNA-seq analysis, making it a versatile tool for various RNA-seq applications. To demonstrate its usability, a dual RNA-seq dataset was analyzed as part of this chapter, showcasing inDAGO capabilities.

3.3 Overview of the interface

The inDAGO software is fully implemented in the R programming language (R Core Team, 2024), and its GUI is developed using the Shiny application framework (Chang *et al.*, 2024). The software is compatible with GNU-Linux, Windows or MacOS operating systems.

As illustrated in Figure 3.1, inDAGO provides a seamlessly integrated analysis framework, offering users the flexibility to choose between the sequential and combined approaches during key steps - namely indexing, mapping, and read discrimination (which is required only for the combined approach). This flexibility ensures that users can tailor the workflow to their specific dual RNA-seq experiment needs.

Another key strength of inDAGO is its dedicated modules for performing bulk RNA-seq analysis, from the genome indexing step through the final stages of the analysis. This feature not only enables users to conduct bulk RNA-seq but also offers tools to handle unmapped reads—those that are not assigned to any specific genomic region. These unmapped reads can be further analyzed by mapping them to other reference sequences, which helps in detecting potential contaminants (reads originating from organisms not included in the experiment). Additionally, users can remap unmapped reads adjusting the mapping stringency (e.g., modifying the number of allowed mismatches or handling multiple mapped reads) to improve read assignment accuracy. This flexibility enhances inDAGO utility for both dual and bulk RNA-seq workflows.

inDAGO comprises seven distinct modules: (1) quality control, (2) filtering, (3) genome indexing, (4) mapping, (5) summarization, (6) exploratory data analysis, and (7) DEG calling. The layout and functionality of all modules are illustrated in Figure 3.2. Scientists can easily access the desired module by selecting it from the tab-panels and drop-down menus. Additionally, any ongoing analysis can be terminated immediately and restarted as needed, providing users with greater flexibility and control over their workflow. The quality control module requires raw data from the sequencing machine in FASTQ format (which can be compressed or uncompressed using the GNU-Linux *gzip* utility) (Cock *et al.*, 2010) and provides a comprehensive quality overview. The filtering module utilizes the same input data as the quality control, outputting only high-quality reads by removing low-quality bases and adapter sequences. The resulting unpaired reads are not included in the final output.

The genome indexing module begins with reference genomes in FASTA format (also available in compressed or uncompressed forms via the GNU-Linux *gzip* utility) and creates a data structure that facilitates rapid access to the genome. The mapping module then aligns high-quality reads to the previously generated genome indices, producing various alignment files in SAM/BAM format that are assigned to each interacting organism, along with unmapped reads retained in FASTQ format.

The summarization module processes SAM/BAM files to generate separate raw counts of genes for each replicate. Following this, the exploratory data analysis module takes these raw counts and provides a comprehensive overview of the mapped reads through a variety of statistical analyses and graphical representations. The DEG analysis module uses the same input data as the exploratory data analysis module and concludes the overall analysis by producing a tabular output that lists the differentially expressed genes. This structured approach ensures a seamless transition from data processing to detailed analysis and final reporting of results.

The source code for inDAGO is available on <https://github.com/CarmineFruggiero/inDAGO>.

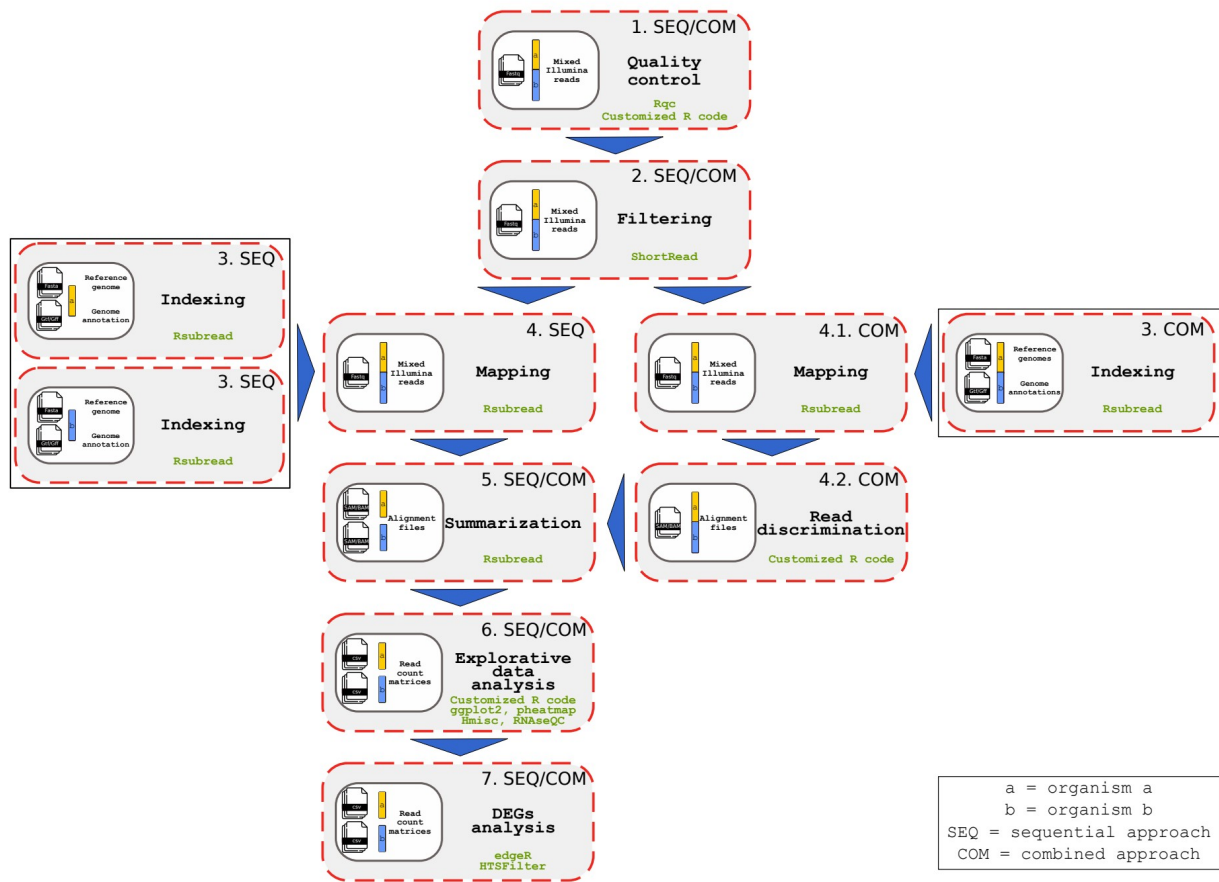


Figure 3.1. inDAGO pipeline for performing dual RNA-seq using either the sequential or combined approach. The pipeline consists of seven steps, where steps 1, 2, 5, 6 and 7 are common to both approaches, while steps 3 and 4 are approach-specific. Step 1: quality check of raw mixed reads (organisms A + organism B in FASTQ format) using the Rqc package. The graphical representation of results is generated through in-house developed R scripts; step 2: filtering of raw mixed reads is performed using the ShortRead package; step 3: indexing of reference sequences (FASTA format) is done with the Rsubread package. In the sequential approach, separate sequences are indexed for each organism, while in the combined approach, a single concatenated genome is indexed; step 4: alignment of filtered reads to the reference genomes is conducted using Rsubread. In the sequential approach, a double mapping step is performed (one for each organism), while in the combined approach, a single mapping is followed by *in silico* read discrimination; step 5: mapped reads (in SAM/BAM format) are assigned to each organism and summarized using the Rsubread package; step 6: summarized reads (in CSV format) are explored through statistical analysis and visualizations using customized R code, along with ggplot2, pheatmap, Hmisc, and RNAseQC packages; step 7: Differentially expressed genes (DEGs) are identified using the edgeR and HTSFilter packages.

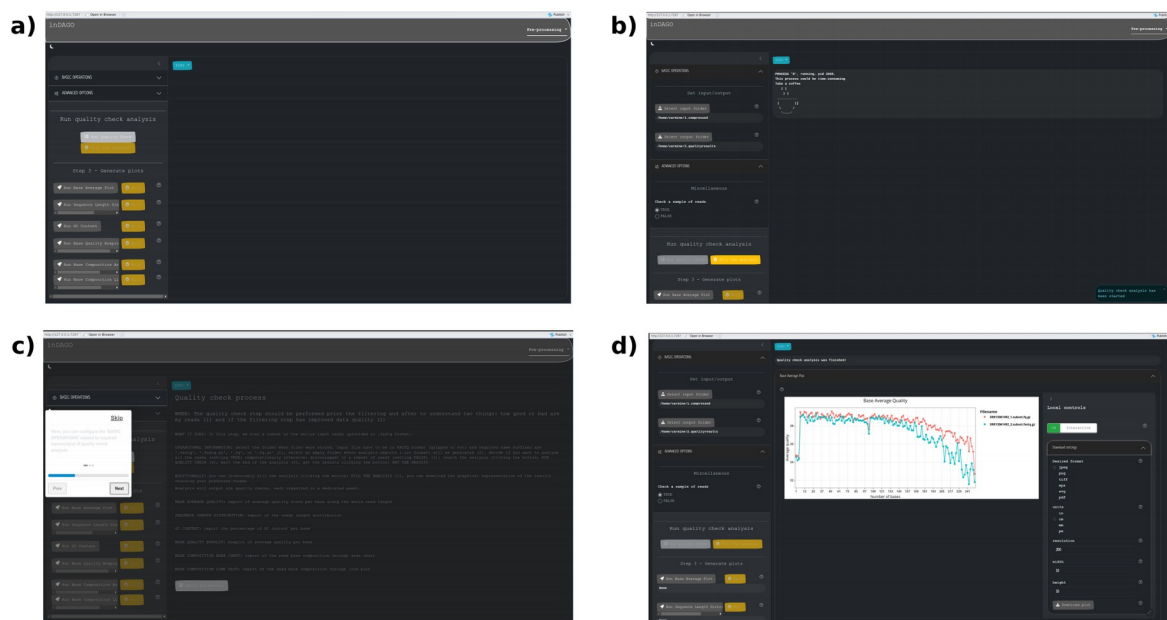


Figure 3.2. Graphical interface of inDAGO. This figure presents screen captures from various modules of inDAGO during different steps of the analysis. The images illustrate the following: (a) the opening screen; (b) access to dynamic documentation; (c) the module execution process, and (d) the results displayed at the end of the analysis. These captures provide a comprehensive overview of the user interface and workflow within *inDAGO*.

3.4 Methods

3.4.1 Quality control

The module is designed to process raw reads obtained from the Illumina sequencing platform in FASTQ format. It extracts critical information from these files and should be executed both before and after the filtering step to assess the reliability of the reads and to determine whether the filtering has improved data quality.

To use the module, two inputs are required:

- a folder containing the raw reads,
- an empty folder where analysis reports will be returned in CSV (comma separated values) format.

Users can choose to analyze all reads- though this option is computationally intensive or to examine only a random subset of one million of record from each input FASTQ file. In addition to providing results in tabular format, the module offers options for visualizing and downloading graphical representations of the data. Users can select from three different formats and can choose to visualize all samples or a specific selection.

The module provides users with an overview of several key metrics:

- **base average quality:** reports the average quality score per base across the entire read length, visualized through both a line plot and a boxplot;
- **sequence length distribution:** displays the distribution of read lengths via a line plot;
- **GC content:** reports the percentage of GC content per base, illustrated with a line plot;
- **base composition:** visualizes the percentage of each base in the reads using an area chart and a line plot.

This comprehensive analysis aids users in understanding the quality and characteristics of their sequencing data.

3.4.2 Filtering

The filtering module processes and cleans raw sequencing data stored in FASTQ format. Specifically, it performs the following tasks: trims low-quality bases from the ends of each read, removes adapter sequences, and applies a sliding window analysis to filter out sequence with a high proportion of low-quality bases. Additionally, after trimming, the module can filter out reads shorter than a specified minimum length. The module requires two inputs: (i) the folder containing raw reads, and (ii) an empty folder where the output will be saved. Users can also configure additional parameters, including:

- **Minimum read length:** specifies the minimum allowable read length after trimming.
- **Quality threshold:** defines the minimum quality score required for base retention.
- **Buffer size:** specifies the number of bytes or characters to be read at a time. Lower values reduce memory usage but increase computing time.

- **Trimming trigger:** specifies the number of consecutive low-quality bases required to initiate end trimming.
- **Window size:** defines the size of the sliding window for quality analysis.
- **Position of low-quality bases:** defines whether to trim only the first succession of low-quality bases or allow their removal from anywhere within the sequence.
- **Trim location:** specifies whether trimming should occur at the left end, right end, or both ends of the read.
- **Adapter sequence removal:** Inserts adapter sequences to be removed from the reads.
- **Adapter mismatches:** specifies the allowable number of mismatches when aligning adapter sequences to reads.
- **Output file compression:** selects whether to compress the output files using Lempel-Ziv coding (with a ".gz" extension).
- **Parallel processing:** defines the number of parallel jobs to divide the analysis into, for faster processing.

At the end of the analysis, the module returns output files in FASTQ format, containing the reads after filtering and trimming have been applied.

3.4.3 Genome indexing

A hash-table data structure was generated from the reference genome file provided in FASTA format. This is essential for the subsequent alignment step and significantly reduces processing time. This module consists of three sub-modules, allowing users to choose between conducting dual RNA-seq analysis in either sequential or combined mode or, to perform a classical bulk RNA-seq analysis. Specifically, the module creates two distinct indexes: one for the target genomes in sequential mode, and a single index for the combined mode, which is generated concatenating both genomes of the interacting organisms.

The required inputs include the two genomes for dual RNA-seq or a single genome for bulk RNA-seq that need to be indexed, along with the folder(s) where the index(es) will be stored. For the combined dual RNA-seq approach, users must specify a different prefix for each genome to label each genome sequence clearly.

Additionally, users have the option to select advanced parameter such as:

- **Index type:** users can choose between a gapped index and a full index. A gapped index requires less time and disk space than a full index, but it may decrease the speed of the subsequent mapping step.
- **Index block size:** the index can be configured as a single block if the user sets a maximum size (in MB) that exceeds the index actual size. Otherwise, the index will consist of multiple blocks, each equal to the selected maximum size. The size of each block is limited by the available RAM, as each block must be loaded into RAM during the mapping step.

- **Removal of highly repetitive subreads:** removing uninformative subreads from the hash table helps reduce mapping ambiguity. Users can set a threshold to determine the maximum number of times a subread is allowed to occur in the reference genome to be included in the hash table.
- **Indexing options:** users can choose to index one genome at a time or both genomes together. This parameter is available only in sequential mode, where two separate indexes must be generated.
- **Temporary file management:** In combined mode, users can select a folder for temporary files. This feature is useful for managing temporary files generated during the analysis, allowing users to decide whether to store or delete them afterward.

The output will consist of two binary index files for each index block, saved in the selected folders, along with three plain text documents.

3.4.4 Mapping

As part of the indexing module, the mapping process differs for each of the three supported analyses. In general, it aligns reads to the provided reference sequences, producing alignment files in SAM/BAM format. Additionally, unmapped reads are output in FASTQ format. Specifically for combined approach, aligned sequences are assigned to each reference sequence from the previous module and, cross-mapped reads are also output in BAM format.

For the sequential approach, two indexes are required, while a single index is sufficient for the combined approach and bulk RNA-seq. These indexes, along with reads in FASTQ format, and two empty folders (for the sequential approach) or one empty folder (for the other two approaches), are needed as input. The user can also configure the following additional parameters:

- **Number of cores** to be used in parallel processing.
- **Number of threads** for the alignment process.
- **Alignment output format:** either uncompressed (SAM) or compressed (BAM).
- **Phred score type** for base-calling (33 or 64).
- **Number of subreads** (short reads used to find the optimal mapping location) extracted from each input read. Increasing subreads can enhance mapping stringency but may increase runtime.
- **Consensus threshold:** the minimum number of consensus subreads required to report a hit. Consensus subreads are those that align to the same location in the reference sequence. Increasing this threshold may reduce the number of mapped reads.
- **Maximum mismatches** allowed in the alignment. Mismatches in soft-clipped bases are excluded from this count.
- Option to **include only uniquely mapped reads** in the alignment file.
- **Maximum number of multiple mapped reads** to report in the alignment file.
- **Maximum number of indels** allowed in the alignment.
- **Maximum and minimum fragment length** for paired-end reads.

- **Read orientation** for paired-end reads: three options are available—(i) the first read on the forward strand, second on the reverse; (ii) the first on the reverse, second on the forward; (iii) both reads on the same strand.
- **Order of reads** in the output: aligned reads can either follow the input order or, for BAM output, be sorted by mapping location.
- **Reporting of all discovered junctions**, including exon-exon junctions and gene fusions.
- Option to use **gene annotation** from a GTF (General Transfer Format) or SAF (Simple Annotation Format) file, which may be gzip-compressed. For GTF, users can specify the feature type (e.g., "exon") and attribute (e.g., "gene_id") for the alignment step.
- **Temporary folder selection** for storing intermediate files during analysis, which will be removed after the process is complete

The results include two BAM file, one for each reference sequence in the dual RNA-seq approach, and one BAM file for bulk RNA-seq. Additionally, unmapped reads are provided in FASTQ format. Three text files are also generated detailing the discovered junctions, indels, and a summary of the analysis. For the combined approach, an additional BAM file is produced, containing only cross-mapped reads.

3.4.5 Summarization

The module processes SAM/BAM files by assigning each read to specific genomic/transcriptomic feature and then quantifying gene expression for each sample. It generates a count table, where the number of reads mapped to each selected feature (e.g. "exon") is reported for sample. To function, the module requires three inputs: (i) a folder containing the alignment files; (ii) an empty folder to store the output files; and (iii) an annotation file in GTF/SAF format. These inputs allow the module to efficiently perform feature assignment and generate precise gene expression data for downstream analysis.

Additional parameters provided by this module:

- **Gene annotation format:** specifies the required format of the gene annotation file (e.g., GTF or SAF).
- **Feature and attribute selection:** selects the feature type (e.g., "exon") and attribute (e.g., "gene_id") for quantification when using a GTF file. Custom features or attributes can also be added, if needed.
- **Quantification level:** selects whether to perform quantification at the feature level (e.g., "exons") or the meta-feature level (e.g., "genes").
- **Overlap threshold:** specifies the minimum number of overlapping bases required to assign a read to a feature or meta-feature.
- **Assignment criteria:** specifies whether the assigned feature/meta-feature should be the one with the largest number of overlapping bases.
- **Multi-mapping reads and chimeric fragments:** specifies whether to include or exclude multi-mapped reads and chimeric fragments (i.e., paired reads aligned to different chromosomes).
- **Fractional counting of multi-mapped reads:** if multi-mapped reads are included, specifies that each read will be divided by the total number of reported alignments, assigning the resulting fraction to each.

- **Minimum mapping quality score:** specifies a threshold for the minimum mapping quality score to filter low-quality reads.
- **Primary vs. secondary alignments:** specifies whether to include only primary alignments (the best alignment for multi-mapped reads) or both primary and secondary alignments.
- **Duplicate read handling:** Specifies that reads flagged as duplicates (e.g., those resulting from PCR artifacts) should be excluded.
- **Strand-specific RNA-seq:** specifies the handling of strand-specific RNA-seq data, if applicable.
- **Paired-end read alignment:** specifies whether paired-end reads should be counted only if both paired reads are successfully aligned.
- **Read sorting:** selects to sort reads by name.
- **Length threshold for paired-end reads:** specifies the minimum length threshold for paired-end reads to be considered in the analysis.
- **CIGAR string criteria:** defines the number of match operations allowed in a CIGAR string to control the number of gaps during alignment.
- **Temporary folder:** selects a temporary folder for storing intermediate files, which will be removed after the analysis is complete.
- **Parallel processing:** specifies the number of threads and the number of samples to analyze in parallel to accelerate the process.

At the end of the analysis, four output files are generated for each sample: (i) a CSV file containing annotation information for the assigned features; (ii) the quantification file in CSV format; (iii) a summary of the quantification process in CSV format; and (iv) a log file documenting the process.

3.4.6 Exploratory data analysis

The primary goal of a Data Exploration Analysis (DEA) is to examine and reason about the data in relation to the specific biological question being investigated. This step allows the user to refine the research question or, if data are insufficient, decide whether to collect new data. This module generates a series of exploratory plots, which are crucial for identifying patterns, detecting potential outliers, and producing informative visuals. These plots are not only helpful for understanding the data but also serve as effective communication tools when presenting the analysis in a published article.

The only required input is the count matrices generated during the summarization step for each sample. Additionally, a matrix indicating the group assignments of each sample can be uploaded. Alternatively, group assignments can be manually defined through the interface.

User can explore and modify all plots interactively, with options to customize the colour schemes from a wide range of palettes. Plot downloads are also highly customizable, allowing users to specify the output path, file format (options include "jpeg", "png", "tiff", "eps", "svg", and "pdf"), and resolution in terms of DPI (dots per inch). Additionally, users can set the height and width of each plot and select the unit of measurement (inches, cm, mm or pixels) for these dimensions.

Plots are generated twice: one before and one after normalization, allowing for a visual comparison of the normalization effect. This is because the module also performs count normalization. The DEA plots include the following:

- **Multi-Dimensional Scaling (MDS)** visualizes differences in expression profiles between samples by using the "leading fold change" (the root-mean-square average of the log-fold-changes) as the distance metric. This metric is applied to the genes that best distinguish each pair of samples. The resulting sample clustering can offer insights into the design matrix construction for downstream differential expression analysis.
- **Principal Component Analysis (PCA)**, like MDS, helps interpret large datasets by reducing dimensionality while preserving as much variability as possible. PCA identifies new variables (principal components) that are linear combinations of the original data, maximizing variance while remaining uncorrelated. These components simplify complex data (Jolliffe and Cadima, 2016).
- **Boxplots** provide a compact visualization of the distribution of transformed count data (log-transformed counts). They display the median, with the lower and upper hinges representing the first and third quartiles (IQR = interquartile range). The whiskers extend to the smallest and largest values within 1.5 times the IQR, with any points beyond considered outliers. Users can control the order of the boxes in the plot.
- **Barplots** visually represent the library size of each sample. Users can arrange the bars in any desired order.
- **Gene expression heatmaps** display the expression levels of genes across samples, with colors representing individual data points. Users can select the number of most representative genes (up to 1,000) to include. Hierarchical clustering groups both samples and genes, and users can define the number of clusters for rows (genes) and columns (samples). Users can also opt to skip clustering for rows and/or columns, choose whether to center and scale data by row or column, and control whether row/column names are displayed.
- **Correlation plots** (correlation heatmaps) represent the correlation matrix between samples. Users can choose between Pearson's correlation (r) or Spearman's rank correlation (ρ) and adjust the same parameters available for the gene expression heatmap, including rounding correlation coefficients.
- **Saturation plots** depict the saturation of gene detection, showing the number of detected genes as a function of sequencing depth for each sample. Users can select between two estimation methods for gene detection: "sampling" or "division."

3.4.7 DEGs analysis

The final module of the pipeline begins by utilizing the same count data from the DEA module and concludes by generating a list of differentially expressed genes along with key statistical parameters such as $\log_2$ fold change, p-value, and adjusted p-value which are output in CSV format.

In detail, the process starts with importing the raw count data, after which a data model matrix is built to represent the experimental design. The data is then filtered to remove low-abundance genes, as these genes provide limited statistical information for

distinguishing between the null and alternative hypotheses. Following this, the data is normalized to account for variations such as sequencing depth, and the dispersion parameters calculated. Dispersion is a crucial measure in RNA-seq data analysis, essential for determining the DEGs. Once the data is prepared, hypothesis tests are performed based on the contrasts provided by the user. These contrasts define the comparisons to be made between the experimental groups and allow the module to identify differentially expressed genes.

For the module to run, the user must provide a folder containing raw count data a text file that specifies the group assignment of the samples, and an empty folder where output will be saved. The user also needs to input the desired contrasts that will guide the comparisons between the groups. Additionally, the module offers the flexibility to assign sample groups directly through the interface without needing to upload a file.

Advanced parameters are available for users who wish to customize the analysis further. These options allow for fine-tuning of the filtering thresholds, normalization methods, and other statistical settings as needed.

First, two different filtering methods are available. The first method retains genes that meet a minimum read count threshold across samples, as detailed in the study (Chen, Lun and Smyth, 2016). The second method calculates a threshold based on the pairwise Jaccard similarity index between replicates within each experimental condition, as outlined in Rau *et al.*, (2013).

For statistical testing, users can select from three methods: exact tests, quasi-likelihood F-tests (QLFTest), or genewise likelihood ratio tests (LRT). The exact test, described in (Robinson and Smyth, 2007), assesses differences in the means between two groups with negative-binomially distributed counts. Both QLFTest and LRT, based on the quasi-likelihood (QL) methods of Lund *et al.*, (2012), perform tests for a specific coefficient or coefficient contrast. However, QLFTest replaces likelihood ratio tests with empirical Bayes quasi-likelihood F-tests, offering a different approach from LRT.

Users can also choose between several normalization methods. These include TMM (trimmed mean of M-values), as described in Robinson *et al* (2010), TMMwsp (a variant of TMM that performs better with datasets containing a high proportion of zeros), RLE (relative log expression, proposed in Anders *et al.*, 2010), and the upperquartile method (Bullard *et al.*, 2010). If desired, users can bypass these normalization methods by setting the normalization factors to 1.

For p-value adjustments in multiple testing comparisons, several correction methods are available: Bonferroni, Holm, Hochberg, Hommel, and FDR (false discovery rate). The Bonferroni correction multiplies p-values by the number of comparisons, while Holm, Hochberg, Hommel, and FDR methods offer fewer conservative alternatives (Holm, 1979; Hochberg, 1988; Hommel, 1988; Benjamini and Hochberg, 1995).

Users can also specify thresholds for both the absolute log₂ fold change (log₂ FC) and the adjusted p-value to filter the list of DEGs. Additionally, the option to generate a volcano plot for each comparison is available. These plots can be customized with the same settings as those used in the DEA module, including the ability to set the maximum number of overlaps for text labels, adjust label size, apply filtering based on log₂ FC and adjusted p-value thresholds, and choose how many DEGs to display.

Lastly, users can generate an upset plot for their analysis, with the same customizable settings as those offered for the plot in the DEA module.

In addition to the DEGs list, two plots are provided to visualize the estimated dispersion: one representing the dispersion based on the negative binomial distribution, and the other showing the quasi-likelihood dispersion.

3.4.8 Graphical structure and dynamic documentation

Each module is organized into a sidebar panel and main panel. The sidebar panel guides users in inputting data and setting both basic and advanced parameters. It is further divided into two sub-panel: basic parameters, which accepts the mandatory inputs required for analysis, and advanced parameters, which allows users to control more technical aspects of the analysis.

The main panel displays notifications about the process status, including any warnings and errors. If the module produces graphical outputs, these will be shown in the main panel, along with options for interactive modifications and downloads of the plots. To enhance user experience, some modules feature a main panel organized into several expandable cards.

Additionally, each module is accompanied by concise documentation that users can easily access.

Specifically, the main panel includes a brief documentation section divided into clear segments:

- **"WHEN"**: explains the appropriate circumstances for performing the analysis using this module.
- **"WHAT IT DOES"**: provides a brief description of the analysis conducted by the module.
- **"OPERATIONAL INFORMATION"**: offers a step-by-step guide on how to conduct the analysis.
- **"ADDITIONALLY"**: lists supplementary options available to users.
- **"RESULTS"**: describes the expected results and their corresponding formats.

At the end of this documentation, users will find the "ABOUT STEPS" button. Clicking this button activates a dynamic documentation feature that underlines and provide explanations for each sidebar box.

To further enhance the module documentation, each sidebar box includes a question mark icon. Clicking this icon displays additional details relevant to that specific box.

3.4.9 Dataset

To evaluate the reliability and performance of InDAGO , we utilized data from Chapter 2, focusing on two interacting species: the model organism *Arabidopsis thaliana* and the parasitic weed *Cuscuta campestris*. The reference genomes for these species were obtained from the European Nucleotide Archive (ENA) repository: *A. thaliana* (GCF_000001735.4) and *C. campestris* (GCA_900332095.2). Detailed information regarding this data can be found in Table 2.1. In order to test the DEG analysis module, it is necessary to consider other samples under a different condition from the same *A. thaliana* ecotype in order to detect differentially expressed genes between the two conditions. In details, this additional RNA-seq data, sampled during the vegetative stage from *A. thaliana* Columbia ecotype (Col-

0) were downloaded from the ENA repository (ENA accession numbers: SRR309145, SRR309146, SRR309147, SRR309148; project: PRJNA154687).

3.5 Results

Preprocessing analysis. The three biological replicates of *A. thaliana* and *C. campestris* were initially pre-processed using the quality control module. Raw reads were then trimmed to remove low-quality bases and filtered based on the default quality threshold of 20. Figure 3.3 presents the quality control results produced by inDAGO, which randomly sampled a subset of the raw reads. For the filtering module, default advanced parameters were applied, including a minimum read length of 40 bp and a quality threshold of 20. The pre-processing step yielded satisfactory results, discarding an average of 5,000 reads per sample, consistent with the pre-processing analysis performed in Chapter 2 on the same dataset.

Alignment to reference sequence. Both the sequential and combined approaches were applied. For the sequential mode, two separate indexed genomes were generated, while a single concatenated genome was created and indexed for the combined mode. In both cases, subreads (16 bp mers) that were repeated more than 100 times were excluded from the genome indexing process. The mapping step was performed using default advanced parameters: 14 subreads were extracted from each read, a consensus threshold of 1 was used, up to 3 mismatches were allowed, 1 multi-mapped read was permitted, the indel length was set to 5, and the fragment length ranged from 50 to 600 bp. More than 90% of high-quality reads successfully aligned to the reference sequences. The read discrimination process showed results consistent with those in Chapter 2 for both the sequential and combined approach. Figure 3.4 presents the mapping statistics for both methods.

From quantification to differential expression analysis. Gene expression profiles were obtained using default parameters, selecting "exon" as the feature, "gene_id" as the attribute, 1 bp as the minimum number of overlapping bases, and allowing 10 match operations for each CIGAR string. Data exploration analyses were conducted separately for each organism, generating MDS plots, PCA plots, boxplots, expression heatmaps, correlation plots and saturation curves both pre- and post-data normalization. All graphical outputs are presented in Figure 3.5, which illustrates the quantification of mixed reads mapped to the *A. thaliana* genome using the sequential mode after data normalization. The raw gene expression counts were then used as input for the differential expression analysis module. The default advanced parameters included: a filtering method as described by Chen et al. (2016), quasi-likelihood F-tests (QLFTest) as the statistical test, TMM as the normalization method, FDR for p-value adjustment, a $\log_2$ FC threshold of 0.58, and an adjusted p-value threshold of 0.05 for classifying up-regulated and down-regulated genes. Figure 3.6 displays the results of the comparison between three biological replicates of *A. thaliana* from Chapter 2 and four biological replicates used specifically to set this contrast for testing the module. Out of 16,839 quantified genes, 3,706 were found to be down-regulated, while 3,738 were identified as up-regulated.

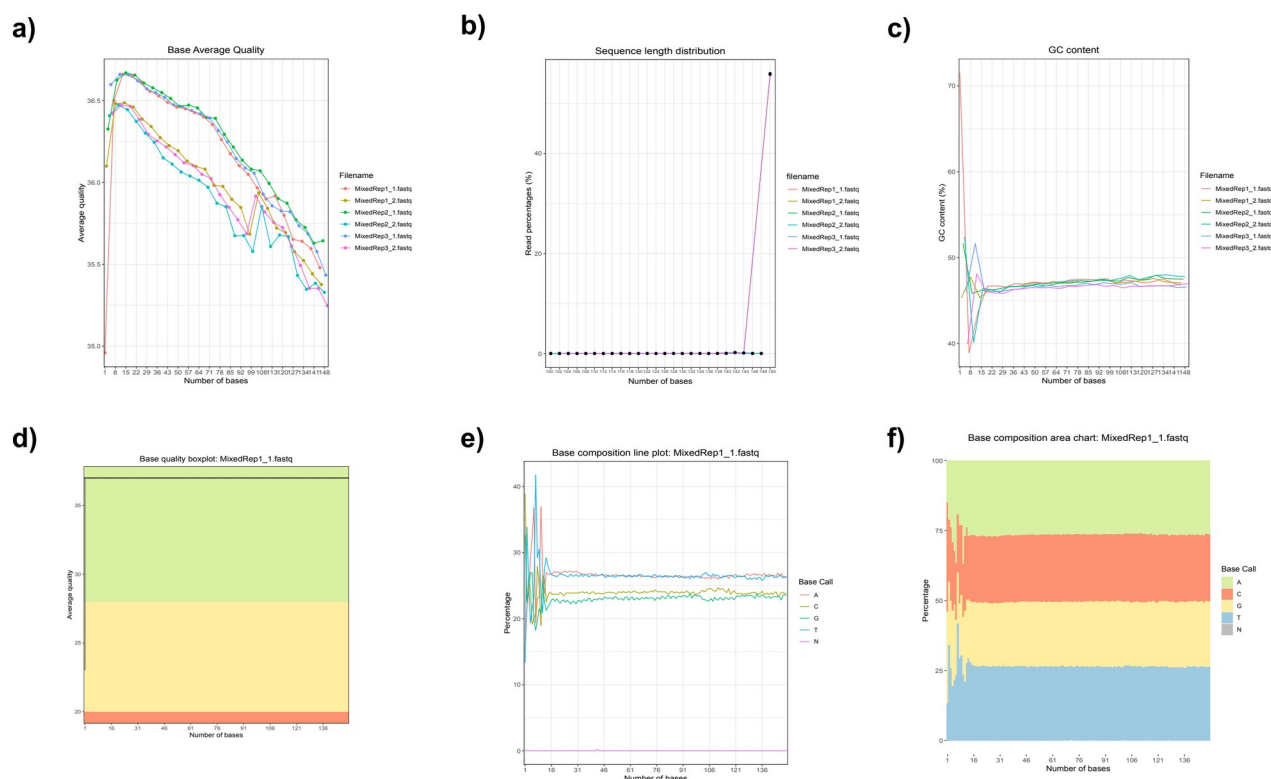


Figure 3.3. Quality control module results. The figure presents the graphical outcomes of the quality control module, including the following plots: (a) base average quality line plot and (d) base average quality boxplot (d), reporting the average quality score per base along the entire read length; (b) sequence length distribution line plot; (c) GC content line plot, reporting the percentage of GC content per base; (e) base composition line plot and (f) area chart plot, reporting the percentage read base composition.

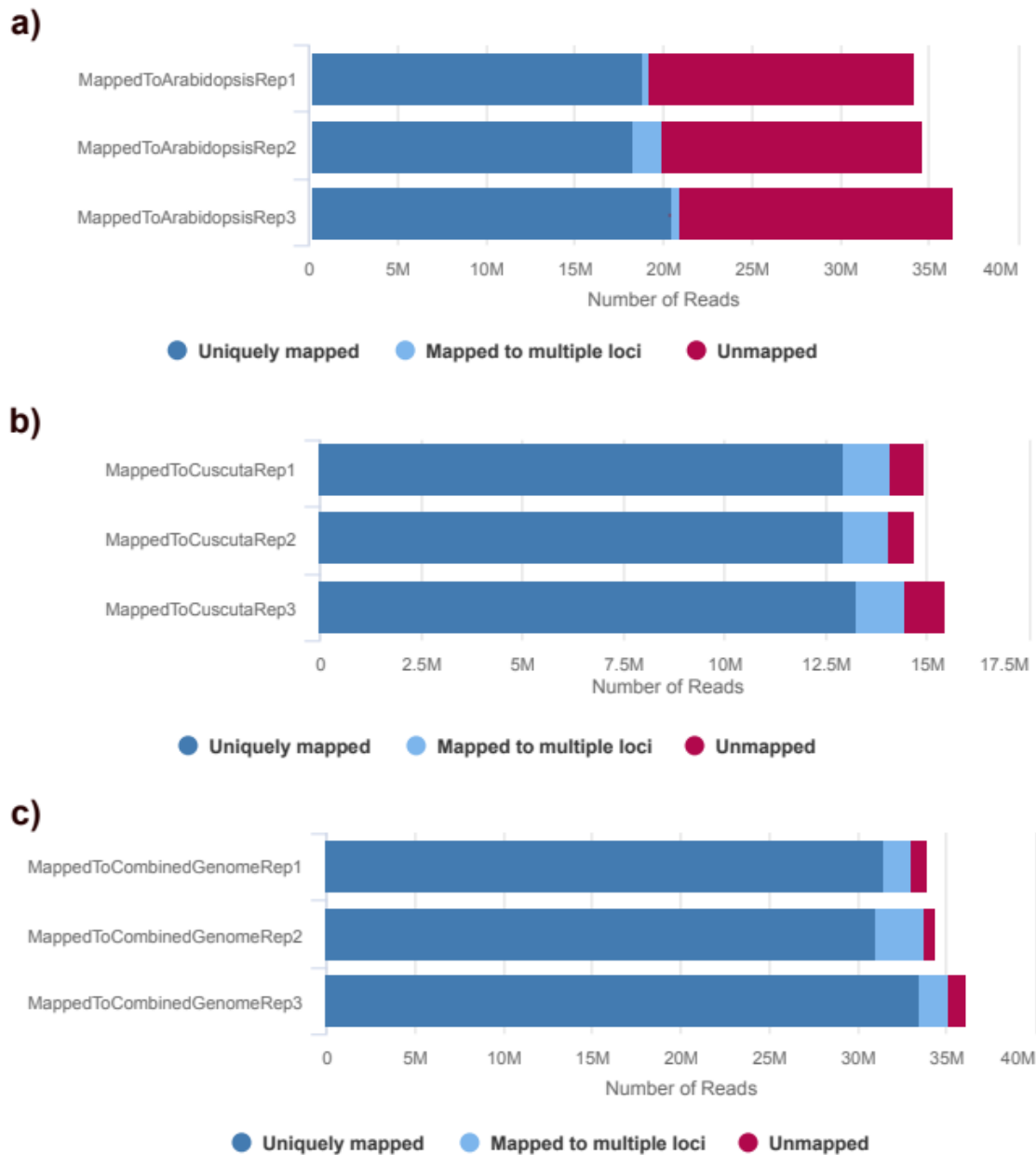


Figure 3.4. Results of the mapping module. This figure presents the mapping statistics obtained from the mapping module, illustrating the results of three distinct stages of the mapping process. (a) First step of sequential mapping: this panel displays the mapping of mixed reads to the *Arabidopsis thaliana* genome.; second step of sequential mapping: this panel shows the mapping of unmapped reads from the first step to the *Cuscuta campestris* genome; (c) this panel illustrates the mapping of mixed reads to the combined genome of *A. thaliana* and *C. campestris*.

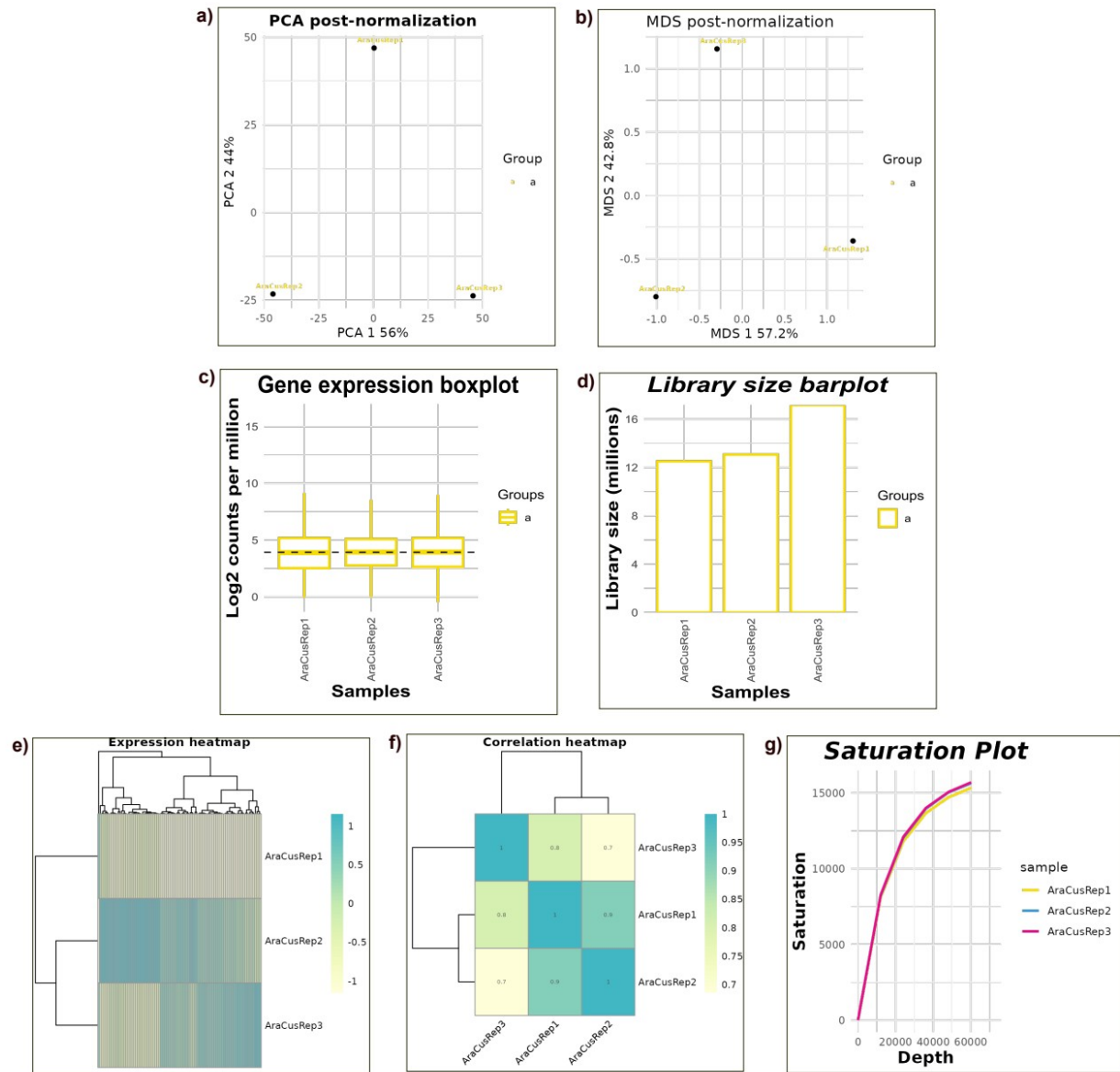


Figure 3.5. Results of the data exploration module. The figures present the graphical outputs generated by the data exploration module, which include: (a) PCA plot; (b) MDS plot; (c) gene expression boxplot; (d) library size bar plot; (e) gene expression heatmap; (f) correlation heatmap (f); and (g) saturation plot. All graphical outputs are based on the quantification of mixed reads mapped to *A. thaliana* genome using the sequential mode.

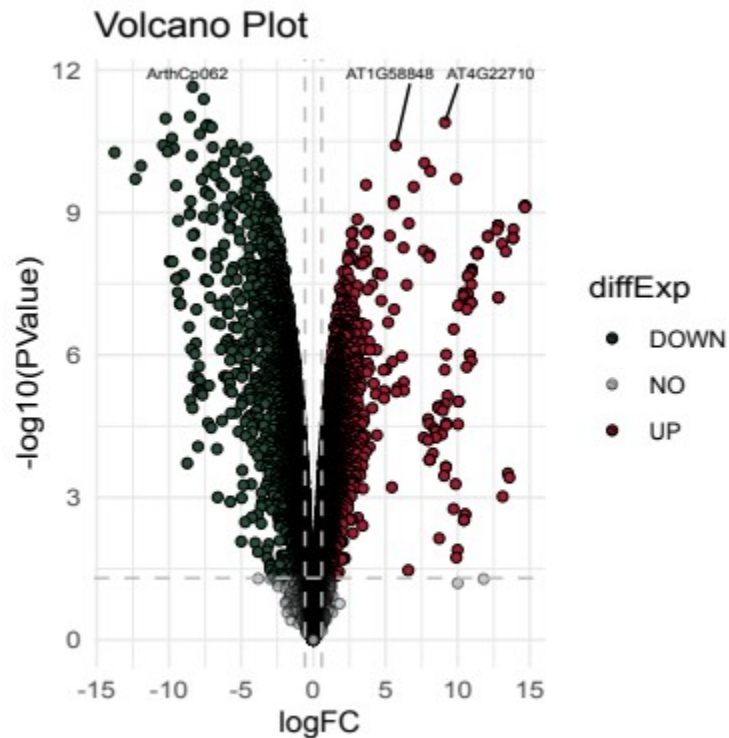


Figure 3.6. Results of the differential gene expression analysis module. The volcano plot provides a comprehensive graphical overview of differentially expressed genes, highlighting both up-regulated and down-regulated genes. The data analyzed includes three biological replicates of *Arabidopsis thaliana* from chapter 2, along with four additional biological replicates, retrieved to set the contrast in this module. Each point on the plot represents a gene, with the x-axis indicating the log fold change in expression and the y-axis representing the negative logarithm of the p-value. Genes exhibiting significant differential expression are distinctly marked, allowing for a quick visual assessment of gene regulation trends.

3.6 Conclusion

This chapter introduces inDAGO, a standalone, cross-platform software tool featuring a user-friendly GUI designed to eliminate the need for programming knowledge.

inDAGO allows scientists to conduct two types of transcriptomic analyses: dual RNA-seq, which can be approached using either sequential or combined methodologies, and bulk RNA-seq, starting from raw data generated by sequencing platforms.

The software integrates a suite of R packages to facilitate various analytical steps. In cases where existing packages do not meet specific needs, we have developed *ad hoc* R code (i.e. the *in silico* read discrimination step was entirely developed in-house).

Future versions of inDAGO are expected to include additional modules for downstream analyses, such as gene enrichment analysis and gene co-expression network construction. We also plan to introduce functionality for investigating non-coding RNA (ncRNA) and extending capabilities to analyze single-cell RNA-sequencing data.

To evaluate the performance and reliability of inDAGO, we tested it using public datasets and simulated a host-parasite interaction between two plant species. The results demonstrated that all modules performed satisfactorily. Overall, inDAGO emerges as a reliable and powerful tool for guiding scientists in the analysis of dual RNA-seq data.

3.7 Code and data availability

The source code for inDAGO is available on GitHub at <https://github.com/CarmineFruggiero/inDAGO>.

Chapter 4. Investigating molecular mechanisms of pre-attachment defence against *P. ramosa*: identifying polymorphisms in strigolactone pathway genes between *S. lycopersicum* and its wild relative *S.pennellii*

4.1 Abstract

The germination rate of *P. aegyptiaca* parasitic seeds is significantly lower when in contact with root exudates of *Solanum pennellii* compared to those from *S. lycopersicum*. Strigolactones (SLs) act as germination stimulants for these parasitic seeds, and their biosynthetic pathway involves several key genes. SL-related genes were retrieved from public databases and their orthologous were identified in both *S. lycopersicum* and *S. pennellii*. Each gene from *S. lycopersicum* was aligned with its ortholog in *S. pennellii* to detect sequence polymorphisms. PCR experiments and Sanger sequencing were conducted to validate the *in silico* predictions.

Two interesting polymorphisms were identified in the first two genes of the SL pathway. The D27 (DWARF 27) gene exhibits an 81-nucleotide indel, while the *S. pennellii* CCD7 gene features a premature stop codon immediately following the ATG start codon. PCR and Sanger sequencing confirmed the presence of the D27 indel, while attempts to amplify the CCD7 stop codon yielded multiple non-specific amplicons. These polymorphisms may play a role in the defense response of host plants. This preliminary analysis serves as a foundation for understanding the genetic bases of *S. pennellii* tolerance mechanisms.

4.2 Introduction

Parasitic plants from the *Orobanchaceae* family and the *Cuscuta* genus pose a significant threat to global crop yields. Effective management is challenging due to their prolific seed production, with each plant capable of generating a vast number of small, long-lived seeds that persist in the soil. Additionally, the use of herbicides is complicated by the fact that, parasitic plants share the same taxonomic kingdom as their hosts, making it difficult to target the parasite without harming the crop itself (Fernández-Aparicio, Delavault and Timko, 2020).

One of the most promising strategies to combat parasitic weeds is the development of resistant cultivars, which requires a thorough understanding of the molecular mechanisms underlying host resistance. Host resistance can be categorized into three distinct stages: pre-attachment, during-attachment and post-attachment.

Pre-attachment resistance mechanisms work by reducing or altering the production of germination stimulants, thereby minimizing the germination of parasitic seeds. During attachment, resistance strategies aim to prevent or inhibit the parasite from penetrating the host tissues. This is primarily achieved by reinforcing the host's root cell walls through the deposition of structural barriers such as suberin, callose and lignin. Post-attachment defense involves two primary mechanisms. The first is a localized hypersensitive response, which triggers programmed cell death at the infection site to prevent further parasite invasion. The second mechanism involves the accumulation of mucilage within the vascular tissues which obstructs the translocation of nutrients, effectively starving the parasite (Fishman and Shirasu, 2021; Albanova *et al.*, 2023).

Pre-attachment resistance primarily depends on changes in the composition of chemical signals exuded from the host roots, which can disrupt the development of parasitic plants. Notably, *Solanum pennellii*, a wild relative of *Solanum lycopersicum*, exhibits a reduced emergence of shoots from *Phelipanche aegyptiaca*, a species closely related to *P. ramosa*. Research has demonstrated that the root exudates of *S. pennellii* can significantly inhibit the seed germination of *P. aegyptiaca* (Minuto *et al.*, 2006; Cuccurullo, Nicolia and Cardì, 2022).

Seeds of *Striga* spp., *Orobanche* spp., and *Phelipanche* spp. require specific molecules, known as strigolactones, for germination. There is considerable variability among SLs, which also contributes to host specificity. Acting as both plant hormones and chemical signals, SLs regulate numerous physiological aspects of plant development, including secondary growth, leaf senescence, inhibition of shoot branching, and root architecture formation (F. Wu *et al.*, 2022). When released into the soil, SLs function as signaling molecules, indicating the presence of a potential host for both symbiotic fungi and root-parasitic plants (García-Garrido *et al.*, 2009). Their name derives from their discovery, as first SLs were identified for their ability to induce seed germination in parasitic plants of the genus *Striga* (commonly known as witchweeds) (Cardoso, Ruyter-Spira and Bouwmeester, 2011). Throughout the process of co-evolution, root parasitic plants have adapted to utilize SLs to recognize their hosts effectively.

Strigolactones play a crucial role in stimulating the germination of parasites belonging to the genera *Striga*, *Orobanche*, *Alectra*, and *Phelipanche*. Additionally, SLs enhance symbiosis with arbuscular mycorrhizal fungi, particularly in nutrient-deficient environments where

nutrients like phosphate are poorly mobile in the soil. In such conditions, the extended network of fungal hyphae can effectively recover these nutrients, benefiting both the host plant and the mycorrhizal fungi (Ruyter-Spira *et al.*, 2013).

SLs are derived from carotenoids, which are plant isoprenoid pigments (Walter and Strack, 2011). Consequently, the biosynthesis of SLs represents a branch of the plant carotenoid biosynthetic pathway, beginning with β -carotene (Jia, Baz and Al-Babili, 2018). The SL biosynthetic pathway involves various enzymes that facilitate the conversion of carotenoids into these important signaling molecules. The biosynthetic pathway of SLs begins with the enzyme trans/9-cis- β -carotene isomerase DWARF27 (D27), followed by the actions of carotenoid cleavage dioxygenase 7 (CCD7) and carotenoid cleavage dioxygenase 8 (CCD8). These initial three enzymes are localized in the plastid and lead to the formation of a key molecule known as carlactone. From carlactone, various types of SLs are synthesized, each involving specific enzymes tailored to produce specific SLs (Mashiguchi, Seto and Yamaguchi, 2021). One such enzyme is More AXillary growth 1 (MAX1), a member of the CYP711 family and a cytochrome P450 enzyme. MAX1 converts carlactone into 5-deoxystrigol, which is thought to be a precursor for all other SLs (Ruyter-Spira *et al.*, 2013). After synthesis, SLs are translocated to specific districts within the plant and released into the soil, through an active transport process. The transporter identified as Pleiotropic Drug Resistance 1 (PDR1) in petunia has been shown to facilitate the exudation of SLs into the rhizosphere (Kretschmar *et al.*, 2012) (Figure 4.1).

This chapter presents a preliminary investigation comparing the SL biosynthetic pathway genes between two species: *S. lycopersicum* and *S. pennellii*. The primary objective of the analysis is to identify polymorphisms that may be associated with the molecular mechanisms underlying *S. pennellii* tolerance to *Phelipanche spp.* infection.

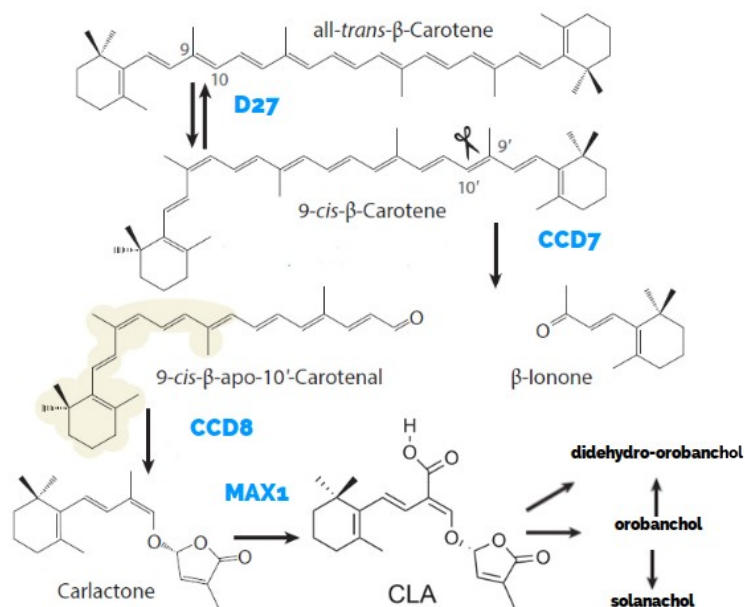


Figure 4.1. Strigolactone biosynthetic pathway. D27 catalyzes 9-cis/all-trans isomerization of β -carotene as first step of the pathway. Successively, CCD7 then mediates the cleavage of 9-cis- β -carotene at the C9'-C10' double bond. In one step, CCD8 converts 9-cis- β -apo-10'-carotenal into the SL carlactone. From carlactone, various types of SLs are synthesized, each involving specific enzymes specialized to produce specific SLs, such as didehydro-orobanchol, orobanchol, and solanachol.

4.3 Materials and methods

4.3.1 Strigolactone pathway genes

The genes involved in the strigolactone biosynthetic pathway were sourced from key literature references (Kretzschmar *et al.*, 2012; Ruyter-Spira *et al.*, 2013; Al-Babili and Bouwmeester, 2015; Jia, Baz and Al-Babili, 2018). The corresponding transcript and protein sequences were retrieved from the NCBI Nucleotide and NCBI Protein database, with their respective accession numbers listed in Table 4.1 for reference.

Genes	NCBI transcript accession number	NCBI protein accession number	Species	Enzyme function
Beta-carotene isomerase DWARF27 (D27)	XM_026032835.1	XP_025888620.1	<i>Solanum lycopersicum</i>	Converting all-trans-beta-carotene into 9-cis-beta-carotene
Carotenoid Cleavage Dioxygenase 7 (CCD7)	NM_001247504.1	NP_001234433.1	<i>Solanum lycopersicum</i>	Cleaving 9-cis-beta-carotene into a 9- cis- β -apo-10'-carotenal
Carotenoid Cleavage Dioxygenase 8 (CCD8)	NM_119434.4	NP_001329787.1	<i>Arabidopsis thaliana</i>	Converting 9-cis- β -apo-10'-carotenal into carlactone
More AXillary growth 1 (MAX1)	XM_004245037.4	XP_004245085.1	<i>Solanum lycopersicum</i>	Converting carlactone into 5- deoxystrigol
Pleiotropic Drug Resistance 1 (PDR1)	XM_006353593.2	XP_006353655.1	<i>Petunia hybrida</i>	Primary active transporter of strigolactones

Table 4.1. Strigolactone pathway genes. The first 4 genes listed are involved in the biosynthesis of strigolactones (SLs), while the final gene, PDR1, encodes a transporter responsible for SL exudation. Gene, transcript, and protein accession numbers were retrieved from the NCBI database and are provided alongside the corresponding species for each entry.

4.3.2 Identifying orthologous in *Solanum lycopersicum*

The reference genome assembly of *S. lycopersicum* (GCF_000188115.5) was obtained from the European Nucleotide Archive (ENA) repository. To identify orthologous sequences, the five protein sequences of SL pathway genes (Table 4.1) were aligned against the *S. lycopersicum* protein complement using the BLASTp program from the BLAST+ package (Camacho *et al.*, 2009) with default parameters. Initial filtering of the alignments was based on a minimum identity percentage of identity of 65%. Subsequently, a manual selection process was performed to identify the best alignment by evaluating key metrics such as "alignment length", "mismatches", "query start", "e-value" and "bit score". The resulting five *S. lycopersicum* protein sequences were then used as queries to obtain their corresponding transcript sequences using the tBLASTn program from BLAST+ package with default settings. Similarly, genomic sequences were identified using the magicBLAST tool (Boratyn *et al.*, 2019). The same workflow was applied to retrieve both transcript and protein sequences from the *S. lycopersicum* genome annotation, version SL4.0, which was accessed through the Solgenomics database (Fernandez-Pozo *et al.*, 2015).

4.3.3 Identifying orthologous in *Solanum pennellii*

The *S. pennellii* reference genome assembly (GCF_001406875.1) was retrieved from the European Nucleotide Archive (ENA) repository. To identify the corresponding orthologous genes in *S. pennellii*, the previously identified protein sequences from *S. lycopersicum* were aligned against the *S. pennellii* protein complement, using the BLASTp tool from the BLAST+ package with default parameters. The best alignments were initially filtered by selecting those with greater than 90% identity and ensuring the alignment covered the full length of the query sequence. The resulting five protein sequences were then used as queries to obtain their corresponding transcript sequences via the tBLASTn program from the BLAST+ package with default parameters. Similarly, the genomic sequences were identified using the magicBLAST tool. The same workflow was applied to retrieve both transcript and protein sequences from the *S. pennellii* genome annotation, accessed through the Solgenomics database.

4.3.4 Comparison of SL pathway genes between *S. lycopersicum* and *S. pennellii*

Genomic sequences and transcripts from *S. lycopersicum* were compared to their *S. pennellii* counterparts using the Needleman-Wunsch global alignment algorithm. Alignments were performed using the Biostrings R package (H. Pagès, 2017) using the "pairwiseAlignment" function. This analysis was performed for all the three sequence type: genome, transcript and protein sequences. Single nucleotide polymorphisms (SNPs) identified in *S. pennellii* through these alignments were further analyzed using the

PredictSNP (Bendl *et al.*, 2016) tool to assess their potential functional impact. Dot plot of nucleotide sequences was generated using the EMBOSS dotmatcher tool (Madeira *et al.*, 2024) with the BLOSUM62 as substitution matrix, a window size of 150 and a threshold score of 23 to visualize sequence similarities and differences.

4.3.5 Primer design

PCR (polymerase chain reaction) primers were designed using Primer-BLAST (10.1186/1471-2105-13-134) based on the following criteria: primer length between 16 and 28 bp, with no more than 3 bp difference in length, melting temperatures between 52 °C and 62 °C, and GC content ranging from 40% and 60%. Primer specificity was assessed, and the absence of hairpin structures and dimer formation was verified. The sequences of the designed primers are listed in Table 4.2.

Gene	Sequence	Length	Melting temp.	GC(%)	Amplicon size	NCBI_ID: start-end	Solgenomic_ID: start-end
CCD7_penn_fw	GGGGTTGAGGGTGATGTAAAC	21	58,56	52,38	161	NC_028637.1:	Spenn-ch01:
CCD7_penn_rv	GGTGGCAAGAGTTTTGGAGG	20	59,04	55,00		84183259 - 84186993	94505240-94506270
D27_lyco_fw	TCCTTGGCTAGTTGGACCTT	20	58,26	50,00	405	XM_026032835.1:	Solyc09g065750:
D27_lyco_rv	CTGTTGCAATCTGCTTGGTG	20	57,94	50,00		1 - 1052	1 - 998
D27_penn_fw	TCCTTGGCTAGTTGGACCTT	20	58,26	50,00	324	XM_015232090.2:	NOT ANNOTATED
D27_penn_rv	CTGTTGCAATCTGCTTGGTG	20	57,94	50,00		1 - 1019	

Table 4.2. PCR primer. PCR primer sequences and their key features, along with NCBI and SolGenomics IDs. IDs are equipped with start and end coordinates indicated in the following fashion: "ID:start_coordinate:end_coordinate". lyco= *Solanum lycopersicum*; penn= *Solanum pennellii*.

4.3.6 Nucleic acid isolation, PCR amplification and Sanger sequencing

Total genomic DNA (gDNA) was extracted as described by Edwards et al. (1991) from young leaves of *S. lycopersicum* and *S. pennellii* genotypes. Total RNA was isolated from young leaves of *S. lycopersicum* and *S. pennellii* genotypes by using the ISOLATE II RNA Plant Kit (Meridian bioscience, Cincinnati, Ohio, USA) according to the method reported by the manufacturer. Total RNA (1 µg) was treated by the Tetro cDNA Synthesis Kit (Meridian bioscience) and cDNA was stored at -20°C until PCR analysis. PCR amplifications were performed, and the relative amplicons were sequenced by Eurofins Genomics (Edsberg, Germany). The sequences were assembled with CAP3 (10.1101/gr.9.9.868) and aligned with MAFFT v7.520 (<https://doi.org/10.1093/bib/bbn013>).

4.4 Results

4.4.1 SL orthologous genes in *S. lycopersicum* and *S. pennellii*

SL pathway genes were identified in both *S. lycopersicum* and *S. pennellii*, and their corresponding protein, transcripts, and genomic sequences were obtained (Table 4.3).

Gene	Genomic NCBI ID: start:end	Transcript NCBI ID	Protein NCBI ID	Genomic SolGenomic ID: start:end	SolGenomic ID	Species
CCD7	NC_015438.3: 84183259-84186993	NM_001247504.1	NP_001234433.1	SL4.0ch01: 76546455-76549956	Solyc01g090660	<i>S. lycopersicum</i>
CCD8	NC_015445.3: 55474195-55477901	XM_019215023.2	XP_019070568.1	SL4.0ch08: 53507655-53510795	Solyc08g066650	<i>S. lycopersicum</i>
D27	NC_015446.3: 64258154-64263652	XM_026032835.1	XP_025888620.1	SL4.0ch09: 60124430-60129781	Solyc09g065750	<i>S. lycopersicum</i>
MAX1	NC_015445.3: 52439564-52442178	XM_004245037.4	XP_004245085.1	SL4.0chr08: 50478269-50480262	Solyc08g062950	<i>S. lycopersicum</i>
PDR1	NC_015443.3: 41186221-41195012	XM_010324437.3	XP_010322739.1	SL4.0ch08: 54715872-54722366	Solyc08g067610	<i>S. lycopersicum</i>
CCD7	NC_028637.1: 94505241-94509008	XM_015231903.2	XP_015087389.1	Spenn-ch01: 94505240-94506270	Sopen01g03639 0	<i>S. pennellii</i>
CCD8	NC_028644.1: 59492005-59495008	XM_015229165.2	XP_015084651.1	Spenn- ch08:59491880- 59495008	Sopen08g02052 0	<i>S. pennellii</i>
D27	NC_028645.1: 75349892-75355970	XM_015232090.2	XP_015087576.1	Spenn-ch09: 75351082-75356078	Not Annotated	<i>S. pennellii</i>
MAX1	NC_028644.1: 56330361-56332855	XM_015227839.2	XP_015083325.1	Spenn-ch08: 56330293-56332587	Sopen08g01850 0	<i>S. pennellii</i>
PDR1	NC_028642.1: 51655533-51664462	XM_015223928.2	XP_015079414.1	Spenn-ch08: 60823616-60830158	Sopen08g02156 0	<i>S. pennellii</i>

Table 4.3. Accession number regarding orthologous sequences of SLs genes. Accession number regarding orthologous protein, transcripts and genomic sequences of SLs genes. Genomic sequences were reported indicating their chromosome of belonging and the relative coordinates through the following fashion: "ID:start_coordinate:end_coordinate".

4.4.2 Searching for polymorphisms between *S. lycopersicum* and *S. pennellii*

Alignments based on the annotation retrieved from NCBI database did not reveal significant polymorphisms among the SL pathway genes of the two species, consistent with the results predicted by PredictSNP (Supplementary Table 4.1, Supplementary Table 4.2). In contrast, alignments performed using the annotation retrieved from SolGenomic database highlighted an intriguing 81-nucleotide insertion in the D27 transcript sequence of *S. pennellii* (Figure 4.2), along with several additional polymorphisms. The D27 gene presents a unique case in this analysis, as it is not annotated in the genome retrieved by Solgenomics. As a consequence, for this gene, transcript and protein alignments between the two *Solanum spp.* were performed using the annotation retrieved from SolGenomic for *S. lycopersicum* and the annotation retrieved from NCBI for *S. pennellii*. Moreover, among the other polymorphisms, only one was deemed significant: a premature stop codon located immediately after the ATG start codon in the CCD7 gene sequence of *S. pennellii* (Figure 4.3).

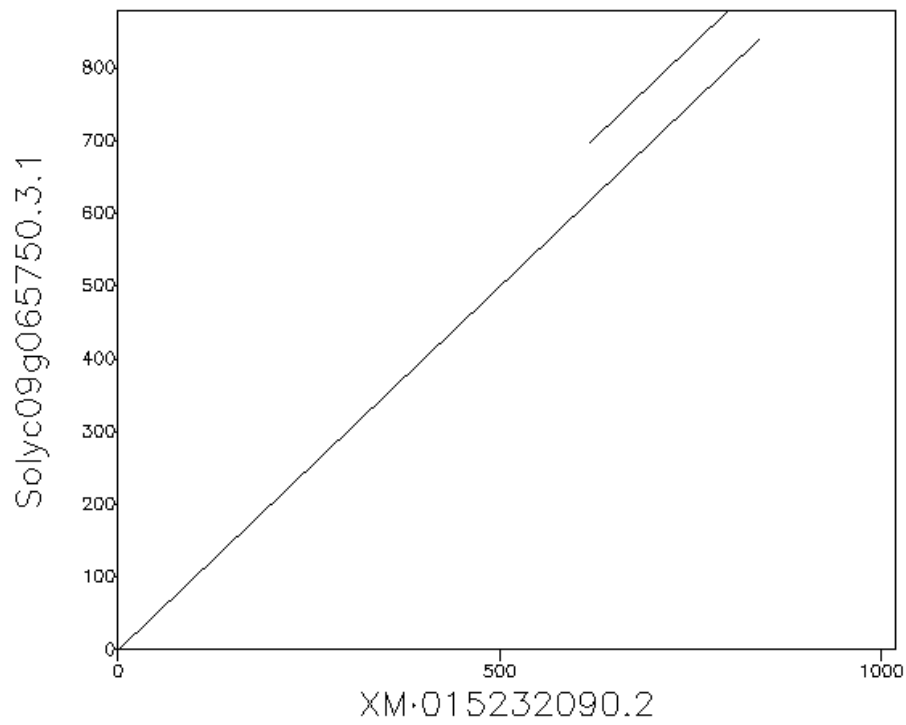


Figure 4.2. Dot plot between *S.lycopersicum* D27 gene (Solyc09g065750) and *S. pennellii* D27 gene (XM_015232090.2). The plot clearly shows the presence of an 81-nucleotide insertion/deletion (indel) at the end of the *S. pennellii* sequence.

>Sopen01g036390.1

MDL*FVSLPPNSKTKAKMQAKACHNINNIPPKLLPPAKLPSTVAMSPSQLTLPSHVARAI
TITTSPTHEVYTPEIDDTVTAYWDYQFLFVSQRSEATEPVSLRVVEGSIPSDFPSGTYYL
TGPGLFADDHGSTVHPLDGHGYLRTFEIDGSTGQVKFMARYIETEAQTEERDPVSGKWRF
THRGPFVSLKGGKMVGNTKVMKNVANTSVLQWGGRLFCLEGGDPYEIDSKTLNLTGKYE
LIKNSDQVLEDKKISHSDFLDVAAQILKPILYGVFKMSPKRLLSHYKIDTRRNRLIMSC
NAEDMLLPRSNFTFYEFDSNFQLLQSQEFEPDHLMIHDWAFTDTHYILFGNRIKLDIPG
SMTAVCGLSPMISALSVNPSKSTSPIYLLPRFPNNNVERDWRKPIEAPSQIWLHVGNFAF
EEIDEQNGNLNIQIQASGCSYQWFNFQKMGYDWQSGKLDPSMMNVEEGEEKLLPHLVQV
CINLDDKKGNCTKCSVNDLNPKNKAADFAMNPEFSGRKNRYIYAATCTGSRQALPHFPF
DTVVKLNAVDKSVQKWSAGRRRFIPEPVFI PRGTNKEDDGYLLVVEYAVSTQRCYLVILD
AQKIGEKNEVVARLEVPRHLNFPLGFHGFWAPTNSSLANLQKIESKCKNSWSMMKDNMVK
LGQ

Figure 4.3. Protein sequence of the *S. pennellii* CCD7 gene (Sopen01g036390.1). The premature stop codon, located immediately after the ATG start codon, is highlighted in red.

4.4.3 *In vivo* validation of detected polymorphisms

PCR experiments confirmed the presence of an 81-nucleotide indel in the *S. pennellii* D27 gene, resulting in two amplicons of different lengths (less than 100 bp). However, the premature stop codon immediately following the ATG codon in the *S. pennellii* CCD7 gene could not be confirmed due to non-specific amplification. Sequencing of the D27 amplicons further validated the 81-bp indel (Figure 4.4).

	780	790	800	810	820	830	840	850	860	870	880	890	900	910	920	930	940
Solyc09g065750.3.1/1-988	GATCCTCCTCCACTAGATAAAGATCCAGCATTTAGGCAACCTTGCTACAAACTATGTAACCTTTTTCATCTACCATAATTATTTTCAGAAAAAACATATTATATTAGTCATGCCCTGAGTATGTTCTTCTTCATGCAGGCAAATTAACAATAAACACCAAGCAGA1																
XM_015232090.2_mRNA/1-1019	GATCCTCCTCCACTAGATACAGATCCAGCATTTATGCAACCTTGCTACAAACTAT-----GCAAATTAACAATAAACACCAAGCAGA1																
D27_23A5ZAA019_23A5ZAA018.fa.cap.contigs:710-1028/1-318	GATCCTCCTCCACTAGATACAGATCCAGCATTTATGCAACCTTGCTACAAACTAT-----GCAAATTAACAATAAACACCAAGCAGA1																
D27_23A5ZAA017_23A5ZAA017.fa.cap.contigs:710-1028/1-318	GATCCTCCTCCACTAGATACAGATCCAGCATTTATGCAACCTTGCTACAAACTAT-----GCAAATTAACAATAAACACCAAGCAGA1																
D27_23A5ZAA015_23A5ZAA014.fa.cap.contigs:720-1038/1-318	GATCCTCCTCCACTAGATACAGATCCAGCATTTATGCAACCTTGCTACAAACTAT-----GCAAATTAACAATAAACACCAAGCAGA1																

Figure 4.4. Multiple alignment of D27 gene sequences. This figure presents a multiple alignment of the *S. lycopersicum* D27 gene (Solyc09g065750), the *S. pennellii* D27 gene (XM_015232090.2), and three sequenced amplicons, focusing on the 3' region of the D27 gene. The aligned sequences confirm the in silico predicted 81-nucleotide insertion/deletion (indel) within the gene.

4.5 Discussions

Cultivated *S. lycopersicum* is highly susceptible to infection by parasitic plants from the *Orobanchaceae* family, particularly, those in the *Phelipanche* genus, which is a serious threat in the Mediterranean region. Interestingly, one of *S. lycopersicum* wild relatives, *S. pennellii*, exhibits tolerance to *P. aegyptiaca* infection (Minuto *et al.*, 2006; Cuccurullo, Nicolia and Cardi, 2022). To dissect the genetic basis of *S. pennellii* tolerance to parasitic plants, a promising approach is the analysis of polymorphisms in SL biosynthetic pathway genes between *S. pennellii* and *S. lycopersicum*.

SLs are crucial plant hormones that regulate various physiological processes, and throughout the co-evolution of host plants and *Orobanchaceae* parasitic weeds, they have taken on the additional role of germination stimulants for *Orobanchaceae* seeds (Waters *et al.*, 2017). As a result, SLs play a pivotal role in the first line of plant defense against *Phelipanche* spp., i.e. acting before the attachment of the parasite's haustorium organ to the plant roots (Fishman and Shirasu, 2021; Albanova *et al.*, 2023). Understanding the genetic variations in SL biosynthesis between these species may provide valuable insights into the mechanisms of parasite tolerance.

SL biosynthetic pathway is a branch of carotenoid pathway, beginning from the β -carotene. Four key enzymes are involved in the conversion of β -carotene into the first SL molecule, 5-deoxystrigol: D27, CCD7, CCCD8 and MAX1 (Seto *et al.*, 2012; Seto and Yamaguchi, 2014; Mashiguchi, Seto and Yamaguchi, 2021). Another crucial protein, PDR1, is involved in the transport of SLs into the rhizosphere (Kretzschmar *et al.*, 2012).

Interestingly, seed of *P. aegyptiaca* exhibit low germination rate when exposed to root exudates from *S. pennellii* (Bai *et al.*, 2020; Cuccurullo, Nicolia and Cardi, 2022). Therefore, this chapter presents a preliminary investigation focused on identifying nucleotide polymorphisms in the genes involved in the biosynthesis and translocation of SLs molecules between *S. pennellii* and cultivated *S. lycopersicum*. Understanding these polymorphisms may offer insights into the mechanisms of *S. pennellii* tolerance to *P. aegyptiaca* infection. The key findings of this investigation focus on the first two genes in the strigolactone biosynthetic pathway: D27 and CCD7. In *S. pennellii*, the D27 gene contains an indel of 81 nucleotides, potentially resulting in the production of a protein that is 27 amino acids shorter than its ortholog in *S. lycopersicum*. This variation was predicted *in silico* and subsequently confirmed through PCR amplification, followed by Sanger sequencing of the target region.

Sequence variations in proteins can lead to a range of outcomes, such changes may either confer a different function to the altered protein or leave its function intact, provided that the chemical and thermodynamic stability of the protein structure is maintained. However, if the variation destabilizes the protein structure, it may render the protein non-functional (Tokuriki and Tawfik, 2009; Soskine and Tawfik, 2010).

This indel in *D27* could thus have significant implications for the protein role in the strigolactone biosynthesis pathway, potentially contributing to *S. pennellii*'s tolerance to parasitic plants. Further research is needed to explore the functional impact of this structural change. As for the *CCD7* gene, a premature stop codon was predicted to occur immediately after the ATG start codon. However, PCR amplification produced a series of non-specific amplicons, likely due to the low specificity of the forward primer used. This primer targets a low complexity region upstream the ATG codon. To improve the specificity of amplification and confirm the presence of this premature stop codon, further experimental design and primer optimization are needed.

If confirmed, this premature stop codon would result in the production of a truncated protein, which would likely be non-functional and unable to carry out the same enzymatic role as its orthologs in *S. lycopersicum*. This could have significant implications for strigolactone biosynthesis and the associated tolerance mechanisms in *S. pennellii*.

The results of this preliminary study shed light on previously unknown polymorphisms that may play a role in tolerance mechanisms of *S. pennellii*. Understanding the spectrum of polymorphisms in SL pathway genes could provide valuable insights into the underlying resistance mechanism. This knowledge has the potential to contribute significantly to expanding the pool of useful alleles for future breeding and mutagenesis programs, ultimately enhancing the development of crop varieties with improved resistance to parasitic plants.

Previous studies have demonstrated that root exudates of *S. pennellii* inhibit the germination of *P. ramosa* seeds (El-Halmouch, Benharrat and Thalouarn, 2006; Bai *et al.*, 2020; Cuccurullo, Nicolia and Cardi, 2022). However, this is the first investigation where genomic, transcript and protein sequences of SL pathway genes from *S. pennellii* and *S. lycopersicum* have been retrieved, compared, and analyzed using both available genome annotations. The study employs a combination of in silico prediction methods and experimental validation to explore these orthologous.

The identified polymorphisms could lead to variations in the final enzymes, potentially resulting in lower SL levels exuded into the soil. This outcome has both positive and negative implications. On the positive side, the germination rate of parasitic plants may decrease, reducing yield losses in cultivated *Solanum spp.* However, as collateral effect, the symbiosis with mycorrhizae may be compromised, limiting the plant's access to nutrients that are difficult to reach with roots alone (García-Garrido *et al.*, 2009). To fully understand the connection between these findings and *S. pennellii* resistance mechanisms, further research is needed to investigate the functional impact of polymorphisms in the protein products of the *D27* and *CCD7* genes.

This study represents an important first step in uncovering the complex defense mechanisms of *S. pennellii*, paving the way for more detailed analyses of the genetic basis of its tolerance to parasitic plants.

4.6 Conclusions

The first two genes in the SL biosynthetic pathway, namely *D27* and *CCD7*, in *S. pennellii* exhibit notable polymorphisms compared to their orthologs in *S. lycopersicum*. Specifically, an 81-nucleotide indel in *D27* was confirmed through PCR experiment and Sanger sequencing. This indel may result in alterations to the gene coding sequence, potentially affecting its function within the strigolactone biosynthetic pathway and consequently influencing plant growth and development.

In the case of *CCD7*, a premature stop codon occurs immediately after the ATG start codon, likely resulting in a non-functional *CCD7* protein. This impairment could disrupt strigolactone production and subsequently affect plant physiology and interactions. This chapter underscores the presence of polymorphisms in the SL pathway genes of *S. pennellii*, which may serve as a foundation for investigating the genetic bases that underpin the molecular of tolerance to *Phelipanche spp.* infection.

Chapter 5. Concluding Remarks

Dual RNA-seq is a powerful technique for studying interaction between organisms, particularly in regions of close contact established during the infection process. One of the primary difficulties, in this approach, is managing cross-mapped reads, where reads from one organism map well to the genome of the other or to both. Dual RNA-seq can be applied to “plant – plant” interactions, giving the opportunity to separate mixed reads *in silico*. In this work, to assess the feasibility of this technique, a simulation was conducted using artificial datasets that mimicked two distinct host-pathogen interactions involving different hosts and the same parasitic plant. Specifically, we simulated interactions between *A. thaliana*-*C. campestris* and *S. lycopersicum*-*C. Campestris*. This assessment revealed the following:

- dual RNA-seq can be applied to “host plant – parasitic plant” interactions, allowing for reliable *in silico* separation of mixed reads;
- cross-mapped reads, arising from homologous sequences between the two genomes, do not introduce significant misleading information;
- both the sequential and combined approaches are equally trustworthy. However, the combined approach is less time-consuming and slightly reduces the cross-mapping phenomenon.

Given the absence of existing packages for dual RNA-seq analysis, in this work, it was developed inDAGO, a GUI that facilitates dual RNA-seq analysis and includes an extension for bulk RNA-seq. This GUI automates the analysis process and provide step-by-step guidance for end users. No prior programming knowledge is required, allowing users to easily navigate between different analysis steps and download intermediate results.

The software integrates a suite of R packages to facilitate various analytical steps. In cases where existing packages do not meet specific needs, we have developed *ad hoc* R code.

To evaluate the performance and reliability of inDAGO, it was tested using public datasets simulating a host-parasite interaction between two plant species. The results demonstrated that all modules performed satisfactorily.

Lastly, a preliminary investigation into polymorphisms of strigolactone pathway genes was conducted between *Solanum lycopersicum* and *Solanum pennellii* to dissect the genetic basis of tolerance mechanisms. As a result, we identified two interesting polymorphisms into 2 key genes of the pathway: D27 and CCD7. The D27 gene contains an indel of 81 nucleotides, while the CCD7 gene has a premature STOP located immediately after the ATG codon.

Looking ahead, it is important to acknowledge that plant-plant interactions, such as those between *Solanum* species infected by *Orobanchaceae* parasitic weeds, can be effectively studied using dual RNA-seq. The new application, inDAGO, offers an accessible platform for such investigations, broadening its use within the scientific community. With recent advancements, future improvements to inDAGO may include extensions for single-cell RNA-seq analysis and a module to distinguish between coding and non-coding RNA-molecules. These enhancements would enable deeper transcriptomic analysis, providing a more detailed understanding of the parasitism phenomenon from a molecular biology perspective.



Supplementary tables

Species	Order	Classification	NCBI Protein accession number
<i>Allium cepa</i>	<i>Asparagales</i>	Crop	YP_009057159.1
<i>Allium sativum</i>	<i>Asparagales</i>	Crop	YP_009317125.1
<i>Asparagus officinalis</i>	<i>Asparagales</i>	Crop	YP_009370029.1
<i>Capsicum annuum</i>	<i>Solanales</i>	Crop	YP_006666038.1
<i>Carthamus tinctorius</i>	<i>Asterales</i>	Crop	YP_009271914.1
<i>Cucumis melo</i>	<i>Cucurbitales</i>	Crop	YP_010167036.1
<i>Cucumis sativus</i>	<i>Cucurbitales</i>	Crop	YP_247607.1
<i>Daucus carota</i>	<i>Apiales</i>	Crop	YP_740125.1
<i>Ipomoea batatas</i>	<i>Solanales</i>	Crop	YP_009128399.1
<i>Solanum lycopersicum</i>	<i>Solanales</i>	Crop	YP_008563096.1
<i>Solanum melongena</i>	<i>Solanales</i>	Crop	YP_009254086.1
<i>Solanum tuberosum</i>	<i>Solanales</i>	Crop	YP_635647.1
<i>Arabidopsis thaliana</i>	<i>Brassicales</i>	ModelOrganism	NP_051067.1
<i>Campsis radicans</i>	<i>Scrophulariales</i>	Ornamental	YP_010743242.1
<i>Catharanthus roseus</i>	<i>Gentianales</i>	Ornamental	YP_008081273.1
<i>Chrysanthemum indicum</i>	<i>Asterales</i>	Ornamental	YP_007474737.1
<i>Foeniculum vulgare</i>	<i>Apiales</i>	Ornamental	YP_009235887.
<i>Geranium Maderense</i>	<i>Geraniales</i>	Ornamental	YP_009250234.1
<i>Hedera helix</i>	<i>Apiales</i>	Ornamental	YP_010633192.1
<i>Impatiens davidii</i>	<i>Ericales</i>	Ornamental	YP_010209555.1
<i>Mentha longifolia</i>	<i>Lamiales</i>	Ornamental	YP_009327395.1
<i>Plectranthus Scutellarioides</i>	<i>Lamiales</i>	Ornamental	YP_010174095.1
<i>Cuscuta australis</i>	<i>Solanales</i>	Parasite	YP_009727865.1
<i>Cuscuta campestris</i>	<i>Solanales</i>	Parasite	YP_009996226.1

<i>Cuscuta europaea</i>	<i>Solanales</i>	Parasite	YP_010621274.1
<i>Cuscuta japonica</i>	<i>Solanales</i>	Parasite	YP_010273552.1
<i>Cuscuta pentagona</i>	<i>Solanales</i>	Parasite	YP_009532296.1
<i>Cuscuta reflexa</i>	<i>Solanales</i>	Parasite	YP_001430120.1
<i>Amaranthus blitoides</i>	<i>Caryophyllales</i>	Weeds	YP_010464873.1
<i>Chenopodium album</i>	<i>Caryophyllales</i>	Weeds	YP_009380220.1
<i>Convolvulus arvensis</i>	<i>Convolvulus</i>	Weeds	YP_010042688.1
<i>Fallopia convolvulus</i>	<i>Caryophyllales</i>	Weeds	YP_010941064.1
<i>Portulaca oleracea</i>	<i>Caryophyllales</i>	Weeds	YP_009441448.1
<i>Solanum nigrum</i>	<i>Solanales</i>	Weeds	YP_009171876.1

Supplementary Table 2.1. Protein sequences accession numbers involved in the phylogenetic analysis. List of NCBI accession numbers for the *rbcL* proteins of *Cuscuta* spp. and their host plants. Six *Cuscuta* species were selected, along with 28 hosts spanning twelve different orders, categorized as crops, ornamentals, weeds, and model organisms.

Interaction	Sequence origin	Replicate	One side cross	Two side cross	Total cross mapping
<i>Arabidopsis-Cuscuta</i>	<i>Arabidopsis</i>	Rep1	39	57600	57639
<i>Arabidopsis-Cuscuta</i>	<i>Arabidopsis</i>	Rep2	756	1335770	1336526
<i>Arabidopsis-Cuscuta</i>	<i>Arabidopsis</i>	Rep3	294	84324	84618
<i>Arabidopsis-Cuscuta</i>	<i>Cuscuta</i>	Rep1	48495	64729	113224
<i>Arabidopsis-Cuscuta</i>	<i>Cuscuta</i>	Rep2	58147	79993	138140
<i>Arabidopsis-Cuscuta</i>	<i>Cuscuta</i>	Rep3	56618	74735	131353
<i>Solanum-Cuscuta</i>	<i>Solanum</i>	Rep1	156187	165748	321935
<i>Solanum-Cuscuta</i>	<i>Solanum</i>	Rep2	110334	121976	232310
<i>Solanum-Cuscuta</i>	<i>Solanum</i>	Rep3	129273	149853	279126
<i>Solanum-Cuscuta</i>	<i>Cuscuta</i>	Rep1	42288	40665	82953
<i>Solanum-Cuscuta</i>	<i>Cuscuta</i>	Rep2	51084	50276	101360
<i>Solanum-Cuscuta</i>	<i>Cuscuta</i>	Rep3	50082	46957	97039

Supplementary Table 2.2. Sequential approach cross-mapping. Summary of both one-side and two-side cross-mapped reads in the sequential approach.

Reference genome	Replicate	Order mapping	Mapped reads	<i>A. thaliana</i> mapped reads	<i>C. campestris</i> mapped reads
<i>A. thaliana</i>	Replicate 1	Host; parasite	19013844	18900620 (99.4045%)	113224 (0.5955%)
<i>A. thaliana</i>	Replicate 2	Host; parasite	19720924	19582784 (99.2995%)	138140 (0.7005%)
<i>A. thaliana</i>	Replicate 3	Host; parasite	20694428	20563075 (99.3653%)	131353 (0.6347%)
<i>C. campestris</i>	Replicate 1	Host; parasite	14113749	39 (0.0003%)	14113710 (99.9997%)
<i>C. campestris</i>	Replicate 2	Host; parasite	14080490	739 (0.0052%)	14079751 (99.9948%)
<i>C. campestris</i>	Replicate 3	Host; parasite	14479566	288 (0.0020%)	14479278 (99.9980%)
<i>A. thaliana</i>	Replicate 1	Parasite; host	18891535	18843058 (99.7434%)	48477 (0.2566%)
<i>A. thaliana</i>	Replicate 2	Parasite; host	18305310	18247182 (99.6825%)	58128 (0.3175%)
<i>A. thaliana</i>	Replicate 3	Parasite; host	20535446	20478852 (99.7244%)	56594 (0.2756%)
<i>C. campestris</i>	Replicate 1	Parasite; host	14236067	57639 (0.4049%)	14178428 (99.5951%)
<i>C. campestris</i>	Replicate 2	Parasite; host	15496252	1336526 (8.6248%)	14159726 (91.3752%)
<i>C. campestris</i>	Replicate 3	Parasite; host	14638652	84618 (0.5780%)	14554034 (99.4220%)

Supplementary Table 2.3. *A. thaliana* and *C. campestris* cross-mapping in sequential approach. Statistics on the number of correctly assigned reads and cross-mapped reads (both one-side and two-side) in the sequential approach applied to *A. thaliana* and *C. campestris*.

Reference genome	Replicate	Order mapping	Mapped reads	<i>S. lycopersicum</i> mapped reads	<i>C. campestris</i> mapped reads
<i>S. lycopersicum</i>	Replicate 1	Host; parasite	35901311	35818358 (99.77%)	82953 (0.23%)
<i>S. lycopersicum</i>	Replicate 2	Host; parasite	35666032	35564672 (99.72%)	101360 (0.28%)
<i>S. lycopersicum</i>	Replicate 3	Host; parasite	40998162	40901123 (99.76%)	97039 (0.24%)
<i>C. campestris</i>	Replicate 1	Host; parasite	14294028	156334 (1.09%)	14137694 (98.91%)
<i>C. campestris</i>	Replicate 2	Host; parasite	14219604	110439 (0.78%)	14109165 (99.22%)
<i>C. campestris</i>	Replicate 3	Host; parasite	14636927	129337 (0.88%)	14507590 (99.12%)
<i>S. lycopersicum</i>	Replicate 1	Parasite; host	35695469	35652944 (99.88%)	42525 (0.12%)
<i>S. lycopersicum</i>	Replicate 2	Parasite; host	35494027	35442951 (99.86%)	51076 (0.14%)
<i>S. lycopersicum</i>	Replicate 3	Parasite; host	40801144	40751577 (99.88%)	49567 (0.12%)
<i>C. campestris</i>	Replicate 1	Parasite; host	14500240	321935 (2.22%)	14178305 (97.78%)
<i>C. campestris</i>	Replicate 2	Parasite; host	14391747	232310 (1.61%)	14159437 (98.39%)
<i>C. campestris</i>	Replicate 3	Parasite; host	14833676	279126 (1.88%)	14554550 (98.12%)

Supplementary Table 2.4. *S. lycopersicum* and *C. campestris* cross-mapping in sequential approach. Statistics on the number of correctly assigned reads and cross-mapped reads (both one-side and two-side) in the sequential approach applied to *S. lycopersicum* and *C. campestris*.

Interaction	Sequence origin	Replicate	One side cross	Two side cross	Total cross mapping
<i>Arabidopsis-Cuscuta</i>	<i>Arabidopsis</i>	Rep1	128	730	858
<i>Arabidopsis-Cuscuta</i>	<i>Arabidopsis</i>	Rep2	1517	12249	13766
<i>Arabidopsis-Cuscuta</i>	<i>Arabidopsis</i>	Rep3	772	1390	2162
<i>Arabidopsis-Cuscuta</i>	<i>Cuscuta</i>	Rep1	3065	979	4044
<i>Arabidopsis-Cuscuta</i>	<i>Cuscuta</i>	Rep2	5187	1334	6521
<i>Arabidopsis-Cuscuta</i>	<i>Cuscuta</i>	Rep3	5147	1441	6588
<i>Solanum-Cuscuta</i>	<i>Solanum</i>	Rep1	241173	6468	247641
<i>Solanum-Cuscuta</i>	<i>Solanum</i>	Rep2	169824	4408	174232
<i>Solanum-Cuscuta</i>	<i>Solanum</i>	Rep3	201261	5676	206937
<i>Solanum-Cuscuta</i>	<i>Cuscuta</i>	Rep1	3345	1724	5069
<i>Solanum-Cuscuta</i>	<i>Cuscuta</i>	Rep2	5634	2592	8226
<i>Solanum-Cuscuta</i>	<i>Cuscuta</i>	Rep3	5650	2631	8281

Supplementary Table 2.5. Cross-mapping in sequential approach. Summary of both one-side and two-side cross-mapped reads in the combined approach.

Reference genome	Replicate	Unambiguously mapped reads	Host mapped reads	Parasite mapped reads
<i>A. thaliana</i>	Replicate 1	18902574	18899509 (99.984%)	3065 (0.016%)
<i>A. thaliana</i>	Replicate 2	19573198	19568011 (99.973%)	5187 (0.027%)
<i>A. thaliana</i>	Replicate 3	20565759	20560612 (99.975%)	5147 (0.025%)
<i>C. campestris</i>	Replicate 1	14176961	128 (0.001%)	14176833 (99.999%)
<i>C. campestris</i>	Replicate 2	14158715	1517 (0.011%)	14157198 (99.989%)
<i>C. campestris</i>	Replicate 3	14552459	772 (0.005%)	14551687 (99.995%)
<i>S. lycopersicum</i>	Replicate 1	35713672	35710327 (99.991%)	3345 (0.009%)
<i>S. lycopersicum</i>	Replicate 2	35504641	35499007 (99.984%)	5634 (0.016%)
<i>S. lycopersicum</i>	Replicate 3	40837378	40831728 (99.986%)	5650 (0.014%)
<i>C. campestris</i>	Replicate 1	14415956	241173 (1.673%)	14174783 (98.327%)
<i>C. campestris</i>	Replicate 2	14324477	169824 (1.186%)	14154653 (98.814%)
<i>C. campestris</i>	Replicate 3	14750971	201261 (1.364%)	14549710 (98.636%)

Supplementary Table 2.6. Cross-mapping in combined approach. Statistics on the number of correctly assigned reads and cross-mapped reads (both one-side and two-side) in the combined approach.

Interaction	Reference genome	Labeled reads	Replicate	Library	Input reads	Uniquely mapped	Multiple mapped	Unmapped
<i>Arabidopsis-Cuscuta</i>	<i>A. thaliana</i>	S-FNh2	Replicate 1	SRR22559142	39	3 (7.7%)	25 (64.1%)	11 (28.2%)
<i>Arabidopsis-Cuscuta</i>	<i>A. thaliana</i>	S-FNh2	Replicate 2	SRR22559143	739	75 (10.1%)	98 (13.3%)	566 (76.6%)
<i>Arabidopsis-Cuscuta</i>	<i>A. thaliana</i>	S-FNh2	Replicate 3	SRR22559144	288	67 (23.3%)	136 (47.2%)	85 (29.5%)
<i>Arabidopsis-Cuscuta</i>	<i>C. campestris</i>	S-FNp2	Replicate 1	SRR12763776	48477	64 (0.1%)	821 (1.7%)	47592 (98.2%)
<i>Arabidopsis-Cuscuta</i>	<i>C. campestris</i>	S-FNp2	Replicate 2	SRR12763787	58128	110 (0.9%)	748 (1.3%)	57270 (98.9%)
<i>Arabidopsis-Cuscuta</i>	<i>C. campestris</i>	S-FNp2	Replicate 3	SRR12763788	56594	82 (0.1%)	1166 (2.1%)	55346 (97.8%)
<i>Solanum-Cuscuta</i>	<i>S. lycopersicum</i>	S-FNh2	Replicate 1	SRR25558913	156334	74898 (47.9%)	18183 (11.6%)	63253 (40.5%)
<i>Solanum-Cuscuta</i>	<i>S. lycopersicum</i>	S-FNh2	Replicate 2	SRR25558914	110439	54383 (49.2%)	12370 (11.2%)	43686 (39.6%)
<i>Solanum-Cuscuta</i>	<i>S. lycopersicum</i>	S-FNh2	Replicate 3	SRR25558915	129337	63779 (49.3%)	13338 (10.3%)	52220 (40.4%)
<i>Solanum-Cuscuta</i>	<i>C. campestris</i>	S-FNp2	Replicate 1	SRR12763776	42525	60 (0.1%)	839 (2.0%)	41626 (97.9%)
<i>Solanum-Cuscuta</i>	<i>C. campestris</i>	S-FNp2	Replicate 2	SRR12763787	51076	107 (0.2%)	911 (1.8%)	50058 (98.0%)
<i>Solanum-Cuscuta</i>	<i>C. campestris</i>	S-FNp2	Replicate 3	SRR12763788	49567	82 (0.2%)	749 (1.5%)	48736 (98.3%)

Supplementary Table 2.7. One-side cross-mapping in sequential approach. Remapping result of one-side cross-mapped read to their respective genomes for the sequential approach involving *A. thaliana*-*C. campestris* and *S. lycopersicum*-*C. campestris* interactions.

Interaction	Reference genome	Labeled reads	Replicate	Library	Input reads	Uniquely mapped	Multiple mapped	Unmapped
<i>Arabidopsis-Cuscuta</i>	<i>A. thaliana</i>	C-FNh	Replicate 1	SRR22559142	128	44 (34.4%)	78 (60.9%)	6 (4.7%)
<i>Arabidopsis-Cuscuta</i>	<i>A. thaliana</i>	C-FNh	Replicate 2	SRR22559143	1517	549 (36.2%)	599 (39.5%)	369 (24.3%)
<i>Arabidopsis-Cuscuta</i>	<i>A. thaliana</i>	C-FNh	Replicate 3	SRR22559144	772	297 (38.5%)	404 (52.3%)	71 (9.2%)
<i>Arabidopsis-Cuscuta</i>	<i>C. campestris</i>	C-FNp	Replicate 1	SRR12763776	3065	477 (15.6%)	21 (0.7%)	2567 (83.8%)
<i>Arabidopsis-Cuscuta</i>	<i>C. campestris</i>	C-FNp	Replicate 2	SRR12763787	5187	780 (15.0%)	43 (0.8%)	4364 (84.1%)
<i>Arabidopsis-Cuscuta</i>	<i>C. campestris</i>	C-FNp	Replicate 3	SRR12763788	5147	713 (13.9%)	37 (0.7%)	4397 (85.4%)
<i>Solanum-Cuscuta</i>	<i>S. lycopersicum</i>	C-FNh	Replicate 1	SRR25558913	241173	147538 (61.2%)	37029 (15.4%)	56606 (23.5%)
<i>Solanum-Cuscuta</i>	<i>S. lycopersicum</i>	C-FNh	Replicate 2	SRR25558914	169824	104429 (61.5%)	25480 (15.0%)	39915 (23.5%)
<i>Solanum-Cuscuta</i>	<i>S. lycopersicum</i>	C-FNh	Replicate 3	SRR25558915	201261	125992 (62.6)	28100 (14.0%)	47169 (23.4%)
<i>Solanum-Cuscuta</i>	<i>C. campestris</i>	C-FNp	Replicate 1	SRR12763776	3345	526 (15.7%)	32 (1.0%)	2787 (83.3%)
<i>Solanum-Cuscuta</i>	<i>C. campestris</i>	C-FNp	Replicate 2	SRR12763787	5634	865 (15.4%)	42 (0.7%)	4727 (83.9%)
<i>Solanum-Cuscuta</i>	<i>C. campestris</i>	C-FNp	Replicate 3	SRR12763788	5650	818 (14.5%)	44 (0.8%)	4788 (84.7%)

Supplementary Table 2.8. One-side cross-mapping in combined approach. Remapping result of one-side cross-mapped read to their respective genomes for combined approach involving *A. thaliana*-*C. campestris* and *S. lycopersicum*-*C. campestris* interactions.

Chr	Parent	Product
A. thaliana chloroplast (NC_000932.1)	ArthCr086	16S ribosomal RNA
A. thaliana chloroplast (NC_000932.1)	ArthCr088	16S ribosomal RNA
A. thaliana chloroplast (NC_000932.1)	ArthCr087	23S ribosomal RNA
A. thaliana chloroplast (NC_000932.1)	ArthCr089	23S ribosomal RNA
A. thaliana chloroplast (NC_000932.1)	ArthCr091	5S ribosomal RNA
A. thaliana chloroplast (NC_000932.1)	ArthCp068	NADH dehydrogenase subunit 2
A. thaliana chloroplast (NC_000932.1)	ArthCp086	NADH dehydrogenase subunit 2
A. thaliana chromosome 1 (NC_003070.9)	AT1G28710	Nucleotide-diphospho-sugar transferase family protein
A. thaliana chromosome 1 (NC_003070.9)	AT1G28720	tRNA-Pro
A. thaliana chromosome 2 (NC_003071.7)	AT2G01010	18S ribosomal RNA
A. thaliana chromosome 2 (NC_003071.7)	AT2G01020	5S ribosomal RNA
A. thaliana chromosome 2 (NC_003071.7)	AT2G03865	uncharacterized misc_RNA
A. thaliana chromosome 2 (NC_003071.7)	AT2G03875	uncharacterized misc_RNA
A. thaliana chromosome 2 (NC_003071.7)	AT2G01180	phosphatidic acid phosphatase 1
A. thaliana chromosome 2 (NC_003071.7)	AT2G01275	RING/FYVE/PHD zinc finger superfamily protein

A. thaliana chromosome 2 (NC_003071.7)	AT2G01275	RING/FYVE/PHD zinc finger superfamily protein_142610
A. thaliana chromosome 2 (NC_003071.7)	AT2G01570	GRAS family transcription factor family protein
A. thaliana chromosome 2 (NC_003071.7)	AT2G01820	Leucine-rich repeat protein kinase family protein
A. thaliana chromosome 2 (NC_003071.7)	AT2G20410	RNA-binding ASCH domain protein
A. thaliana chromosome 2 (NC_003071.7)	AT2G07733	pseudogene of NADH dehydrogenase 2B~AT2G07733.1
A. thaliana chromosome 2 (NC_003071.7)	AT2G07709	pseudogene of NADH dehydrogenase subunit 7~AT2G07709.1
A. thaliana chromosome 2 (NC_003071.7)	AT2G07717	pseudogene of NADH dehydrogenase subunit 4~AT2G07717.1
A. thaliana chromosome 2 (NC_003071.7)	AT2G07731	pseudogene of NADH dehydrogenase 6~AT2G07731.1
A. thaliana chromosome 2 (NC_003071.7)	AT2G07711	pseudogene of NADH dehydrogenase 5B~AT2G07711.1
A. thaliana chromosome 3 (NC_003074.8)	AT3G33075	uncharacterized proteintrue_13640060
A. thaliana chromosome 3 (NC_003074.8)	AT3G41768	18S ribosomal RNA
A. thaliana chromosome 3 (NC_003074.8)	AT3G06365	uncharacterized misc_RNA
A. thaliana chromosome 3 (NC_003074.8)	AT3G33136	uncharacterized proteintrue_13934223
A. thaliana chromosome 3 (NC_003074.8)	AT3G39230	uncharacterized proteintrue_14160795
A. thaliana chromosome 3 (NC_003074.8)	AT3G42260	uncharacterized proteintrue_14440090
A. thaliana chromosome 3 (NC_003074.8)	AT3G57765	U2.3%3B snRNA

A. thaliana chromosome 3 (NC_003074.8)	AT3G57770	Protein kinase superfamily protein_21397542
A. thaliana chromosome 3 (NC_003074.8)	AT3G42650	uncharacterized protein true_14739356
A. thaliana chromosome 4 (NC_003075.7)	AT4G37930	serine transhydroxymethyltransferase 1
A. thaliana chromosome 5 (NC_003076.8)	AT5G56030	heat shock protein 81-2
A. thaliana chromosome 5 (NC_003076.8)	AT5G60390	GTP binding Elongation factor Tu family protein
A. thaliana chromosome 5 (NC_003076.8)	AT5G61455	U2.7%3B snRNA
A. thaliana chromosome 5 (NC_003076.8)	AT5G44635	minichromosome maintenance (MCM2/3/5) family protein
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	DA397_mgr02	18S ribosomal RNA
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	DA397_mgr01	26S ribosomal RNA
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	GeneID:38088376	cosRNA M127
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	GeneID:38088290	cosRNA M198
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	GeneID:38088291	cosRNA M120
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	GeneID:38088391	cosRNA M197
A. thaliana ecotype Col-0 mitochondrion	GeneID:38088298	cosRNA M36

(NC_037304.1)

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

GeneID:38088302

cosRNA M42

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

DA397_mgp30

cosRNA M15

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

GeneID:38088395

cosRNA M29

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

DA397_mgp28

cosRNA M93

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

GeneID:38088417

cosRNA M166

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

DA397_mgp37

cosRNA M75

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

GeneID:38088375

cosRNA M128

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

GeneID:38088423

cosRNA M44

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

DA397_mgp38

NADH dehydrogenase subunit 9

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

DA397_mgp02

cytochrome c biogenesis B

A. thaliana ecotype Col-0
mitochondrion

(NC_037304.1)

DA397_mgp31

NADH dehydrogenase subunit 2

A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	DA397_mgp12	cytochrome c biogenesis FN2
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	DA397_mgp03	ATPase subunit 4

Supplementary Table 2.9. *A. thaliana* and *C. campestris* two-side cross-mapping in sequential approach. List of *A. thaliana* loci identified investigating on two-side cross-mapped reads from the *A. thaliana*-*C. campestris* interaction via the sequential approach. Chromosome (Chr), locus identifier (Parent) and description of the gene function (Product) were extracted from *A. thaliana* genome annotation (gff3 file) were reported.

Chr	Parent	Product
S. lycopersicum chloroplast (NC_007898.3)	LyesC2t021	tRNA-Leu
S. lycopersicum chloroplast (NC_007898.3)	LyesC2r001	16S ribosomal RNA
S. lycopersicum chloroplast (NC_007898.3)	LyesC2r002	23S ribosomal RNA
S. lycopersicum chloroplast (NC_007898.3)	LyesC2p069	photosystem I P700 apoprotein A1
S. lycopersicum chloroplast (NC_007898.3)	LyesC2p053	photosystem II cytochrome b559 beta subunit
S. lycopersicum chloroplast (NC_007898.3)	LyesC2p084	ATP synthase CF1 alpha subunit
S. lycopersicum chloroplast (NC_007898.3)	LyesC2p062	ATP synthase CF1 beta subunit
S. lycopersicum chloroplast (NC_007898.3)	LyesC2p047	ribosomal protein S18
S. lycopersicum chloroplast (NC_007898.3)	LyesC2p007	ribosomal protein S7
S. lycopersicum chloroplast (NC_007898.3)	LyesC2p070	photosystem I P700 apoprotein A2
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101255723	microtubule-associated protein 70-2-like %2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101266854	digalactosyldiacylglycerol synthase 1%2C chloroplastic-like
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101055539	Hop-interacting protein THI140
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101248602	7-dehydrocholesterol reductase
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101253713	uncharacterized LOC101253713

S. lycopersicum chromosome 1 (NC_015438.3)	LOC112940845	5.8S ribosomal RNA
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101257502	importin subunit alpha-2%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC100736538	mitogen-activated protein kinase 13%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101261746	DNA-directed RNA polymerase I subunit 2%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101267059	eukaryotic translation initiation factor 3 subunit B
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101268226	transcription factor UNE10
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101246962	NAD-dependent malic enzyme 59 kDa isoform%2C mitochondrial
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101253532	C-terminal binding protein AN
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101262050	ubiquitin-60S ribosomal protein L40
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101250341	proline--tRNA ligase%2C chloroplastic/mitochondrial
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101248142	uncharacterized LOC101248142
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101262960	topless-related protein 3%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101249410	ADP-ribosylation factor 1%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101262258	probable receptor-like protein kinase At5g18500%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101268807	flap endonuclease 1%2C
S. lycopersicum chromosome 1 (NC_015438.3)	Rab1B	small GTP-binding protein%2C

S. lycopersicum chromosome 1 (NC_015438.3)	Rab1B	small GTP-binding protein
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101265073	shaggy-related protein kinase kappa%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101265681	stromal 70 kDa heat shock-related protein%2C chloroplastic
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101255143	ankyrin repeat domain-containing protein 2B
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101248429	endoplasmic reticulum-Golgi intermediate compartment protein 3
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101267066	gamma-glutamyl peptidase 5
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101263943	coatomer subunit gamma
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101267748	SUMO-activating enzyme subunit 2%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101268622	tRNA-dihydrouridine(47) synthase [NAD(P)(+)]-like
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101244534	PCI superfamily protein
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101257812	alpha-1%2C3-mannosyl-glycoprotein 2- beta-N-acetylglucosaminyltransferase %2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101263540	callose synthase 9
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101263837	callose synthase 3-like%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101258397	COP9 signalosome complex subunit 2
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101252107	serrate RNA effector molecule
S. lycopersicum chromosome 1	LOC112940756	5S ribosomal RNA

(NC_015438.3)		
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101257698	CTP synthase%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101246863	cellulose synthase A catalytic subunit 3 [UDP-forming]
S. lycopersicum chromosome 1 (NC_015438.3)	HDR	4-hydroxy-3-methylbut-2-enyl diphosphate reductase
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101265778	phosphatidylinositol/phosphatidylcholine transfer protein SFH8
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101250243	ADP-ribosylation factor 2
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101252106	UDP-galactose/UDP-glucose transporter 3
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101253117	ruBisCO large subunit-binding protein subunit beta%2C chloroplastic
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101260571	histone H3.2-like
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101245796	ATP-dependent helicase BRM%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101267560	E3 ubiquitin-protein ligase KEG
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101247163	elongation factor 1-delta 1%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LeMA-1	26S protease regulatory subunit 6A homolog
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101244419	zinc finger matrin-type protein 2%2C
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101244419	zinc finger matrin-type protein 2
S. lycopersicum chromosome 1 (NC_015438.3)	TRNAP-CGG	tRNA-Pro
S. lycopersicum chromosome 1	TRNAP-CGG-2	tRNA-Pro

(NC_015438.3)		
S. lycopersicum chromosome 1 (NC_015438.3)	TRNAP-CGG-3	tRNA-Pro
S. lycopersicum chromosome 1 (NC_015438.3)	TRNAP-CGG-4	tRNA-Pro
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101245293	dihydrolipoyl dehydrogenase 2%2C chloroplastic
S. lycopersicum chromosome 1 (NC_015438.3)	LOC101268620	serine/threonine-protein kinase SRK2E %2C
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101249122	nuclear cap-binding protein subunit 1
S. lycopersicum chromosome 10 (NC_015447.3)	TRNAF-GAA-22	tRNA-Phe
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101253324	ras-related protein RABB1c
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101245103	pre-mRNA-splicing factor syf2%2C
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101245103	pre-mRNA-splicing factor syf2
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101251637	DNA-directed RNA polymerase III subunit 2%2C
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101266576	transcription initiation factor IIB-2%2C
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101252540	transcription initiation factor IIB-2
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101254347	thioredoxin reductase NTRC
S. lycopersicum chromosome 10 (NC_015447.3)	Tom52	Actin
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101268442	casein kinase 1-like protein HD16%2C
S. lycopersicum chromosome	LOC101254437	uncharacterized LOC101254437

10 (NC_015447.3)		
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101256436	40S ribosomal protein S14-2
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101253952	tubulin beta-2 chain
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101254748	mitochondrial import inner membrane translocase subunit TIM14-1
S. lycopersicum chromosome 10 (NC_015447.3)	hsc70-2	heat shock cognate 70 kDa protein 2
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101250168	polygalacturonate 4-alpha- galacturonosyltransferase
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101253033	histone acetyltransferase GCN5%2C
S. lycopersicum chromosome 10 (NC_015447.3)	RPS25	ribosomal protein S25
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101256637	DNA polymerase delta catalytic subunit
S. lycopersicum chromosome 10 (NC_015447.3)	ER69	5-methyltetrahydropteroyltriglutamate-- homocysteine methyltransferase
S. lycopersicum chromosome 10 (NC_015447.3)	ER69	5-methyltetrahydropteroyltriglutamate-- homocysteine methyltransferase%2C
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101245012	S-adenosylmethionine synthase 3-like
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101252449	40S ribosomal protein S23
S. lycopersicum chromosome 10 (NC_015447.3)	TRNAP-CGG-6	tRNA-Pro
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101256039	polyubiquitin
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101267272	cytochrome c oxidase copper chaperone 1-like%2C
S. lycopersicum chromosome	cTPT	triose phosphate/phosphate translocator

10 (NC_015447.3)		%2C chloroplastic
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101263060	ribose-phosphate pyrophosphokinase 1%2C
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101267068	NADH dehydrogenase [ubiquinone] iron- sulfur protein 8%2C mitochondrial
S. lycopersicum chromosome 10 (NC_015447.3)	LOC101252935	calcium-dependent protein kinase 4%2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101255850	GDT1-like protein 3
S. lycopersicum chromosome 11 (NC_015448.3)	LOC112940382	5.8S ribosomal RNA
S. lycopersicum chromosome 11 (NC_015448.3)	LOC112940417	5.8S ribosomal RNA
S. lycopersicum chromosome 11 (NC_015448.3)	LOC104644593	uncharacterized LOC104644593
S. lycopersicum chromosome 11 (NC_015448.3)	LOC112940383	28S ribosomal RNA
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101253748	kinesin-like protein KIN-7D%2C mitochondrial%2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101260396	DNA replication licensing factor MCM2
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101262684	ribosomal protein S12%2C mitochondrial
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101265892	vacuolar protein sorting-associated protein 41 homolog
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101247099	protein TIC 100
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101251648	histone deacetylase 9
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101253156	peptidyl-prolyl cis-trans isomerase CYP71
S. lycopersicum chromosome	LOC101256841	NF-kappa-B-activating protein

11 (NC_015448.3)		
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101243958	calmodulin 4
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101256147	pyruvate kinase 1%2C cytosolic-like%2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101246780	O-fucosyltransferase family protein%2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101259801	ras-related protein RABA5a%2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101264480	dihydrolipoyllysine-residue acetyltransferase component 5 of pyruvate dehydrogenase complex%2C chloroplastic
S. lycopersicum chromosome 11 (NC_015448.3)	LOC112940391	5.8S ribosomal RNA
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101265610	syndetin%2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101264178	protein TPLATE
S. lycopersicum chromosome 11 (NC_015448.3)	ARF4	auxin response factor 4
S. lycopersicum chromosome 11 (NC_015448.3)	HDS	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase%2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101265798	cell division cycle protein 48 homolog %2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101256340	probable complex I intermediate- associated protein 30%2C
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101268826	replication factor C subunit 2
S. lycopersicum chromosome 11 (NC_015448.3)	LOC101256545	phosphoinositide phosphatase SAC1
S. lycopersicum chromosome	LOC112940394	5.8S ribosomal RNA

11 (NC_015448.3)		
S. lycopersicum chromosome 11 (NC_015448.3)	MAPKKKe	MAP3K epsilon protein kinase%2C
S. lycopersicum chromosome 11 (NC_015448.3)	MAPKKKe	MAP3K epsilon protein kinase
S. lycopersicum chromosome 11 (NC_015448.3)	LOC109118967	ribosomal protein S14%2C mitochondrial
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101266099	splicing factor U2af large subunit B%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101245436	polypyrimidine tract-binding protein homolog 3
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101245924	general negative regulator of transcription subunit 3%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101260909	actin-related protein 4
S. lycopersicum chromosome 12 (NC_015449.3)	CD4B	ATP-dependent protease
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101246802	homeobox-leucine zipper protein ATHB- 15
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101251658	ubiquitin carboxyl-terminal hydrolase 6
S. lycopersicum chromosome 12 (NC_015449.3)	LOC112939957	U6 spliceosomal RNA
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101268171	exocyst complex component SEC3A
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101268752	ATP-dependent 6-phosphofructokinase 3- like%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101261210	troponin T%2C fast skeletal muscle
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101247506	S-adenosylmethionine synthase 2
S. lycopersicum chromosome	LOC101248858	ubiquitin-40S ribosomal protein S27a-like

12 (NC_015449.3)		%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101246230	ADP-ribosylation factor 2%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101256573	PH%2C RCC1 and FYVE domains- containing protein 1%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101259027	cytochrome B5-like protein%2C
S. lycopersicum chromosome 12 (NC_015449.3)	HO1	heme oxygenase 1
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101259713	60S ribosomal protein L11-1
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101267777	protein FAM133%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101243968	diacylglycerol kinase 1-like%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC544156	54-kD signal recognition particle (SRP) specific protein
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101250973	acetyl-CoA carboxylase 1-like
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101263674	pyrophosphate-energized vacuolar membrane proton pump-like
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101266100	heterogeneous nuclear ribonucleoprotein 1
S. lycopersicum chromosome 12 (NC_015449.3)	TBG1	beta-galactosidase precursor
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101248085	5'-3' exoribonuclease 3-like%2C
S. lycopersicum chromosome 12 (NC_015449.3)	LOC101259420	adenosylhomocysteinase
S. lycopersicum chromosome 12 (NC_015449.3)	Pti1	Pto kinase interactor 1
S. lycopersicum chromosome	Pti1	Pto kinase interactor 1%2C

12 (NC_015449.3)

S. lycopersicum chromosome 2 (NC_015439.3)	LOC101260524	calumenin-B
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101258313	katanin p80 WD40 repeat-containing subunit B1 homolog%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101244538	AP2-like ethylene-responsive transcription factor At2g41710
S. lycopersicum chromosome 2 (NC_015439.3)	LOC100301945	auxin response factor 8%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101258617	T-complex protein 1 subunit zeta 1
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101246688	calcium-transporting ATPase 2%2C plasma membrane-type-like
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101261290	ribosomal protein S25
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101261290	ribosomal protein S25%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC104645853	uncharacterized LOC104645853
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101263666	protein MOR1%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101261570	lysine-specific demethylase JMJ25%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101260688	uncharacterized LOC101260688%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101256833	ATP-dependent zinc metalloprotease FTSH 12%2C chloroplastic
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101243849	pyrophosphate--fructose 6-phosphate 1- phosphotransferase subunit beta-like
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101250651	probable splicing factor 3A subunit 1%2C
S. lycopersicum chromosome 2	LOC112941003	U2 spliceosomal RNA

(NC_015439.3)		
S. lycopersicum chromosome 2 (NC_015439.3)	LOC778257	DNA-directed RNA polymerase II subunit 1
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101263071	26S proteasome non-ATPase regulatory subunit 4 homolog%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101267759	CRM-domain containing factor CFM2%2C chloroplastic
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101266287	alpha-mannosidase 2
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101263958	UDP-glucose 6-dehydrogenase 2
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101262273	50S ribosomal protein L7/L12
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101243842	CLK4-associating serine/arginine rich protein
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101259995	uncharacterized LOC101259995%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101263559	actin-related protein 2
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101263274	autophagy-related protein 101
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101267959	3-oxoacyl-[acyl-carrier-protein] synthase I%2C chloroplastic
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101245725	uncharacterized LOC101245725%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC112940999	U2 spliceosomal RNA
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101256541	small nuclear ribonucleoprotein SmD3b
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101254439	serine hydroxymethyltransferase%2C mitochondrial%2C
S. lycopersicum chromosome 2	LOC101261474	serine/threonine-protein kinase

(NC_015439.3)		STY13%2C
S. lycopersicum chromosome 2 (NC_015439.3)	TRNAP-CGG-5	tRNA-Pro
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101255935	methyl-CpG-binding domain protein 4- like protein%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC109119483	uncharacterized LOC109119483
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101255157	ubiquitin domain-containing protein DSK2b%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101256641	succinate dehydrogenase [ubiquinone] flavoprotein subunit 1%2C mitochondrial
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101261294	auxin transport protein BIG
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101246582	glycosyltransferase family protein
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101248827	CDPK-related protein kinase%2C
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101247774	uncharacterized LOC101247774
S. lycopersicum chromosome 2 (NC_015439.3)	LOC101267666	heterogeneous nuclear ribonucleoprotein A0
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101266391	3-isopropylmalate dehydratase large subunit%2C chloroplastic
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101250369	uncharacterized LOC101250369%2C
S. lycopersicum chromosome 3 (NC_015440.3)	ARF8-1	auxin response factor 8-1
S. lycopersicum chromosome 3 (NC_015440.3)	ARF8-1	auxin response factor 8-1%2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101247792	small nuclear ribonucleoprotein SmD3b %2C
S. lycopersicum chromosome 3	LOC101248750	transmembrane and coiled-coil domain-

(NC_015440.3)		containing protein 4%2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC112941219	5.8S ribosomal RNA
S. lycopersicum chromosome 3 (NC_015440.3)	LOC112941208	5.8S ribosomal RNA
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101265404	probable galacturonosyltransferase 13%2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101245038	protein ESMERALDA 1-like%2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101244464	ubiquitin-60S ribosomal protein L40
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101251760	uncharacterized LOC101251760
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101259924	dolichyl-diphosphooligosaccharide-- protein glycosyltransferase subunit STT3B
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101264202	ribonuclease J%2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101260712	tubulin gamma-1 chain
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101253271	60S ribosomal protein L44
S. lycopersicum chromosome 3 (NC_015440.3)	LOC112941110	uncharacterized LOC112941110
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101263008	O-fucosyltransferase 34-like
S. lycopersicum chromosome 3 (NC_015440.3)	SIVP2	H+-inorganic pyrophosphatase 2
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101262797	stromal processing peptidase%2C chloroplastic%2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101257548	homeobox-leucine zipper protein ATHB- 15

S. lycopersicum chromosome 3 (NC_015440.3)	LOC101246329	serine/threonine protein phosphatase 2A 55 kDa regulatory subunit B beta isoform %2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101244750	casein kinase II subunit alpha-2
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101267099	probable NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 12
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101244477	uncharacterized ATP-dependent helicase YprA%2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101248281	RWD domain-containing protein 1%2C
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101261490	mediator of RNA polymerase II transcription subunit 36a-like
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101244348	rac-like GTP-binding protein RHO1
S. lycopersicum chromosome 3 (NC_015440.3)	ACT	actin-7
S. lycopersicum chromosome 3 (NC_015440.3)	LOC101246035	alanine aminotransferase 2%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101246043	probable splicing factor 3A subunit 1%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101264814	microtubule-associated protein 70-2%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101259528	proteasome subunit beta type-3-A
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101263507	UDP-glucuronic acid decarboxylase 6%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC543565	14-3-3 family protein
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101248472	serine/threonine-protein kinase VPS15%2C

S. lycopersicum chromosome 4 (NC_015441.3)	LOC101249451	translation factor GUF1 homolog%2C chloroplastic
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101249451	translation factor GUF1 homolog%2C chloroplastic%2C
S. lycopersicum chromosome 4 (NC_015441.3)	gluII	alpha glucosidase II
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101253276	regulator of nonsense transcripts UPF2%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101261810	F-box protein At1g78280
S. lycopersicum chromosome 4 (NC_015441.3)	TFT3	14-3-3 protein homologue
S. lycopersicum chromosome 4 (NC_015441.3)	GPI1	glucose-6-phosphate isomerase 1
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101252774	AP-4 complex subunit mu
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101244864	tubulin alpha chain
S. lycopersicum chromosome 4 (NC_015441.3)	dem	dem protein
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101259629	26S proteasome non-ATPase regulatory subunit 14 homolog
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101264622	protein S-acyltransferase 21%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101260126	ADP-ribosylation factor 1%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101256582	vacuolar protein sorting-associated protein 53 A%2C
S. lycopersicum chromosome 4 (NC_015441.3)	VDE	violaxanthin de-epoxidase
S. lycopersicum chromosome 4 (NC_015441.3)	IDI1	isopentenyl diphosphate isomerase

S. lycopersicum chromosome 4 (NC_015441.3)	LOC101262912	serine decarboxylase%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101265636	protein FAM91A1-like%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101261514	GTPase HflX
S. lycopersicum chromosome 4 (NC_015441.3)	hsc70.3	heat shock protein 70-3%2C
S. lycopersicum chromosome 4 (NC_015441.3)	hsc70.3	heat shock protein 70-3
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101249170	ATP synthase subunit beta%2C mitochondrial
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101254189	outer envelope pore protein 37%2C chloroplastic
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101258940	NAD-dependent protein deacylase SRT2%2C
S. lycopersicum chromosome 4 (NC_015441.3)	ide	insulin degrading enzyme
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101263205	putative potassium transporter 12
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101248382	cellulose synthase A catalytic subunit 2 [UDP-forming]
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101260724	putative leucine-rich repeat-containing protein DDB_G0290503%2C
S. lycopersicum chromosome 4 (NC_015441.3)	LOC101245941	probable E3 ubiquitin-protein ligase ARI2
S. lycopersicum chromosome 4 (NC_015441.3)	LOC112941414	U2 spliceosomal RNA
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101246135	ATP-citrate synthase alpha chain protein 2
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101247318	ADP-ribosylation factor 2

S. lycopersicum chromosome 5 (NC_015442.3)	LOC101244581	uncharacterized LOC101244581
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101256777	uncharacterized LOC101256777
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101255293	conserved oligomeric Golgi complex subunit 7
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101266720	60S ribosomal protein L34
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101261919	ATP synthase subunit beta%2C mitochondrial
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101247914	3-isopropylmalate dehydrogenase%2C chloroplastic
S. lycopersicum chromosome 5 (NC_015442.3)	ASS	argininosuccinate synthase
S. lycopersicum chromosome 5 (NC_015442.3)	GAPC2	glyceraldehyde-3-phosphate dehydrogenase
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101263315	microtubule-associated protein 70-1
S. lycopersicum chromosome 5 (NC_015442.3)	Rab1C	small GTP-binding protein
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101259158	clathrin heavy chain 1
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101248484	dnaJ protein homolog
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101260819	basic leucine zipper and W2 domain- containing protein 2
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101250118	proteasome subunit beta type-5%2C
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101264824	calcium-dependent protein kinase SK5
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101252391	ultraviolet-B receptor UVR8

S. lycopersicum chromosome 5 (NC_015442.3)	LOC101259750	26S proteasome non-ATPase regulatory subunit 1 homolog A%2C
S. lycopersicum chromosome 5 (NC_015442.3)	LOC101261521	60S acidic ribosomal protein P0
S. lycopersicum chromosome 6 (NC_015443.3)	LeEF-1	elongation factor 1-alpha
S. lycopersicum chromosome 6 (NC_015443.3)	LOC104644287	uncharacterized LOC104644287
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101267998	uncharacterized LOC101267998
S. lycopersicum chromosome 6 (NC_015443.3)	ACI29	60S ribosomal protein L10
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101261236	60S ribosomal protein L19-3
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101244084	elongation factor 1-alpha
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101246434	uncharacterized LOC101246434%2C
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101264431	60S ribosomal protein L36-2
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101250125	clathrin heavy chain 1
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101267413	ISWI chromatin-remodeling complex ATPase CHR11-like
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101262927	26S proteasome regulatory subunit 7 homolog A
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101247034	T-complex protein 1 subunit eta
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101261336	DEAD-box ATP-dependent RNA helicase 56%2C
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101245555	mitogen-activated protein kinase 9%2C

S. lycopersicum chromosome 6 (NC_015443.3)	LOC101256497	uncharacterized LOC101256497
S. lycopersicum chromosome 6 (NC_015443.3)	LHA2	plasma membrane H ⁺ -ATPase
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101248017	basic proline-rich protein%2C
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101245851	UDP-D-apiose/UDP-D-xylose synthase 2
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101266129	cell division cycle protein 48 homolog
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101263523	uncharacterized LOC101263523
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101253792	probable histone H2B.1
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101252889	uncharacterized LOC101252889
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101245368	pumilio homolog 2-like
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101262728	tubulin beta chain
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101255504	DExH-box ATP-dependent RNA helicase DExH12
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101254604	uncharacterized LOC101254604
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101247535	transcriptional corepressor LEUNIG_HOMOLOG-like
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101260445	guanylate-binding protein 5-like
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101264125	DExH-box ATP-dependent RNA helicase DExH11%2C
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101254496	putative ABC1 protein At2g40090

S. lycopersicum chromosome 6 (NC_015443.3)	LOC101245551	importin subunit beta-1-like
S. lycopersicum chromosome 6 (NC_015443.3)	LOC112941765	U2 spliceosomal RNA
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101244963	eukaryotic peptide chain release factor GTP-binding subunit ERF3A%2C
S. lycopersicum chromosome 6 (NC_015443.3)	LOC778225	isochorismate synthase%2C
S. lycopersicum chromosome 6 (NC_015443.3)	LOC778225	isochorismate synthase
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101262318	60S ribosomal protein L12
S. lycopersicum chromosome 6 (NC_015443.3)	hsc70	heat shock protein cognate 70
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101257778	probable NAD(P)H dehydrogenase (quinone) FQR1-like 3%2C
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101249736	golgin candidate 5%2C
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101268694	V-type proton ATPase subunit c''1
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101247435	uncharacterized protein RAB5IF homolog
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101260243	auxilin-related protein 2-like
S. lycopersicum chromosome 6 (NC_015443.3)	LOC101257576	fimbrin-2
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101256892	60S ribosomal protein L39-3
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101250421	RNase L inhibitor protein 2%2C
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101263723	methionine--tRNA ligase%2C cytoplasmic

S. lycopersicum chromosome 7 (NC_015444.3)	LOC101255810	protein CTR9 homolog
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101261134	ketol-acid reductoisomerase%2C chloroplastic
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101266739	ATP-dependent zinc metalloprotease FTSH 2%2C chloroplastic%2C
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101254504	translation initiation factor IF-2%2C chloroplastic
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101252198	ras-related protein RABB1c
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101247235	protein MEI2-like 2
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101258282	polyubiquitin
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101254921	tryptophan synthase beta chain 1-like
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101250426	polyadenylate-binding protein RBP45
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101248594	uncharacterized LOC101248594
S. lycopersicum chromosome 7 (NC_015444.3)	HSC80	heat shock cognate protein 80
S. lycopersicum chromosome 7 (NC_015444.3)	SIWRKY2	WRKY transcription factor 2%2C
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101266344	ADP-ribosylation factor 1
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101264027	coatamer subunit alpha-1-like
S. lycopersicum chromosome 7 (NC_015444.3)	NIM1	non-inducible immunity 1%2C
S. lycopersicum chromosome 7 (NC_015444.3)	NIM1	non-inducible immunity 1

S. lycopersicum chromosome 7 (NC_015444.3)	LOC101259860	myosin-17-like
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101250424	ATP-dependent DNA helicase Q-like 3%2C
S. lycopersicum chromosome 7 (NC_015444.3)	PP5	serine/threonine-protein phosphatase 5%2C
S. lycopersicum chromosome 7 (NC_015444.3)	LOC100134910	trehalose-phosphate synthase 1%2C
S. lycopersicum chromosome 7 (NC_015444.3)	LOC100134910	trehalose-phosphate synthase 1
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101259647	suppressor of mec-8 and unc-52 protein homolog 2
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101258674	phospholipid:diacylglycerol acyltransferase 1-like
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101268775	60S ribosomal protein L11-1
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101252794	DNA polymerase I B%2C chloroplastic/mitochondrial%2C
S. lycopersicum chromosome 7 (NC_015444.3)	LOC101250908	putative DUF21 domain-containing protein At3g13070%2C chloroplastic%2C
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101254013	tubulin alpha-3 chain
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101258879	regulator of nonsense transcripts 1 homolog%2C
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101256200	calcium-dependent protein kinase 1
S. lycopersicum chromosome 8 (NC_015445.3)	IPMS2	isopropylmalate synthase
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101264039	3-oxoacyl-[acyl-carrier-protein] synthase I%2C chloroplastic
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101264235	dnaJ homolog subfamily C GRV2%2C

S. lycopersicum chromosome 8 (NC_015445.3)	LOC101264448	hydroxyproline O-galactosyltransferase HPGT3%2C
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101244892	40S ribosomal protein S3-3-like
S. lycopersicum chromosome 8 (NC_015445.3)	LOC104648840	uncharacterized LOC104648840
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101256793	N6-adenosine-methyltransferase MT- A70-like
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101252014	conserved oligomeric Golgi complex subunit 6
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101258089	dynammin-related protein 1E
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101253514	ubiquitin-conjugating enzyme E2 5
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101251406	general transcription and DNA repair factor IIH helicase subunit XPB1
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101264240	ubiquitin-ribosomal fusion protein
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101246271	tryptophan--tRNA ligase%2C cytoplasmic
S. lycopersicum chromosome 8 (NC_015445.3)	DAD1	dad-1 protein
S. lycopersicum chromosome 8 (NC_015445.3)	LOC112942107	U2 spliceosomal RNA
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101247146	mitochondrial pyruvate carrier 1%2C
S. lycopersicum chromosome 8 (NC_015445.3)	MPK2	mitogen-activated protein kinase 2
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101261642	aspartate aminotransferase%2C chloroplastic%2C
S. lycopersicum chromosome 8 (NC_015445.3)	LOC101251706	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 8%2C

		mitochondrial
S. lycopersicum chromosome 8 (NC_015445.3)	BiP/grp78	glucose-regulated protein 78
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101243737	phosphatidate phosphatase PAH1-like %2C
S. lycopersicum chromosome 9 (NC_015446.3)	PGH1	enolase
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101262246	uncharacterized LOC101262246%2C
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101251215	signal recognition particle 54 kDa protein %2C chloroplastic
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101247151	heat shock cognate 70 kDa protein 2-like
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101249951	oligoribonuclease%2C
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101259877	vesicle-associated membrane protein
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101267840	myosin-6-like
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101246377	small nuclear ribonucleoprotein Sm D2
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101265769	60S ribosomal protein L19-3
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101254034	serine/threonine-protein kinase tricornet %2C
S. lycopersicum chromosome 9 (NC_015446.3)	SAHH	S-adenosyl-l-homocysteine hydrolase
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101264663	double-strand break repair protein MRE11%2C
S. lycopersicum chromosome 9 (NC_015446.3)	LOC112942231	uncharacterized protein DDB_G0271670- like
S. lycopersicum chromosome 9	LOC101259578	nucleolar GTP-binding protein 1%2C

(NC_015446.3)		
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101247369	UDP-glucuronate 4-epimerase 6
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101248035	ATP-dependent zinc metalloprotease FTSH 8%2C mitochondrial
S. lycopersicum chromosome 9 (NC_015446.3)	LOC100301935	glutathione reductase%2C
S. lycopersicum chromosome 9 (NC_015446.3)	LOC101268614	kinesin-like protein KIN-12B
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgr05	26S ribosomal RNA
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgt16	tRNA-Ser
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgt17	tRNA-Phe
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgr03	18S ribosomal RNA
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgr02	18S ribosomal RNA
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgt14	tRNA-Tyr
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp40	NADH dehydrogenase subunit 3
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp39	ribosomal protein S12
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp38	ribosomal protein L5
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp36	ATP synthase subunit 4
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp34	ribosomal protein S1
S. lycopersicum mitochondrion	CQW56_mgp33	ATP synthase subunit 8

(NC_035963.1)		
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp32	cytochrome c oxidase subunit 3
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp30	apocytochrome b
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp24	SecY-independent transporter protein
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp23	ATP synthase subunit 1
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp13	NADH dehydrogenase subunit 4
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp11	cytochrome c oxidase subunit 1
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp10	cytochrome c maturation protein CcmFN
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp09	NADH dehydrogenase subunit 7
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp08	ribosomal protein S13
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp35	NADH dehydrogenase subunit 4L
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp16	maturase R
S. lycopersicum unplaced genomic scaffold (NW_020442489.1)	LOC101245746	uncharacterized LOC101245746true
S. lycopersicum unplaced genomic scaffold (NW_020442490.1)	TAL	transaldolase
S. lycopersicum unplaced genomic scaffold (NW_020442490.1)	LOC104645248	uncharacterized LOC104645248

S. lycopersicum unplaced genomic scaffold (NW_020442490.1)	LOC101249808	peptidyl-prolyl cis-trans isomerase FKBP19%2C chloroplastic%2C
S. lycopersicum unplaced genomic scaffold (NW_020442524.1)	LOC101245341	L-arabinokinase-like%2C
S. lycopersicum unplaced genomic scaffold (NW_020442529.1)	LOC101246232	uncharacterized LOC101246232
S. lycopersicum unplaced genomic scaffold (NW_020442534.1)	LOC112940610	NADH-ubiquinone oxidoreductase chain 5-like
S. lycopersicum unplaced genomic scaffold (NW_020443272.1)	LOC101254183	uncharacterized LOC101254183_493
S. lycopersicum unplaced genomic scaffold (NW_020443395.1)	LOC104645327	uncharacterized LOC104645327_2824
S. lycopersicum unplaced genomic scaffold (NW_020444325.1)	ARF6	auxin response factor 6_107

Supplementary Table 2.10. *S. lycopersicum* and *C. campestris* two-side cross-mapping in sequential approach. List of *S. lycopersicum* loci identified investigating on two-side cross-mapped reads from the *S. lycopersicum*-*C. campestris* interaction via the sequential approach. Chromosome (Chr), locus identifier (Parent) and description of the gene function (Product) were extracted from *S. lycopersicum* genome annotation (gff3 file) and reported.

Chr	Parent	Product
A. thaliana chloroplast (NC_000932.1)	ArthCr088	16S ribosomal RNA
A. thaliana chloroplast (NC_000932.1)	ArthCp086	NADH dehydrogenase subunit 2
A. thaliana chromosome 2 (NC_003071.7)	AT2G01010	18S ribosomal RNA
A. thaliana chromosome 2 (NC_003071.7)	AT2G07731	pseudogene of NADH dehydrogenase
A. thaliana chromosome 3 (NC_003074.8)	AT3G41768	18S ribosomal RNA
A. thaliana chromosome 3 (NC_003074.8)	AT3G06365	uncharacterized misc_RNA
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	DA397_mgr01	26S ribosomal RNA
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	DA397_mgp38	NADH dehydrogenase subunit 9
A. thaliana ecotype Col-0 mitochondrion (NC_037304.1)	GeneID:38088291	cosRNA M120

Supplementary Table 2.11. *A. thaliana* and *C. campestris* two-side cross-mapping in sequential approach. List of *A. thaliana* loci identified investigating on two-side cross-mapped reads from *A. thaliana*-*C. campestris* interaction via the combined approach. Chromosome (Chr), locus identifier (Parent) and description of the gene function (Product) were extracted from *A. thaliana* genome annotation (gff3 file) and reported.

Chr	Parent	Product
S. lycopersicum chromosome 11 (NC_015448.3)	LOC11294038 3	28S ribosomal RNA
S. lycopersicum chromosome 3 (NC_015440.3)	LOC11294120 8	5.8S ribosomal RNA
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgr05	26S ribosomal RNA
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgr03	18S ribosomal RNA
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgr02	18S ribosomal RNA
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp39	ribosomal protein S12
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp32	cytochrome c oxidase subunit 3
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp30	apocytochrome b
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp23	ATP synthase subunit 1
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp33	ATP synthase subunit 8
S. lycopersicum mitochondrion (NC_035963.1)	CQW56_mgp11	cytochrome c oxidase subunit 1
S. lycopersicum unplaced genomic scaffold (NW_020442489.1)	LOC10124574 6	uncharacterized LOC101245746
S. lycopersicum unplaced genomic scaffold (NW_020442534.1)	LOC11294061 0	NADH-ubiquinone oxidoreductase chain 5- like

Supplementary Table 2.12. *S. lycopersicum* and *C. campestris* two-side cross-mapping in combined approach. List of *S. lycopersicum* loci identified investigating on two-side cross-mapped reads from *S. lycopersicum*-*C. campestris* interaction via combined approach.

Chromosome (Chr), locus identifier (Parent) and description of the gene function (Product) were extracted from *S. lycopersicum* genome annotation (gff3 file) and reported.

SolGenomic ID	Enzyme	AA position	Wild residue (<i>S. pennellii</i>)	Target residue (<i>S. lycopersicum</i>)	predictSNP prediction	predictSNP expected accuracy
Sopen01g036390	CCD7	239	Y	F	NEUTRAL	0.82622462
Sopen01g036390	CCD7	266	I	L	NEUTRAL	0.82622462
Sopen01g036390	CCD7	382	S	P	NEUTRAL	0.82622462
Sopen01g036390	CCD7	393	P	R	DELETERIOUS	0.54946365
Sopen01g036390	CCD7	411	I	M	NEUTRAL	0.82622462
Sopen01g036390	CCD7	501	K	E	NEUTRAL	0.82622462
Sopen01g036390	CCD7	542	T	A	NEUTRAL	0.73834499
Sopen08g020520	CCD8	4	F	L	NEUTRAL	0.82622462
Sopen08g020520	CCD8	69	N	S	NEUTRAL	0.82622462
Sopen08g020520	CCD8	148	E	D	NEUTRAL	0.82622462
Sopen08g020520	CCD8	155	S	I	NEUTRAL	0.73688811
Sopen08g020520	CCD8	170	A	V	NEUTRAL	0.6025641
Sopen08g020520	CCD8	259	I	M	NEUTRAL	0.82622462
Sopen08g018500	MAX1	206	E	D	NEUTRAL	0.82622462
Sopen08g018500	MAX1	281	M	I	NEUTRAL	0.82622462
Sopen08g021560	PDR1	76	K	T	NEUTRAL	0.82622462
Sopen08g021560	PDR1	128	I	V	NEUTRAL	0.82622462
Sopen08g021560	PDR1	129	E	D	NEUTRAL	0.82622462
Sopen08g021560	PDR1	157	L	I	NEUTRAL	0.82622462

Sopen08g021560	PDR1	404	E	K	NEUTRAL	0.73688811
Sopen08g021560	PDR1	452	H	D	NEUTRAL	0.82622462
Sopen08g021560	PDR1	511	T	A	NEUTRAL	0.82622462
Sopen08g021560	PDR1	703	I	T	NEUTRAL	0.6025641
Sopen08g021560	PDR1	794	A	V	NEUTRAL	0.82622462
Sopen08g021560	PDR1	796	K	R	NEUTRAL	0.82622462
Sopen08g021560	PDR1	797	N	I	NEUTRAL	0.82622462
Sopen08g021560	PDR1	822	P	A	NEUTRAL	0.82622462
Sopen08g021560	PDR1	881	Q	R	NEUTRAL	0.82622462
Sopen08g021560	PDR1	982	S	N	NEUTRAL	0.82622462
Sopen08g021560	PDR1	1175	K	E	NEUTRAL	0.82622462
Sopen08g021560	PDR1	1249	V	I	NEUTRAL	0.82622462
Sopen08g021560	PDR1	1314	T	A	NEUTRAL	0.82622462
Sopen08g021560	PDR1	1381	I	V	NEUTRAL	0.82622462
Sopen08g021560	PDR1	1402	R	K	NEUTRAL	0.82622462
Sopen08g021560	PDR1	1448	S	A	NEUTRAL	0.82622462

Supplementary Table 4.1. PredictSNP results related to SolGenomic annotation of *S. lycopersicum* and *S. pennellii*. PredictSNP analysis of polymorphisms of strigolactone pathway genes belonging to *S. pennellii* using the annotation retrieved from SolGenomic database. Amino acidic position in *S. pennellii*, residues in both *S. pennellii* and *S. lycopersicum*, predictSNP and its accuracy are reported.

NCBI ID	Enzyme	AA position	Wild residue	Target residue	predictSNP prediction	predictSNP expected accuracy
			(S. pennellii)	(S. lycopersicum)		
XP_015087389.1	CCD7	4	Q	V	NEUTRAL	0.62529138
XP_015087389.1	CCD7	4	Q	S	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	I	NEUTRAL	0.61421911
XP_015087389.1	CCD7	4	Q	K	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	F	NEUTRAL	0.62529138
XP_015087389.1	CCD7	4	Q	P	NEUTRAL	0.6025641
XP_015087389.1	CCD7	4	Q	Y	NEUTRAL	0.61421911
XP_015087389.1	CCD7	4	Q	E	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	H	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	W	NEUTRAL	0.6025641
XP_015087389.1	CCD7	4	Q	G	NEUTRAL	0.61421911
XP_015087389.1	CCD7	4	Q	A	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	R	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	C	NEUTRAL	0.6025641
XP_015087389.1	CCD7	4	Q	N	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	L	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	M	NEUTRAL	0.82622462

XP_015087389.1	CCD7	4	Q	D	NEUTRAL	0.82622462
XP_015087389.1	CCD7	4	Q	T	NEUTRAL	0.61421911
XP_015087389.1	CCD7	239	F	Y	NEUTRAL	0.82622462
XP_015087389.1	CCD7	266	L	I	NEUTRAL	0.82622462
XP_015087389.1	CCD7	382	P	S	NEUTRAL	0.82622462
XP_015087389.1	CCD7	393	R	P	NEUTRAL	0.62831343
XP_015087389.1	CCD7	411	M	I	DELETERIOUS	0.50595948
XP_015087389.1	CCD7	501	E	K	NEUTRAL	0.82622462
XP_015087389.1	CCD7	542	A	T	NEUTRAL	0.82622462
XP_015084651.1	CCD8	4	L	K	NEUTRAL	0.63151762
XP_015084651.1	CCD8	69	S	Y	DELETERIOUS	0.50595948
XP_015084651.1	CCD8	69	S	T	NEUTRAL	0.82622462
XP_015084651.1	CCD8	148	D	I	DELETERIOUS	0.78903456
XP_015084651.1	CCD8	155	I	S	NEUTRAL	0.82622462
XP_015084651.1	CCD8	170	V	P	NEUTRAL	0.82622462
XP_015084651.1	CCD8	259	M	I	NEUTRAL	0.82622462
XP_015087576.1	D27	19	N	D	NEUTRAL	0.82622462
XP_015087576.1	D27	52	C	G	NEUTRAL	0.73834499
XP_015087576.1	D27	230	K	T	NEUTRAL	0.68365861
XP_015087576.1	D27	235	R	M	NEUTRAL	0.63151762
XP_015083325.1	MAX1	206	D	E	NEUTRAL	0.82622462

XP_015083325.1	MAX1	281	I	M	NEUTRAL	0.82622462
XP_015079414.1	PDR1	10	N	S	NEUTRAL	0.82622462
XP_015079414.1	PDR1	57	Q	E	NEUTRAL	0.82622462
XP_015079414.1	PDR1	61	G	V	NEUTRAL	0.62529138
XP_015079414.1	PDR1	67	N	S	NEUTRAL	0.73688811
XP_015079414.1	PDR1	194	K	T	NEUTRAL	0.82622462
XP_015079414.1	PDR1	234	Q	H	NEUTRAL	0.82622462
XP_015079414.1	PDR1	1097	A	T	NEUTRAL	0.82622462
XP_015079414.1	PDR1	1262	I	V	NEUTRAL	0.82622462

Supplementary Table 4.2. PredictSNP results related to NCBI annotation of *S. lycopersicum* and *S. pennellii*. PredictSNP analysis of polymorphisms of strigolactone pathway genes belonging to *S. pennellii* using the annotation retrieved from NCBI database. Amino acidic position in *S. pennellii*, residues in both *S. pennellii* and *S. lycopersicum*, predictSNP and its accuracy are reported.

Bibliography

- Adams, M.D. *et al.* (1992) 'Sequence identification of 2,375 human brain genes', *Nature*, 355(6361), pp. 632–634. Available at: <https://doi.org/10.1038/355632a0>.
- Afgan, E. *et al.* (2018) 'The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update', *Nucleic Acids Research*, 46(W1), pp. W537–W544. Available at: <https://doi.org/10.1093/nar/gky379>.
- Al-Babili, S. and Bouwmeester, H.J. (2015) 'Strigolactones, a Novel Carotenoid-Derived Plant Hormone', *Annual Review of Plant Biology*, 66(1), pp. 161–186. Available at: <https://doi.org/10.1146/annurev-arplant-043014-114759>.
- Albanova, I.A. *et al.* (2023) 'Host Resistance to Parasitic Plants—Current Knowledge and Future Perspectives', *Plants*, 12(7), p. 1447. Available at: <https://doi.org/10.3390/plants12071447>.
- Albert, M., Axtell, M.J. and Timko, M.P. (2020) 'Mechanisms of resistance and virulence in parasitic plant–host interactions', *Plant Physiology*, 185(4), pp. 1282–1291. Available at: <https://doi.org/10.1093/plphys/kiaa064>.
- Alseekh, S. *et al.* (2013) 'Resolution by recombination: breaking up *Solanum pennellii* introgressions', *Trends in Plant Science*, 18(10), pp. 536–538.
- Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data', *Genome Biology*, 11(10), p. R106. Available at: <https://doi.org/10.1186/gb-2010-11-10-r106>.
- Ari, N. and Mamatnazarova, N. (2014) 'Programming languages', in *2014 11th International Conference on Electronics, Computer and Computation (ICECCO). 2014*, Abuja, Nigeria: IEEE, pp. 1–8. Available at: <https://doi.org/10.1109/ICECCO.2014.6997548>.
- Assis, A.F. *et al.* (2022) 'What Is the Transcriptome and How It Is Evaluated', in G.A. Passos (ed.) *Transcriptomics in Health and Disease*. Cham: Springer International Publishing, pp. 3–50. Available at: https://doi.org/10.1007/978-3-030-87821-4_1.
- Au, K.F. *et al.* (2012) 'Improving PacBio Long Read Accuracy by Short Read Alignment', *PLoS ONE*. Edited by Y. Xing, 7(10), p. e46679. Available at: <https://doi.org/10.1371/journal.pone.0046679>.
- Bai, J. *et al.* (2020) 'Exploration of resistance to *PHELIPANCHE AEGYPTIACA* in tomato', *Pest Management Science*, 76(11), pp. 3806–3821. Available at: <https://doi.org/10.1002/ps.5932>.

- Bawin, T. *et al.* (2022) 'A host-free transcriptome for haustoriogenesis in *Cuscuta campestris*: Signature gene expression identifies markers of successive development stages', *Physiologia Plantarum*, 174(2), p. e13628. Available at: <https://doi.org/10.1111/ppl.13628>.
- Bayerlová, M. *et al.* (2015) 'Comparative study on gene set and pathway topology-based enrichment methods', *BMC Bioinformatics*, 16(1), p. 334. Available at: <https://doi.org/10.1186/s12859-015-0751-5>.
- Bendl, J. *et al.* (2016) 'PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions', *PLOS Computational Biology*. Edited by P.P. Gardner, 12(5), p. e1004962. Available at: <https://doi.org/10.1371/journal.pcbi.1004962>.
- Benesova, S., Kubista, M. and Valihrach, L. (2021) 'Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis', *Diagnostics*, 11(6), p. 964. Available at: <https://doi.org/10.3390/diagnostics11060964>.
- Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1), pp. 289–300. Available at: <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Birschwilks, M. *et al.* (2006) 'Transfer of phloem-mobile substances from the host plants to the holoparasite *Cuscuta* sp.', *Journal of Experimental Botany*, 57(4), pp. 911–921. Available at: <https://doi.org/10.1093/jxb/erj076>.
- Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. Available at: <https://doi.org/10.1093/bioinformatics/btu170>.
- Boratyn, G.M. *et al.* (2013) 'BLAST: a more efficient report with usability improvements', *Nucleic Acids Research*, 41(W1), pp. W29–W33. Available at: <https://doi.org/10.1093/nar/gkt282>.
- Boratyn, G.M. *et al.* (2019) 'Magic-BLAST, an accurate RNA-seq aligner for long and short reads', *BMC Bioinformatics*, 20(1), p. 405. Available at: <https://doi.org/10.1186/s12859-019-2996-x>.
- Bullard, J.H. *et al.* (2010) 'Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments', *BMC Bioinformatics*, 11(1), p. 94. Available at: <https://doi.org/10.1186/1471-2105-11-94>.

- Cai, G. *et al.* (2012) 'Accuracy of RNA-Seq and its dependence on sequencing depth', *BMC Bioinformatics*, 13(S13), p. S5. Available at: <https://doi.org/10.1186/1471-2105-13-S13-S5>.
- Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. Available at: <https://doi.org/10.1186/1471-2105-10-421>.
- Cardoso, C., Ruyter-Spira, C. and Bouwmeester, H.J. (2011) 'Strigolactones and root infestation by plant-parasitic Striga, Orobanche and Phelipanche spp.', *Plant Science*, 180(3), pp. 414–420. Available at: <https://doi.org/10.1016/j.plantsci.2010.11.007>.
- Casadesús, A. and Munné-Bosch, S. (2021) 'Holoparasitic plant–host interactions and their impact on Mediterranean ecosystems', *Plant Physiology*, 185(4), pp. 1325–1338. Available at: <https://doi.org/10.1093/plphys/kiab030>.
- Chang, W. *et al.* (2024) *shiny: Web Application Framework for R*. Available at: <https://CRAN.R-project.org/package=shiny>.
- Chaudron, C. *et al.* (2021) 'Introducing ecosystem engineers for grassland biodiversity conservation: A review of the effects of hemiparasitic Rhinanthus species on plant and animal communities at multiple trophic levels', *Perspectives in Plant Ecology, Evolution and Systematics*, 52, p. 125633. Available at: <https://doi.org/10.1016/j.ppees.2021.125633>.
- Chen, Y., Lun, A.T.L. and Smyth, G.K. (2016) 'From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline', *F1000Research*, 5, p. 1438. Available at: <https://doi.org/10.12688/f1000research.8987.2>.
- Clarke, C.R. *et al.* (2019) 'Molecular Dialog Between Parasitic Plants and Their Hosts', *Annual Review of Phytopathology*, 57(1), pp. 279–299. Available at: <https://doi.org/10.1146/annurev-phyto-082718-100043>.
- Cock, P.J.A. *et al.* (2010) 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Research*, 38(6), pp. 1767–1771. Available at: <https://doi.org/10.1093/nar/gkp1137>.
- Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*, 17(1), p. 13. Available at: <https://doi.org/10.1186/s13059-016-0881-8>.
- Cuccurullo, A., Nicolai, A. and Cardi, T. (2022) 'Resistance against broomrapes (Orobanche and Phelipanche spp.) in vegetables: a comprehensive view on classical and innovative breeding efforts', *Euphytica*, 218(6), p. 82. Available at: <https://doi.org/10.1007/s10681-022-03035-7>.

- Dafaallah, A.B. (2019) 'Biology and physiology of witchweed (*Striga* spp.): A review', *Int J Acad Multidiscip Res*, 3(10), pp. 42–51.
- Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*, 10(2), p. giab008. Available at: <https://doi.org/10.1093/gigascience/giab008>.
- Dayhoff, M.O. and Ledley, R.S. (1962) 'Comproteins: a computer program to aid primary protein structure determination', in *Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall). the December 4-6, 1962, fall joint computer conference*, Philadelphia, Pennsylvania: ACM Press, pp. 262–274. Available at: <https://doi.org/10.1145/1461518.1461546>.
- Delavault, P. (2015) 'Knowing the Parasite: Biology and Genetics of Orobanche', *Helia*, 38(62), pp. 15–29. Available at: <https://doi.org/10.1515/helia-2014-0030>.
- Deschamps-Francoeur, G., Simoneau, J. and Scott, M.S. (2020) 'Handling multi-mapped reads in RNA-seq', *Computational and Structural Biotechnology Journal*, 18, pp. 1569–1576. Available at: <https://doi.org/10.1016/j.csbj.2020.06.014>.
- Di, Y. *et al.* (2011) 'The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq', *Statistical Applications in Genetics and Molecular Biology*, 10(1). Available at: <https://doi.org/10.2202/1544-6115.1637>.
- Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. Available at: <https://doi.org/10.1093/bioinformatics/bts635>.
- Dong, Z. and Chen, Y. (2013) 'Transcriptomics: Advances and approaches', *Science China Life Sciences*, 56(10), pp. 960–967. Available at: <https://doi.org/10.1007/s11427-013-4557-2>.
- van Dorst, H.J.M. and Peters, D. (1974) 'Some biological observations on pale fruit, a viroid-incited disease of cucumber', *Netherlands Journal of Plant Pathology*, 80(3), pp. 85–96. Available at: <https://doi.org/10.1007/BF01980613>.
- Du, H. *et al.* (2022) 'Dual RNA-seq Reveals the Global Transcriptome Dynamics of *Ralstonia solanacearum* and Pepper (*Capsicum annuum*) Hypocotyls During Bacterial Wilt Pathogenesis', *Phytopathology*®, 112(3), pp. 630–642. Available at: <https://doi.org/10.1094/PHYTO-01-21-0032-R>.
- Edman, P. and Begg, G. (1967) 'A Protein Sequenator', *European Journal of Biochemistry*, 1(1), pp. 80–91. Available at: <https://doi.org/10.1111/j.1432-1033.1967.tb00047.x>.

- Edwards, K., Johnstone, C. and Thompson, C. (1991) 'A simple and rapid method for the preparation of plant genomic DNA for PCR analysis', *Nucleic Acids Research*, 19(6), pp. 1349–1349. Available at: <https://doi.org/10.1093/nar/19.6.1349>.
- El-Halmouch, Y., Benharrat, H. and Thalouarn, P. (2006) 'Effect of root exudates from different tomato genotypes on broomrape (*O. aegyptiaca*) seed germination and tubercle development', *Crop Protection*, 25(5), pp. 501–507. Available at: <https://doi.org/10.1016/j.cropro.2005.08.005>.
- Engreitz, J.M. *et al.* (2014) 'RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites', *Cell*, 159(1), pp. 188–199. Available at: <https://doi.org/10.1016/j.cell.2014.08.018>.
- Espindula, E. *et al.* (2019) 'The combined analysis as the best strategy for Dual RNA-Seq mapping', *Genetics and Molecular Biology*, 42(4), p. e20190215. Available at: <https://doi.org/10.1590/1678-4685-gmb-2019-0215>.
- Farajollahi, S. and Maas, S. (2010) 'Molecular diversity through RNA editing: a balancing act', *Trends in Genetics*, 26(5), pp. 221–230. Available at: <https://doi.org/10.1016/j.tig.2010.02.001>.
- Fernández-Aparicio, M., Delavault, P. and Timko, M.P. (2020) 'Management of Infection by Parasitic Weeds: A Review', *Plants*, 9(9), p. 1184. Available at: <https://doi.org/10.3390/plants9091184>.
- Fernández-Aparicio, M., Reboud, X. and Gibot-Leclerc, S. (2016) 'Broomrape Weeds. Underground Mechanisms of Parasitism and Associated Strategies for their Control: A Review', *Frontiers in Plant Science*, 7. Available at: <https://doi.org/10.3389/fpls.2016.00135>.
- Fernandez-Pozo, N. *et al.* (2015) 'The Sol Genomics Network (SGN)—from genotype to phenotype to breeding', *Nucleic Acids Research*, 43(D1), pp. D1036–D1041. Available at: <https://doi.org/10.1093/nar/gku1195>.
- Fishman, M.R. and Shirasu, K. (2021) 'How to resist parasitic plants: pre- and post-attachment strategies', *Current Opinion in Plant Biology*, 62, p. 102004. Available at: <https://doi.org/10.1016/j.pbi.2021.102004>.
- Galise, T.R., Esposito, S. and D'Agostino, N. (2021) 'Guidelines for Setting Up a mRNA Sequencing Experiment and Best Practices for Bioinformatic Data Analysis', in P. Tripodi (ed.) *Crop Breeding*. New York, NY: Springer US (Methods in Molecular Biology), pp. 137–162. Available at: https://doi.org/10.1007/978-1-0716-1201-9_10.

- García-Garrido, J.M. *et al.* (2009) 'Strigolactones, signals for parasitic plants and arbuscular mycorrhizal fungi', *Mycorrhiza*, 19(7), pp. 449–459. Available at: <https://doi.org/10.1007/s00572-009-0265-y>.
- Gatahi, D.M. (2020) 'Challenges and Opportunities in Tomato Production Chain and Sustainable Standards', *International Journal of Horticultural Science and Technology*, 7(3). Available at: <https://doi.org/10.22059/ijhst.2020.300818.361>.
- Gauthier, J. *et al.* (2019) 'A brief history of bioinformatics', *Briefings in Bioinformatics*, 20(6), pp. 1981–1996. Available at: <https://doi.org/10.1093/bib/bby063>.
- GNU, P. (2007) 'Free Software Foundation. Bash (3.2. 48)[Unix shell program]'.
- Gondane, A. and Itkonen, H.M. (2023) 'Revealing the History and Mystery of RNA-Seq', *Current Issues in Molecular Biology*, 45(3), pp. 1860–1874. Available at: <https://doi.org/10.3390/cimb45030120>.
- H. Pagès, P.A. (2017) 'Biostrings'. Bioconductor. Available at: <https://doi.org/10.18129/B9.BIOC.BIOSTRINGS>.
- Hartenstein, M., Albert, M. and Krause, K. (2023) 'The plant vampire diaries: a historic perspective on *Cuscuta* research', *Journal of Experimental Botany*. Edited by J. Lunn, 74(10), pp. 2944–2955. Available at: <https://doi.org/10.1093/jxb/erad082>.
- Haupt, S. *et al.* (2001) 'Macromolecular trafficking between *Nicotiana tabacum* and the holoparasite *Cuscuta reflexa*', *Journal of Experimental Botany*, 52(354), pp. 173–177. Available at: <https://doi.org/10.1093/jexbot/52.354.173>.
- Hays, J.B. (2002) 'Arabidopsis thaliana, a versatile model system for study of eukaryotic genome-maintenance functions', *DNA Repair*, 1(8), pp. 579–600. Available at: [https://doi.org/10.1016/S1568-7864\(02\)00093-9](https://doi.org/10.1016/S1568-7864(02)00093-9).
- Hegenauer, V., Körner, M. and Albert, M. (2017) 'Plants under stress by parasitic plants', *Current Opinion in Plant Biology*, 38, pp. 34–41. Available at: <https://doi.org/10.1016/j.pbi.2017.04.006>.
- Hochberg, Y. (1988) 'A sharper Bonferroni procedure for multiple tests of significance', *Biometrika*, 75(4), pp. 800–802. Available at: <https://doi.org/10.1093/biomet/75.4.800>.
- Holm, S. (1979) 'A simple sequentially rejective multiple test procedure', *Scandinavian journal of statistics*, pp. 65–70.

- Hommel, G. (1988) 'A stagewise rejective multiple test procedure based on a modified Bonferroni test', *Biometrika*, 75(2), pp. 383–386.
- Honaas, L.A. *et al.* (2013) 'Functional genomics of a generalist parasitic plant: Laser microdissection of host-parasite interface reveals host-specific patterns of parasite gene expression', *BMC Plant Biology*, 13(1), p. 9. Available at: <https://doi.org/10.1186/1471-2229-13-9>.
- Hosford, R.M. (1967) 'Transmission of plant viruses by dodder', *The Botanical Review*, 33(4), pp. 387–406. Available at: <https://doi.org/10.1007/BF02858742>.
- Hu, T. *et al.* (2021) 'Next-generation sequencing technologies: An overview', *Human Immunology*, 82(11), pp. 801–811. Available at: <https://doi.org/10.1016/j.humimm.2021.02.012>.
- Hudzik, C. *et al.* (2020) 'Exchange of Small Regulatory RNAs between Plants and Their Pests1 [OPEN]', *Plant Physiology*, 182(1), pp. 51–62. Available at: <https://doi.org/10.1104/pp.19.00931>.
- Huizinga, S. and Bouwmeester, H.J. (2023) 'Role of Strigolactones in the Host Specificity of Broomrapes and Witchweeds', *Plant And Cell Physiology*, 64(9), pp. 936–954. Available at: <https://doi.org/10.1093/pcp/pcad058>.
- Ikeue, D. *et al.* (2015) 'A bioinformatics approach to distinguish plant parasite and host transcriptomes in interface tissue by classifying RNA-Seq reads', *Plant Methods*, 11(1), p. 34. Available at: <https://doi.org/10.1186/s13007-015-0066-6>.
- Jhu, M.-Y. *et al.* (2022) 'Investigating Host and Parasitic Plant Interaction by Tissue-Specific Gene Analyses on Tomato and *Cuscuta campestris* Interface at Three Haustorial Developmental Stages', *Frontiers in Plant Science*, 12, p. 764843. Available at: <https://doi.org/10.3389/fpls.2021.764843>.
- Jhu, M.-Y. and Sinha, N.R. (2022) 'Cuscuta species: Model organisms for haustorium development in stem holoparasitic plants', *Frontiers in Plant Science*, 13. Available at: <https://doi.org/10.3389/fpls.2022.1086384>.
- Jia, K.-P., Baz, L. and Al-Babili, S. (2018) 'From carotenoids to strigolactones', *Journal of Experimental Botany*, 69(9), pp. 2189–2204. Available at: <https://doi.org/10.1093/jxb/erx476>.
- Jia, L. *et al.* (2022) 'Development of interactive biological web applications with R/Shiny', *Briefings in Bioinformatics*, 23(1), p. bbab415. Available at: <https://doi.org/10.1093/bib/bbab415>.

- Jiang, L. *et al.* (2011) 'Synthetic spike-in standards for RNA-seq experiments', *Genome Research*, 21(9), pp. 1543–1551. Available at: <https://doi.org/10.1101/gr.121095.111>.
- Jiang, Z. *et al.* (2015) 'Whole transcriptome analysis with sequencing: methods, challenges and potential solutions', *Cellular and Molecular Life Sciences*, 72(18), pp. 3425–3439. Available at: <https://doi.org/10.1007/s00018-015-1934-y>.
- Johnson, N.R. and Axtell, M.J. (2019) 'Small RNA warfare: exploring origins and function of *trans*-species microRNAs from the parasitic plant *Cuscuta*', *Current Opinion in Plant Biology*, 50, pp. 76–81. Available at: <https://doi.org/10.1016/j.pbi.2019.03.014>.
- Jolliffe, I.T. and Cadima, J. (2016) 'Principal component analysis: a review and recent developments', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), p. 20150202. Available at: <https://doi.org/10.1098/rsta.2015.0202>.
- Kado, T. and Innan, H. (2018) 'Horizontal Gene Transfer in Five Parasite Plant Species in Orobanchaceae', *Genome Biology and Evolution*. Edited by T. Dagan, 10(12), pp. 3196–3210. Available at: <https://doi.org/10.1093/gbe/evy219>.
- Kamińska, M. and Korbin, M. (1999) 'Graft and dodder transmission of phytoplasma affecting lily to experimental hosts', *Acta Physiologiae Plantarum*, 21(1), pp. 21–26. Available at: <https://doi.org/10.1007/s11738-999-0023-y>.
- Katoh, K. (2002) 'MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research*, 30(14), pp. 3059–3066. Available at: <https://doi.org/10.1093/nar/gkf436>.
- Khatri, P., Sirota, M. and Butte, A.J. (2012) 'Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges', *PLoS Computational Biology*. Edited by C.A. Ouzounis, 8(2), p. e1002375. Available at: <https://doi.org/10.1371/journal.pcbi.1002375>.
- Kienle, H.M. (2010) 'It's About Time to Take JavaScript (More) Seriously', *IEEE Software*, 27(3), pp. 60–62. Available at: <https://doi.org/10.1109/MS.2010.76>.
- Kim, G. *et al.* (2014) 'Genomic-scale exchange of mRNA between a parasitic plant and its hosts', *Science*, 345(6198), pp. 808–811. Available at: <https://doi.org/10.1126/science.1253122>.
- Klaren, C.H. and Janssen, G. (1978) 'Physiological Changes in the Hemiparasite *Rhinanthus serotinus* before and after Attachment', *Physiologia Plantarum*, 42(1), pp. 151–155. Available at: <https://doi.org/10.1111/j.1399-3054.1978.tb01556.x>.

- Koch, C.M. *et al.* (2018) 'A Beginner's Guide to Analysis of RNA Sequencing Data', *American Journal of Respiratory Cell and Molecular Biology*, 59(2), pp. 145–157. Available at: <https://doi.org/10.1165/rcmb.2017-0430TR>.
- Krebs, J.E., Goldstein, E.S. and Kilpatrick, S.T. (2017) *Lewin's Genes Twelve*. Jones & Bartlett Learning.
- Kretschmar, T. *et al.* (2012) 'A petunia ABC protein controls strigolactone-dependent symbiotic signalling and branching', *Nature*, 483(7389), pp. 341–344. Available at: <https://doi.org/10.1038/nature10873>.
- La Ferlita, A. *et al.* (2021) 'RNAdetector: a free user-friendly stand-alone and cloud-based system for RNA-Seq data analysis', *BMC Bioinformatics*, 22(1), p. 298. Available at: <https://doi.org/10.1186/s12859-021-04211-7>.
- Lambers, H. and Oliveira, R.S. (2019) 'Biotic Influences: Parasitic Associations', in H. Lambers and R.S. Oliveira (eds) *Plant Physiological Ecology*. Cham: Springer International Publishing, pp. 597–613. Available at: https://doi.org/10.1007/978-3-030-29639-1_15.
- LeBlanc, M., Kim, G. and Westwood, J.H. (2012) 'RNA trafficking in parasitic plant systems', *Frontiers in Plant Science*, 3. Available at: <https://doi.org/10.3389/fpls.2012.00203>.
- Li, J. and Tibshirani, R. (2013) 'Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data', *Statistical Methods in Medical Research*, 22(5), pp. 519–536. Available at: <https://doi.org/10.1177/0962280211428386>.
- Li, W.V. and Li, J.J. (2018) 'Modeling and analysis of RNA-seq data: a review from a statistical perspective', *Quantitative Biology*, 6(3), pp. 195–209. Available at: <https://doi.org/10.1007/s40484-018-0144-7>.
- Li, X., Yang, L. and Chen, L.-L. (2018) 'The Biogenesis, Functions, and Challenges of Circular RNAs', *Molecular Cell*, 71(3), pp. 428–442. Available at: <https://doi.org/10.1016/j.molcel.2018.06.034>.
- Li, Z., Qin, F. and Li, H. (2018) 'Chimeric RNAs and their implications in cancer', *Current Opinion in Genetics & Development*, 48, pp. 36–43. Available at: <https://doi.org/10.1016/j.gde.2017.10.002>.
- Liao, Z.-X. *et al.* (2019) 'Dual RNA-seq of *Xanthomonas oryzae* pv. *oryzicola* infecting rice reveals novel insights into bacterial-plant interaction', *PLOS ONE*. Edited by Y.-W. He, 14(4), p. e0215039. Available at: <https://doi.org/10.1371/journal.pone.0215039>.

- Lipman, D.J. and Pearson, W.R. (1985) 'Rapid and Sensitive Protein Similarity Searches', *Science*, 227(4693), pp. 1435–1441. Available at: <https://doi.org/10.1126/science.2983426>.
- Liu, P. and Si, Y. (2014) 'Cluster Analysis of RNA-Sequencing Data', in S. Datta and D. Nettleton (eds) *Statistical Analysis of Next Generation Sequencing Data*. Cham: Springer International Publishing, pp. 191–217. Available at: https://doi.org/10.1007/978-3-319-07212-8_10.
- Liu, Y., Zhou, J. and White, K.P. (2014) 'RNA-seq differential expression studies: more sequence or more replication?', *Bioinformatics*, 30(3), pp. 301–304. Available at: <https://doi.org/10.1093/bioinformatics/btt688>.
- Liu, Yuansheng *et al.* (2023) 'Sequence Alignment/Map format: a comprehensive review of approaches and applications', *Briefings in Bioinformatics*, 24(5), p. bbad320. Available at: <https://doi.org/10.1093/bib/bbad320>.
- Lovén, J. *et al.* (2012) 'Revisiting Global Gene Expression Analysis', *Cell*, 151(3), pp. 476–482. Available at: <https://doi.org/10.1016/j.cell.2012.10.012>.
- Lund, S.P. *et al.* (2012) 'Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunk Dispersion Estimates', *Statistical Applications in Genetics and Molecular Biology*, 11(5). Available at: <https://doi.org/10.1515/1544-6115.1826>.
- Madeira, F. *et al.* (2024) 'The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024', *Nucleic Acids Research*, 52(W1), pp. W521–W525. Available at: <https://doi.org/10.1093/nar/gkae241>.
- Maizel, A. *et al.* (2020) 'To move or not to move: roles and specificity of plant RNA mobility', *Current Opinion in Plant Biology*, 57, pp. 52–60. Available at: <https://doi.org/10.1016/j.pbi.2020.05.005>.
- Maleki, F. *et al.* (2020) 'Gene Set Analysis: Challenges, Opportunities, and Future Research', *Frontiers in Genetics*, 11, p. 654. Available at: <https://doi.org/10.3389/fgene.2020.00654>.
- Mamedov, I.Z. *et al.* (2013) 'Preparing Unbiased T-Cell Receptor and Antibody cDNA Libraries for the Deep Next Generation Sequencing Profiling', *Frontiers in Immunology*, 4. Available at: <https://doi.org/10.3389/fimmu.2013.00456>.
- Manhart, J.R. (1994) 'Phylogenetic Analysis of Green Plant rbcL Sequences', *Molecular Phylogenetics and Evolution*, 3(2), pp. 114–127. Available at: <https://doi.org/10.1006/mpev.1994.1014>.

- Marasco, L.E. and Kornblihtt, A.R. (2023) 'The physiology of alternative splicing', *Nature Reviews Molecular Cell Biology*, 24(4), pp. 242–254. Available at: <https://doi.org/10.1038/s41580-022-00545-z>.
- Marsh, J.W. *et al.* (2017) 'Bioinformatic analysis of bacteria and host cell dual RNA-sequencing experiments', *Briefings in Bioinformatics* [Preprint]. Available at: <https://doi.org/10.1093/bib/bbx043>.
- Martinez, W.L. (2011) 'Graphical user interfaces', *WIREs Computational Statistics*, 3(2), pp. 119–133. Available at: <https://doi.org/10.1002/wics.150>.
- Mashiguchi, K., Seto, Y. and Yamaguchi, S. (2021) 'Strigolactone biosynthesis, transport and perception', *The Plant Journal*, 105(2), pp. 335–350. Available at: <https://doi.org/10.1111/tpj.15059>.
- Mateus, I.D. *et al.* (2019) 'Dual RNA-seq reveals large-scale non-conserved genotype × genotype-specific genetic reprogramming and molecular crosstalk in the mycorrhizal symbiosis', *The ISME Journal*, 13(5), pp. 1226–1238. Available at: <https://doi.org/10.1038/s41396-018-0342-3>.
- Mauromicale, G., Monaco, A.L. and Longo, A.M.G. (2008) 'Effect of Branched Broomrape (*Orobanche ramosa*) Infection on the Growth and Photosynthesis of Tomato', *Weed Science*, 56(4), pp. 574–581. Available at: <https://doi.org/10.1614/WS-07-147.1>.
- Mercer, T.R. *et al.* (2014) 'Targeted sequencing for gene discovery and quantification using RNA CaptureSeq', *Nature Protocols*, 9(5), pp. 989–1009. Available at: <https://doi.org/10.1038/nprot.2014.058>.
- Minh, B.Q. *et al.* (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*. Edited by E. Teeling, 37(5), pp. 1530–1534. Available at: <https://doi.org/10.1093/molbev/msaa015>.
- Minuto, A. *et al.* (2006) 'Control of soilborne pathogens of tomato using a commercial formulation of *Streptomyces griseoviridis* and solarization', *Crop Protection*, 25(5), pp. 468–475. Available at: <https://doi.org/10.1016/j.cropro.2005.08.001>.
- Mortazavi, A. *et al.* (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*, 5(7), pp. 621–628. Available at: <https://doi.org/10.1038/nmeth.1226>.
- Mubeen, S. *et al.* (2022) 'On the influence of several factors on pathway enrichment analysis', *Briefings in Bioinformatics*, 23(3), p. bbac143. Available at: <https://doi.org/10.1093/bib/bbac143>.

- Mutuku, J.M. *et al.* (2019) 'The Structural Integrity of Lignin Is Crucial for Resistance against *Striga hermonthica* Parasitism in Rice', *Plant Physiology*, 179(4), pp. 1796–1809. Available at: <https://doi.org/10.1104/pp.18.01133>.
- Naidoo, S. *et al.* (2018) 'Dual RNA-Sequencing to Elucidate the Plant-Pathogen Duel', *Current Issues in Molecular Biology*, pp. 127–142. Available at: <https://doi.org/10.21775/cimb.027.127>.
- Nickrent, D. and Musselman, L. (2004) 'Introduction to Parasitic Flowering Plants', *The Plant Health Instructor*, 13. Available at: <https://doi.org/10.1094/PHI-I-2004-0330-01>.
- Nickrent, D.L. (2020) 'Parasitic angiosperms: how often and how many?', *Taxon*, 69(1), pp. 5–27.
- O'Keeffe, K.R. and Jones, C.D. (2019) 'Challenges and solutions for analysing dual RNA -seq data for non-model host–pathogen systems', *Methods in Ecology and Evolution*. Edited by S. Kembel, 10(3), pp. 401–414. Available at: <https://doi.org/10.1111/2041-210X.13135>.
- Park, I. *et al.* (2019) 'Cuscuta Species Identification Based on the Morphology of Reproductive Organs and Complete Chloroplast Genome Sequences', *International Journal of Molecular Sciences*, 20(11), p. 2726. Available at: <https://doi.org/10.3390/ijms20112726>.
- Park, S.-Y. *et al.* (2022) 'Mobile Host mRNAs Are Translated to Protein in the Associated Parasitic Plant *Cuscuta campestris*', *Plants*, 11(1), p. 93. Available at: <https://doi.org/10.3390/plants11010093>.
- Patterson, J. *et al.* (2019) 'Impact of sequencing depth and technology on de novo RNA-Seq assembly', *BMC Genomics*, 20(1), p. 604. Available at: <https://doi.org/10.1186/s12864-019-5965-x>.
- Pérez-De-Luque, A. *et al.* (2009) 'Understanding *Orobanche* and *Phelipanche* –host plant interactions and developing resistance', *Weed Research*, 49(s1), pp. 8–22. Available at: <https://doi.org/10.1111/j.1365-3180.2009.00738.x>.
- Petitot, A. *et al.* (2016) 'Dual RNA-seq reveals *Meloidogyne graminicola* transcriptome and candidate effectors during the interaction with rice plants', *Molecular Plant Pathology*, 17(6), pp. 860–874. Available at: <https://doi.org/10.1111/mpp.12334>.
- R Core Team (2024) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.

- Rau, A. *et al.* (2013) 'Data-based filtering for replicated high-throughput transcriptome sequencing experiments', *Bioinformatics*, 29(17), pp. 2146–2152. Available at: <https://doi.org/10.1093/bioinformatics/btt350>.
- Robinson, M.D. and Oshlack, A. (2010) 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biology*, 11(3), p. R25. Available at: <https://doi.org/10.1186/gb-2010-11-3-r25>.
- Robinson, M.D. and Smyth, G.K. (2007) 'Small-sample estimation of negative binomial dispersion, with applications to SAGE data', *Biostatistics*, 9(2), pp. 321–332. Available at: <https://doi.org/10.1093/biostatistics/kxm030>.
- Russo, F. and Angelini, C. (2014) 'RNASeqGUI: a GUI for analysing RNA-Seq data', *Bioinformatics*, 30(17), pp. 2514–2516. Available at: <https://doi.org/10.1093/bioinformatics/btu308>.
- Ruyter-Spira, C. *et al.* (2013) 'The biology of strigolactones', *Trends in Plant Science*, 18(2), pp. 72–83. Available at: <https://doi.org/10.1016/j.tplants.2012.10.003>.
- Schurch, N.J. *et al.* (2016) 'How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?', *RNA*, 22(6), pp. 839–851. Available at: <https://doi.org/10.1261/rna.053959.115>.
- Seto, Y. *et al.* (2012) 'Recent Advances in Strigolactone Research: Chemical and Biological Aspects', *Plant and Cell Physiology*, 53(11), pp. 1843–1853. Available at: <https://doi.org/10.1093/pcp/pcs142>.
- Seto, Y. and Yamaguchi, S. (2014) 'Strigolactone biosynthesis and perception', *Current Opinion in Plant Biology*, 21, pp. 1–6. Available at: <https://doi.org/10.1016/j.pbi.2014.06.001>.
- Shahid, S. *et al.* (2018) 'MicroRNAs from the parasitic plant *Cuscuta campestris* target host messenger RNAs', *Nature*, 553(7686), pp. 82–85. Available at: <https://doi.org/10.1038/nature25027>.
- Sisou, D. *et al.* (2021) 'Biological and Transcriptomic Characterization of Pre-Haustorial Resistance to Sunflower Broomrape (*Orobanche cumana* W.) in Sunflowers (*Helianthus annuus*)', *Plants*, 10(9), p. 1810. Available at: <https://doi.org/10.3390/plants10091810>.
- Slovin, S. *et al.* (2021) 'Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview', in E. Picardi (ed.) *RNA Bioinformatics*. New York, NY: Springer US (Methods in Molecular Biology), pp. 343–365. Available at: https://doi.org/10.1007/978-1-0716-1307-8_19.

- Smith, D.R. (2013) 'The battle for user-friendly bioinformatics', *Frontiers in Genetics*, 4. Available at: <https://doi.org/10.3389/fgene.2013.00187>.
- Solomon, N. *et al.* (2015) 'A Review of Host-Parasite Relationships', *Annual Research & Review in Biology*, 5(5), pp. 372–384. Available at: <https://doi.org/10.9734/ARRB/2015/10263>.
- Soltis, D.E. *et al.* (2011) 'Angiosperm phylogeny: 17 genes, 640 taxa', *American Journal of Botany*, 98(4), pp. 704–730. Available at: <https://doi.org/10.3732/ajb.1000404>.
- Soskine, M. and Tawfik, D.S. (2010) 'Mutational effects and the evolution of new protein functions', *Nature Reviews Genetics*, 11(8), pp. 572–582. Available at: <https://doi.org/10.1038/nrg2808>.
- Spies, D. and Ciaudo, C. (2015) 'Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis', *Computational and Structural Biotechnology Journal*, 13, pp. 469–477. Available at: <https://doi.org/10.1016/j.csbj.2015.08.004>.
- Srivastava, A. *et al.* (2020) 'Alignment and mapping methodology influence transcript abundance estimation', *Genome Biology*, 21(1), p. 239. Available at: <https://doi.org/10.1186/s13059-020-02151-8>.
- Subhankar, B. *et al.* (2021) 'Trans-species small RNAs move long distances in a parasitic plant complex', *Plant Biotechnology*, 38(2), pp. 187–196. Available at: <https://doi.org/10.5511/plantbiotechnology.21.0121a>.
- Susatya, A. (2020) 'The growth of flower bud, life history, and population structure of *Rafflesia arnoldii* (Rafflesiaceae) in Bengkulu, Sumatra, Indonesia', *Biodiversitas Journal of Biological Diversity*, 21(2).
- Tarazona, S. *et al.* (2011) 'Differential expression in RNA-seq: A matter of depth', *Genome Research*, 21(12), pp. 2213–2223. Available at: <https://doi.org/10.1101/gr.124321.111>.
- Teichman, G. *et al.* (2023) 'RNAlysis: analyze your RNA sequencing data without writing a single line of code', *BMC Biology*, 21(1), p. 74. Available at: <https://doi.org/10.1186/s12915-023-01574-6>.
- Tokuriki, N. and Tawfik, D.S. (2009) 'Stability effects of mutations and protein evolvability', *Current Opinion in Structural Biology*, 19(5), pp. 596–604. Available at: <https://doi.org/10.1016/j.sbi.2009.08.003>.
- Twyford, A.D. (2018) 'Parasitic plants', *Current Biology*, 28(16), pp. R857–R859. Available at: <https://doi.org/10.1016/j.cub.2018.06.030>.

- Valihrach, L. *et al.* (2024) 'A practical guide to spatial transcriptomics', *Molecular Aspects of Medicine*, 97, p. 101276. Available at: <https://doi.org/10.1016/j.mam.2024.101276>.
- Walker, P.L. *et al.* (2024) 'Dual RNA-sequencing of Fusarium head blight resistance in winter wheat', *Frontiers in Plant Science*, 14. Available at: <https://doi.org/10.3389/fpls.2023.1299461>.
- Walter, M.H. and Strack, D. (2011) 'Carotenoids and their cleavage products: Biosynthesis and functions', *Natural Product Reports*, 28(4), p. 663. Available at: <https://doi.org/10.1039/c0np00036a>.
- Wang, L.-G. *et al.* (2020) 'Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data', *Molecular Biology and Evolution*. Edited by S. Kumar, 37(2), pp. 599–603. Available at: <https://doi.org/10.1093/molbev/msz240>.
- Waters, M.T. *et al.* (2017) 'Strigolactone Signaling and Evolution', *Annual Review of Plant Biology*, 68(1), pp. 291–322. Available at: <https://doi.org/10.1146/annurev-arplant-042916-040925>.
- Westermann, A.J., Barquist, L

- . and Vogel, J. (2017) 'Resolving host–pathogen interactions by dual RNA-seq', *PLOS Pathogens*. Edited by J.B. Bliska, 13(2), p. e1006033. Available at: <https://doi.org/10.1371/journal.ppat.1006033>.
- Westermann, A.J., Gorski, S.A. and Vogel, J. (2012) 'Dual RNA-seq of pathogen and host', *Nature Reviews Microbiology*, 10(9), pp. 618–630. Available at: <https://doi.org/10.1038/nrmicro2852>.
- Westwood, J.H. *et al.* (2010) 'The evolution of parasitism in plants', *Trends in Plant Science*, 15(4), pp. 227–235. Available at: <https://doi.org/10.1016/j.tplants.2010.01.004>.
- Westwood, J.H. and Kim, G. (2017) 'RNA mobility in parasitic plant – host interactions', *RNA Biology*, 14(4), pp. 450–455. Available at: <https://doi.org/10.1080/15476286.2017.1291482>.
- Wicke, S. and Naumann, J. (2018) 'Molecular Evolution of Plastid Genomes in Parasitic Flowering Plants', in *Advances in Botanical Research*. Elsevier, pp. 315–347. Available at: <https://doi.org/10.1016/bs.abr.2017.11.014>.
- Wilhelm, B.T. and Landry, J.-R. (2009) 'RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing', *Methods*, 48(3), pp. 249–257. Available at: <https://doi.org/10.1016/j.ymeth.2009.03.016>.
- Wolf, T. *et al.* (2018) 'Two's company: studying interspecies relationships with dual RNA-seq', *Current Opinion in Microbiology*, 42, pp. 7–12. Available at: <https://doi.org/10.1016/j.mib.2017.09.001>.
- Wu, F. *et al.* (2022) 'Biological Functions of Strigolactones and Their Crosstalk With Other Phytohormones', *Frontiers in Plant Science*, 13, p. 821563. Available at: <https://doi.org/10.3389/fpls.2022.821563>.
- Wu, Y. *et al.* (2022) 'Bidirectional lncRNA Transfer between Cuscuta Parasites and Their Host Plant', *International Journal of Molecular Sciences*, 23(1), p. 561. Available at: <https://doi.org/10.3390/ijms23010561>.
- Xie, C., Jauhari, S. and Mora, A. (2021) 'Popularity and performance of bioinformatics software: the case of gene set analysis', *BMC Bioinformatics*, 22(1), p. 191. Available at: <https://doi.org/10.1186/s12859-021-04124-5>.
- Yang, Z. *et al.* (2015) 'Comparative Transcriptome Analyses Reveal Core Parasitism Genes and Suggest Gene Duplication and Repurposing as Sources of Structural Novelty', *Molecular Biology and Evolution*, 32(3), pp. 767–790. Available at: <https://doi.org/10.1093/molbev/msu343>.

- Yang, Z. *et al.* (2019) 'Convergent horizontal gene transfer and cross-talk of mobile nucleic acids in parasitic plants', *Nature Plants*, 5(9), pp. 991–1001. Available at: <https://doi.org/10.1038/s41477-019-0458-0>.
- Ying Sha, Phan, J.H. and Wang, M.D. (2015) 'Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data', in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan: IEEE, pp. 6461–6464. Available at: <https://doi.org/10.1109/EMBC.2015.7319872>.
- Yoder, J.I. and Scholes, J.D. (2010) 'Host plant resistance to parasitic weeds; recent progress and bottlenecks', *Current opinion in plant biology*, 13(4), pp. 478–484.
- Young, M.D. *et al.* (2012) 'Differential Expression for RNA Sequencing (RNA-Seq) Data: Mapping, Summarization, Statistical Analysis, and Experimental Design', in N. Rodríguez-Ezpeleta, M. Hackenberg, and A.M. Aransay (eds) *Bioinformatics for High Throughput Sequencing*. New York, NY: Springer New York, pp. 169–190. Available at: https://doi.org/10.1007/978-1-4614-0782-9_10.
- Yu, G. (2020) 'Using ggtree to Visualize Data on Tree-Like Structures', *Current Protocols in Bioinformatics*, 69(1), p. e96. Available at: <https://doi.org/10.1002/cpbi.96>.
- Zambelli, F. and Pavesi, G. (2015) 'RIP-Seq Data Analysis to Determine RNA–Protein Associations', in E. Picardi (ed.) *RNA Bioinformatics*. New York, NY: Springer New York (Methods in Molecular Biology), pp. 293–303. Available at: https://doi.org/10.1007/978-1-4939-2291-8_18.
- Zhang, B. and Horvath, S. (2005) 'A General Framework for Weighted Gene Co-Expression Network Analysis', *Statistical Applications in Genetics and Molecular Biology*, 4(1). Available at: <https://doi.org/10.2202/1544-6115.1128>.
- Zhao, S. *et al.* (2015) 'Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap', *BMC Genomics*, 16(1), p. 675. Available at: <https://doi.org/10.1186/s12864-015-1876-7>.
- Zyprych-Walczak, J. *et al.* (2015) 'The Impact of Normalization Methods on RNA-Seq Data Analysis', *BioMed Research International*, 2015, pp. 1–10. Available at: <https://doi.org/10.1155/2015/621690>.

Author's publication

Cuomo, P., Medaglia, C., Casillo, A., Gentile, A., **Fruggiero, C.**, Corsaro, M. M., Capparelli, R. (2024) Phage-resistance alters Lipid A reactogenicity: a new strategy for LPS-based conjugate vaccines against Salmonella Rissen. *Frontiers in Immunology*, 15, 1450600. DOI: 10.3389/fimmu.2024.1450600.

Fruggiero C.*, Aufiero G.*, D'Angelo D., Pasolli E., D'Agostino N. (2024) Refining dual RNA-seq mapping: sequential and combined approaches in host-parasitic plant dynamics. *Frontiers in Plant Science*, 15:1483717. DOI: 10.3389/fpls.2024.1483717.

*Contributed equally

Aufiero G.*, **Fruggiero C.***, D'Angelo D., D'Agostino N. (2024) Homoeologs in allopolyploids: navigating redundancy as both an evolutionary opportunity and a technical challenge - A transcriptomics perspective. *Genes*, 15(8):977. DOI: 10.3390/genes15080977.

*Contributed equally

Fruggiero C., Aufiero G., D'Agostino N. (2024) Bioinformatics: Analyzing Data From Living Things. *Front. Young Minds* 12:1266091. DOI: 10.3389/frym.2024.1266091.

