ABSTRACT

CHA, SOYEON. Is the Biology of *M. chitwoodi* and *M. hapla* Reflected in Their Genomes? (Under the direction of Dr. David McK. Bird).

The plant parasitic root-knot nematodes (RKN; *Meloidogyne* spp.) are responsible for substantial crop losses. In this dissertation, I present an optimized *do novo* assembly of RKN (*M. chitwoodi*) genome and an effective way for its automated annotation. Based on a variant of simple grid search, it is demonstrated that empirical optimization of key parameters from the assembler software has a disproportionate influence on *de novo* assembly of the genome. The RKN and plant transcriptomes were collected throughout their parasitic developmental stages and at several time-points during a 24-hour period to study their spatio-temporal interactions. Based on differential expression analysis, it was demonstrated that overall the expression of RKN (*M. hapla*) genes was likely to increase throughout the infection period, it being more active at night. In contrast, parasitic plant genes showed a decreasing trend of expression after RKN infection, with increased expression observed during daytime. Furthermore, the expressions of RKN genes exhibited oscillation in accordance with the circadian rhythm of plants. Lastly, to inspect how the genomes of the two RKN species, *M. hapla* and *M. chitwoodi*, have evolved along evolutionary history for adaptation, their genomes were compared to another closely related RKN species, *M. incognita*, and a free-living nematode, *C. elegans*. With the featured synteny, varied types of conserved regions and sources of genomic arrangement were identified. Besides other advantages, observing synteny at micro level could enhance the missing annotation of *M. hapla* genome. In addition, the automated annotation of *M. chitwoodi* genome could be reinforced by identifying orthologous *M. hapla* genes in conjunction with the evaluation of their transcriptome profiles.

Is the Biology of *M. chitwoodi* and *M. hapla* Reflected in Their Genomes?

by
Soyeon Cha

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2016

APPROVED BY:

_____
Dr. David McK. Bird
Committee Chair

_____
Dr. Dahlia M. Nielsen

_____
Dr. Charles Opperman

_____
Dr. Eric A. Stone

_____
Dr. Colleen Doherty

DEDICATION

To my parents and brother.

BIOGRAPHY

Personal History

- Born in Masan, Republic of Korea, Autumn 1987
- Grew up in Changwon for childhood years, ~ Spring 1998
- Educated in Masan for middle and high school, ~ Winter 2005
- Spent early 20's in Seoul, ~ Spring 2011
- Spent mid and late 20's in Raleigh, ~ Summer 2016

Education

- B.S. in Applied Statistics, double major in Economics,
  Yonsei University, Spring 2006 ~ Spring 2011
- Ph.D. in Bioinformatics,
  North Carolina State University, Autumn 2011 ~ Summer 2016

Experiences

- Masan Union Youth Volunteers, vice president, 2003 ~ 2005
- Yonsei Amateur Orchestra, viola member, 2006 ~ 2010
- Internship Program, Research International Korea Inc., research assistant, Spring 2008
- Exchange Student Program, Montana State University, Spring 2009
- NCSU Korean Student Association, vice president, Autumn 2014 ~ Spring 2016

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. David McK. Bird who supported me mentally, emotionally, and academically with his thoughtful patience and considerate understanding. He led me to build up a fundamental attitude as a researcher and to develop academic competence in my career. Also, I would like to thank Dr. Dahlia M. Nielsen for her timely advice, analytical critiques, and open-mindedness especially when I was a novice in my early years of Ph.D. studies. I am also grateful to Dr. Colleen Doherty for her personal and professional support, encouraging scientific communication of a broad range of biological questions. I would like to thank Dr. Charles Opperman for always being willing to share his knowledge and for working one-on-one to give helpful advice on my dissertation. I am thankful to Dr. Eric A. Stone for his enthusiastic and inspirational guidance, time, support, and commitment.

I am sincerely thankful to Chris Smith for helping to resolve technical problems whenever I was often faced with computing problems. I would like to thank Dr. Peter DiGennaro for helpful discussions and assistance. Additionally, I would like to extend my thanks to Dr. Yuelong Guo who shared insightful knowledge to answer my many questions and was a great team mate of our group projects. I further extend my thanks to Dr. Qiwen Hu for her advice and support, and other classmates and BRC friends, Dr. Zhongkai Liu, Cai Li, Yiwen Luo, Yunbo Cai, Guozhu Zhang, Shuping Ruan, Dr. Kwangyu Wang, and Dr. Elizabeth Scholl for their assistance.

I would like to acknowledge Dr. Zhao-Bang Zeng who offered admission for my entrance to the Bioinformatics program. I am indebted to Dr. Hakbae Lee, Dr. Chuleung

Kim, Dr. Seungho Kang, Dr. Yongho Jeon, and Dr. Junil Kim who encouraged me to pursue an academic career in the first place. I would like to express appreciation to Dr. Sangun Park, Dr. Sunku Hahn, Dr. Chaehyoung Ahn, Dr. Hyewon Byun, and Eunyoung Heo who shared their wisdom and valuable time as mentors in my early 20's. I further thank all my precious friends in Raleigh and in South Korea.

Finally, I would like to highlight my warmest gratitude to my parents and brother for their everlasting love throughout my whole life. And thanks to God.

TABLE OF CONTENTS

[1] Reprinted with partial reproduction by permission of Cha and Bird. 2016. Optimizing k-mer size using a variant grid search to enhance *de novo* genome assembly. Bioinformation. 12(2): 36-40.

[2] Portion of this chapter has been submitted for publication. Authors: Cha, DiGennaro, and Bird.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

1. INTRODUCTION

1.1 Nematoda

Nematodes originated 650–1,200 million years ago and are the most abundant and speciose metazoans (Blaxter 1998). The phylum Nematoda is estimated to be composed of around 10^6 to 10^8 species; however, only 25,000 of these have been described (Zhang 2013), which account for up to 80% of the members of kingdom Animalia (Decraemer and Hunt 2006). It is believed that over the long course of nematode evolution, some free-living nematodes evolved into parasites (Blaxter et al. 1998; Blaxter and Koutsovoulos 2015). Phylogenetic analysis indicates that parasitism evolved through multiple origins, at least four times with animal parasites and three times with plant parasites. Particularly, each plant parasitic lineage of Tylenchomorpha, Triplonchida, and Dorylaimida has arisen from three disparate clades. Altough the orders Dorylaimida and Triplonchida are phylogenetically distant, nematode species in each of these orders share an extraordinary parasitic strategy of transmitting plant viruses. This process by which distantly-related organisms acquire a similar trait by adaptation is called convergent evolution. Each parasitic species has evolved its own lifestyle for survival. To occupy diverse ecological habitats, including sea, hot springs, freshwater, and soil (Perry and Moens 2011) as well as regions with climate as varied as that of tropics, poles, and deserts, the body of parasites is specifically adapted (Shannon et al. 2005).

Among 25,000 described species in phylum Nematoda, about 10,000 are free-living (Poulin and Morand 2000; Hugot et al. 2001), feeding on microbes such as bacteria and fungi, rotting organic matter, or other nematodes (Ferris 2010). Some of them have beneficial impact on ecosystem by decomposing organic materials, delivering nutrients and energy to plants and animals and attacking insect pests (Freckman 1988; Lacey and Georgis 2012). In contrast, there are some parasitic nematodes, which comprise more than half of the phylum, including around 4,000 plant, 3,500 invertebrate, and 8,400 vertebrate parasites (Poulin and Morand 2000; Hugot et al. 2001). They infect human or livestock, causing blindness, anemia, damages to intestinal and respiratory systems, and even transmit infections from animals to humans (Jasmer et al. 2003).

Moreover, plant parasitic nematodes (PPNs) cause an estimated $130 billion annual crop losses worldwide (Elling 2013). To counter their devastating effects on crop yields, development of nematode control as well as resistant cultivars has been an important area of research in plant pathology over several decades (Williamson 1999; Chitwood 2002; Chitwood 2003; Thurau et al. 2010). More effective combat strategies are expected to be formulated from a deeper understanding of PPN biology.

1.2 Plant Parasitic Nematodes

Plant parasitic nematodes are broadly categorized to possess one of the four lifestyles: sedentary or migratory endoparasitic and sedentary or migratory ectoparasitic. Of the PPNs possessing these lifestyles, the sedentary endoparasites are considered the most damaging; these include the root-knot nematodes (RKN, *Meloidogyne* spp.) and cyst nematodes (CN,

Globodera spp., *Heterodera* spp.). In general, RKN has a wide host range and occurs in almost all vascular plants, whereas CN has a narrow host range and primarily infects soybean and other legume crops, such as peas, vetch, and clovers.

Life Cyle. All nematodes undergo six lifecycle stages: the egg stage, four juvenile (or larval) stages, and an adult stage. Throughout their parasitic lifecycle, RKN and CN show subtle distinctions. Mature female RKN can lay hundreds of eggs in a gelatinous matrix outside the plant root, which molt to developmentally arrested second-stage juveniles (J2; Bartlem et al. 2014). Unlike RKN, most of the eggs laid by CN are retained within the body of dead females, in which cuticles are transformed to a hardened structure called cyst. After hatching, the infective J2 penetrates the host plant near the root tip in apical meristematic zone of elongation. Whereas the RKN J2 intercellularly migrates between the cortical cells, CN J2 intracellularly migrates through the cortical cells. After one to two days of penetration, upon reaching the developing vascular cylinder, a sedentary RKN J2 induces three to six giant cells (GC), which are formed by plant nuclear division without cytokinesis. In contrast, CN J2 induces syncytium, which is formed by dissolution of hundreds of nighboring cells. These specialized structures can serve as intermediaries between the phloem stream and the feeding worm (Bird 1996). After the feeding site formation, J2 undergoes three molts within the cuticle of the second stage, without feeding, to become an adult. Under favorable environmental conditions, primarily depending on temperature, the life cycle of nematode is usually completed in one to two months of reproduction (Niebel et al. 1994). CN eggs, retained in the cyst, can withstand adverse environmental conditions for several years (Gundy 1965; Perry 1989). Although RKN eggs are unable to tolerate such conditions for

long, they can still remain intact for months by remaining unhatched in a dormant state called diapause (Guiran 1979). Moreover, the dormant dauer juveniles, J2 through J4, are tolerant to unfavorable conditions of moisture, temperature, and chemicals. Female adults lay eggs over their entire lifespan of several months (Gems 2000).

Plant-RKN Interaction. During nematode parasitism, a variety of morphological and physiological changes occur in the host plant and parasites. The noticeable pathogenic symptoms in the above-ground plant parts include wilt, chlorosis, and stunting, which are caused due to nutrient deficiency (Mai and Abawi 1987). The symptoms in the parts below the ground include morphological changes such as induction of syncytium by CN and swollen galls surrounding the giant cells formed by RKN (Williamson and Gleason 2003). These structures function as feeding sites for the respective nematode. For feeding site formation, PPNs are believed to secrete products central to plant parasitism through a hollow mouth spear called a stylet. The secretions originate from three pharyngeal gland cells, two subventral glands, and one dorsal gland (Hussey and Mims 1990). For many years, the molecules secreted from these organs have been the subject of studies for identification of key protein effectors that aid in parasitism (Goverse et al. 1999; Wang et al. 2001; Olsen and Skriver 2003; Huang et al. 2006; Patel et al. 2010; Rehman et al. 2009; Replogle et al. 2011). Subventral gland cells become enlarged and active, especially in the early stage of parasitism during nematode penetration and migration in the host plant, and secrete cell wall modifying enzymes (Bird 1968; Bird and Kaloshian 2002). These cell wall- degrading or modifying enzymes were the first discovered protein effectors, that include β-1,4-endoglucanases, pectate lyases, polygalacturonases, and expansins (Quentin et al. 2013). In the later stage of

parasitism, the dorsal gland cell becomes more predominantly enlarged, producing secretory proteins for feeding site development and maintenance (Bird 2004). The first secretory peptide discovered from nematode dorsal gland cell was a novel protein with high sequence similarity to plant peptide hormone, CLAVATA3/Embryo surrounding region (CLV3/ESR or CLE; Wang et al. 2001; Olsen and Skriver 2003). Subsequently, other plant peptide analogs were also identified, which include SPRY domain-containing protein (SPRYSEC; Rehman et al. 2009) and C-terminally encoded peptide (CEP)-like genes (Bobay et al. 2013). Growing evidence points to the mimicking of plant molecules involved in differentiation of meristem cells and lateral root development by RKN-encoded CLE or CEP and to the redirection of plant developmental signals for triggering the formation of feeding site (Lu et al. 2009; Replogle et al. 2011; Imin et al. 2013; Ogilvie et al. 2014; Rehman et al. 2016). Furthermore, PPNs may secrete other protein effectors including chorismate mutase, peroxidases, glutathione S-transferase (GST), superoxide dismutase, and other detoxifying molecules to avoid, inhibit, and suppress plant signaling pathways involved in eliciting the immune responses in plants.

Plant Immunity. To counter pathogen infection, plants have evolved their own immune system (Jones and Dangl 2006; Denancé et al. 2013; Hewezi and Baum 2013). This involves two compoenets; in the first, known as “pattern-triggered immunity (PTI)”, receptor-like transmembrane proteins known as pattern recognition receptors (PRRs) detect pathogen-associated molecular patterns (PAMPs) whereas the second, designated as “effector-triggered immunity (ETI), is a stronger immune system, which has coevolved with nematode secretome effectors that suppress PTI responses. Through ETI, plants activate many

resistance (R) genes and immune receptors with nucleotide-binding domain and leucine-rich domainS. On elicitation of these immune responses, plants activate defensive phytohormone signaling pathways including the salicylic acid (SA)- and jasmonic acid (JA)/ethylene (ET)-dependent networks. Moreover, other pathways involving phytohormones, such as auxin, gibberellins, and cytokinins may work together in coordinated networks to backbone the SA- and JA-signaling pathways as well as to regulate plant development in response to biotic or abiotic stimuli.

In SA-mediated regulation, chorismate is converted to salicylic acid which functions as a defense activator of common systemic mechanisms. This leads to callose deposition in cell walls for their thickening, oxygen species accumulation for hypersensitive cell death, and defense response signaling for pathogenesis-related gene expression. SA-mediated resistance is more commonly associated with defense against biotrophs, whereas JA, in concert with ET, is supposed to be involved in defense against necrotrophs and functions antagonistically to SA. In JA/ET-dependent signaling pathways, linoleic and linolenic acids from membrane lipids are metabolized by lipoxygenase-mediated peroxidation, to promote the production of hydroperoxides and protease inhibitors and to offer the resistance to pathogenic nematodes (Brenner et al. 1998; Gao et al. 2008; Karajeh 2008).

However, despite of of these multi-layered plant immune systems, RKNs might have evolved at such fast pace that they could avoid or inhibit plant defense responses. It has been shown that ttrigger or suppression of SA/JA/ET-dependent plant signaling pathways by RKN infection depend on the number of days since the pathogen invaded (Bhattarai 2008; Gao et al. 2008; Nahar et al. 2011; Naher et al. 2013; Nguyen et al. 2014; Kumari et al. 2016). In

particular, it was observed that *Meloidogyne incognita* secrete the calreticulin Mi-CRT, suppressing the JA- and SA-pathways as well as the PTI (Jaouannet et al. 2013; Nguyen 2014). The Msp40 effector of *M. incognita* was shown to suppress plant cell death associated with PTI and ETI signals (Niu et al. 2016). On the other hand, cyst nematode was found to secrete effector 10A06, which disrupts SA-dependent signaling and increases the succeptibility of plants to pathogens (Wubben et al. 2008). Furthermore, the sp1A/ryanodine receptor domain (SPRYSEC)-containing proteins, secreted by CN, were identified to target disease resistance proteins of the host plant, inhibiting ETI response (Rehman et al. 2009; Postma et al. 2012). As described, biotrophic PPNs normally suppress the plant defense-induced cell death at the feeding site (Kyndt et al. 2012; Kyndt et al. 2014). They activate metabolic processes for nutrient transport to the feeding site by inducing gibberellin pathways involved in plant development.

1.3 Pre- and Post- Genome Era

Studies to unravel the mechanism behind the plant-RKN interplay have progressed owing to advancements in molecular and genetic technologies. Construction of a genetic linkage map on the basis of Mendel's theory of nineteenth-century involves inspection of hundreds of genes, individually, over several generations. Using such maps, a gene called '*Mi*', which provides host plants with resistance to *Meloidogyne* spp. was discovered (Smith 1944). Since the advent of molecular mapping, several types of markers have been used to identify other closely linked traits in the region surrounding the *Mi* gene and to associate these genetic markers with phenotypes; such efforts have contributed to the development of effective

RKN-resistant cultivars. These markers include restriction fragment length polymorphism (RFLP) markers (Messeguer et al. 1991; Klein-Lankhorst et al. 1991; Ho et al. 1992; Brown et al. 1996), random amplified polymorphic DNA (RAPD) markers (Burow et al. 1996), and amplified fragment length polymorphism (AFLP) markers (Lu et al. 1998). Accelerated pace of growth of sequencing technology has been instrumental in the initiation of nematode genome projects. *Caenorhabditis elegans* was the first species among multicellular organisms that was completely sequenced (*C. elegans* Sequencing Consortium 1998). As an entrée into the genome era, a large number of expressed sequence tags (ESTs) of nematodes derived from cDNA libraries have been deposited (Parkinson et al. 2003). In 2005, as the next-generation high throughput sequencing emerged, Sanger sequencing got replaced by automated sequencing performed on Roche454, Illumina, and Applied Biosystems platforms. These technologies were further facilitated by transition from microarray chips to RNA-Seq profiles, thus allowing more sophisticated analysis of temporal and spatial gene expressions. Using these modern techniques, more genomes of RKN species have been released including those of *M. hapla*, *M. incognita*, *M. floridensis*, and *M. chitwoodi*. Along with the progress in bioinformatics, a whole-genome approach within and across *Meloidogyne* species has been adopted for phylogenetic analysis (Scholl and Bird 2011), functional characterization (Opperman et al. 2008, Bird et al. 2013), and expression quantitative trait loci (eQTL) analysis (Dahlia M. Nielsen personal commucation). Knock-down of genes involved in parasitism by RNA interference (RNAi) has been demonstrated, but this process is not fully understood as yet (Dalzell et al. 2011). Nevertheless, extensive comparative genomics has been shedding light on the evolution of parasitism, uncovering novel parasitism genes

specific to nematodes, which share no similarity to the existing sequences in the database. In particular, these presence of genes exclusively identified in PPNs among the metazoans is explained by the horizontal gene transfer (HGT) from bacteria or fungi (Smant et al. 1998; Haegeman et al. 2011; Mayer et al. 2011; Scholl and Bird 2011). These include cell wall modifying enzymes (CWDE), chorismate mutase, and NodL. Overall, ongoing research in plant pathology constitutes essential steps towards expanding our knowledge of host-pathogen interactions, identification of major PPN gene targets, and development of sustainable nematode control methodologies.

In the present dissertation, we outline the bioinformatics approaches dealing with genomic information of *Meloidogyne* species. The subsequent chapters describe the research performed for this dissertation, starting with the description of *de novo* assembly of a complete *M. chitwoodi* genome in Chapter 2. We develop an array of protocols for optimized *de novo* assembly, gene annotation, and functional predictions. In Chapter 3, it is shown that manifold approaches for analyzing temporal and spatial transcriptome expression data of *M. hapla* and host plants reveal a far-reaching dynamic interaction throughout the nematode developmental and diurnal period. Lastly, based on the assembly of different nematode genomes, they are compared to find common or distinct genomic features and to suggest a refined annotation combing the transcriptome expression profiles in Chapter 4. The results from these studies are expected to provide directions for future studies.

Common Name	Genus	Lifestyle	Importance Ranking
Root-knot	*Meloidogyne*	**Sedentary Endo**	1375
Root-lesion	*Pratylenchus*	Migratory Endo	782
Cyst	*Heterodera*	**Sedentary Endo**	606
Stem & Bulb	*Ditylenchus*	Diverse	251
Potato Cyst	*Globodera*	**Sedentary Endo**	244
Citrus	*Tylenchulus*	Sedentary Ecto	233
Dagger	*Xiphinema*	Migratory Ecto	205
Burrowing	*Radopholus*	Migratory Endo	170
Reniform	*Rotylenchulus*	Sedentary Ecto	142
Spiral	*Helicotylenchus*	Migratory Ecto	122

Table 1. This table was reproduced based on a questionnaire conducted by the International *Meloidogyne* Project (IMP) in 1987. Most damaging genera of PPN were ranked by total weighted votes. Generally it is agreed that these rankings still represent the importance of nematode pest at present.

CHAPTER 2

de novo Assembly of *Meloidogyne chitwoodi* Genome and Automated Annotation.

- This work has been published as Cha and Bird. 2016. Optimizing k-mer size using a variant grid search to enhance *de novo* genome assembly. Bioinformation. 12(2): 36-40. For this chapter, we represented with additional data and figures.

ABSTRACT

Largely driven by huge reductions in per-base costs, sequencing nucleic acids has become a near-ubiquitous technique in laboratories performing biological and biomedical research. Most of the effort goes to re-sequencing, but assembly of *de novo*-generated, raw sequence reads into contigs that span as much of the genome as possible is central to many projects. Although truly complete coverage is not realistically attainable, maximizing the amount of sequence that can be correctly assembled into contigs contributes to coverage. Here we compare three commonly used assembly algorithms (ABySS, Velvet and SOAPdenovo2), and show that empirical optimization of k-mer values has a disproportionate influence on *de novo* assembly of a eukaryotic genome, the nematode parasite *Meloidogyne* chitwoodi. Each assembler was challenged with ~40 million Iluumina II paired-end reads, and assemblies performed under a range of k-mer sizes. In each instance, the optimal k-mer was 127, although based on N50 values, ABySS was more efficient than the others. That the assembly was not spurious was established using the "Core Eukaryotic Gene Mapping Approach", which indicated that 98.79% of the *M. chitwoodi* genome was accounted for by the assembly. Subsequent gene finding and annotation are consistent with this and suggest that k-mer optimization contributes to the robustness of assembly.

INTRODUCTION

The progression of technology from Sanger sequencing to the current "next-generation" platforms has heralded striking reductions in the cost of generating data. Sequencing nucleic acids has become a near-ubiquitous technique in laboratories performing biological and bio-medical research. Sequencing comes in two forms, distinguished by their needs for assembly into a contiguous reconstruction of a larger molecule. Most prevalent are various forms of "re-sequencing" in which the sequencing reads are aligned with a reference genome to reveal bases polymorphic between samples. Computationally, this is not a difficult undertaking. The other mode is the assembly of *de novo*-generated, raw sequence reads into contigs that are, as close as possible a full accounting of the genome of the organism in question. In practice, except for the smallest of genomes, complete coverage is neither attainable nor usually needed. None-the-less, maximizing the amount of sequence that can be correctly assembled into contigs is desirable. Reference-free assembly is based on stacking overlapping sequences of genomic fragments of a defined size (the k-mer), generated by breaking each read into k-mer size. Here we examined three commonly used assembly platforms, and showed that optimization of k-mer values has a disproportionate influence on *de novo* assembly of a eukaryotic genome.

Genome assembly algorithms permit adjustment of k-mer size, and also of the related feature coverage (or depth) of the k-mer assembly. The k-mer optimizing tool "Velvet advisor" (available as web service or scripts; Torsten 2012), for example, estimates a theoretically optimal k-mer size as follows:

$$\text{k-mer size} = 1 + \text{read length} - \frac{\text{k-mer coverage x read length}}{\text{genome coverage}}$$

$$\text{where genome coverage} = \frac{\text{a total number of reads x read length}}{\text{estimated genome size}}$$

Thus, k-mer size and k-mer coverage approximate an inverse relationship (Figure 1). Because k-mer size and coverage impact the assembly, methods to predict optimal k-mer size have been proposed. In particular, Chikhi and Medvedev (2013) developed the KmerGenie algorithm (Chikhi and Medvedev 2014) to guide selection of k-mer size, and demonstrated its utility with the assembly tools Velvet (Zerbino and Birney 2008) and SOAPdenovo2 (Luo R et al. 2012).

In our lab, we study plant-parasitic, root-knot nematodes (*Meloidogyne* spp.), which are responsible for annual crop losses approaching USD 80 billion worldwide. These pathogens have genomes in the 50 Mbp to 150 Mbp range, with marked differences in gene number between species. In cool climates, two species (*M. hapla* and *M. chitwoodi*) predominate and appear to occupy the same niche (i.e., are sympatric). Whole genome comparison would likely shed much light on the basis for sympatry. A well-annotated draft sequence is available for *M. hapla* (Opperman et al. 2008; Guo et al. 2014), and we recently sequenced the *M. chitwoodi* genome.

Prior to assembly of the *M. chitwoodi* reads, we queried "Velvet advisor" and "KmerGenie" to compute a value for k-mer size (247 and 260 respectively). Although similar, these values are not identical, and led us to explore empirical optimization of k-mer size. In this study, we show that a 'Simple Grid Search', a widely used optimization

algorithm, achieves the best k-mer value for assembly. Our proposed method has three steps. Firstly, we explicitly specified an equally-spaced interval including the k-mer size predicted by 'Velvet advisor' or 'KmerGenie'. Those k-mers were evaluated according to N50. Secondly, we selected a next set of k-mers in a more narrow interval around those k-mers with the largest N50 from the first evaluation. Lastly, we chose the best k-mer by assessing the second set of k-mers by taking into account N50 as well as other statistics. We found that assembly size is much more sensitive to k-mer size than has been theoretically estimated (Figure 2). Importantly, we found that our empirical approach yielded an assembly with an N50 of 70,023, compared to best N50 values of 46,442 (Velvet advisor) or 42,333 (KmerGenie).

Alone, the N50 value provides no information about the quality of the assembly, which needs to be verified by some independent means. One useful metric is to detect the presence or absence of a set of genes encoding proteins established to be crucial for eukaryotes. The "Core Eukaryotic Gene Mapping Approach" (CEGMA) tool performs such an analysis using a defined database of 458 core proteins (Parra et al. 2007). The percent of that protein set identified serves as a surrogate for the percentage genome coverage by the assembly. Additionally, the highly defined CEGMA proteins identify a reference set from which to unambiguously deduce elements of gene structure, including translation start/stop sites and intron/exon boundaries. Such gene models represent a reliable training set for gene prediction algorithms such as AUGUSTUS (Stanke et al. 2004). In our project, we further seeded the gene finders with EST data. Finally, genomic features were elucidated using RepeatMasker (Smit 2007), and functional domains were predicted using InterProScan (Quevillon 2005)

and Blast2GO (Conesa 2005) as functional annotation. The logic and algorithms we used are graphically represented as a cartoon (Figure 3). The results we present here indicate that the assembly based on empirically-determined k-mers yields not just a larger N50, but also a useful genome assembly.

MATERIALS AND METHODS

Data Generation and Processing. Total genomic DNA was isolated from *Meloidogyne* chitwoodi collected in a potato field in Washington State, and shipped as an ethanol precipitate to NCSU. Libraries with an average insert size of 700 bp were constructed to facilitate 300 bp paired-end reads, and sequences determined on an Illumina MiSeq II instrument. Low quality reads (Phred values ≤30) were rejected, and the remainder used for assembly. Because it is likely that different assembly algorithms will give different results in a genome-specific (and an a priori unpredictable) manner, we performed k-mer optimization on three commonly-used assembly algorithms, viz., ABySS (version 1.3.7), Velvet (version 1.2.10) and SOAPdenovo (version 2.01).

***de novo* Assembly.** In advance of fine-tuning parameters, we estimated the recommended k-mer size using "KmerGenie" to be a 260-mer. We trimmed this to 259-mers to suppress palindromes. The "Velvet advisor" (Torsten 2012) recommended an optimal k-mer to be a 247-mer (coverage cut-off 15). We performed assemblies by employing a variant of a 'Simple Grid Search' where we used k-mers ranging from 259-mer to 63-mer, with intervals of 36. Other attributes were set to the default setting of coverage-cut-off for each algorithm. To assess if the largest N50 is observed at one particular k-mer size among the previously

tested set, k1 (k-mer with the largest N50) and k2 (k-mer with the second largest N50) were averaged to k3, (k1+k2)/2. More values of k-mer surrounding k3 at intervals of 2 were performed to identify the optimal k-mer value (Table 1-1). The total number of reads aligned to contigs was also taken into account. In addition, for further parameter fine-tuning, results from different coverage-cut-offs other than default settings were compared (Table 1-2). SOAPdenovo was run under k-mer sizes equal to or less than 127-mer as it is the maximum k-mer size available in this program. To further validate our method, we arbitrarily selected two organisms for evaluation: the bacteria Neisseria gonorrhoeae (assembled genome size 2.15 Mbp) and Camelpox virus (assembled genome size 0.20 Mbp). FASTQ files were obtained from the European Nucleotide Archive (ENA) and were *de novo* assembled by ABySS using k-mer sizes chosen by KmerGenie and Velvet advisor as well as by our empirical methods.

Gene Prediction and Automated Annotation. To generate an initial training set, we queried our assemblies using CEGMA (version 2.4). To expand the training set, we incorporated cDNAs as evidence obtained from nematode.net (Martin et al. 2015) and NCBI. These sets were processed using the AUGUSTUS web server (Hoff and Stanke 2013) for predicting genes in genome *ab initio*. Additionally, gene annotations generated by AUGUSTUS were searched by InterProScan and Blast2GO to identify GO terms and gene families. We investigated DNA elements and repeat regions using RepeatMasker (version 4.0.5) (Smit 2007), and GC contents using a tool set of Biopieces (http://www.biopieces.org).

Database Construction. By using MySQL server and PHP scripts, database for *M. chitwoodi* was developed on https://brcwebportal.cos.ncsu.edu/haplatome/1-0_chitwoodi.php

where *M. chitwoodi* gene functions could be searchable with any key words relevant to itself. The elements incorporated to the database contain ‘contig number, gene position on the contig, the results of BLAST searches queried to a set of *M. chitwoodi* cDNAs (obtained from nematode.net and NCBI), InterProScan predictions, GO terms, KEGG pathways, E-value or Bit score, and CDS sequences on each contig. Also, based on the AGUSTUS annotations for gene model, GBrowse (Stein et al. 2002) was constructed on https://brcwebportal.cos.ncsu.edu/cgi-bin/gb2/gbrowse/chitwoodi/ for genome viewer.

RESULTS AND DISCUSSION

Library Features. Illumina sequencing yielded a total of 42,011,068 paired-end sequence reads (21,005,534 from each end), occupying 27.5 gigabytes in FASTQ format. The average of quality score is about 34. The reads were empirically optimized for *de novo* assembly.

***de novo* Assembly.** Under default settings of coverage-cut-off, the overall trend of N50 was concave, peaking at 127-mer (Figure 1, Table 1-1). We observed that the decrease of N50 within 127-mer and the increase of N50 within 247-mer across all the software we tested as we increased coverage-cut-offs within each k-mer size (Figure 1, Table 1-2). The largest N50 within 127-mer was still larger than the largest N50 within 247-mer in ABySS. On the other hand, the largest N50 within 127-mer was smaller than the largest N50 within 247-mer in velvet. When compared across software, the largest N50 of 70,023 was achieved by ABySS tool at optimized k-mer size of 127 at the coverage threshold of 4.6. Thus, our empirical optimization achieved better assemblies than the commonly-used k-mer predictors. For the

following further analysis, we elected to use our strategy to optimize the *M. chitwoodi* genome. The genome size of this selected assembly is 152,604,382 (150Mb).

Gene Prediction and Annotation. At the protein level, CEGMA predicted, in the *M. chitwoodi* genome, 245 (98.79%) of the 248 core proteins, implying near 100% genome coverage. In addition, it identified 2.23 average number of orthologs per CEG and 94.29% had more than one potential ortholog (Table 2). This was supported by blasting CEGMA proteins as a query against the assembled contigs as a database, resulting in one protein hit with more than two contigs. This would imply genome duplication or a genome with high heterozygosity, as has been established for *M. incognita* (Abad et al. 2008) but not for *M. hapla* (Opperman et al. 2008).

With an initial set of CEGMA annotation, AUGUSTUS predicted gene structure against 5,614 contigs (summed length 147 Mbp) out of all 128,239 assembled contigs (total length 169 Mbp). On these annotated contigs, 26,365 genes and 160,203 CDS were identified. For functional prediction, the AUGUSTUS-annotated genes were scanned by InterProScan and Blast2GO. To see overall constituents of *M. chitwoodi* genome, the protein domains and pathways with a significance of E-05 were categorized (Figure 4, 5). Also, gene families of *M. chitwoodi* genome (assembled genome size of 152 Mbp) were compared to *M. hapla* (assembled genome size of 54 Mbp) and *C. elegans* (genome msize 100 Mbp). In all the four genes of G protein-coupled receptor (GPCR; one of the largest groups of membrane proteins in *C. elegans*), nuclear hormone receptor (NHR), collagen, and glyceraldehyde-3-phosphate (GPD), the count number was largely reduced in *M. hapla* and *M. chitwoodi* (Figure 6). This implies that gene loss in *M. hapla* and *M. chitwoodi* or gene expansion in *C. elegans* might

have occurred during their adaptation to specific niche environment. Furthermore, larger genome size and higher number of genes of *M. chitwoodi* than *M. hapla* reinforces the CEGMA result interpretation as the *M. chitwoodi* genome duplication event (Table 3).

Validation with Other Organism Models. The broad applicability of our approach was demonstrated in diverse species, including a bacterium and a virus (Figure 7). For Neisseria gonorrhoeae, the k-mers predicted by Velvet advisor and KmerGenie were 275-mer and 198-mer, respectively, yielding N50s of 28,848 and 44,552. In contrast, our method returned an N50 of 48,678 using a 155-mer. On Camelpox, the Velvet advisor k-mer of 301 resulted in a failed assembly. KmerGenie recommended a k-mer of 58, resulting in an assembly with an N50 of 179,206. By contrast, our method yielded an assembly with an N50 of 190,481.

Database Applications. By integrating all predictions carried out in this study, it is launched online. For genome browser, gene models predicted by AUGUSTUS were visualized through GBrowse tool. In addition, for locating genes of which function is of interest, they could be searchable via that function as a keyword. For example, if a user would like to find genes encoding ‘chorismate mutase’, then the gene predicted by InterProScan will be shown, along with other related information including the contig on which the gene is contained, the gene position on the contig, KEGG pathways, or GO terms. Further details of applicability refer to Chapter 5 of this thesis.

CONCLUSION

In assembling a whole genome, it is desirable to achieve a balance between computational costs and the trade-off relationships between k-mer size and its coverage; namely large k-mer

size with low coverage or a small k-mer size with deep coverage. Tools “Velvet advisor” and “KmerGenie” were developed to resolve these problems. As seen in our study, however, those tools cannot be directly applied to the experimental data. Their predicted k-mer sizes gave *de novo* assembly quite different from our empirically optimized assembly of *M. chitwoodi*. This was confirmed by our experiments with two other organisms of bacteria and virus. To overcome this, we showed that our approach, using a variant of a ‘Simple Grid Search’ to identify optimal k-mer size and coverage, led to a more complete assembly. The quality of assembly was confirmed by CEGMA, predicting 98.79% core proteins in the *M. chitwoodi* genome.

By integrating different tools of CEGMA and AUGUSTUS, more reliable gene models could be generated. This could also improve the completeness of subsequent analyses, for example, functional analysis or comparative genomics approach. In future studies, we aim to examine the evolutionary history of the genus *Meloidogyne* and how that relates to, or is derived from, attributes germane to parasitism. For example, because *M. chitwoodi* and *M. hapla* are sympatric, they presumably have similar gene compliments.

	Kmer Size	Cov. Cut Off	Reads On Contigs	# of Contigs	Total Lgth	Reads /Contig	Avg. Lgth	Longest Contig	N50
ABySS	63	5.6	40,630,414	297,634	169,228,606	137	569	403,332	42,265
	99	4.9	38,878,837	185,458	170,576,726	210	920	528,011	60,946
	125	4.6	37,224,400	132,597	169,195,886	280	1,276	758,109	69,778
	127	4.6	37,088,213	128,239	168,988,320	289	1,318	758,111	**70,023**
	129	4.5	36,955,274	126,133	169,001,206	292	1,339	758,113	70,751
	131	4.5	36,815,722	123,000	168,764,165	299	1,372	758,114	69,968
	135	4.4	36,534,686	118,212	168,707,383	309	1,427	733,018	69,506
	161	4.0	34,563,084	92,400	167,205,432	374	1,809	515,211	68,049
	197	3.5	31,610,306	55,721	162,550,195	567	2,917	328,230	55,555
	233	3.0	28,370,399	42,096	159,159,885	673	3,780	243,375	30,450
	247	2.6	27,070,127	44,302	155,691,515	611	3,514	243,375	13,486
	259	2.2	25,107,148	64,765	147,149,476	387	2,272	243,375	4,552
	261	2.2	24,517,414	70,074	143,110,673	349	2,042	243,375	3,690

Table 1-1. Comparison of *de novo* assembly over different k-mer sizes, setting other parameters at default. We performed assemblies using k-mer values 63, 99, 161, 197, 233, 247, and 261. The value 247-mer was predicted "Velvet advisor, and 261-mer by "KmerGenie". At the default k-mer coverage-cut-offs, 5.6, 4.9, 4.0, 3.5, 3.0, 2.6, 2.2, and 2.2 respectively, ABySS resulted in gradual increase in N50 from 63-mer to 161-mer and gradual decrease from 161-mer to 261-mer. To investigate more narrow ranges of k-mer, the averaged value of k-mer sizes which resulted in two largest N50 (161-mer: 68,049; 99mer: 60,946), 130-mer, was chosen. Surrounding 130-mer, we increased or decreased k-mer size by 2 (125, 127, 129, 131, and 135), resulting in increasing and decreasing N50s (69 778, 70 023, 70 751, 69 968, 69 506). Though 129-mer resulted in a slightly higher N50 (70,751), it wasted about 20 percentages reads from one-end (unaligned reads: 975,453; singleton: 3,104,888; total one-end reads on contigs: 16,925,193 = 21,005,534 - 975,453 - 3,104,888). Thus, to keep more than 80 percentages of reads, we determined to cut at 127-mer which achieved the second largest N50 (70,023) as well as enough amount of information on reads (unaligned reads: 936,337; singleton: 3,050,181; total one-end reads on contigs: 17,019,016 = 21,005,534 - 936,337 - 3,050,181).

	Kmer Size	Cov. Cut Off	Reads On Contigs	# of Contigs	Total Lgth	Reads /Contig	Avg. Lgth	Longest Contig	N50
ABySS	63	5.6	40,630,414	297,634	169,228,606	137	569	403,332	42,265
	99	4.9	38,878,837	185,458	170,576,726	210	920	528,011	60,946
	127	4.6	37,088,213	128,239	168,988,320	289	1,318	758,111	**70,023**
		10	36,951,014	66,039	158,776,887	560	2,404	344,946	46,837
		15	35,995,149	64,949	145,747,066	554	2,244	344,995	9,625
	247	2.6	27,070,127	44,302	155,691,515	611	3,514	243,375	13,486
		10	17,222,135	10,028	50,018,879	1,717	4,988	243,375	41,713
		15	16,818,494	6,379	48,832,892	2,637	7,655	241,800	46,442
	259	2.2	25,107,148	64,765	147,149,476	387	2,272	243,375	4,552
		10	16,436,784	7,557	49,225,842	2,175	6,513	197,242	42,333
		15	2,264,807	3,573	20,68,931	633	579	22,094	668
Velvet	63	5.6	38,385,172	344,938	167,922,256	111	487	154,606	15,745
	99	4.9	34,464,770	344,938	180,340,024	100	523	254,652	13,703
	127	4.6	31,625,371	193,070	173,230,698	164	897	238,111	27,066
		10	31,925,142	115,193	160,122,474	277	1,390	137,609	13,257
		15	31,155,644	121,249	145,123,478	257	1,197	109,145	3,417
	247	2.6	24,473,404	93,236	159,178,884	262	1,707	159,323	2,917
		10	15,327,828	17,643	49,107,159	869	2,783	159,323	27,978
		15	15,088,514	11,075	45,945,149	1,362	4,149	159,323	30,258
Soap	63	5.6	39,536,184	92,498	150,923,645	427	1,632	206,215	18,340
	99	4.9	37,912,050	265,563	175,094,770	143	659	161,631	18,424
	127	4.6	36,453,910	83,451	157,061,370	437	1,882	144,722	18,720
		10	36,094,908	126,143	151,515,830	286	1,201	109,147	2,419
		15	31,333,799	127,446	103,225,180	246	810	109,147	1,018

Table 1-2. With the selected k-mer sizes, different coverage-cut-offs were compared across three software tools. Empirical optimization of k-mer sizes enhances genome assembly across different software platforms.

	# Prots	% Completeness	# Total	Average	% Ortho
Complete	245	98.79	546	2.23	94.29
Group 1	65	98.48	140	2.15	89.23
Group 2	55	98.21	124	2.25	96.36
Group 3	61	100.00	142	2.33	96.72
Group 4	64	98.46	140	2.19	95.31
Partial	245	98.79	577	2.36	95.92
Group 1	65	98.48	148	2.28	92.31
Group 2	55	98.21	129	2.35	98.18
Group 3	61	100.00	148	2.43	96.72
Group 4	64	98.46	152	2.38	96.88

Table 2. Statistics of the completeness of the genome based on 248 CEGs (# Prots, number of 248 ultra-conserved CEGs present in genome; % Completeness, percentage of 248 ultra-conserved CEGs present; Total, total number of CEGs present including putative orthologs; Average, average number of orthologs per CEG; %Ortho, percentage of detected CEGS that have more than 1 ortholog; Complete, those predicted proteins with 70% alignment length of protein length; Partial, those proteins which exceed minimum alignment score).

	M. chitwoodi	*M. hapla*	*C. elegans*
genome size	152 Mbp	54 Mbp	100 Mbp
contigs	5,614	3,452	245
genes	26,365	14,230	21,733
CDS	160,203	93,794	22,844
G+C (%)	22.19	27.4	35.4

Table 3. Comparison of *M. chitwoodi* genome statistics with *M. hapla* and *C. elegans*

sequences:	128,239		
total length:	168,988,320 bp (168,952,146 bp excl N/X-runs)		
GC level:	22.19%		
bases masked:	14,087,756 bp (8.34 %)		
	number of elements*	length occupied	percentage of sequence
Retroelements	201	151,353 bp	0.09%
LTR elements:	201	151,353 bp	0.09%
BEL/Pao	111	84,060 bp	0.05%
Gypsy/DIRS1	90	67,293 bp	0.04%
DNA transposons	90	38,123 bp	0.02%
Tc1-IS630-Pogo	26	17,508 bp	0.01%
Unclassified:	7	1,806 bp	0.00%
Total interspersed repeats:		191,282 bp	0.11%
Small RNA	486	115,925 bp	0.07%
Satellites:	37,625	4,698,342 bp	2.78%
Simple repeats:	134,596	6,251,509 bp	3.7%
Low complexity:	56,136	2,836,860 bp	1.68%

Table 4. Result of RepeatMasker.

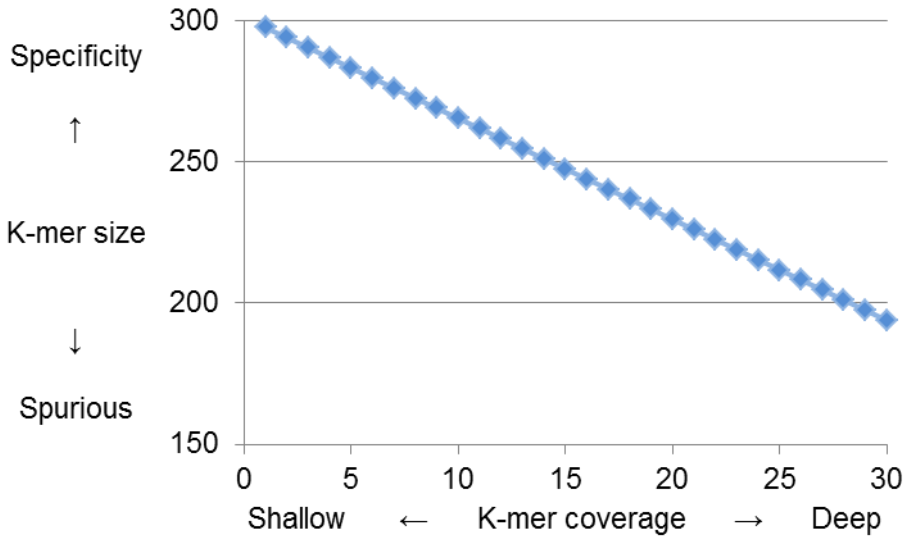


Figure 1. The inverse relationship between k-mer size and k-mer coverage. They were plotted according to the equation:

$$\text{k-mer size} = 1 + \text{read length} - \frac{\text{k-mer coverage x read length}}{\text{genome coverage}}$$

$$\text{where genome coverage} = \frac{\text{a total number of reads x read length}}{\text{estimated genome size}}$$

Our data values were substituted for: a total number of reads 42 M, read length 300, and estimated genome size 150 M. Not only as k-mer size and coverage are in trade-off relationship, but also as theoretical estimation are not fitted to the real data, we have to find optimal point somewhere between k-mer size and coverage.

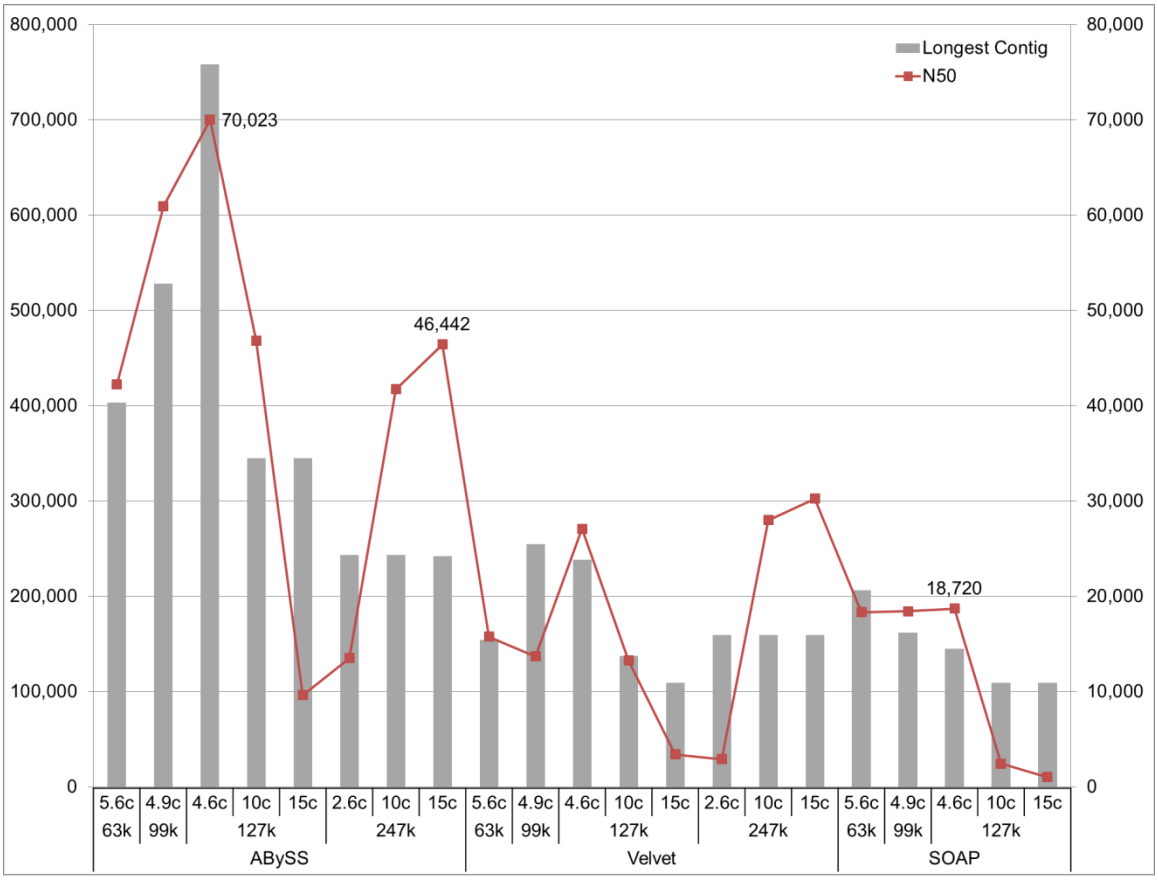


Figure 2. Empirical optimization of k-mer sizes enhances genome assembly across three software platforms (For details, see Table 1-2). X-axis indicates software, k-mer size, and coverage cut off. Y-axis on the left side indicates the length of longest contig (bp) as a function of x-axis, corresponding to grey bars. Y-axis on the right side indicates N50 length (bp), corresponding to red lines. During optimization process, to assess assemblies by N50 (red edges), it is compared of *de novo* assembly of ABySS, Velvet, and SOAPdenovo using different k-mer sizes and coverage cut offs. The larger N50, the more contiguous assembly. At the default coverage thresholds, when k-mer sizes were increased, N50 was overall concave, peaking at 127-mer. When coverage threshold was increased within the same k-mer size, N50 was decreased within 127-mer whereas increased within 247-mer. The length of longest contig (grey bar), though not exactly identical, shows similar pattern as N50. Among the selected k-mers, the largest numbers of N50 and the length of the longest contig were achieved at 127-mer and 4.6 coverage-cut-off by ABySS.

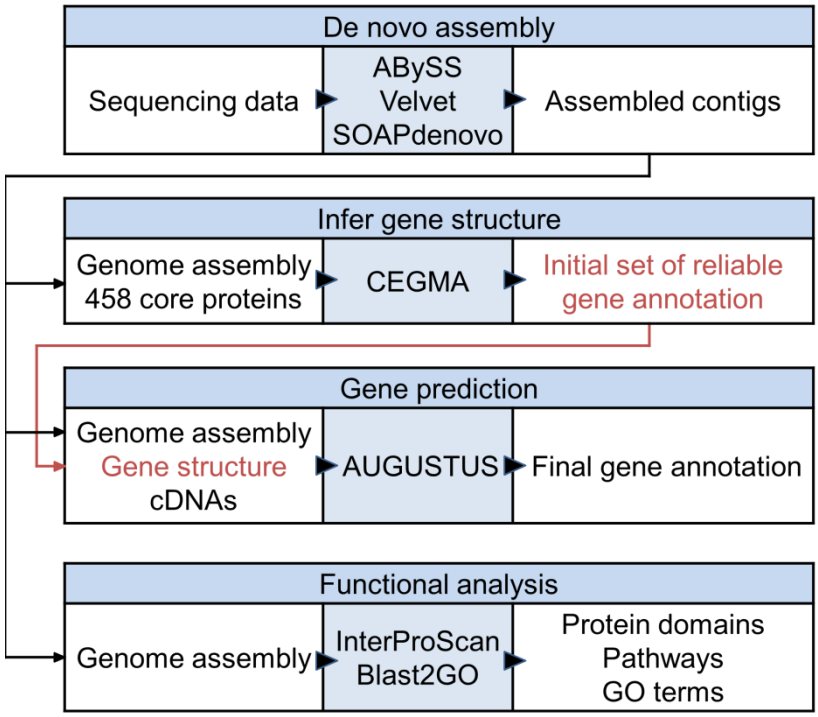


Figure 3. Pipeline protocol from a whole genome *de novo* assembly to functional analysis. With genomic DNA sequencing data, parameters for *de novo* assembly were optimized. To establish reliable gene models, CEGMA panel was used as a training-set, supported by EST data as evidence, to seed AUGUSTUS. Gene identity was predicted using InterProScan and Blast2GO as functional annotation. Finally, genomic features, such as type and distribution of transposons, were elucidated using RepeatMasker.

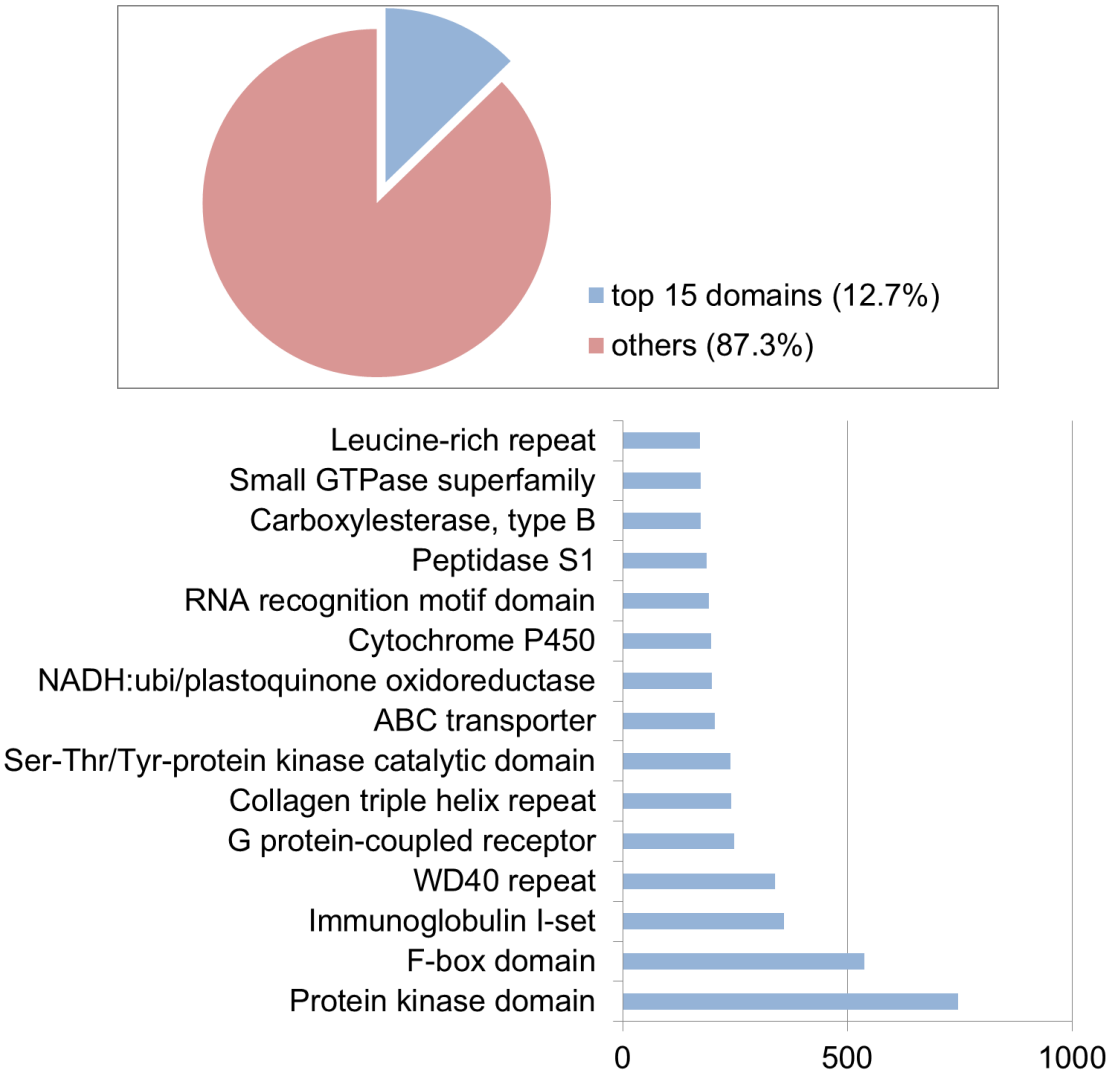


Figure 4. The top 15 matches with a significance of E-05 among protein domains found in *M. chitwoodi* based on a search of Pfam by InterProScan.

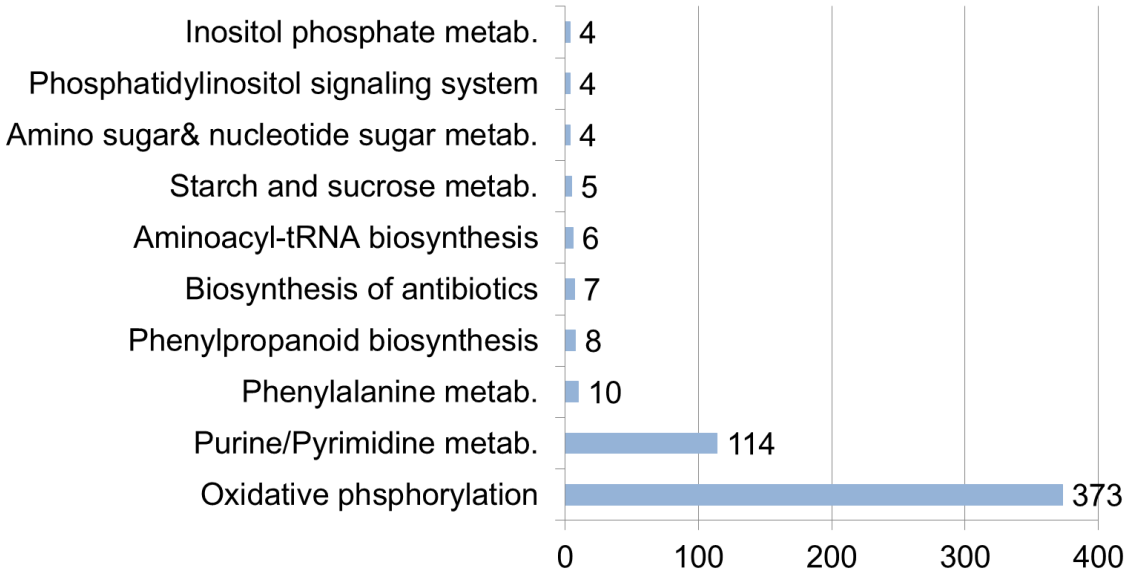


Figure 5. The 10 most common KEGG pathways found in *M. chitwoodi* by Blast2GO.

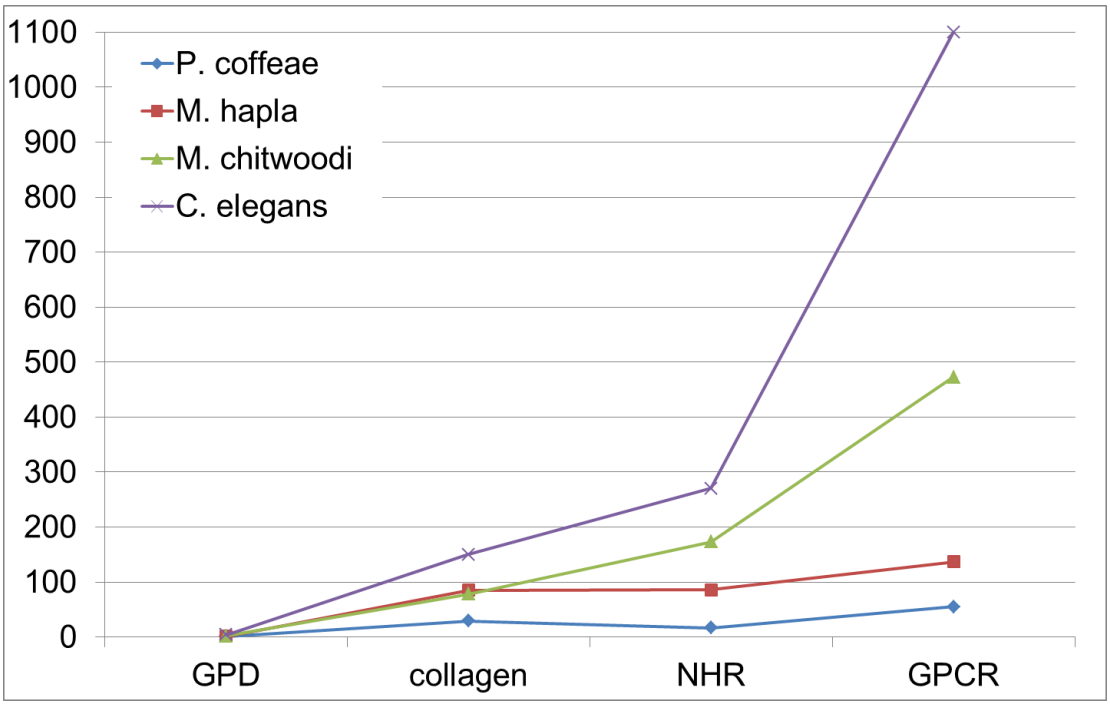


Figure 6. Gene family comparisons of *M. chitwoodi*, *M. hapla*, *C. elegans*, and *P. coffeae* (GPD, glyceraldehyde-3-phosphate; NHR, nuclear hormone receptor; GPCR, G-protein coupled receptor). Statistics of *P. coffeae*, *M. hapla*, and *C. elegans* refers to Opperman et al. 2008.

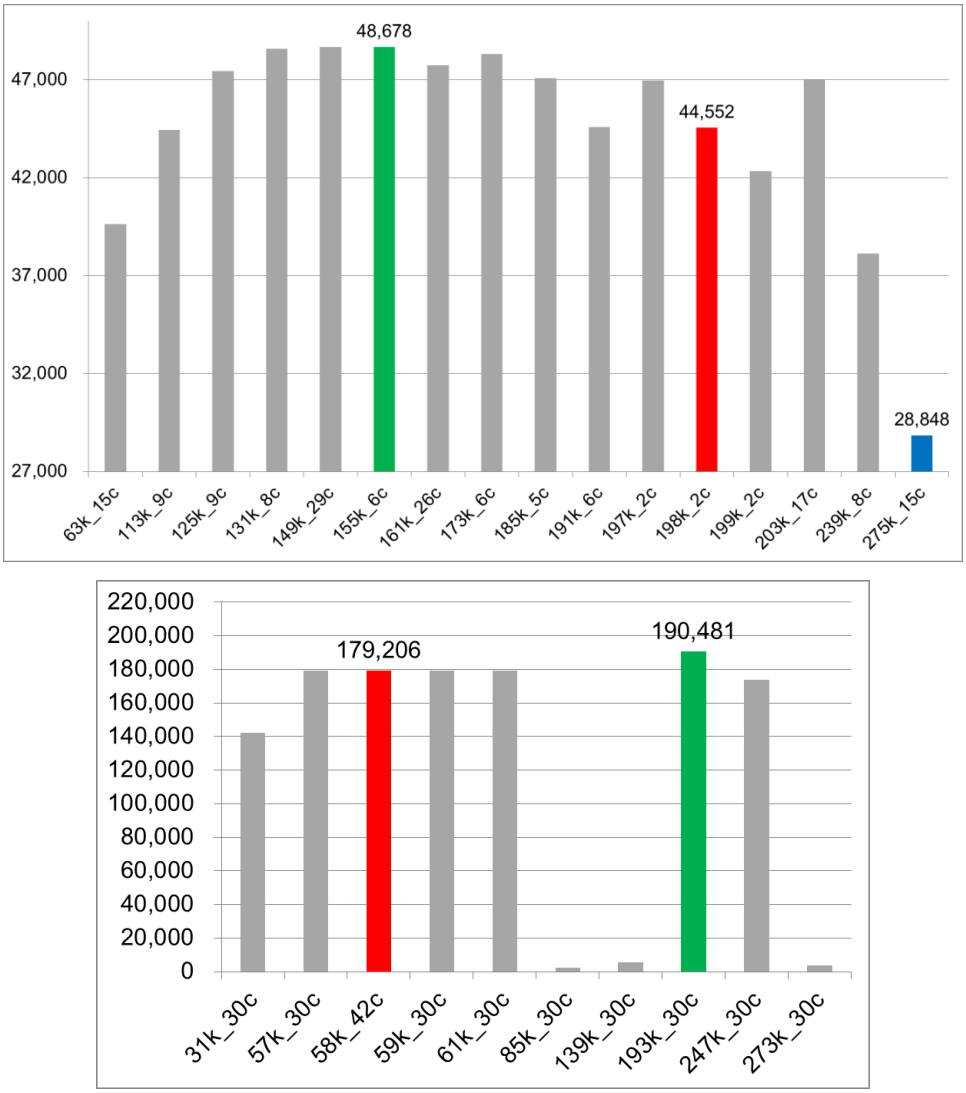


Figure 7. Comparison of N50s of assembled genome of *Neisseria gonorrhoeae* (up) and *Camelpox* (bottom) under different settings of k-mer sizes and coverage thresholds (y-axis, N50; x-axis, a pair of k-mer size and coverage cut-off; red, KmerGenie; blue, Velvet advisor). For *Neisseria gonorrhoeae*, the predicted k-mers were 275-mer (Velvet advisor) and 198-mer (KmerGenie), yielding N50s of 28,848 and 44,552. In contrast, our method returned an N50 of 48,678 using a 155-mer. On Camelpox, the Velvet advisor k-mer of 301 resulted in a failed assembly. KmerGenie recommended a k-mer of 58, resulting in an assembly with an N50 of 179,206. By contrast, our method yielded an assembly with an N50 of 190,481.

CHAPTER 3

Small, but significant, changes in *Meloidogyne* and *Medicago* transcriptome profiles reflect tight coupling of host and parasite biology.

- Portion of this chapter has been submitted for publication. Authors: Cha, DiGennaro, and Bird.

ABSTRACT

Root-knot nematode (RKN: *Meloidogyne* spp.) is an obligate parasite of the Tracheophytes, causing severe crop loss world-wide. Complex and stereotypic changes in the morphology of host and parasite over the course of infection suggest a dynamic exchange of materials between both the organisms. Because the physiology of plants is primarily regulated in a circadian manner, we hypothesized that the RKN transcriptome might also reflect a diurnal bias. To test presumption, we obtained statistically-robust transcriptional profiles of the model legume *Medicago truncatula* infected with *Meloidogyne hapla* J2. Shoot and root tissues were collected from the uninfected and infected plants, as well as from the *M. hapla* eggs and pre-penetration J2. Plants were sampled at six time points during a 24 hour day. To generate statistical power, a total of 46 independent libraries were constructed for independent flow cell analysis. Differentially expressed genes were clustered into pattern groups, disclosing key *M. hapla* genes with expressions that increased throughout the infection period, the expression being more at night. In contrast, expression of *M. truncatula* defense response genes showed a decreasing trend, beginning a day after inoculation; these genes were more active during daytime. Diurnal transcription profiles were consistent with those of *M. hapla* genes exhibiting oscillation related to plant circadian rhythm.

INTRODUCTION

Nematode Biology. Plant-parasitic nematodes (PPN) cost an estimated $173 billion in annual crop damages worldwide (Elling. 2013). The most damaging nematodes are sedentary endoparasitic forms, including cyst nematodes (CN; *Globodera* and *Heterodera* spp.) and root-knot nematode (RKN, *Meloidogyne* spp.; Williamson and Hussey 1996). All nematodes have six life stages: the egg stage, four juvenile (or larval) stages, and an adult stage (Bird et al. 2009). Female RKN can lay hundreds of eggs. Embryonic development progresses to the formation of first-stage juvenile (J1), which molts inside the egg to produce the arrested second-stage juvenile (J2). The J2 penetrates the root epidermis in the zone of elongation, and migrates intercellularly between the cortical cells to the vascular cylinder, where a feeding site is established. The morphology of feeding sites is characterized by the induction of a number of "giant cells" (GC). Although this process has been extensively researched (Bird 1996), the full role of GC has not been established. One presumed function of GC is the rebalancing of amino acid-to-sugar ratio for effective detoxification of the phloem stream. Genetic (Weerasinghe et al. 2005) and structural (Bobay et al. 2013) evidences indicate that the invading nematode communicate with the host at least in part *via* the trans-membrane receptors. Ligands for these receptors include the nematode-encoded replicas of host peptide hormones (zenomones), nematode-encoded cytokinins (Lohar et al. 2004), and probably a suite of ascarosides (Choe et al. 2012). This composition is consistent with the niche of the RKN J2, which is present entirely in the apoplast.

Penetration and migration of a nematode in its host are accompanied by secretion of proteins into the apoplast. Collectively, these (and presumably other) nematode secretions

have been termed the “secretome”. Scretome is constituted of specific “effector proteins” believed to be necessary for parasitism. The sum total of the effectors has been dubbed the “parasitome” (Gao et al. 2003; Quentin et al. 2013; Mitchum et al. 2013). Predominant in the parasitome are phytolytic enzymes which include various cellulases, β-1,4-endoglucanases, pectinases, pectate lyase, xylanase, polygalacturonase, and glucosidase (Smant et al. 1998; Giebel 1974; Hussey et al. 2002). To enhance cell wall degradation, expansin may be secreted to increase the access of cellulolytic enzymes by disrupting hydrogen bonds between the cellulose chains (Qin et al. 2004). Furthermore, the presence of an increased number of organelles and an enlarged cytoplasm during the induction of GC point to the increased metabolic activity and activation of plant growth regulators (for example, auxin, cytokinin, actin, ethylene; Caillaud et al. 2008).

In response to RKN infection, plants up-regulate diverse functions, activating peroxidase, chitinase, lipoxygenase, extensin, glyceollin, and proteinase inhibitors, and producing callose and lignin (Gheysen and Fenoll 2002). Signaling pathways involving jasmonic acid (JA), salicylic acid (SA), and ethylene (ET) are also are activated (Kunkel and Brooks 2002, Glazebrook 2005, Nahar et al. 2011). However, RKN in turn suppresses plant defenses by expressing detoxifiers including glutathione-S-transferase, glutathione peroxidase, and peroxiredoxin, or manipulators of plant pathways (e.g. chorismate mutase).

These complex host-pathogen interactions of “attack-defense-defensive attack” have been studied over several decades (Williamson and Kumar 2006). Recent advances in sequencing technology have empowered quantitative loci (QTL) analysis (Dahlia M. Nielsen personal commucation), RNAi-mediated gene knock down (Dalzell et al. 2011), functional

characterization at whole genome level (Opperman et al. 2008, Bird et al. 2013), development of genetic markers from expressed sequence tags (ESTs; Parkinson et al. 2004), and multi-locus phylogenetic analysis (Scholl and Bird 2011). Although the functions of some protein "effectors" have been predicted, a comprehensive picture of the biology and the temporal dynamics of plant–pathogen interactions remain to be established. Thus, in the present study, we performed a comprehensive transcriptome analysis of *M. hapla* and the model legume *Medicago truncatula*, from a temporal dynamics point of view.

Experimental Design. Based on nematode biology, we hypothesized that throughout the nematode life cycle, individual genes will exhibit differential expression. As a plant parasite, it is inconceivable that RKN would not be acutely "aware" of the metabolic status of the host. As plant metabolism is highly circadian, we hypothesized that RKN might exhibit diurnal fluctuation *pari passu* with the host, following its own, or the host's circadian clock.

To test these hypotheses, we designed two broad experiments: a diurnal study and a longitudinal study. In the diurnal study (Figure 1-1), the roots of infected *M. truncatula* were harvested at six time points through a day, with four replicates for each time point (for construction of 24 sequencing libraries) to achieve sufficient statistical power. In the longitudinal study (Figure 1-2), sampling was performed at seven developmental stages of parasitism, from egg and J2 of *M. hapla*, and 1, 2, 4, 5, and 7 day after inoculation (DAI) from root and shoot tissues of infected or uninfected *M. truncatula*, generating additional 22 sequencing libraries. Tissue sampling as described above was performed with the objective of achieving a balance among the following: 1) budget and logistical scale, 2) statistical power, and 3) read depth, which is an obstacle from a statistical perspective (Bar-Joseph et al.

2012). It is pertinent to mention here that American Statistical Association (ASA) has recently recommended that the results of statistical analyses should not be concluded relying on p-value alone but should rather be interpreted in an integrative context (Wasserstein and Lazara 2016). Thus, to extricate most of the intrinsic biological information from data in agreement with the recommendations of ASA as well as to maximize the statistical power of data from our 46 flow cells, we carefully framed the questions to permit different permutations of the tissue samples to serve as biological replicates. For example, in one experiment we compared the day-time transcriptome with that at night. Here we consider the data from all day-time points to be biological replicates, and all the night-time samples to be separate replicates. In another experiment in which transcriptomes at the day-night boundary were compared with those at the night-day boundary, transcriptomes from the transitions were treated as replicates.

To discover genes that might exhibit diurnal regulation, we obtained RNA-Seq data from six different time points, and compared the night vs. day data by three tests. To establish the overall changes in the expression pattern during night and day, we compared all data from time points sampled at night and all time points sampled during day. To detect genes which were activated or deactivated in the middle of night or day, we directly compared one time point from central night and one time point from central day. Finally, to identify genes differentially expressed after the host and nematode adapted to the darkness or light, sampling was done at two time points each at late night and late in the day.

For analyzing longitudinal RNA-Seq data from different samples, we used a two-layered approach that we later compared. First, we compared the pre- vs. post- feeding site

formation by treating the first three time points and the latter four time points as replicates for each condition, respectively. For *M. truncatula*, we compared infected root (IR) vs. uninfected root (UR), shoot from infected plant (IS) vs. shoot from uninfected plant (US), IR vs. IS, and UR vs. US by treating five time points as replicates for each tissue type. By explicitly making time points replicates, we could increase statistical power, at the cost of limiting the scope of questions we might pose. In the second approach of longitudinal study, for both *M. hapla* and *M. truncatula*, we conducted all the possible pairwise comparisons for each tissue type to determine the direction of fold changes. For display purposes, patterns were sorted into three bins: significantly up, significantly down, and no significant change. Discretizing the data in this manner helped dampen the noise of expression. Furthermore, an emphasis was made to take into account the critical fold changes between non-consecutive time points. If only genes differentially expressed between consecutive time points are considered, then genes with differences in expression between consecutive time points not large enough to be identified as statistically significant but for which the sum of expression changes across more than two time points are detectable, might be missed (Supplementary Figure 1).

As biological pathways are not always abruptly switched on or off, we designed a unique experimental approach to capture the gradual changes in transcript levels. It involved generation of all the possible expression pattern groups encapsulating expression changes along all time points. We observed that many of the genes could be classified into a few bins of expression pattern groups. For this classification of *M. hapla* genes, we considered differentially expressed genes (DEGs) in the test "pre-vs. post-feeding site formation"

(approach 1) and DEGs in the pairwise comparisons at different time points within the infected root (approach 2; our most interest being in the parasitism of root). For *M. truncatula*, DEGs in the test "IR vs. UR" (approach 1) and DEGs in the pairwise comparisons at different time points within the infected root (approach 2) were considered for classification of gene expression and the subsequent functional analysis. Throughout the study, we have arbitrarily used the term of "up- or down-regulation" to indicate "positive or negative fold changes" of DEGs. Strictly speaking, such usage might be incorrect as gene expression is regulated not only at transcriptional level but also through RNA processing and mRNA decay; we, however, measure steady-state RNA levels. Moreover, expression of up- or down-regulation "during the day" was used to indicate fold changes between night and daytime; this pertains to the changes not during the daytime only but those over a 24-hour day.

Our goal was to obtain a set of candidate genes central to parasitism through differential gene expression analysis, and to show how a majority of genes temporally behaves along different stages of parasitic life cycles as well as diurnally. Additionally, we present ways of comparing gene expressions at different time points and classifying them as one of the all-inclusive expression patterns.

MATERIALS AND METHODS

Experimental Design. To characterize the parasitic interaction between the nematode, throughout its lifecycle, and the host, we infected the model legume *Medicago truncatula* with *Meloidogyne hapla* and harvested tissue samples over a time course (Figure 1-2). RNA

was isolated from *M. hapla* eggs and pre-penetration juveniles (J2) and also from samples of *M. truncatula* infected with *M. hapla* J2 collected at five time points, 1, 2, 4, 5, and 7 DAI. Root (local) and shoot (global) tissues from infected and uninfected plants were sampled, and four biological replicates collected for each time point were pooled into one sample. Collectively, this provided 22 samples. In addition, to investigate the parasitic interaction throughout a 24-hour day, *M. truncatula* roots inoculated with *M. hapla* were sampled at six time-points viz., 22:30, 02:00, 05:00, 06:30, 14:00, and 21:00 h (Figure 1-1). Lights were turned off at 22:00 h and turned on at 06:00 h. Four biological replicates were taken for each time point, providing a total of 24 samples. Transcriptome sequencing was performed on each of the 46 samples described using the Illumina RNA-Seq method.

Sample Collection, Library Construction, and Sequencing. The plants used in the experiments were grown under greenhouse conditions. In all the experiments, *M. hapla* were maintained on the root system of tomato (cv. Rutgers) plants. An egg mass of *M. hapla* isolated by the standard bleach procedure was purified on a 40% sucrose density gradient as described previously (Byrd et al. 1996). For sterile *M. hapla* egg hatches, hatcheries were prepared with mesh-based sieves in a shallow bowl. The collection of viable infective J2 was obtained after three days of hatching. Host plant, *M. truncatula*, was synchronously inoculated with *M. hapla* J2. The acquisition of homogenous J2 mixture was supported by collection over three days after egg isolation and hatching, which diluted the potential effects of post-hatch age of J2 on larval penetration variance.

RKN and host plant materials were harvested at the indicated time points and immediately frozen in liquid nitrogen. RNA was extracted by Dynabeads® RNA DirectTM

Purifciation Kit as per the manufacturer's protocol. The quality and concentration of RNA were checked on Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA) and RNA Pico Chip, respectively. RNA-Seq libraries were prepared using a TruSeq RNA library prep kit v1 (Illumina, San Diego, CA, USA). RNA-Seq data are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress; Kolesnikov et al. 2015) under accession number E-MTAB-4724 (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4724/).

For deconvolution of *M. hapla* and *M. truncatula* RNA-Seq reads, the reads of infected root for each library were aligned independently using TopHat version 2.0.10 (Trapnell et al. 2009) to both of the *M. hapla* and *M. truncatula* reference genome. The RNA samples of *M. hapla* eggs and juveniles were mapped against the *M. hapla* genome, whereas those of shoots from infected plants, uninfected root, and shoot from mock-infected plants were mapped to the *M. truncatula* genome.

Counting of transcripts and analysis of differential expression. Transcript counts were made with HTSeq-0.5.4p1 (Anders et al. 2015), using the mode of non-stranded union, the feature type of exon, and the attribute of gene_id to which counts were assigned. Using these counts, the DEGs were identified with bioconductor package EdgeR version 3.0.8 which assumes an over dispersed Poisson distribution for both biological and technical variability across transcripts (Robinson et al. 2010). To determine the significance of DEGs with an adjusted p-value < 0.05 in multiple tests, we used the Benjamini-Hochberg false discovery rate (FDR; Benjamini et al. 1995).

With the longitudinal data, we analyzed DEGs in two ways. In the first approach, to identify *M. hapla* genes differentially expressed between pre- and post-feeding site formation,

certain time points were treated as replicates. Hence, for this experiment, egg, juvenile, and 1 DAI samples were considered to reflect expression prior to the feeding site formation, whereas 2, 4, 5, and 7 DAI samples reflected expression during and after the feeding site formation. For *M. truncatula*, by considering 1, 2, 4, 5, and 7 DAI as replicates, the following four comparisons were made: IR vs. UR, IS vs. US, IR vs. IS, and UR vs. US. In the second approach, to identify DEGs across the different time points, all possible pairwise comparisons were made using the 'exact test' function in edgeR. For *M. hapla*, we used samples collected at seven different time points: egg, juvenile, 1, 2, 4, 5, and 7 DAI (for a total of 21 pairwise comparisons). For *M. truncatula*, samples from five different time points, 1, 2, 4, 5, and 7 DAI were compared (for a total of 10 pairwise comparisons).

To identify *M. hapla* and *M. truncatula* genes differentially expressed between day and night, we considered three time points (22:30, 2:00, and 05:00 h) to represent night and other three points (06:30, 14:00, and 21:00 h) to represent day. Based on this binary (night and day) representation, comparisons were conducted in three ways: (1) test Diurnal 1 (D1): whole night (22:30, 02:00, 05:00 h) vs. whole day (06:30, 14:00, 21:00 h), (2) test Diurnal 2 (D2): central night (2:00 h) vs. central day (14:00 h), and (3) test Diurnal 3 (D3): night-adjusted (14:00, 21:00 h) vs. day-adjusted (02:00, 05:00 h). With DEGs identified in each test, the intersection set ($Di \cap Dj$, Di being a Diurnal i test, where, $i \neq j$, and $i, j = 1, 2, 3$) or the relative complement set ($Di - (Dj \cup Dk)$ or $Di \setminus (Dj \cup Dk)$ or $Di \setminus D$, where, $i \neq j \neq k$) were compared. Furthermore, to identify genes that are differentially expressed not between dusk and dawn but at the day-night transition, we instituted three sets of comparisons: (1) genes expressed at dusk (21:00, 22:30 h) vs. those expressed at dawn (05:00, 06:30 h), (2) genes

expressed immediately before (21:00 h) vs. those expressed after (22:30 h) the lights being turned off, and (3) genes expressed immediately before (05:00 h) vs. those expressed after (06:30 h) the lights being turned on.

Classification of temporal gene expression profiles. To track the dynamics of gene expression along the developmental or diurnal time-course, we generated all possible gene expression patterns and classified each gene expression profile to one of the patterns. Changes in expression ($C_{i,j}$) between any two time points, i and j were represented based on the directions of fold changes (U, up-regulation; D, down-regulation; ─, no significant change). Accordingly, three consecutive time-point changes, $C_{i-1,i} \cdot C_{i,i+1}$, could be among U·U, U·D, U·─, D·U, D·D, D·─, ─·U, ─·D, or ─·─. In addition, for the case where a gene was up-regulated (or down-regulated) from time-point i-1 to i+1 and had no significant fold changes from i-1 to i and i to i+1 ($C_{i-1,i}$ = ─ and $C_{i,i+1}$ = ─), it was represented as $C_{i-1,i+1}$ = U·O (or D·O) and was called as a gradual up-regulation (or down-regulation) from i-1 to i+1. Similarly, this could be generalized for the time points separated by more than two points. If expression changes were revealed between i and j, and no significant changes existed between p and q where $j-i \geq 2$ and $j > q \geq p > i$, then it was represented by $U \cdot O^n$ or $D \cdot O^n$ (n = j - i - 1, O was repeatedly arranged by n). Following this procedure, all possible expression groups which contained the direction of fold changes of all possible pairwise comparisons between any two time points were generated. Consequently, our experiment had 2,131 theoretical patterns involving seven time points (egg, juvenile 2, 1 DAI, 2 DAI, 4 DAI, 5 DAI, and 7 DAI) for *M. hapla* and 153 patterns involving five time points (1 DAI, 2 DAI, 4 DAI, 5 DAI, and 7 DAI) for *M. truncatula*. Additionally, the classification of DEGs in the

diurnal study encompassed 571 patterns involving six time points (22:30, 02:00, 05:00, 06:30, 14:00, and 21:00 h). Each DEG with expression dynamics was classified to one of the groups, which were then examined to deduce commonality of each expression group.

Functional annotation and gene enrichment. Gene ontology (GO) terms, gene families, and pathways of DEGs identified in the differential expression were searched using InterProScan (Jones et al. 2014). By restricting the functional prediction to DEGs alone, computation time was substantially reduced.

Database Construction. Based on all the information generated through this study of *M. hapla* and *M. truncatula*, a search-engine was developed using MySQL server and PHP scripts (located at https://brcwebportal.cos.ncsu.edu/haplatome/1-0_hapla.php and https://brcwebportal.cos.ncsu.edu/haplatome/1-0_medicago.php). The available components included the transcript expression levels in all the experiments, DEG significance, classified expression pattern group, DEG functions predicted by InterProScan and BLAST, GO terms, and pathways. In addition, *M. truncatula* genes orthologous to *Arabidopsis thaliana* from PLAZA database (Proost et al. 2009) were incorporated. Also, CDS (coding DNA sequence) region of particular genes of interest could be extracted with just one click. All of this information in the database was made searchable with any relevant keywords. For more extended applicability, please refer to Chapter 5 of this thesis.

RESULTS

Library features. For the diurnal experiment (Figure 1-1), a total of 20.3 million reads from infected root (day: 7.4 million reads; night: 12.9 million reads) were mapped to the genome

of *M. truncatula* (day: 98%; night: 62%) and the genome of *M. hapla* (day: 12.7 %; night: 6%; Table 1-1). For the longitudinal experiment (Figure 1-2), the Illumina RNA-Seq platform yielded 7.94 million high-quality reads (62,861,541 reads from the eggs and juveniles of *M. hapla*, 237,816,836 reads from the infected root, 172,982,264 reads from the shoot of infected plant, 178,623,532 reads from the uninfected root, 141,870,052 reads from the shoot of uninfected plant; Table 1-2). These reads were mapped to the respective reference genome. A total of 4.61×10^8 reads mapped to the genome of *M. truncatula* (63%), 5.2×10^7 reads from egg and juvenile, and 1.4×10^7reads from the infected root mapped to the genome of *M. hapla* (egg: 83%; juvenile: 6%).

DEGs in Diurnal Clock. We identified 1,301 up- and 793 down-regulated genes for *M. hapla* and 1,333 up- and 2,226 down-regulated genes for *M. truncatula*, based on differential expression in at least one of three comparisons, test Diurnal 1, test Diurnal 2, and test Diurnal 3. More than 95% of *M. hapla* DEGs belonged to D1 - (D2 ∪ D3) where DEGs were identified only in the test Diurnal 1 (i.e., differentially expressed not in the central- or late-night vs. day comparison but only in whole night vs. day comparison), showing an overall expression of most genes (Figure 2-1, (a) left panel, (b) left panel). This implied that a majority of genes of *M. hapla* were highly responsive to lighting stimuli at first, which affected their global expression changes over time; this was followed by mild changes later. In this set, there were more up-regulated than down-regulated genes. However, the ratio of down-regulated genes to the total number of DEGs of one test was increased from the test Diurnal 1 to Diurnal 2 and Diurnal 3 (Figure 2-1 (c) left panel). This implied that when we considered genes under regulation at any given point, down-regulated genes became

predominant than up-regulated genes as time passed from the point of night/day or day/night transition. Among the *M. hapla* DEGs which overlapped between the tests, D1 ∩ D2 (i.e., differentially expressed in the both tests, whole night vs. whole day and central night vs. central day), D2 ∩ D3 (i.e., differentially expressed in the both tests, central night vs. central day and late night vs. late day), D1 ∩ D3 (i.e., differentially expressed in the both tests, whole night vs. whole day and late night vs. late day), and D1 ∩ D2 ∩ D3 (i.e., differentially expressed in the all three tests, whole, central, and late night vs. day), the latter class represented those genes exhibiting differential expression in night vs. day in all the 'whole', 'central', and even 'late' manners. Not unexpectedly, their expression pattern profiles were noticeably distinct between night and day (Figure 2-1 (d)). Of the 39 DEGs belonging to D1 ∩ D2 ∩ D3, 37 down-regulated genes showed concave profiles at night and convex ones during the day. InterProScan predicted that the *M. hapla* genes in D1 ∩ D2 ∩ D3 encoded proteins for a broad range of biological functions including structural proteins like myosin, cuticle collagen, and filament protein (Supplementary Table 1-1). Furthermore, in D1 ∩ D2, up- (or down-) regulated genes at central night showed a subsequent gentle decrease (or increase) till the night ended and remained constantly decreased (or increased) during a day (Figure 2-1 (d)). In other words, these genes were highly up- (or down-) regulated especially in the middle of night but thereafter maintained almost constant level of expression. According to the InterProScan searches, *M. hapla* genes of D1 ∩ D2 had functions as enzymes involved in basic cell biology (Supplementary Table 1-1). Moreover, the DEGs in D2 ∩ D3 represented those genes differentially expressed especially in the late night vs. day comparison. InterProScan-predicted functions included Glutathione S-transferase.

On the other hand, for *M. truncatula*, 53% of genes were differentially expressed only in Diurnal1 test, D1 - (D2 ∪ D3) (Figure 2-1 (a) right panel). In this set, more up-regulated genes (970 DEGs) were responsive to lighting than the down-regulated genes (903 DEGs). Also, 24% of DEGs were identified in all the three tests (D1 ∩ D2 ∩ D3). This higher portion of D1 ∩ D2 ∩ D3 may be reflective of *M. truncatula* encoding more genes than *M. hapla*, with significant expression changes as the time passed. Specifically, in D1 ∩ D2 ∩ D3 set in *M. truncatula*, the down-regulated genes contributed to these active changes (788 down-regulated genes and 64 up-regulated genes). Furthermore, the ratio of up-regulated genes to the total DEGs in each test was convex from the test Diurnal 1 to Diurnal 2 and to Diurnal 3 (Figure 2-1 (c) right panel). This might imply that the portion of up-regulated genes became less as the mid-period of day or night, but increased back as they began to approach the next diurnal transition. One interpretation could be that the changes were not in direct response to the environment, but reflect a regulated, circadian response. Based on an InterProScan search, genes down-regulated during the day (or alternatively, up-regulated during night) were highly enriched for resistance-related proteins including chitinases, extension, peroxidases, lipoxygenases, proteinase inhibitors, oxidoreductase, lipolytic enzymes, leucine rich repeat, and 14-3-3 protein, cell wall modifying proteins including endo-1,3(4)-β-glucanase, pectinesterase, pectin methylesterase, pectin lyase, wall-associated receptor kinase, and lectin, and transcription factors KNOX and WRKY which are known to be highly activated in giant cells (Supplementary Table 1-2). The up-regulated genes during daytime were predicted to be mostly related to general metabolism and included basic helix-

loop-helix domain containing proteins, photosystem and phloem proteins, and ubiquitin, though some protein functions overlapped with those encoded by the down-regulated genes.

DEGs between Tissue Types. We identified 21 DEGs (19 up-regulated and 2 down-regulated) in the comparison pre- vs. post- feeding site formation of *M. hapla*. In the comparisons for *M. truncatula*, we identified 45 DEGs (1 up-regulated and 44 down-regulated) in IR vs. UR, 7,662 DEGs (4,106 up-regulated and 3 556 down-regulated) in IR vs. IS, and 6 393 DEGs (3,372 up-regulated and 3,021 down-regulated) in UR vs. US (Figure 2-2). The number of common genes identified as DEGs in both the two comparisons, IR vs. IS ('I': a set of genes differentially expressed between IR and IS) and UR vs. US ('U': a set of genes differentially expressed between UR and US), was 5,652. This implied that these common DEGs belonging to I ∩ U were differentially expressed between the root and shoot regardless of the infection. On the other hand, by excluding those common DEGs of I ∩ U set, the number of genes in the remained difference set of I, I - (I ∩ U), was 2,010, and the number of genes in the remained difference set of U, U - (I ∩ U), was 741. The difference set of I represented genes that were differentially expressed between roots and shoots only in the infected plant. Similarly, the difference set of U were DEGs between roots and shoots only in uninfected plant.

DEGs between Time Points. For *M. hapla*, the largest number (910) of DEGs was identified in egg vs. J2 comparison (Figure 2-3). A moderate number of DEGs was observed in comparisons of either egg vs. other DAI points (1 to 7 DAI) or J2 vs. other DAI points (1 to 7 DAI). On the other hand, almost no DEGs were identified in comparisons between any numbers of days after infection. This implied that a substantial number of *M. hapla* genes

were actively down- or up- regulated during a period of pre-penetration (infection) to its host. For shoot of uninfected *M. truncatula*, no DEG was identified in comparisons between any time points. However, for all other tissue types (IR, IS, and UR) of *M. truncatula*, in general, DEGs were high in comparisons of 1 DAI vs. 2 DAI, 1 DAI vs. 4 DAI, 2 DAI vs. 5 DAI, 2 DAI vs. 7 DAI, 4 DAI vs. 5 DAI, and 4 DAI vs. 7 DAI. In other words, many genes were identified as differentially expressed between 1 DAI and its close time point, 2 DAI or 4 DAI, but few genes were identified in the 2 DAI vs. 4 DAI comparison. This might be because some genes (re)initiated transcription at 4 DAI and showed differential expression in 4 DAI vs. 5 DAI or 4 DAI vs. 7 DAI comparisons. This indicated that shortly after infection of *M. hapla*, many host plant genes were drastically regulated, starting at 1 DAI, gradually returning to normal phase, and they (re)initiated transcription at 4 DAI. To obtain deeper insight into how gene expression profiles adapt along a time course, we tracked each gene expression profiles and classified them according to its unique pattern as described in the later part of this study.

Two Approaches to Find Common DEGs (tissue type comparisons and time point comparisons). To identify *M. hapla* genes that were differentially expressed not only in a comparison of pre- vs. post- feeding site formation (approach 1), but also along the developmental stages from egg to 7 DAI (approach 2), we found overlapping DEGs in approach 1 and approach 2 (PP ∩ IR_Mh). For *M. truncatula*, to find genes in the infected root that were differentially expressed among the five time points (approach 2) as well as DEGs due to the effect of *M. hapla* infection (approach 1), common DEGs belonging to the intersection of two approaches were identified. Similarly, other sets of DEGs from approach

1 (R, IR vs. UR; I, IR vs. IS; U, UR vs. US) and approach 2 (IR, pairwise comparisons of time points in infected root; IS, in shoot of infected plant; UR, uninfected root) were compared to find intersection set (∩) of both the approaches: R ∩ UR, I ∩ IR, I ∩ IS, U ∩ UR, and U ∩ US (Figure 2-2). Interestingly, out of 21 genes of *M. hapla* that were identified as DEGs in PP (pre vs. post feeding site formation) of approach 1, twenty genes were found in PP ∩ IR_Mh. This means that most of the *M. hapla* genes which were differentially expressed between pre- vs. post feeding site formation also showed differential expression along the time courses in infected root. Furthermore, out of 45 genes of *M. truncatula* which were differentially expressed in R (IR vs. UR) of approach 2, 40 genes were identified in R ∩ IR. This implied that most of the genes which were differentially expressed between the infected root vs. uninfected root also showed differential expression along the time courses in the infected root. In summary, those genes involved in the interaction between pathogens and pathogen-infected plants had dynamic time course gene expression.

Classification of Expression Pattern Profiles of DEGs and, Gene Functions. The number of all possible theoretical patterns was 2,131 for *M. hapla* and 153 for *M. truncatula*. Of the 2,131 groups for *M. hapla*, 989 DEGs identified in approach 2 were sorted into 23 groups (Supplementary Table 4-1). Of these 23 groups, there were 9 groups into which 20 common DEGs in PP ∩ IR_Mh were classified (Figure 3). Interestingly, all of these 9 groups together represented an overall tendency of gradual up-regulation from E or J to any days after infection. Gene functions searched by InterProScan predicted that ─UOO─·─ (gradual up-regulation from J2 to 4 DAI; 5 genes), ─UOOO─ (gradual up-regulation from J2 to 5 DAI; 7 genes), UOOO─·─ (gradual up-regulation from egg to 4 DAI; 2 genes), and UOOOO─

(gradual up-regulation from egg to 5 DAI; 1 gene) included collagen, lectin, chorismate mutase, and papain family cysteine protease. This indicates that important temporal events occur during the establishment of parasitism. In addition, of the 20 common genes in PP ∩ IR_Mh, 4 genes were differentially expressed between night vs. day; all the 4 genes were down-regulated during a day and predicted to function as cuticle collagen and histidine phosphatase (Figure 4).

Furthermore, of all the possible 153 groups for *M. truncatula*, 699 DEGs in infected root over the different time-courses were classified into 33 groups (Supplementary Figure 3-2). Of the 33 groups, there were 6 groups into which the common DEGs in R ∩ IR_Mt (40 genes) were classified (Figure 3). Ninety percent (36 genes) of *M. truncatula* genes in this set were down-regulated from 1 DAI to 2 DAI or from 4 DAI to 5 DAI, and were, thus, classified into either D─·─·─ (down-regulation from 1 to 2 DAI) or D─D─ (down-regulation from 1 to 2 DAI and from 4 to 5 DAI). Interestingly, of these 36 down-regulated genes, 20 were also identified as DEGs of the test Diurnal 1, all of which were up-regulated during a day (Figure 4). This is exactly contrary to *M. hapla* genes which were up-regulated along the life cycle from egg to 7 DAI and down-regulated during a day. For the 36 genes of *M. truncatula*, InterProScan predicted that DEGs in D─·─·─ or D─D─ might have functions including those of hydrolase, ubiquitin-conjugating enzyme, oxidoreductase, cytochrome oxidase, ATP synthase, ATPase, ribosomal protein, and carbonic anhydrase.

Collectively, *M. hapla* genes showed dynamic gene expression over time-courses such that many were gradually up-regulated starting from the juvenile stage. These included DEGs down-regulated during the day. In contrast, *M. truncatula* genes that were differentially

expressed in the root by the effect of *M. hapla* infection showed an overall tendency of decrease starting from 1 DAI, and most of them were up-regulated during a day.

Transcriptional Response of *M. hapla* to Dawn. When genes expressed at dusk (21:00, 22:30 h) and dawn (05:00, 06:30 h) were compared, no DEGs were identified for *M. hapla*, whereas 101 DEGs were identified for *M. truncatula* (Table 2). The non-DEGs identified in the dusk vs. dawn test were further tested at the transition of either dusk or dawn—21:00 vs. 22:30 h and 05:00 vs. 06:30 h, respectively. Interestingly, the DEGs of *M. hapla* were identified not in test 05:00 vs. 06:30 h test (0 DEG) but only in the 21:00 vs. 22:30 h test (35 down- and 70 up-regulated DEGs). In contrast, considerably more *M. truncatula* DEGs (94 up- and 240 down-regulated DEGs) were identified in the 05:00 vs. 06:30 h test than in the 21:00 vs. 22:30 h test (17 DEGs). This might imply that genes of *M. hapla* and *M. truncatula* would be largely under regulation at dusk and dawn, respectively.

Furthermore, DEGs identified from pairwise comparisons between any two time points were classified into one of the expression pattern groups. The largest number of *M. hapla* DEGs (855) were classified into ─U─·─·─ (up-regulation during the central night and no differential expression at other time points) whereas the largest number of *M. truncatula* DEGs (932) was classified into ─D─·─·─ (down-regulation during the central night and no differential expression at other time points; Table 3), each of which showed the opposite expression profile. Additionally, the top four or six expression pattern groups into which the largest number of DEGs were classified were those involving differential expression particularly at night—2,012 DEGs in ─U─·─·─, DU─·─·─ (down-regulation during the early night and up-regulation during the central night), ─D─·─·─, and UD─·─·─ (up-

regulation during the early night and down-regulation during the central night) for *M. hapla*, and 2,640 DEGs in ─D─·─·─, ─DO─·─ (gradual down-regulation during the central night over three time points), ─U─·─·─, UD─·─·─, ─UO─·─ (gradual up-regulation during the central night over three time points), and DU─·─·─ for *M. truncatula*. Collectively, though the expression of some *M. hapla* genes decreased at dusk, many were activated during the period beginning from central night. Conversely, for *M. truncatula*, the expression of a large number of genes decreased at central night. However, their expression might increase in the period immediately preceding the lights being turned on, which might have led to many of these genes being identified as down-regulated at the dawn transition test.

DISCUSSION

In this study involving diurnal and longitudinal experiments, one of our objectives was to effectively analyze the transcriptome data under circumstances of budgetary limitations aloowing the inclusion of too few replicates. Also, a consideration was that the approach should be able to extract the underlying biology of temporally dynamic parasitism. In the analysis of longitudinal RNA-Seq profiles, we generated the all-inclusive gene expression pattern groups which could capture gradually accumulating expression changes over distant time points. Also, by treating time points as *de facto* biological replicates, different tissue types (infected root vs. uninfected root for *M. truncatula*, pre- vs. post- feeding site formation for *M. hapla*) were compared, circumventing the replication problem (unmasking the hidden statistical power). In analyzing the diurnal RNA-Seq profiles, we compared transcript levels in a whole night vs. a whole day, central night vs. central day, and late night vs. late day

comparisons to detect gene expression fluctuations throughout the night or day, as well as the overall gene expression changes between night and day. We found that the 95% of *M. hapla* genes that were expressed differentially between the pre- and post-feeding site formation showed progressive increases from 1 DAI to 7 DAI. Of these, the DEGs overlapped to diurnal study were all down-regulated at night. The predicted functions of these DEGs were of histidine phosphatase, cuticle collagen, papain family cysteine protease, chorismate mutase, and lectin (the first two being common to the longitudinal and diurnal tests). Histidine phosphatase, known to function in sensorimotor neurons (Klumpp et al. 2002), may aid in neuronal functions of larva during the parasitic cycles. Also, as J2 undergoes two more molt stages in the root before growing to adult, it may express cuticle collagen to include it in its surface coat. On the other hand, nematode intestinal proteases including cysteine protease are hypothesized to elicit immune response of the parasitic host (Sun et al. 2013). Also, the differentially up-regulated lectin along the time-courses might contribute to innate immune defenses of the nematodes. To suppress plant defense signaling pathways, nematodes may secrete enzymes from esophageal gland cells, including chorismate mutase (CM; Jones et al. 2003). CM is known to convert chorismate, the shikimate pathway product from plant, to prephenate, thus disrupting plant secondary metabolites, such as auxin and salicylic acid (SA).

In response to nematode penetration, plants react by a series of several defense strategies; these include: (1) recognition of pathogen-associated molecular pattern and activation of pattern recognition receptors; this strategy is, therefore, called “pattern-triggered immunity (PTI)”, (2) induction of “effector-triggered immunity (ETI)”, a stronger immune

response for targeting pathogen effectors which manipulate plant immune system, and (3) amplification of signaling pathways including salicylic acid (SA)- and jasmonic acid (JA)-pathways which could contribute to both the PTI and ETI (Thomma et al. 2011). Regardless of these plant immune systems, RKNs may have evolved at such fast pace that they could avoid or even dampen the plant defense responses. As many turnovers are going back and forth in the parasitic plants during parasitism, some immune responses are activated whereas others are suppressed at different time points, thus, making it more complex to explain the observed changes in plants. Moreover, there are other surrounding pathogenic organisms besides the nematodes like bacteria and fungi, which cause physiological changes in the plants as well. Alhough what is observed could not be interpreted according to a single definitive rule, our study might suggest plausible scenarios occurring in the infected plants. Among the genes differentially expressed between the infected and uninfected roots of *M. truncatula*, 89% showed significant dynamic changes from 1 to 7 DAI in the infected root, a majority of which were classified to D─D─ (down-regulation from 1 to 2 DAI) or D─·─·─ (down-regulation from 1 to 2 DAI and from 4 to 5 DAI) in our designated pattern groups. In addition, the DEGs common to both the longitudinal and diurnal study were all up-regulated during daytime. Their predicted functions were carbonic anhydrase, cytochrome, oxidureductase (OXI), and ATP synthase in the plant photo system. OXI involved in plant defense and stress response is likely to interact with 4F01, the annexin-like nematode effector, increasing the plant susceptibility to pathogen (Patel et al. 2010, Chen et al. 2015). Also cytochrome has been studied to be induced by plant JA- and SA- signaling pathways (Li et al. 2002) and to participate in oxygenation reaction to produce toxic metabolites such as

oxidized fatty acids (e.g., hydrogen peroxide and hydroperoxides; Pinot and Beisson 2011). In the plant SA-pathway, carbonic anhydrase is a SA-binding protein, which promotes plant defenses (Slaymaker et al. 2002). These hypersensitive plant defenses could in turn encounter counteroffensive peroxidases (e.g., thioredoxin peroxidase and glutathione peroxidase) secreted from nematode pharyngeal gland cells. In our study, the three proteins OXI, cytochrome, and carbonic anhydrase showed decreased expression, beginning 1 DAI, implying that *M. hapla* might have successfully suppressed the defense response of *M. truncatula*. Furthermore, ATP synthase (involved in ATP generation required for plant photosynthesis) which was classified to D─D─ in our pattern group and was up-regulated during the day could be explained by such an effect of *M. hapla* that photosynthetic performance was disturbed throughout the period of infection although it was more active during the day than at night to some degree. In summary, there may exist an overall opposite tendency between *M. hapla* and *M. truncatula* such that the crucial genes of *M. hapla* are likely to be up-regulated throughout the developmental stages, being more active at night. In contrast, the immune defense genes of *M. truncatula* would be suppressed throughout the infection period and are more active during daytime (Figure 4).

Our diurnal experiment alone could support the hypothesis that since plants follow highly circadian cycles, their parasites may accordingly do, as well. When the results of diurnal experiments were considered alone, excluding the results of longitudinal development experiments, it became more apparent. The *M. truncatula* genes more active at night (or down-regulated during day) contained a wide range of resistant genes including peroxidases, lipoxygenase, protease inhibitor, 14-3-3 protein, Toll-interleukin1 resistance,

chitinase, extensin, and KNOX1, a gene aiding in feeding site development, with shapes of gene expressions curves being convex at night and concave during the day (Supplementary table 1-2). In *M. hapla*, a high mass of genes showed similar gene expression curve shapes as *M. truncatula*. These *M. hapla* genes, which were more activated at night, included collagen and thioredoxin, lipoprotein receptor, and adrenodoxin reductase, the latter of which contributes to suppressing plant immune response (Supplementary table 3-1). A small number of genes were up-regulated during the daytime including Glutathione S-transferase. Collectively, same expression pattern between *M. truncatula* defense genes and *M. hapla* suppressor genes suggests that more vigorous interaction between nematode and parasitic plants might occur at night; however there might be a few interactions during daytime. In addition to the plant defense proteins, the genes for plant cell wall loosening and feeding site formation (e.g., wall-associated receptor kinase, Endo-1,3(4)-β-glucanase, pectin lyase, auxin efflux carrier) also showed sin curve, that is, convex at night and concave during the day. They were differentially expressed, especially at central or late night vs. day.

Many researches on parasitism and circadian rhythm are underway, but these two aspects are mostly studied in seclusion. The interconnection between circadian clock and plant parasitism is largely unknown, for example, the effects of pathogen infection on circadian rhythms or the control of circadian rhythm on defense responses. In recent years, it was discovered that not only the free-living nematode, *C. elegans* possessed clock-controlled genes (van der Linden et al. 2010), plant defense genes were also temporally controlled by circadian regulator, CIRCADIAN CLOCK-ASSOCIATED 1 (CCA1; Wang et al. 2011). In the plant's response to pathogen, the expression of a plant defense gene PCC1 was observed

to peak at night with circadian rhythm (Sauerbrunn and Schlaich 2004). Also, it has been suggested that plants would be more easily damaged at night than during day (Greenham and McClung 2015). Compared to night, it is at dawn when plant immunity becomes the strongest, through activated defenses. This was possibly reflected in our study as the portion of *M. truncatula* DEGs was increased from the test Diurnal 2 (central night vs. day) to the test Diurnal 3 (late night vs. day) whereas the number of *M. hapla* DEGs was decreased. This implies that plants may start to prepare for susceptible infection right before lighting was imposed. Although other studies about circadian clock and plant-pathogen interactions have proposed possible underlying mechanisms by using bacteria, oomycete, insects, fungus, or herbivores, plant parasitic nematodes could be expected to show similar behaviors as they share common strategic characteristics due to convergent evolution. However, there definitely exist differences among different plant pathogens (Torto-Alalibo et al. 2009). Thus, we believe, that this study could contribute in accelerating the understanding of the cross-link between RKN-host plant interactions and the circadian control.

In conclusion, we adopted various approaches for dissecting a time-course transcriptome data to obtain a deeper insight into dynamically changing plant-pathogen interactions both developmentally and diurnally. We identified several key regulators central to parasitism, by assorting them into particular expression patterns throughout the developmental period and oscillating curves for 24-hour day. This study will provide the basis of future studies where more detailed dynamics of physiological mechanisms could be revealed through biological validation.

	Input	Mapped reads	
		M. hapla	*M. truncatula*
A1	10,707,452	1,615,921 (15.1)	6,097,616 (56.9)
A2	5,446,882	583,037 (10.7)	3,274,061 (60.1)
A3	3,032,806	317,422 (10.5)	1,835,136 (60.5)
A4	1,541,408	98,828 (6.4)	946,047 (61.4)
B1	4,069,794	281,175 (6.9)	2,474,055 (60.8)
B2	2,925,837	185,874 (6.4)	1,828,132 (62.5)
B3	10,286,235	485,548 (4.7)	6,557,870 (63.8)
B4	2,262,043	95,927 (4.2)	1,450,010 (64.1)
C1	8,148,121	656,845 (8.1)	4,643,321 (57.0)
C2	8,311,164	787,852 (9.5)	5,051,573 (60.8)
C3	6,953,178	509,780 (7.3)	4,453,173 (64.0)
C4	10,473,479	3,860,642 (36.9)	4,240,162 (40.5)
D1	15,325,792	897,613 (5.9)	9,852,259 (64.3)
D2	13,558,513	1,118,008 (8.2)	8,408,571 (62.0)
D3	12,347,900	467,891 (3.8)	7,041,884 (57.0)
D4	7,916,954	626,557 (7.9)	5,014,338 (63.3)
E1	8,823,268	660,280 (7.5)	5,566,675 (63.1)
E2	9,618,369	1,612,978 (16.8)	5,428,820 (56.4)
E3	12,684,356	735,548 (5.8)	8,268,195 (65.2)
E4	8,753,623	1,036,088 (11.8)	5,190,961 (59.3)
F1	7,031,379	525,972 (7.5)	4,283,287 (60.9)
F2	9,666,593	594,020 (6.1)	6,193,619 (64.1)
F3	12,881,563	696,360 (5.4)	8,322,534 (64.6)
F4	10,801,848	638,324 (5.9)	6,960,433 (64.4)
Sum	203,568,557	19,088,490 (9.4)	123,382,732 (60.6)

Table 1-1. The total number of RNA-Seq reads extracted at diurnal time-points from *M. hapla* and *M. truncatula*, and the number of reads mapped to the reference genome respectively (A, 10:30 p.m.; B, 2:00 a.m.; C, 5:00 a.m.; D, 6:30 a.m.; E, 2:00 p.m.; F, 9:00 p.m.; numbers in parentheses, the percentage of mapped reads). There were four replicates for each time-point.

(a) *M. hapla*

		E	J			
Pre-penetration	total	28,961,071	33,900,470			
	mapped	25,101,358 (86.7)	27,360,266 (80.7)			
		1DAI	**2DAI**	**4DAI**	**5DAI**	**7DAI**
Infected root	total	76,005,336	40,575,349	55,888,830	33,247,990	32,099,331
	mapped	11,347,307 (14.9)	175,920 (0.4)	2,190,678 (3.9)	92,090 (0.3)	849,187 (2.6)

(b) *M. truncatula*

		1DAI	**2DAI**	**4DAI**	**5DAI**	**7DAI**
IR	total	76,005,336	40,575,349	55,888,830	33,247,990	32,099,331
	mapped	55,598,042 (73.2)	24,058,200 (59.3)	34,926,888 (62.5)	20,965,766 (63.1)	20,897,876 (65.1)
IS	total	52,228,331	34,495,316	36,265,567	27,131,930	22,861,120
	mapped	31,933,444 (61.1)	21,653,796 (62.8)	23,135,409 (63.8)	17,031,546 (62.8)	14,152,820 (61.9)
UR	total	36,531,982	30,794,279	30,048,459	49,258,046	31,990,766
	mapped	23,296,592 (63.8)	19,189,138 (62.3)	18,624,375 (62)	31,164,767 (63.3)	20,430,355 (63.9)
US	total	31,570,648	35,894,555	30,961,593	31,204,149	12,239,107
	mapped	19,943,272 (63.2)	17,731,774 (49.4)	19,802,207 (64)	19,434,043 (62.3)	7,492,444 (61.2)

Table 1-2. The total number of RNA-Seq reads extracted at different life-stages of *M. hapla* and *M. truncatula*, and the number of reads mapped to the reference genome respectively (E, egg; J, juvenile; DAI, days after infection; numbers in parentheses, the percentage of mapped reads; IR, root from infected plant; IS, shoot from infected plant; UR, root from uninfected plant; US, shoot from uninfected plant).

	M. hapla (14230)	*M. truncatula* (45108)		
dusk vs. dawn	14230 non-DEGs	45007 non-DEGs		101 DEGs
Dusk	up (35) + dw (70)	up (12) + dw (5)		
Dawn	0 DEG	dw (4)	up (94) + dw (240)	up (11) + dw (1)

Table 2. Number of differentially expressed genes (DEGs) in night-day boundary tests (dusk, 21:00 vs. 22:30 h test; dawn, 05:00 vs. 06:30 h test). When gene expression at dusk (21:00, 22:30 h) was compared to that at dawn (05:00, 06:30 h), zero DEG and 101 DEGs were identified for *M. hapla* and *M. truncatula*, respectively. Among the non-DEGs in the test, a majority of *M. hapla* DEGs were identified in the dusk test, whereas most DEGs of *M. truncatula* were identified in the dawn test. The four *M. truncatula* DEGs common to both the dusk and dawn tests were down-regulated at the dawn test but up-regulated at the dusk test. Of 101 DEGs of the dusk vs. dawn test, 12 DEGs were identified in the dawn test.

Top 10 groups	Number of Classified *M. hapla* DEGs (%)	Top 10 groups	Number of Classified *M. truncatula* DEGs (%)
─U─·─·─ *	855 (26.8)	─D─·─·─ *	932 (16.6)
DU─·─·─ *	491 (15.4)	─DO─·─ *	661 (11.8)
─D─·─·─ *	351 (11.0)	─U─·─·─ *	528 (9.4)
UD─·─·─ *	315 (9.9)	UD─·─·─ *	519 (9.2)
─UOO─	222 (7.0)	─UO─·─ *	369 (6.6)
─UO─·─ *	159 (5.0)	DU─·─·─ *	351 (6.2)
─DOO─	116 (3.6)	UDO─·─	328 (5.8)
─UOOO	112 (3.5)	─DOO─	270 (4.8)
─·─·─·─U	97 (3.0)	─UOO─	268 (4.8)
─DOOO	72 (2.3)	─·─UO─	108 (1.9)
Other 34 groups	399 (12.5)	Other 92 groups	1286 (22.9)
Sum	3189	Sum	5620

Table 3. Top 10 expression group patterns into which the largest number of DEGs from pairwise comparisons between time-points in the diurnal study were classified. The largest number of DEGs of *M. hapla* and *M. truncatula* were classified into ─U─·─·─ and ─D─·─·─, respectively, which had opposite expression profiles. The expression pattern groups that involved differential expression at night are marked with asterisk; all of these were ranked highly (─U─·─·─, up-regulation during the central night and no differential expression at other time points; DU─·─·─, down-regulation during the early night and up-regulation during the central night; ─D─·─·─, down-regulation during the central night and no differential expression at other time points; UD─·─·─, up-regulation during the early night and down-regulation during the central night; ─UO─·─, gradual up-regulation during the central night over three time points; ─DO─·─, gradual down-regulation during the central night over three time points).

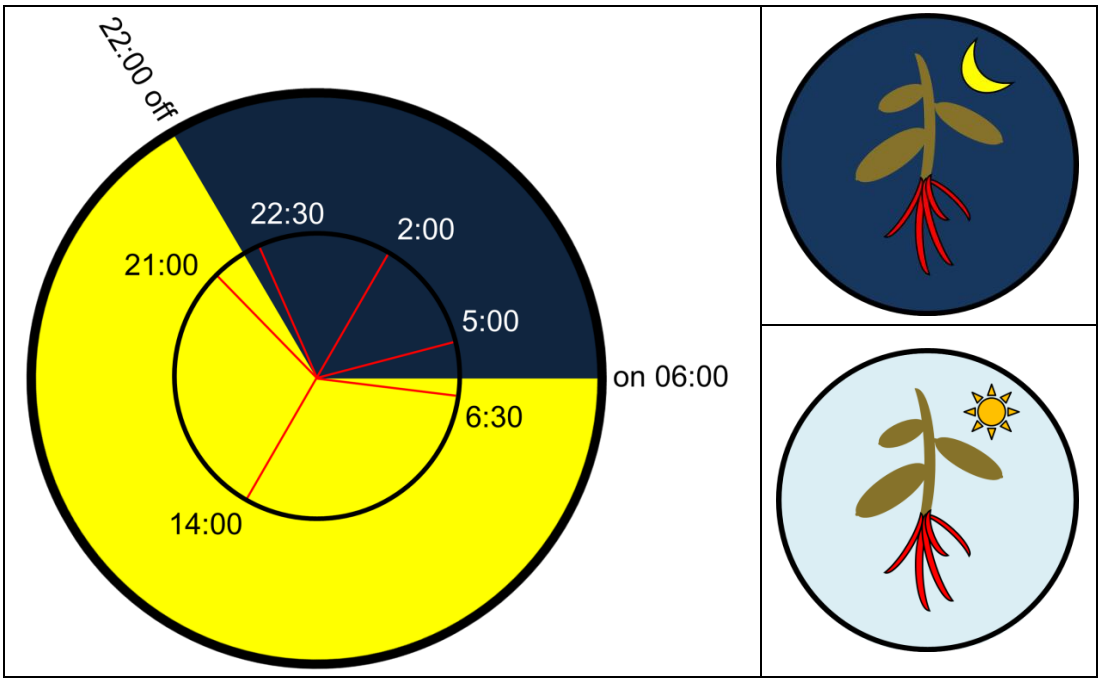


Figure 1-1. Time-points at which samples were collected from the infected root of *M. truncatula* for obtaining the RNA-Seq data and lighting was turned on (or off). Lighting was turned on at 06:00 and turned off at 22:00 h to determine the period of day and night. Thirty minutes after the lights were turned off, samples were collected at 22:30 h. In the middle of the night, samples were collected at 02:00 h. An hour prior to switching on the light, samples were collected at 05:00 h. Samples were collected at 06:30 h, 30 minutes after the lights were turned on. In the middle of day, samples were collected at 14:00 h. Samples were collected at 21:00 h, an hour before the lights were turned off. Four replicates were taken for each time point. We conducted three tests to identify the differentially expressed genes between night vs. day: (1) test Diurnal 1: whole night (22:30, 02:00, 05:00 h) vs. whole day (06:30, 14:00, 21:00 h), (2) test Diurnal 2: central night (02:00 h) vs. central day (14:00 h), and (3) test Diurnal 3: night-adjusted (14:00, 21:00 h) vs. day-adjusted (02:00, 5:00 h).

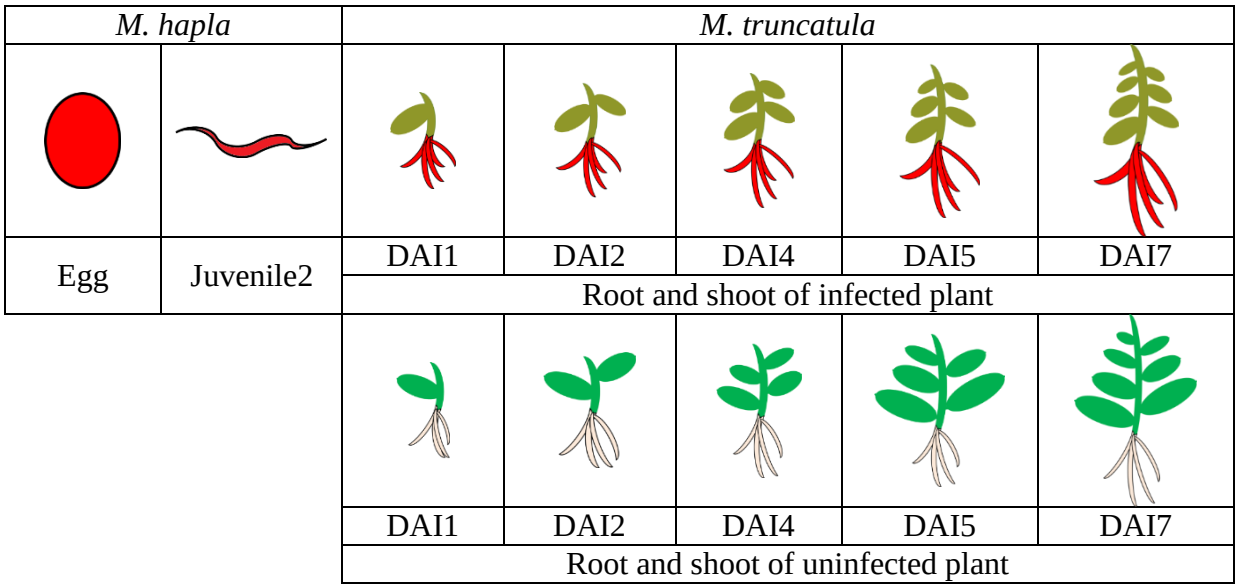
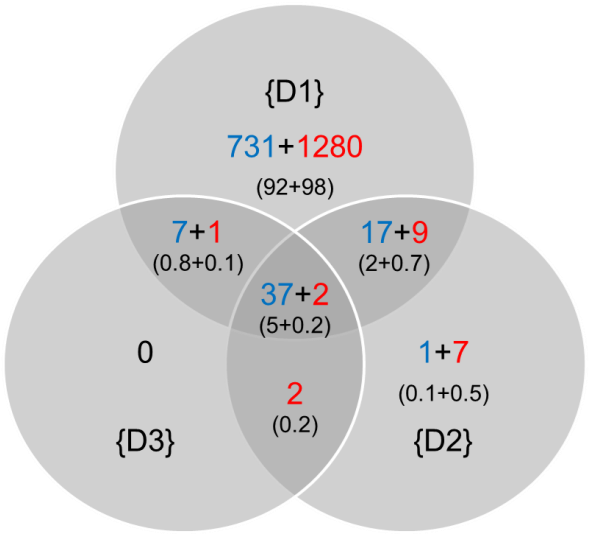
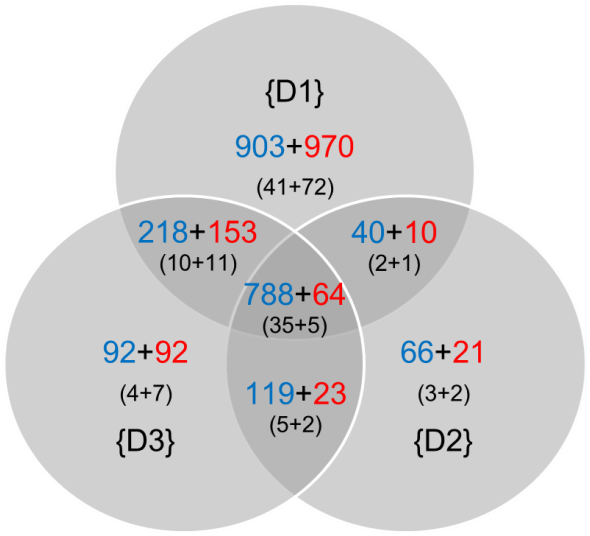


Figure 1-2. Time-points at which samples were collected from egg and juvenile of *M. hapla* and root and shoot tissues of infected or uninfected *M. truncatula* for obtaining the RNA-Seq data. For either of the infected or uninfected *M. truncatula*, data were sampled at five time points: 1, 2, 4, 5, and 7 days after infection (DAI). Root and shoot tissues were separately harvested, thus, providing 20 treatments (five time points × four tissue types) in total. For *M. hapla*, samples were taken from egg and J2 stages. Since the infected root contained mRNAs of *M. hapla*, it had seven treatments (five time points within infected root plus egg and J2) in total. Based on these samples, we compared the RNA expression in infected root (IR) vs. uninfected root (UR), shoot from infected plant (IS) vs. shoot from uninfected plant (US), 'IR vs. IS', and 'UR vs. US' by treating time points as replicates for each tissue type. For *M. hapla*, pre- vs. post-feeding site formation was tested by treating three samples (egg, J2, and 1 DAI) and four samples (2, 4, 5, and 7 DAI) as replicates for each condition, respectively. Furthermore, the time-point pairwise comparisons were conducted across the egg, J2, 1 DAI to 7 DAI samples for *M. hapla* and across the 1 DAI to 7 DAI samples of each each tissue type (IR, IS, UR, and US) for *M. truncatula*.

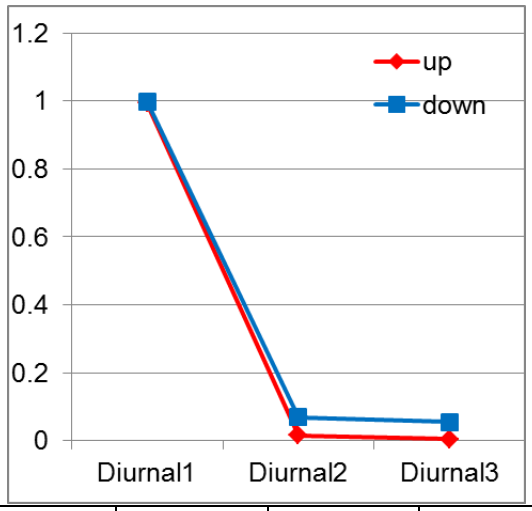
(a) *M. hapla*

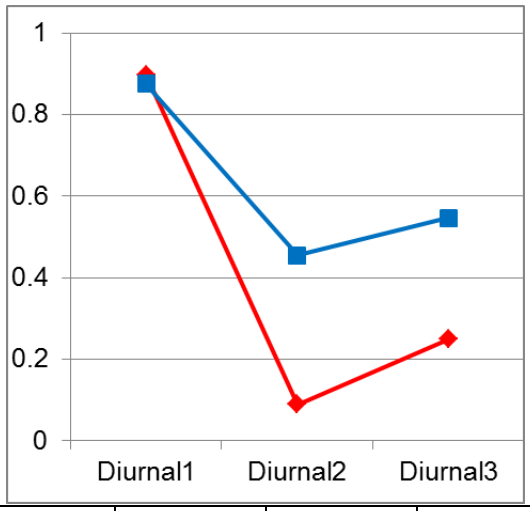


M. truncatula



(b)

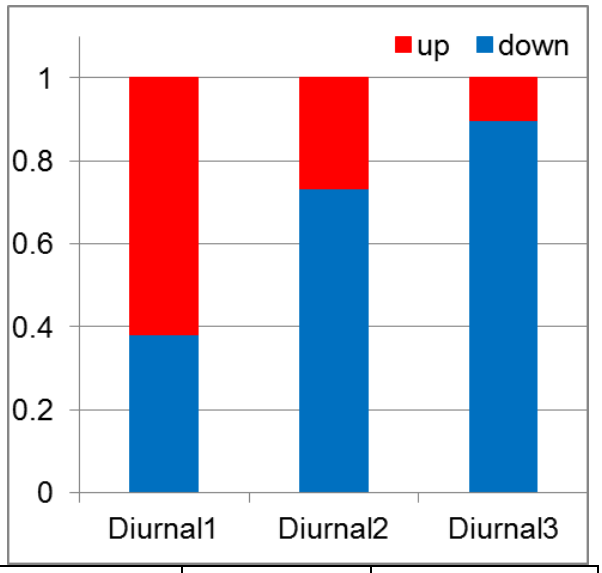


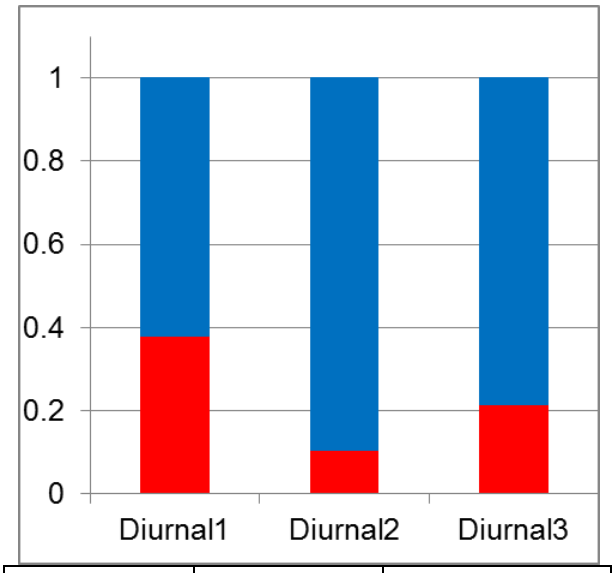


	D1	D2	D3	net
up	1292	20	5	1301
down	792	55	44	793

D1	D2	D3	net
1197	118	332	1333
1949	1013	1217	2226

(c)

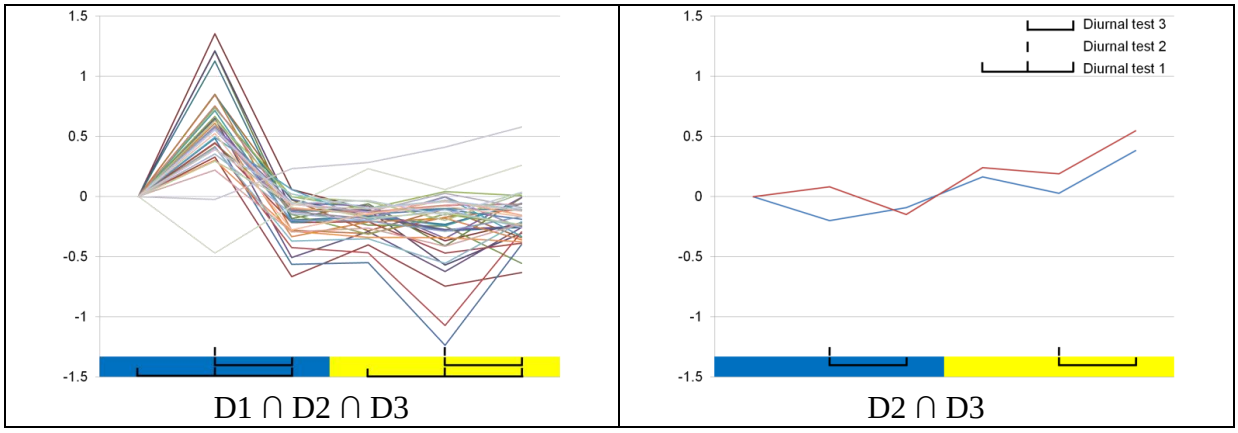




	D1	D2	D3
up	1292	20	5
down	792	55	44
sum	2084	75	49

D1	D2	D3
1197	118	332
1949	1013	1217
3146	1131	1549

(d)



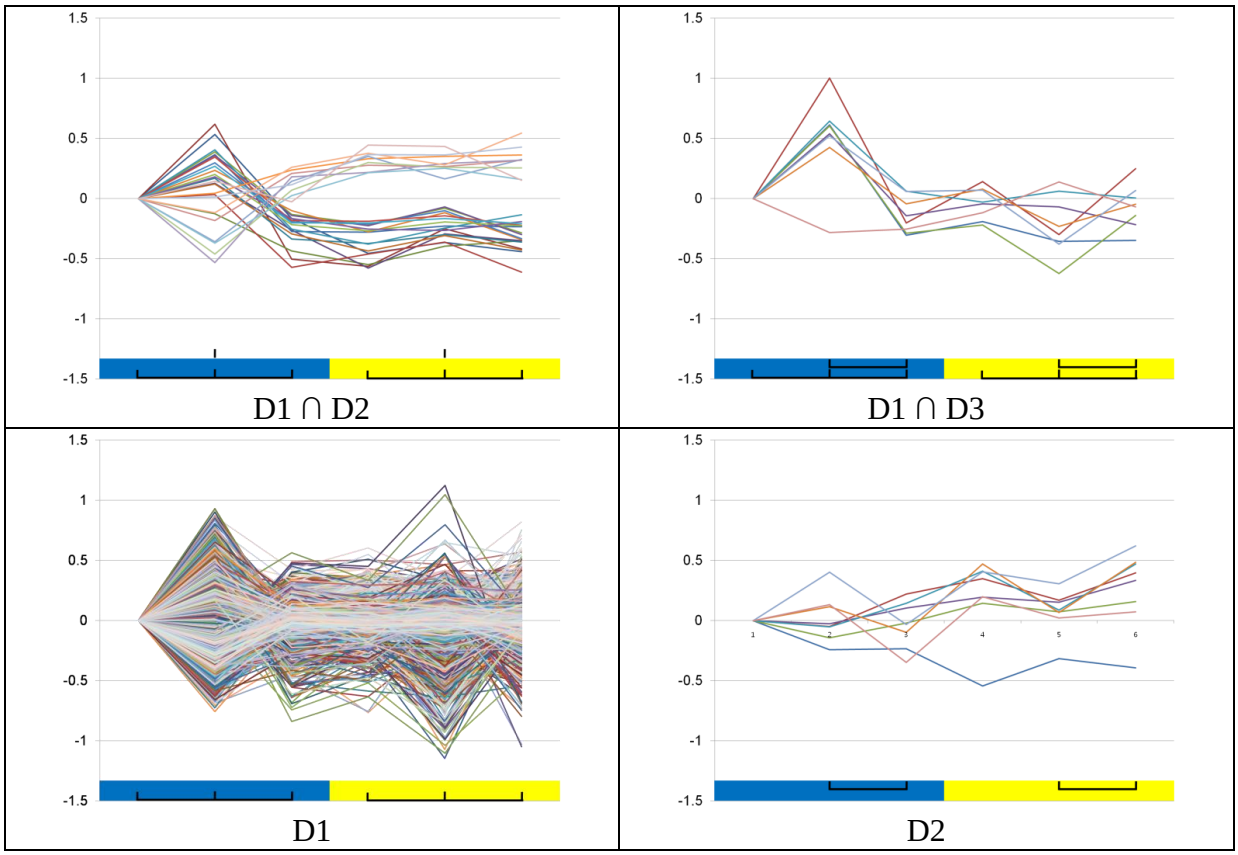


Figure 2-1. Number of differentially expressed genes (DEGs) in diurnal clocks and their expression profiles. (a) The number of DEGs of *M. hapla* (left) and *M. truncatula* (right) (Di, Diurnal test i, i=1, 2, 3; blue, genes down-regulated during a day; red, genes up-regulated during a day; number in parenthesis, percentage of up-regulated genes in a set among the total up-regulated genes across the three tests, percentage of down-regulated genes in a set among the total down-regulated genes across the three tests). About 95% of *M. hapla* DEGs belonged to D1-(D2 ∪ D3). In D1 ∩ D2 ∩ D3 of *M. hapla*, the down-regulated genes overwhelmed the up-regulated genes in number. (b) Ratio of up- or down-regulated genes in a test to the total up- or down- regulated genes across the three tests for *M. hapla* (left) and *M. truncatula* (right). (c) Relative ratio of up- and down-regulated genes to the summed DEGs within each test for *M. hapla* (left) and *M. truncatula* (right). (d) Expression profiles of *M. hapla* DEGs in each test or in overlapping set of tests (see Supplementary Figure 2 for *M. truncatula* expression profiles).

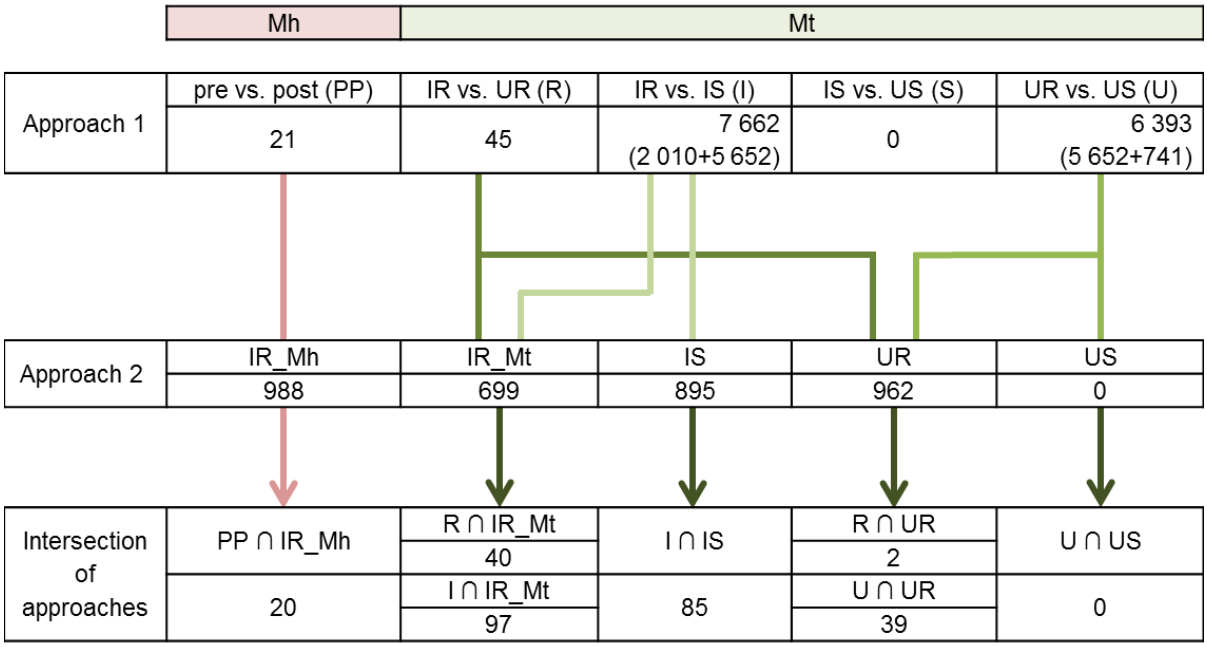
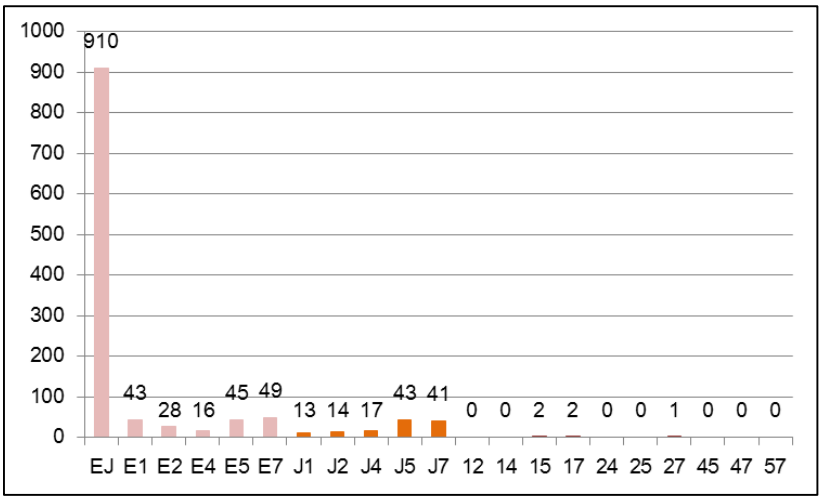


Figure 2-2. Number of differentially expressed genes of (DEGs) (Mh, *M. hapla*; Mt, *M. truncatula*; IR, infected root; UR, uninfected root; IS, shoot of infected plant; US, shoot of uninfected plant; PP, comparison of pre- vs. post-feeding site formation; R, comparison of IR vs. UR; I, comparison of IR vs. IS; S, comparison of IS vs. US; U, comparison of UR vs. US; IR_Mh, infected root of *M. hapla*; IR_Mt, infected root of *M. truncatula*). In approach 1, the number of DEGs between pre- and post-feeding site formation of *M. hapla* was identified as 21. For *M. truncatula*, the number of DEGs was 45, 7,662, 0, and 6,393 in the comparison of different tissue types, R, I, S, and U, respectively. The number of common genes which were identified as DEGs in both the two comparisons, I and U, was 5,652, implying that these common DEGs were differentially expressed between the root and shoot regardless of infection. On the other hand, by excluding the common DEGs, the number of genes in the remaining difference set of I was 2,010 and that of U was 741. In approach 2,988 genes of *M. hapla* were identified as differentially expressed over seven different time points, whereas 699, 895, 962, and 0 genes were found to be differentially expressed over five time points within tissue types IR, IS, UR, and UR, respectively, of *M. truncatula*. In approach 3, common DEGs to both the approaches 1 and 2 were found.

(a) *M. hapla*



(b) *M. truncatula*

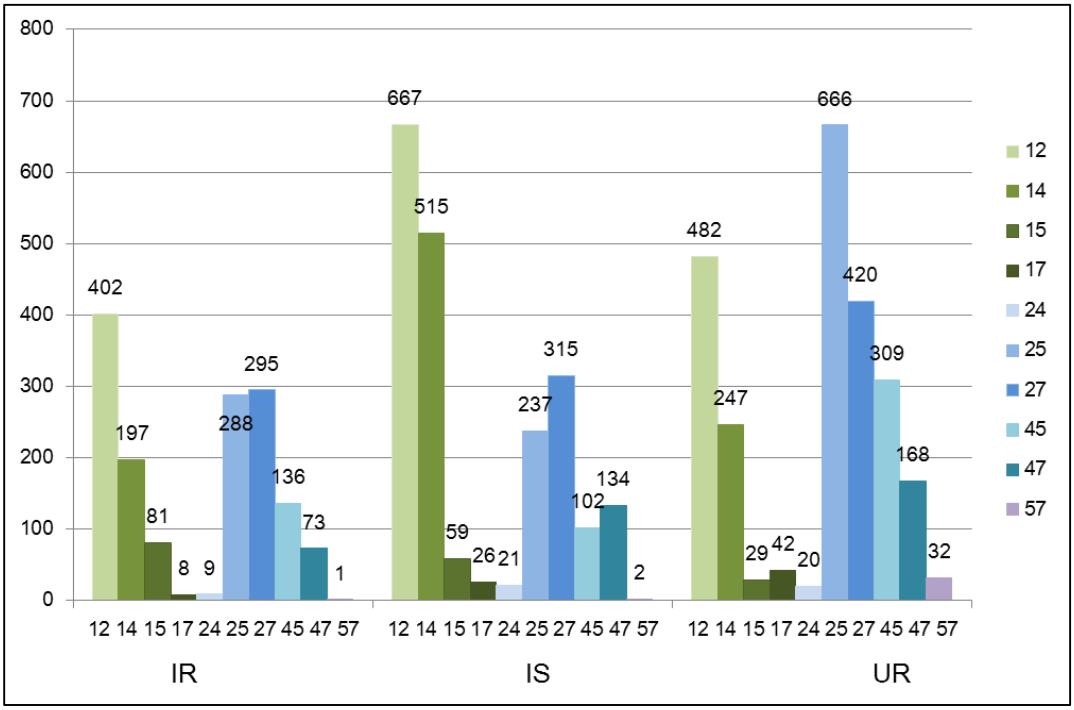


Figure 2-3. Number of DEGs in time-point comparisons. X-axis indicates comparison of two time points (e.g. EJ, egg vs. juvenile; J1, juvenile vs. 1 DAI; 12, 1 DAI vs. 2 DAI). Y-axis indicates the number of DEGs in each comparison. (a) For *M. hapla*, there were 21 comparisons. The total number of DEGs obtained by summing up across all the comparisons was 1,224 (988). (b) For *M. truncatula*, there were 10 comparisons in total for each tissue

type: IR, IS, and UR. The total number of DEGs was 1,490 (699), 2,078 (895), and 2,415 (962) for IR, IS, and UR, respectively. The numbers in parenthesis are those counted by excluding DEGs that were redundantly identified between the different comparisons.

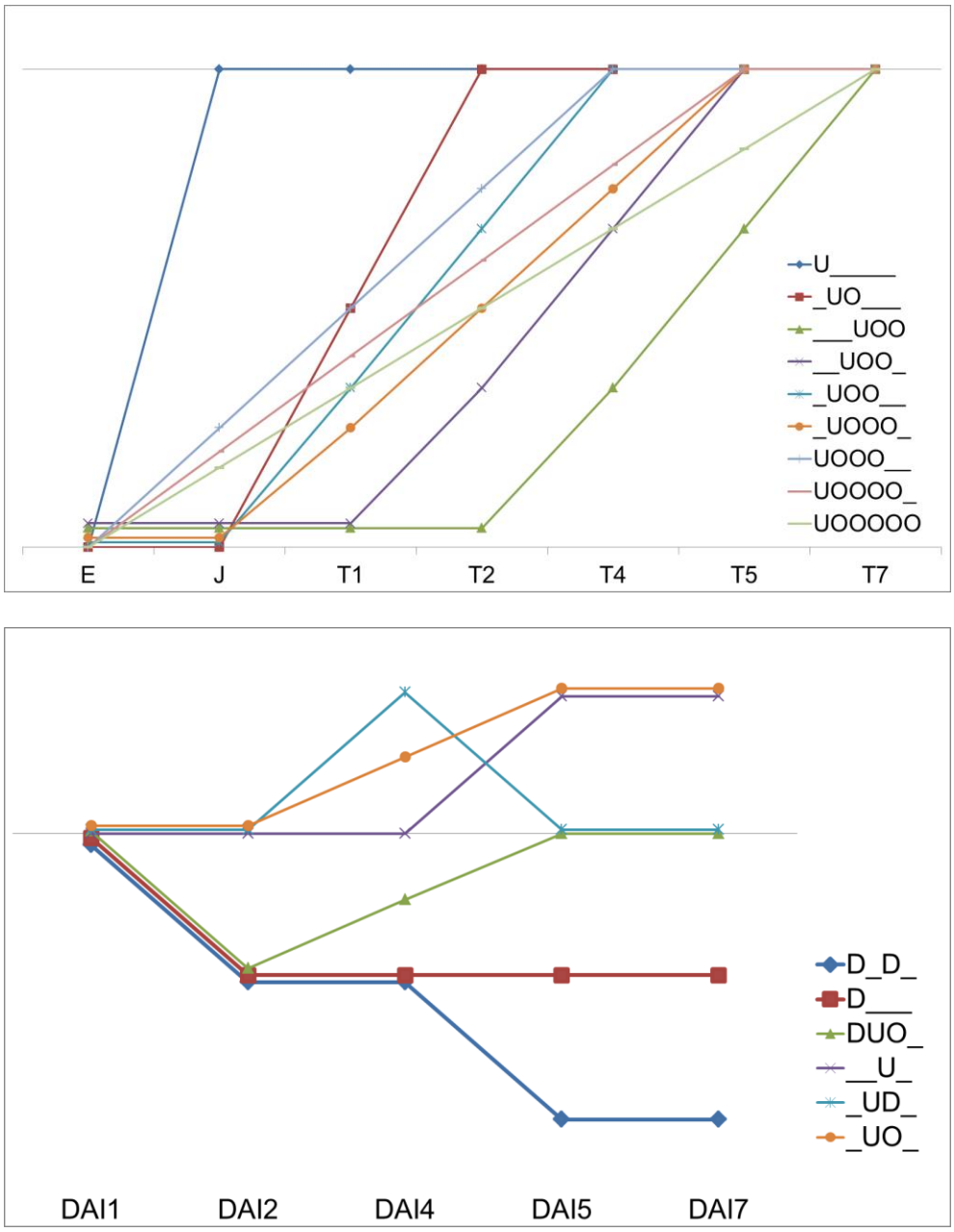
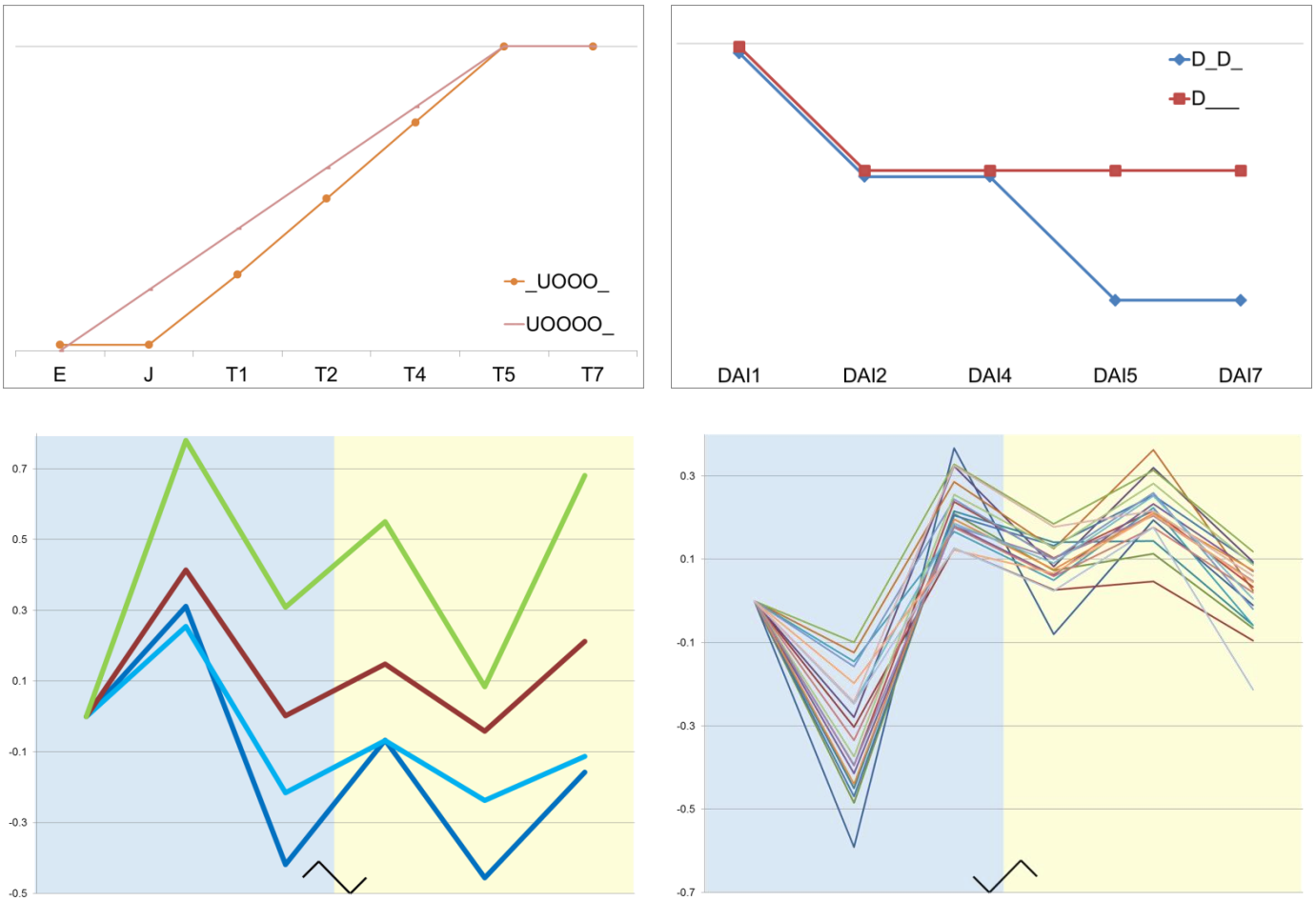


Figure 3. Nine groups in which DEGs of PP ∩ IR_Mh were classified as (up-regulated) and six groups in which DEGs of PP ∩ IR_Mt were classified as (down-regulated). An interesting point to mention here is that for *M. hapla*, all the six groups represented (gradual) up-regulation along the time courses. Also, for *M. truncatula*, 36 genes (90%) were classified into the down-regulated groups, D─D─, or D─·─·─.



M. hapla *M. truncatula*

Figure 4. Classified pattern groups (upper) and RPKM expression profiles in diurnal experiment (bottom) were graphed for DEGs commonly identified in both the longitudinal and diurnal experiments. *M. hapla* genes (left) including histidine phosphatase and collagen may be gradually increasing from egg to 7 DAI and be down-regulated during the day. In contrast, *M. truncatula* genes (right) for carbonic anhydrase, cytochrome, oxidureductase (OXI), and ATP synthase showed an overall decreased activity throughout the infection period and were up-regulated during the daytime.

CHAPTER 4

Comparative genomics: revolving around synteny

ABSTRACT

The root-knot nematodes (RKN, *Meloidogyne* spp.) are responsible for huge annual crop losses worldwide. Over a long evolutionary history, each species has been equipped with its own specific strategic life style for successful parasitism. Tracking the genetic origin of parasites has shed light on how they have functionally adapted, as well as what genomic features differentiate them from the free-living nematodes. Being closely related in phylogeny and being sympatric, *M. hapla* and *M. chitwoodi* could be presumed to exhibit high similarity in genes or alleles central to parasitism. In the present study, we report the featured synteny conserved between the northern RKNs, *M. hapla* and *M. chitwoodi*. They were additionally compared to a phylogenetically closely related southern RKN species *M. incognita* and to *C. elegans* as an outgroup. Using synteny identification, we inferred likely evolutionary events which might have destroyed synteny or collinearity. Furthermore, we propose the advantages achieved through synteny findings as reinforcing the existing annotations and refining the automated annotation of the *de novo* assembled genome. In addition, gene expansions that might have aided in parasitism as deciphered from the genome of *M. hapla* and *M. chitwoodi* are described. The *M. chitwoodi* genome duplication was described by identifying two ortholog on average. Collectively, the results show that genomic structures of *M. hapla* and *M. chitwoodi* are quite similar, compared to that of *M. incognita*. In addition, these RKN genomes contain informative gene elements required for parasitism.

INTRODUCTION

Nematode Diversity. Nematodes are considered as the most abundant animals on earth, with an estimated number of 10^6 to 10^8 species, accounting for 80% of all the metazoans (Sommer and Streit 2011). Not surprisingly, different species have evolved in such a way that they could adapt to a diverse ecological niche with different life style strategies including sedentary endoparasitic, migratory endoparasitic, and migratory ectoparasitic. The most damaging nematodes are sedentary endoparasitic, including cyst nematodes (CN, *Globodera* and *Heterodera* spp.) and root-knot nematode (RKN, *Meloidogyne* spp.) (Williamson and Hussey 1996). They span a wide range of agricultural systems, occurring from tropical to cold climate regions. In their geographic distributions in US, those RKNs, which are tolerant to low extremes of temperature, include *M. hapla* and *M. chitwoodi*. They are mostly found in Northeast and Northwest regions, although *M. hapla* has been found in all states except Alaska. In contrast, Southern species (*M. incognita*, *M. floridensis*, *M. arenaria*, and *M. javanica*) live in (sub-) tropical regions and cannot withstand temperatures below 3 °C. Regardless of their ecological diversity, both RKN and CN go through similar developmental life cycle stages: egg - four juvenile (J) stages - adult. J2 hatches from the egg under favorable conditions and (re) infects near the root tips. After penetration into the plant root, J2 induces the formation of a unique plant cell called the giant cell (GC) which performs a sole function of being a nutrient reserve for the nematode. This structure is one of the notable distinctions between RKN and CN. Whereas RKN forms GC by karyokinesis uncoupled with cytokinesis, CN induces syncytia by coalescing of uni-nuclear cells. Following this feeding site formation and several moltings, RKN females typically lay eggs in a gelatinous matrix

whereas CN females mostly retain eggs within their body. The reproduction modes of nematodes are also distinct among the different species. For example, CN is typically an obligate sexual reproductive species whereas RKN is an asexual parthenogenetic species. In addition, even within RKN species, distinction can be made between *M. hapla* with facultative, meiotic parthenogenesis and *M. incognita*, having obligate, mitotic parthenogenesis. Asexual parthenogenesis as a mode of reproduction in RKN is thought to contribute to retaining the genetic variation and rapid adaptation to unfavorable conditions (Castagnone-Sereno 2006).

To identify the distinctions and commonalities in plant parasitic nematodes, plant–pathogen interactions have been studied over several decades at the molecular level with the aim of overcoming destructive infection by nematodes. Investigations have been conducted to discover morphological novelties that facilitate parasitism, and to identify the genetic basis of parasitism which could then be targeted by chemicals. Furthermore, owing to the lowered cost of sequencing technology, genome-wide and cross-species inspections have been enabled. In the genome comparisons across different species, it could have been easily speculated that species closer in phylogenetic distance might be closely related in ecology, as well, because factors that contribute to the evolutionarily diverged genomes could have largely come from adaptations to the natural environment. However, it has been found that not only could the phylogenetically close dwell in completely different niches but phylogenetically distant species could also share similar ecologies. So far, no general rule has been elucidated from the comparison of genomes that could explain the distinct parasitic life styles or ecological niches. Rather, it has been presumed that each species has evolved

independently at least three times in plant parasitic nematodes and established its own survival strategies to potentiate its parasitic life style within the host, thus exhibiting convergent evolution. Another interesting point in RKN genome evolution is the horizontal gene transfer (HGT), which is extraordinarily widespread in nematodes. Over the past quarter century, accumulating studies have proposed that the enzymes which are absent in other metazoans but are exclusively expressed by plant parasites of RKN and CN might have originated by horizontal acquisition from bacteria or fungi (Smant et al. 1998; Bird and Koltai 2000; Haegeman et al. 2011; Mayer et al. 2011; Scholl and Bird 2011). These include plant cell wall degradation enzymes (e.g. pectate lyase, β-1,4-endoglucanases, and polygalacturonase), host defense suppressor (e.g. chorismate mutase), and NodL, which aid in plant parasitism through diversified gene functions.

As described, the history of nematode evolution should be reflected in their genomes to some extent. To obtain deeper insights and more evolutionary information from the genome, more thorough genome-wide comparisons at both macro- and micro-scales are needed especially for the closely related species which have not yet undergone many evolutionary changes after speciation and thus contain the most recent evolutionarily diverged events. Because the genomes of Southern RKN *M. incognita* (Abad et al. 2008), Northern RKN *M. hapla* (Opperman et al. 2008), and Columbia RKN *M. chitwoodi* (Cha and Bird 2016) have been sequenced and released, the opportunity to compare the genomes of closely related RKN species has expanded. Recently, transposon elements (TE) were broadly compared across 42 nematode species ranging from free living- to animal parasitic- and plant parasitic-nematodes (Szitenberg et al. 2016). These comparisons revealed that the composition of TE

varied over different species, consistent with phylogenetic trees. Notably, *M. hapla* and *M. chitwoodi* share much more similarity having four to six times less TEs than *M. incognita* and *M. floridensis*. Importantly, another study suggested that not only did *M. chitwoodi* and *M. hapla* cause synergistic damages to hosts, the host plant's resistance to these two RKN species was correlated (McCord 2012). These findings are compatible with the evolutionary distance observed in the phylogenetic tree (Figure 1, Scholl and Bird 2005) and the distribution of the nematode niche. Being sympatric with *M. hapla*, *M. chitwoodi* may share similar genes or alleles adapted to parasitism. To test this, we examined the syntenic regions of *M. hapla* and *M. chitwoodi* and observed that they were more conserved, compared to other nematode species. *Meloidogyne incognita* and *C. elegans* were chosen as less close and an outgroup genome, respectively. Moreover, we underlined the advantages that could be achieved through synteny identification, including the identification of annotation errors and reinforcement of existing annotations.

Synteny. What contributes to speciation cannot be explained by one general rule. Rather, it is understood by possible events that are likely to have happened. Assuming that such events would be reflected in the genome, we could infer the plausible scenario by observing the distinct or shared genomic features among the species. If several species have shared the same ancestry, it could be hypothesized that the closer the distance between species is, the more similar their genomes should be. In this sense, one of the most commonly observed characteristics in genome comparisons is described by 'synteny', which is referred to as the conserved block on the chromosome where genes are in the same order between different species. In other words, high abundance of a syntenic region could be an evidence for the

close phylogenetic distance of different species. The syntenic genes probably evolved from a common ancestral gene, and thus have the same function in different species; such genes are called ‘orthologs’. On the other hand, diverse sources of genomic rearrangements contribute to the differences between the genomes. As such, synteny would be broken by inversion, transposition, loss or gain, deletion or insertion, and duplication. When genome duplication occurs, the existing gene may evolve to have a new, related function, called a ‘paralog’. Moreover, as mentioned earlier HGT is responsible for shaping these nematode genomes, as well.

In this study, based on genome alignments, we observed some types of featured synteny in *M. chitwoodi*, *M. hapla*, *M. incognita*, and *C. elegans*. In addition, we demonstrate how an expansion of possibly horizontally acquired gene of *M. hapla* and *M. chitwoodi* might have occurred and how their patterns differ from each other. In addition, based on mapping the core eukaryotic genes conserved across wide ranges of taxa, we postulate that genome duplication might have been caused in *M. chitwoodi*. Finally, by combining RNA-Seq expression profiles of *M. hapla*, we present a possibility for more refined finishing work of gene annotation in *M. chitwoodi*.

RESULTS AND DISCUSSION

A global picture of four genome alignments (Figure 2) with the edges connecting the syntenic regions of any two species implicated that there would be a large amount of syntenies among the four nematode species, *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans*. A typical synteny could be demonstrated by two neighboring genes, *gck-3* and *snx-6*,

on linkage group V (LG. V) of *C. elegans* (*Cel-gck-3* and *Cel-snx-6*; Figure 3). For *M. incognita*, the contig containing *gck-3* and *snx-6* (*Mi-gck-3* and *Mi-snx-6*) was reversed and lacked two genes, whereas it was highly conserved between *M. hapla* and *M. chitwoodi*, encompassing a length of 30 Kb (Appendix B. Figure 1. (A)). Based on the tBLASTx searches for other genes existing on this contig, it was predicted to contain *pak-1* and *ama-1* (*Cel-pak-1* and *Cel-ama-1*), each of which was located on LG. X and LG. IV of *C. elegans*, respectively (Appendix B. Figure 1. (B), (C)). These results imply that during evolutionary process, two neighboring genes, *Cel-gck-3* and *Cel-snx-6* on LG. V and each of the two genes *Cel-pak-1* on LG. X and *Cel-ama-1* on LG. IV were independently concatenated into one contig with other genes. Following the speciation to *M. chitwoodi* and *M. hapla*, there were gene losses at several break points on this conserved region of *M. incognita* genome. Compared to *gck-3* and *snx-6*, the order of which was conserved among *C. elegans*, *M. hapla* and *M. chitwoodi*, *mua-3* and *unc-47* appeared to be interfered by fractionation, so called a broken synteny. The orthologs of two neighboring *C. elegans* genes, namely *Cel-mua-3* and *Cel-unc-47* were found in proximity on the contig of *M. hapla* and *M. chitwoodi* (Figure 4). The order of these two genes was disturbed by 11 other inserted genes in *M. hapla* and 7 other inserted genes in *M. chitwoodi*, over a range of 35 Kb (Appendix B. Figure 2). These contigs of *M. hapla* and *M. chitwoodi* were conserved in reverse orientation with several break points and several gene insertions or deletions.

Several notable observations were made in the syntenic regions while navigating the multiple gene alignments. Firstly, the most frequently observed patterns were the local alignments of conserved sequences between *M. hapla* and *M. chitwoodi*. In the conversed

synteny region of *M. hapla* and *M. chitwoodi*, it was partially conserved or sometimes reversed for *M. incognita*. Additionally, it seems that a majority of neighboring genes of *C. elegans* were probably independently aligned to each gene on separate contigs of other three *Meloidogyne* species. Secondly, the number of coding sequences (CDS) of the gene varied across *M. chitwoodi*, *M. hapla* and *C. elegans*, in decreasing order (Table 1; Appendix B. Table 1). This supposedly implied that *Meloidogyne* species might have had more splicing iso-forms of any particular gene than did *C. elegans*. Thirdly, despite the high sequence conservation between *M. hapla* and *M. chitwoodi*, *M. chitwoodi* was devoid of a gene model within the conserved region whereas, for *M. hapla*, annotation of genes in the corresponding position was available. Thus, synteny identification between the two closely related species could reinforce the existing annotation based on an automated algorithm. Lastly, for some cases, one gene of *C. elegans* was aligned as having been split into two or three genes of *M. hapla* and *M. chitwoodi*. These separately annotated genes of *M. hapla* supposedly originating from one orthologous gene of *C. elegans* could be better understood with the help of transcriptome profiles of *M. hapla*. These observations were not generalized for all the syntenies but represented an overall tendency. To describe the observation in detail, in the following sections, we present some featured syntenies formed by four nematodes, by focusing mainly on *M. hapla* and *M. chitwoodi*.

Synteny along with transcriptome profiles. *Cel-unc-70* and *Cel-unc-68* are located on LG. V of *C. elegans*, at a distance of 10 Kb; three more protein coding genes are present in between those two genes. Although each of these was separately aligned on different contigs of *M. hapla* or *M. chitwoodi*, the contig on which either *Cel-unc-70* or *Cel-unc-68* was

aligned was conserved between *M. hapla* and *M. chitwoodi*, except for one or two gene insertions for *M. hapla* (Figure 5-1). For *Cel-unc-70* on LG. V of *C. elegans*, it was reversely aligned to the contig which was conserved between *M. hapla* and *M. chitwoodi* except for two more gene insertions on the *M. hapla* contig (Appendix B. Figure 3. (A)). These inserted genes were predicted to be uncharacterized hypothetical proteins. Furthermore, *Cel-unc-68* on LG. V of *C. elegans* was aligned on the contig conserved between *M. hapla* and *M. chitwoodi* except for one gene insertion for *M. hapla* (Figure 5-2; Appendix B. Figure 3. (B)). However, it was aligned as three separate genes for *M. hapla* and two separate genes for *M. chitwoodi*, with several CDS being missed or gained. In addition, in between these separated genes, one to three genes were inserted for *M. hapla* and *M. chitwoodi*. The function of these inserted genes was uncharacterized and they were designated as hypothetical proteins in the tBLASTx search. The expression levels of transcripts in the diurnal experiments indicated that among the three *M. hapla* genes to which *Cel-unc-68* aligned in split form, only one gene was identified to be differentially expressed between night vs. day (Figure 5-3). This implies that the three split *Mh-unc-68* might play a role in different ways.

A *C. elegans* gene *pyr-1* on LG. II and another gene, *dpyd-1*, on LG. X were separately aligned on two different contigs of *M. chitwoodi*, namely contig 238736 and contig 238218 (Figure 6-1). However, each of these two *M. chitwoodi* contigs was conserved with each end of the *M. hapla* contig 740, that is, *M. chitwoodi* contig 238736 with left arm of the *M. hapla* contig 740 (Appendix B. Figure 4. (A)) and *M. chitwoodi* contig 238218 with right arm of the *M. hapla* contig 740 (Appendix B. Figure 4. (B)). In addition, *Cel-dpyd-1* was split to be aligned on two separate genes in both the *M. hapla* and *M. chitwoodi* contigs (Figure 6-2).

Based on this, we were able to conceive several scenarios: (1) after *pyr-1* and *dpyd-1* were concatenated with other genes into one conserved region, it might have been fractionated at the point of speciation to *M. chitwoodi*, (2) after speciation to *M. chitwoodi*, it might have been concatenated into one in *M. hapla*, (3) if there were no breaks in either region of *M. hapla* or *M. chitwoodi*, two contigs of *M. chitwoodi* should have been assembled into one contig in the process of *de novo* assembly of the *M. chitwoodi* genome. In the diurnal study of the two *Mh-dpyd-1* genes on which one *Cel-dpyd-1* gene was aligned, only one of them showed differential expression between night and day (Figure 6-3). *Mh-dpyd-1* and *Mh-pyr-1* showed a decrease in transcript levels during the day following the peak observed in the central night.

The contigs of *M. hapla* and *M. chitwoodi* on which *myo-3* on LG. V of *C. elegans* was aligned were conserved except for two more genes insertion in *M. hapla* (Figure 7-1; Appendix B. Figure 5). These two genes were predicted to code for hypothetical proteins. On the *M. hapla* contig, *C. elegans* gene *Cel-myo-3* was aligned by being split into three separate genes, with one gene added in between them. In the diurnal study, all these three genes showed similar expression pattern though there were some differences in the degree of expression (Figure 7-2). This was in contrast to the case of *Mh-unc-68* or *Mh-dpyd-1* in that among the several genes on which *Cel-unc-68* or *Cel-dpyd-1* were aligned, only one was differentially expressed.

In summary, in the diurnal study, of the three genes of *Mh-unc-68* (orthologs of *Cel-unc-68*), only one was a DEG with an expression curve opposite to that of the other two genes. Between the two genes of *Mh-dpyd-1* (orthologs of *Cel-dpyd-1*), only one gene was a

DEG with an expression curve at night opposite to that of the other gene. Although all the three genes of *Mh-myo-3* (orthologs of *Cel-myo-3*) showed similar expression curves, two of them were identified as DEGs. From these examples, the following possible inferences could be drawn: (1) those *M. hapla* orthologs that were annotated as different genes should actually be a single gene and its iso-forms derived by alternative splicing, and (2) the genes inserted in between the *M. hapla* orthologs were likely to be annotated wrongly and no gene model was required at that position. Thus, as described in these three examples, genome alignment along with the transcriptome profiles could help understand the gene annotation and its expression better.

Synteny and a missing annotation. A synteny of about 73 Kb was conserved between *M. hapla* contig 20 and *M. chitwoodi* contig 238000 (Figure 8; Appendix B. Figure 6). On this synteny, the gene *copb-2* on LG. IV of *C. elegans* was aligned. This represented a typical example of synteny widely observed in alignments of the four nematode genomes: the highly conserved syntenic region between *M. hapla* and *M. chitwoodi* could be partially and reversely conserved with *M. incognita*, containing only one conserved gene of *C. elegans*. This implied that after speciation to *M. chitwoodi* and *M. hapla*, there might have been many breaks or rearrangements for the *M. incognita* genome sequences. Yet, in this synteny between *M. hapla* and *M. chitwoodi* with high sequence similarity, one region of *M. hapla* was devoid of gene model though one gene was observed in the corresponding position in *M. chitwoodi*. This might imply a possible missed-model annotation in *M. hapla*.

As another example of a typical synteny diverged with several gene aberrations, the *M. hapla* contig 2 and the *M. chitwoodi* contig 241645 was conserved except for three more

gene insertions in *M. hapla*, or alternatively, except for three more gene deletions on *M. chitwoodi* (Figure 9; Appendix B. Figure 7). On these contigs of *M. hapla* and *M. chitwoodi*, *alg-1* on LG. X of *C. elegans* was reversely aligned, whereas it was aligned in the same direction in *M. incognita*. One function of the *M. hapla* genes that disturbed the complete synteny between *M. hapla* and *M. chitwoodi* was predicted to be peptidylprolyl isomerase (*pinn-1*). On the other hand, though *M. chitwoodi* showed sequence similarity to *M. hapla*, it was without one gene annotation, implying a missing annotation of *M. chitwoodi* genome. Along with *copb-2*, where a possible missing annotation was described, synteny analysis between closely related species could help to enhance each other's annotation.

Other examples of synteny. Contrary to the widely observed syntenies in the four genome alignments, such as the synteny containing *copb-2* and *alg-1* described in the previous section, there were some cases where *M. hapla* and *M. incognita* shared the synteny in high degree but it was partially and reversely conserved with *M. chitwoodi* (Figure 10; Appendix B. Figure 8). For example, the *M. hapla* contig 121 and *M. incognita* contig 257 formed 36 Kb long synteny, in which sequences were reversely conserved with *M. chitwoodi*. Although this synteny of *M. hapla* and *M. incognita* contained *cca-1*, it was deleted on the *M. chitwoodi* contig. This implied that at the point of speciation to *M. chitwoodi* and *M. hapla*, this region was reversed at several broken points for *M. chitwoodi* with *cca-1* having been lost.

Similar to the *cca-1*, the *M. hapla* contig 78 and the *M. incognita* contig 77 formed a dense synteny, but it was partially conserved with *M. chitwoodi* (Figure 11; Appendix B. Figure 9). In addition, there existed one gene that might have been deleted in *M. hapla* but

was present in *M. chitwoodi* and *M. incognita*. In addition, there was one gene deletion for *M. incognita*, as well, which was present in the other two. This implied that at the point of speciation to *M. chitwoodi*, there might have been several gene deletions in this region for *M. chitwoodi* whereas both *M. hapla* and *M. chitwoodi* lost one gene each. Furthermore, this conserved region among *M. hapla*, *M. chitwoodi*, and *M. incognita* was aligned to *asp-4* on LG. X of *C. elegans*. Similar to the *alg-1*, one region of *M. chitwoodi* with high sequence similarity to *M. hapla* lacked gene model, indicating a possible missing annotation of *M. chiwoodi*.

The *M. hapla* contig 1494 was conserved with the *M. chitwoodi* contig 236739 except for two more gene insertions in the *M. hapla* contig and one reversed gene to which the *Cel-spc-1* was aligned (Figure 12; Appendix B. Figure 10). This whole region of *M. hapla* was reversely conserved with a contig of *M. incognita*, except for two gene insertions in *M. hapla*. This implied that at the point of the speciation to *M. chitwoodi*, there should have been the reversion of SPC-1 for *M. chitwoodi*. In addition, whole of this region was reversed in *M. incognita* whereas two more genes were inserted in *M. hapla*.

Collagen Ortholog. The *M. hapla* contig 1040 containing collagen gene formed a synteny with the *M. chitwoodi* contig 240302, although two genes on the left arm of collagen were concatenated to one gene in *M. chitwoodi* (Figure 13-1). The function of this fused gene12 was predicted to be CUTiclin-Like (cutl-6) by tBLASTx. The left arm of this *M. chitwoodi* contig was also conserved with the *M. hapla* contig 78 in reverse direction (Appendix B. Figure 11). When the amino acids of collagen were aligned among *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans*, the three RKN species showed higher similarity compared to *C.*

elegans (Figure 13-2). As nematode undergoes several molt stages in the root before growing into an adult, it may express cuticle collagen to comprise its surface coat. As expected, in our longitudinal study, the *M. hapla* collagen gene showed gradual increase from J2 to 4 DAI (Figure 13-3).

KOG Protein Matches. Of the 248 core proteins, CEGMA predicted 245 proteins with 2.23 matches on an average in the genome of *M. chitwoodi*. However, when we directly blasted 248 KOG proteins against the translated genome database of *M. chitwoodi*, there was discrepancy of the matches between the CEGMA prediction and our BLAST result (Table 2). In BLAST, there were 240 proteins with 2.21 average numbers of matches. In addition, when the KOG proteins were blasted to the genome of *M. hapla*, there were 247 proteins with one match on an average. Among the proteins of *M. hapla* and *M. chitwoodi* hit by the KOG proteins, 124 proteins were found to be within a syntenic region between *M. hapla* and *M. chitwoodi*. In other words, one half of the core proteins were separately located over the genome whereas another half of them were in synteny. This is not surprising as genes are randomly distributed over the genome.

With the predicted KOG proteins on two different contigs of *M. chitwoodi* and one contig of *M. hapla*, we compared the two selected proteins through multiple alignments. For cytochrome c1 (KOG3052), each gene of *C. elegans* and *M. hapla* and two genes of *M. chitwoodi* were conserved, although two CDS at the 5′-end of *M. chitwoodi* cytochrome did not appear to be not conserved with either *M. hapla* or *C. elegans* (Figure 14). When their translated sequences were aligned to amino acid sequences from other species including *Drosophila melanogaster*, *Arabidopsis thaliana*, human, *Schizosaccharomyces pombe*, and

yeast, as expected, there were more gaps at the 5′-end of *M. hapla* and *M. chitwoodi* than at the 5′-end of *C. elegans* (Appendix B. Figure 12).

The gene encoding protein serine-threonine phosphatase (PSP) was within the syntenic region of *M. hapla* and *M. chitwoodi* that was conserved except for two more gene insertions in *M. hapla* (Figure 15). One CDS at the 3′-end of *M. hapla* PSP and each CDS at both the ends of *M. chitwoodi* PSP were not conserved with each end of *C. elegans*. In addition, on different contig of *M. chitwoodi*, there was another gene encoding PSP whose 5′-end was not conserved with that of *C. elegans*. When all these were aligned to the PSP amino acid sequences of other species, the 9–18 starting amino acids of *M. chitwoodi* were not matched (Appendix B. Figure 13).

Horizontally Transferred Genes. Of the candidates known to be originally acquired by horizontal transfer from bacteria or fungi, two proteins were selected as examples in this study, pectate lyase and chorismate mutase. The *M. hapla* contig 338 and the *M. chitwoodi* contig 237519 were conserved syntenies containing tandem and duplicated sequences encoding pectate lyase (Figure 16-1 (A)). Specifically, four genes of *M. hapla* and three genes of *M. chitwoodi* were aligned to each other. This might imply that in the evolution of RKN, pectate lyase was expanded in both *M. hapla* and *M. chitwoodi*, and that after speciation to *M. hapla*, one more gene was duplicated for *M. hapla*. In addition, another conserved synteny between *M. hapla* (contig 418) and *M. chitwoodi* (contig 236335) also contained the gene encoding pectate lyase in two copies for *M. hapla* and in one copy for *M. chitwoodi* (Figure 16-1 (B)). Similarly, it could be inferred that one more pectate lyase was added in *M. hapla*. These two cases of pectate lyases existing in synteny could be considered

to be in orthologous and paralogous relationship. Of all these pectate lyases, two genes in the *M. hapla* contig 338 and one gene in the *M. hapla* contig 418 were identified as differentially expressed in our longitudinal study, with expression patterns ─U─·─·─·─ (up-regulation from J2 to 1 DAI), UDOOO─ (up-regulation from egg to J2, followed by gradual decrease from J2 to 5 DAI), and U─·─·─·─·─ (up-regulation from egg to J2) (Figure 16-2). Thus, even orthologous or paralogous genes encoding pectate lyases might function in different developmental stages. There were other pectate lyases not in syntenic region but independently on the contig in *M. hapla* and *M. chitwoodi*. For example, the gene of pectate lyase on contig 329 and contig 711 of *M. hapla* was aligned with the gene on contig 236792 and contig 241346 of *M. chitwoodi* respectively (Figure 16-1 (C), (D), (E)).

The gene encoding chorismate muase on *M. hapla* contig 1702 and *M. chitwoodi* contig 236308 was aligned, in reversed direction (Figure 17-1). Both of them had three CDS each of which was found to be conserved with other species including *M. arenaria* and *G. pallida* (Appendix B. Figure 14). In our longitudinal study, this chorismate mutase showed a large increase in expression from egg to J2 (Figure 17-2), indicating a critical role in parasitism.

CONCLUSION AND FUTURE STUDIES

In summary, by presenting diverse featured syntenies among closely related RKN species and *C. elegans* as an outgroup, we supported the presumption that *M. hapla* and *M. chitwoodi* share more similar genomic features as sympatric species. Additionally, double hits of core eukaryotic proteins on the *M. chitwoodi* genome could enforce the possible genome duplication events. In addition, through multiple alignments of key parasitism proteins, we

revealed the differences in the inherited orthologs and paralogs. Furthermore, the orthologous gene functions of *M. chitwoodi* could be inferred by combining the mRNA transcript expression profiles of *M. hapla* along the developmental stages and diurnal rhythm. Subsequently, we showed the possibility of reinforcing the partially incomplete automated annotation of *M. chitwoodi*. Furthermore, the identification of synteny could facilitate the gap closure of contigs, which should have been assembled as one contig when previously being assembled *de novo*. Based on the featured synteny presented in this study, our lab will conduct more intensive whole genome comparisons of *M. hapla* and *M. chitwoodi* in the near future. Finally, tracking the pattern in which RKNs have evolved their genomes could broaden our knowledge about their historic survival strategies by which they have successfully adapted to the environments and suppressed the plant's responsive defense. Based on thorough understandings of their genomes, more effective genetic control methods could be envisaged to emasculate the parasites and to strengthen the crops.

MATERIALS AND METHODS

Data Source. The genome and annotation of *M. hapla*, *M. incognita* and *C. elegans* were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/assembly/GCA_000172435.1/, http://www.ncbi.nlm.nih.gov/assembly/GCA_000180415.1/, http://www.ncbi.nlm.nih.gov/assembly/GCA_000002985.3). The genome of *M. chitwoodi* was referred to in the European Nucleotide Archive (Leinonen et al. 2011) under accession number PRJEB12397 (http://www.ebi.ac.uk/ena/data/view/PRJEB12397). We used the annotation file of *M. chitwoodi* generated in a previous study (Cha and Bird 2016).

tBLASTn. According to the CEGMA prediction, in the *M. chitwoodi* genome, 245 of 248 core proteins were matched with 2.23 orthologs on average, implying genome duplication or a genome with high heterozygosity. As an extended study, we blasted a 248-core protein with the translated genome of *M. chitwoodi* and *M. hapla*, respectively, to identify the candidates of core protein orthologs, setting the p-value at $1.0 \times E^{-5}$. Top hit of two or one was used for *M. chitwoodi* or *M. hapla*, respectively.

Multiple Alignment of Amino Acids. Amino acid sequences of the conserved protein domain family were obtained from NCBI resources (http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=KOG#### where '####' is replaced by any KOG protein number).

Software. The Core Eukaryotic Gene Mapping Approach (CEGMA) was used to identify the number of conserved CEGs in the *M. chitwoodi* genome. For overall identification of synteny among *M. hapla*, *M. chitwoodi*, *M. incognita* and *C. elegans*, we implemented MAUVE (Darling et al. 2004). Based on the probable synteny candidates obtained from MAUVE, we deployed CoGe (Lyons et al. 2008) to zoom in the local syntenic regions at higher resolution and to visualize the conserved synteny blocks and gene models.

	M. hapla	*M. chitwoodi*	*C. elegans*
gck-3	9 CDS	11 CDS	8 CDS
snx-6	9 CDS	9 CDS	10 CDS
pak-1	12 CDS	18 CDS	11 CDS
ama-1	24 CDS	27 CDS	12 CDS
mua-3	70 CDS	92 CDS	34 CDS
unc-47	14 CDS	18 CDS	7 CDS
pyr-1	40 CDS	41 CDS	13 CDS
dpyd-1	14 + 9 CDS	7 + 17 CDS	13 CDS
unc-70	32 CDS	36 CDS	13 CDS
unc-68	31 + 36 + 5 CDS	74 + 14 CDS	57 CDS
copb-2	11 CDS	13 CDS	10 CDS
cca-1	39 CDS	CDS	30 CDS
spc-1	32 CDS	32 CDS	13 CDS
asp-4	11 CDS	9 CDS	6 CDS
myo-3	15 + 2 + 10 CDS	32 CDS	7 CDS
alg-1	16 CDS	21 CDS	7 CDS

Table 1. Number of CDS of genes within the synteny across *M. chitwoodi*, *M. hapla* and *C. elegans*, normally in decreasing order

KOG	*M. hapla*	Function
KOG1466	o	Translation initiation factor 2B, alpha subunit (eIF-2Balpha/GCN3)
KOG1760	x	Molecular chaperone Prefoldin, subunit 4
KOG2531	o	Sugar (pentulose and hexulose) kinases
KOG2535	o	RNA polymerase II elongator complex, subunit ELP3/histone acetyltransferase
KOG3013	o	Exosomal 3′-5′ exoribonuclease complex, subunit Rrp4
KOG3318	o	Predicted membrane protein
KOG3479	o	Mitochondrial import inner membrane translocase, subunit TIM9

Table 2. Out of the 248 core proteins, 7 KOG proteins which in BLAST did not match the translated genome of *M. chitwoodi* (x, unmatched with the both *M. chitwoodi* and *M. hapla*; o, unmatched with *M. chitwoodi* but matched with *M. hapla*). There were 241 proteins with 2.21 average numbers of matches in the genome of *M. chitwoodi*. There were 247 proteins with one match on an average in the genome of *M. hapla*.

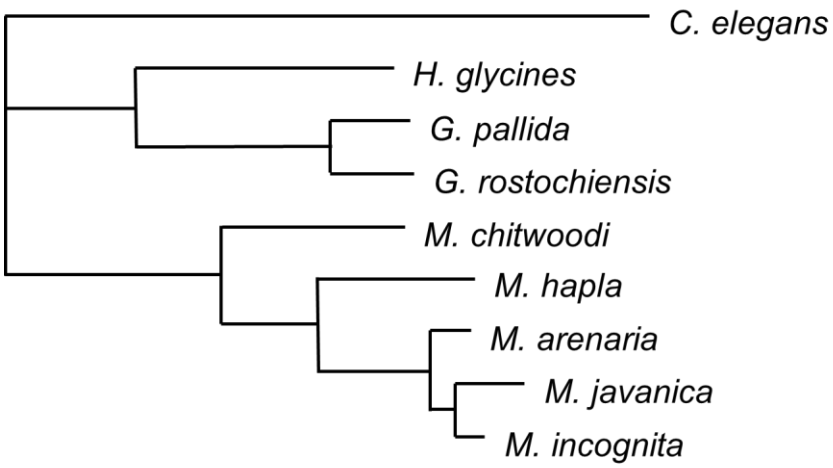


Figure 1. Resolving tylenchid evolutionary relationships through multiple gene analysis (Scholl and Bird 2005)

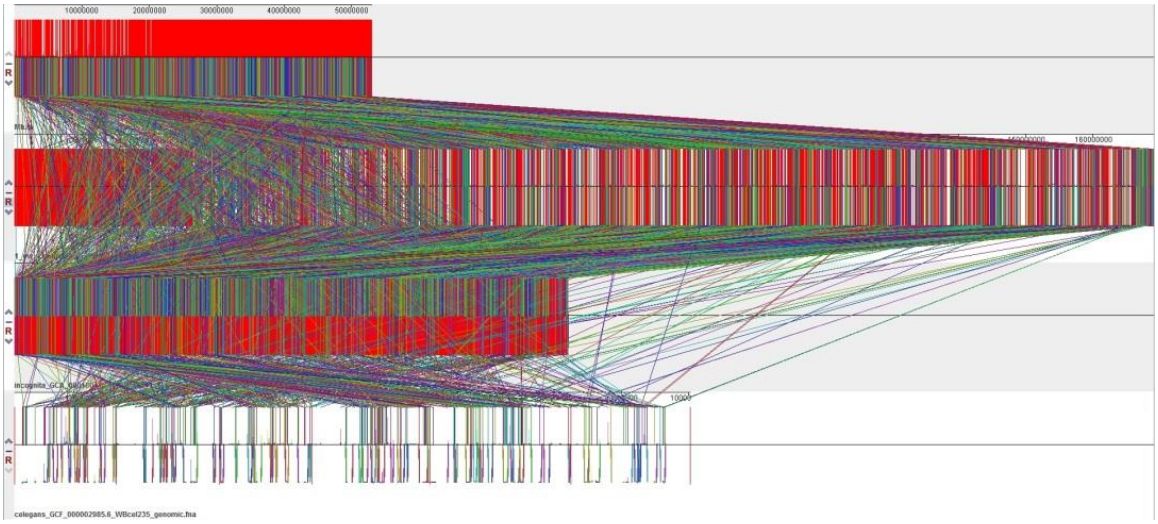


Figure 2. Genome alignment of four nematode species, *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (from top to bottom).

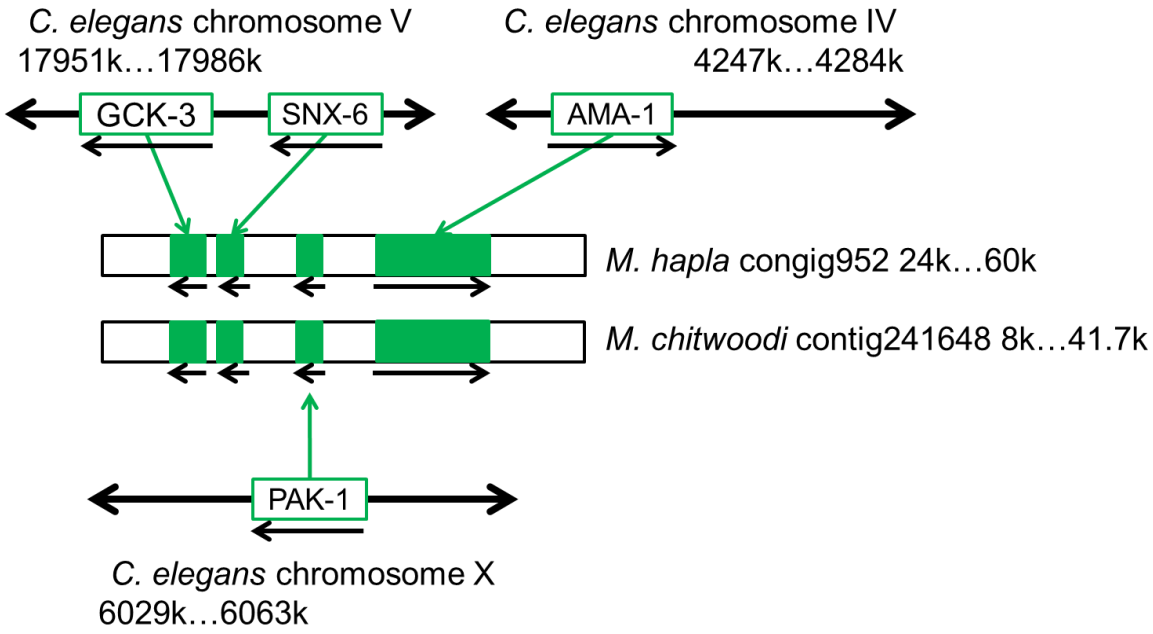


Figure 3. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans*, where *gck-3*, *snx-6*, *ama-1*, and *pak-1* were aligned.

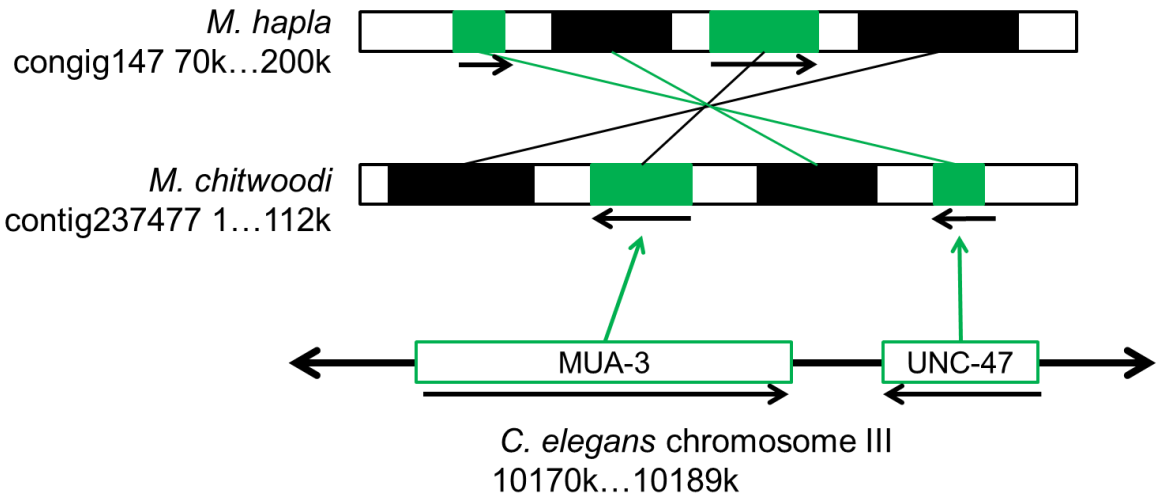


Figure 4. Reversely conserved region between *M. hapla* and *M. chitwoodi* on which the two neighboring *C. elegans* genes, *Cel-mua-3* and *Cel-unc47*, were aligned.

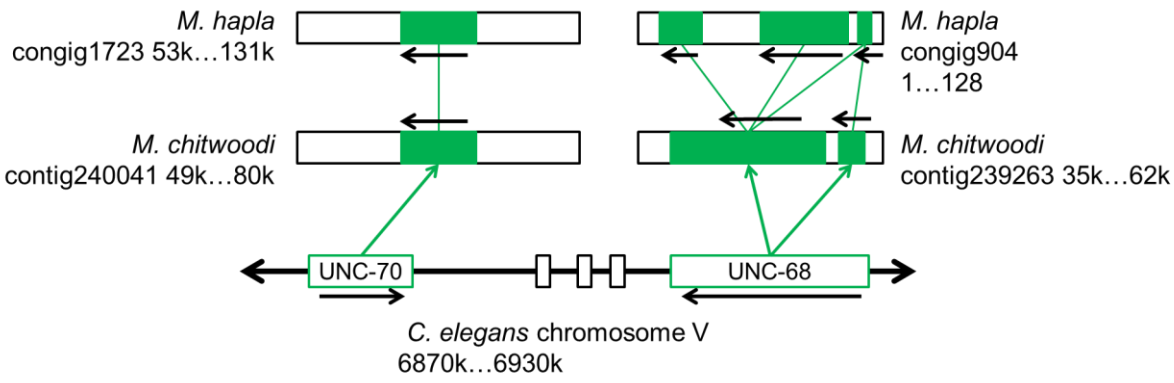


Figure 5-1. Two *C. elegans* genes, *Cel-unc-70* and *Cel-unc-68*, in proximity, separately aligned on different syntenic regions, conserved between *M. hapla* and *M. chitwoodi*. *Cel-unc-68* was separately aligned as the three genes of *M. hapla*.

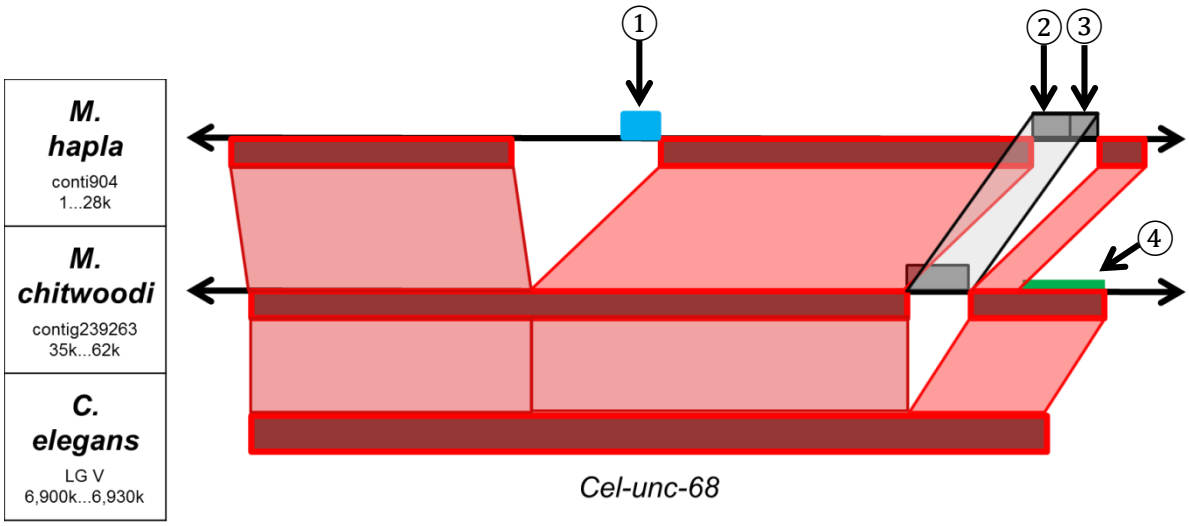


Figure 5-2. *Cel-unc-68* aligned as three separate genes for *M. hapla* and two separate genes for *M. chitwoodi*. In between these separated genes, there were three gene (①–③) insertions for *M. hapla* and one gene insertion for *M. chitwoodi*. The gene insertion that disturbed the conserved region between *M. hapla* and *M. chitwoodi* was indicated by blue box (①). The partial deletion of *unc-68* in *M. hapla* was indicated by green line (④).

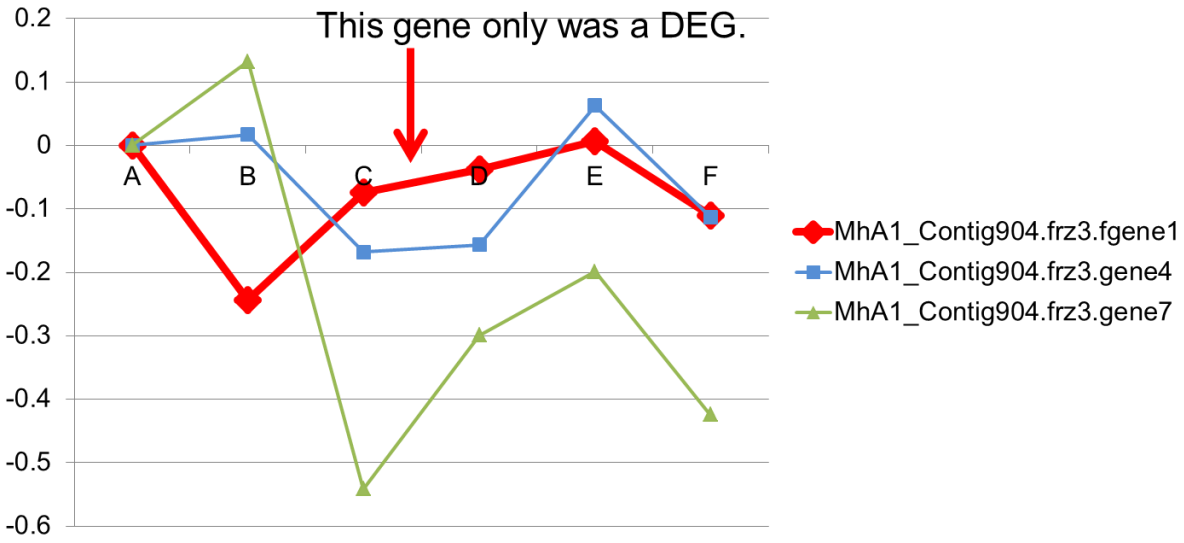


Figure 5-3. Three *M. hapla* genes on which one split *C. elegans* gene, *Cel-unc-68*, was aligned. Of the three *M. hapla* genes, only MhA1_Contig904.frz3.fgene1 was differentially expressed in the test, night vs. day. (A, B, C: time points at night; D, E, F: time points during the day). It is known that *Cel-unc-68* aids in normal body tension and locomotion. Thus, it could be deduced that different isoforms with different functions in body tension might be differentially activated during night and day.

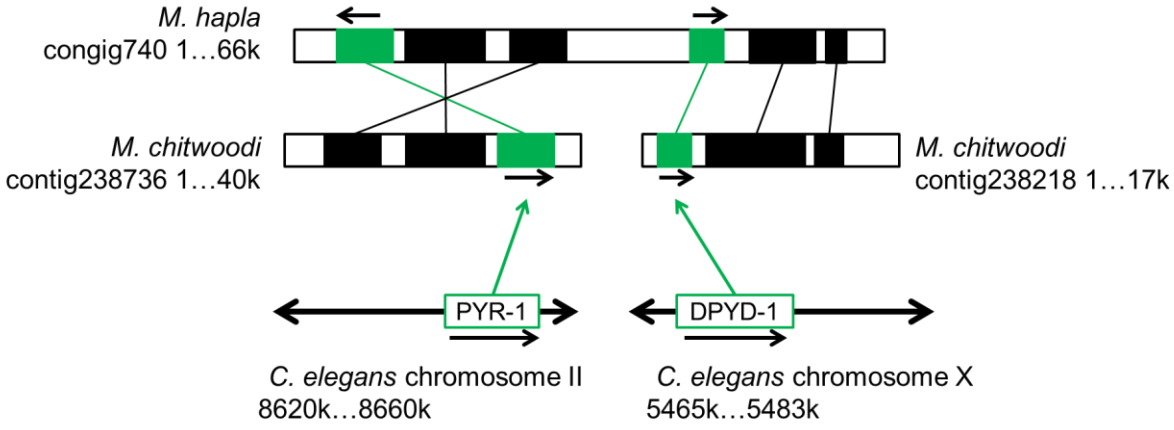


Figure 6-1. Two *C. elegans* genes, *Cel-pyr-1* and *Cel-dpyd-1*, located on separate linkage groups aligned on the same contig, showing three conserved region between *M. hapla* and *M. chitwoodi*.

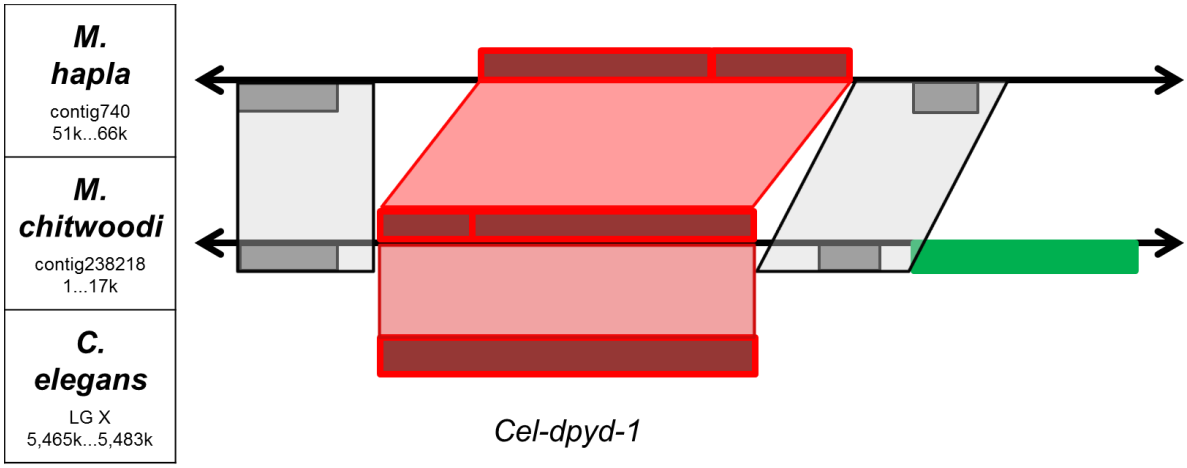


Figure 6-2. *Cel-dpyd-1* aligned on two separate genes in both the *M. hapla* and *M. chitwoodi* contigs. One gene insertion in *M. chitwoodi* or, alternatively, one gene deletion in *M. hapla* was indicated by green box.

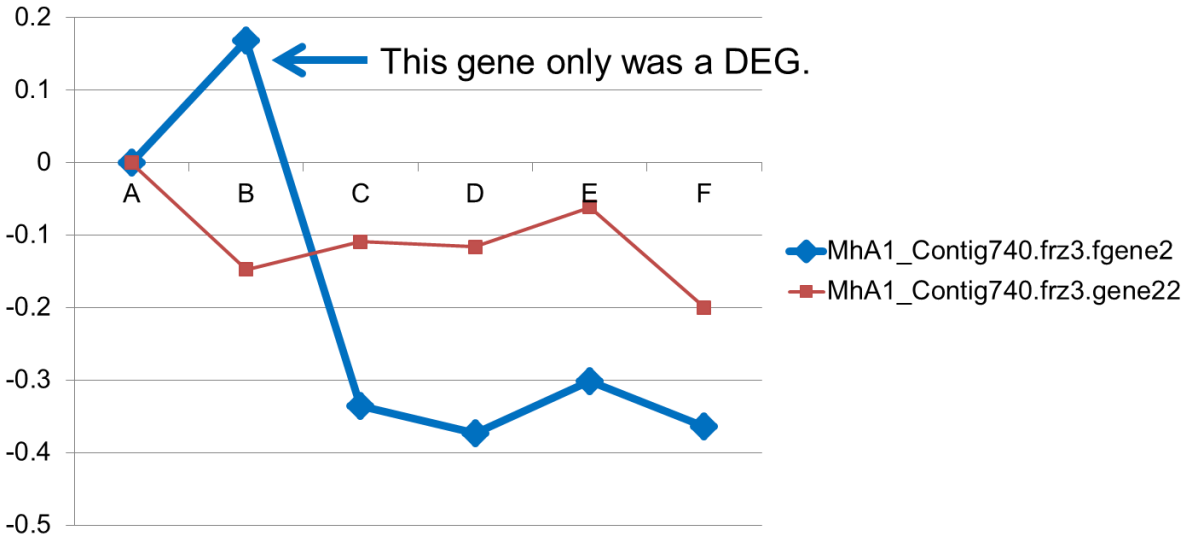


Figure 6-3. Split *Cel-dpyd-1* aligned to two *M. hapla* genes. Of the two *Mh-dpyd-1* genes, only one was differentially expressed and showed a decrease during the day. *Mh-pyr-1* showed same expression pattern as *Mh-dpyd-1*. (A, B, C: time points at night; D, E, F: time points during the day)

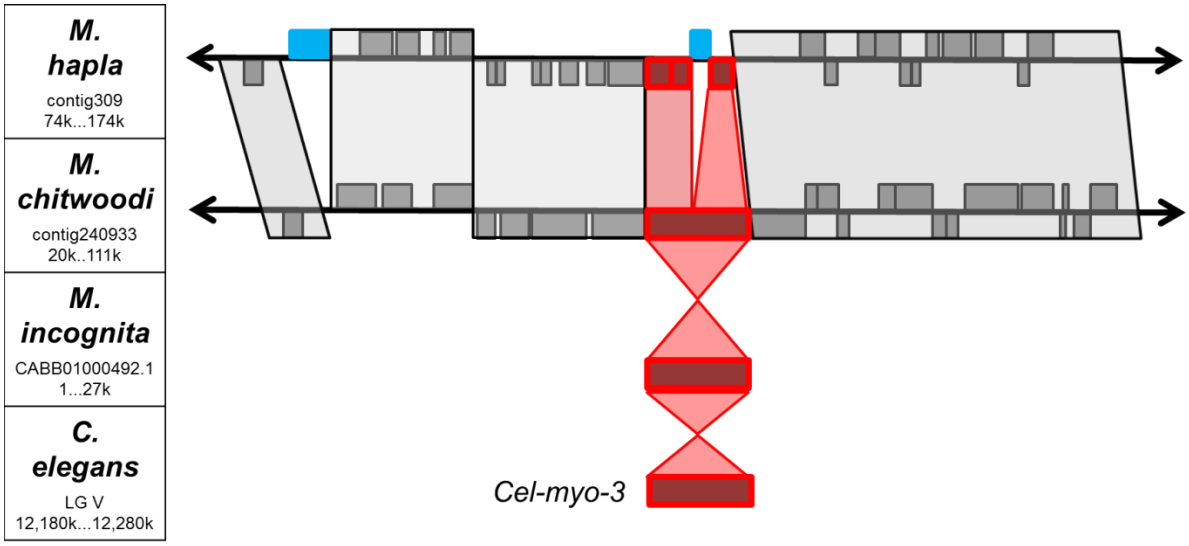


Figure 7-1. Syntenic contig between *M. hapla* and *M. chitwoodi* on which *Cel-myo-3* was aligned. On the *M. hapla* contig, *C. elegans* gene *Cel-myo-3* was aligned by being split into three separate genes, with one gene added in between them.

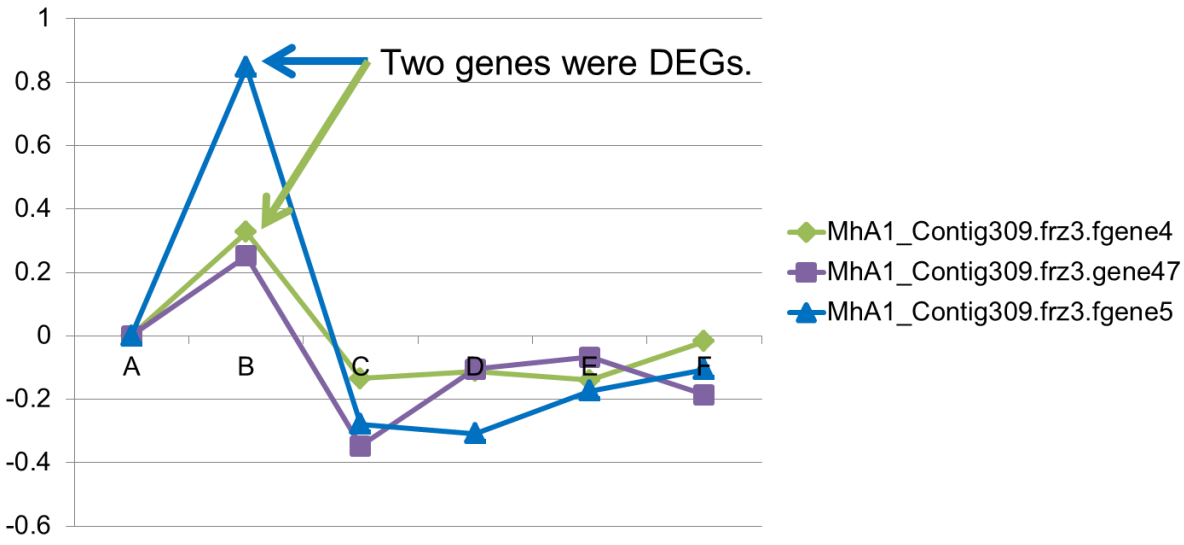


Figure 7-2. Three *M. hapla* genes on which the split *Cel-myo-3* was aligned showing the same expression pattern throughout a day; the expression decreased from night to day. (A, B, C: time points at night; D, E, F: time points during the day)

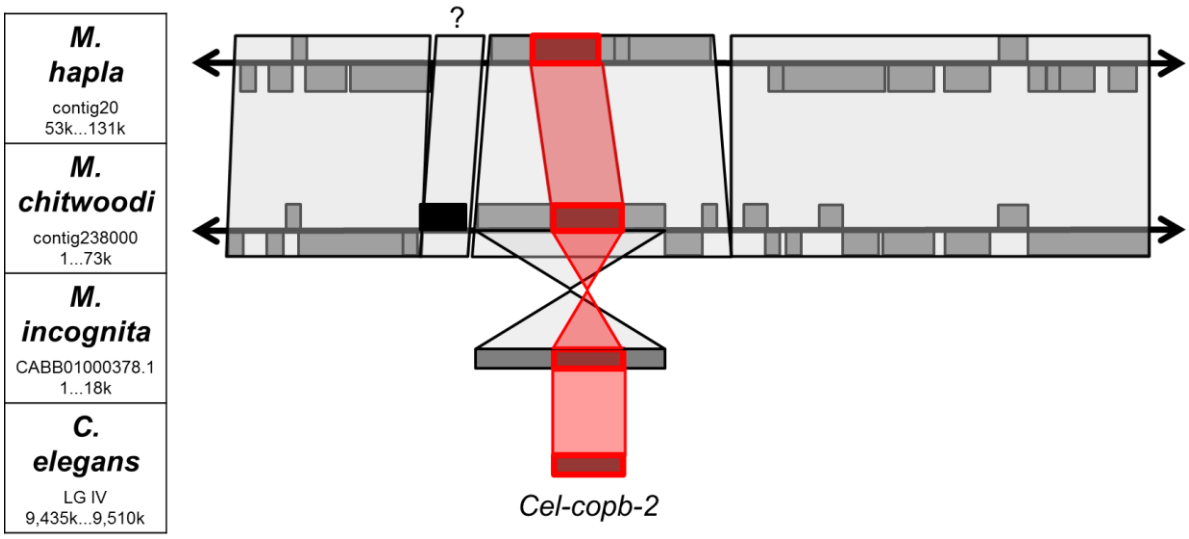


Figure 8. Syntenic contig between *M. hapla* and *M. chitwoodi* on which *Cel-copb-2* was aligned. *M. hapla* was devoid of gene model though one gene was observed in the corresponding position in *M. chitwoodi*, implying a possible missed-model annotation in *M. hapla*. The highly conserved syntenic region between *M. hapla* and *M. chitwoodi* was partially and reversely conserved with *M. incognita*.

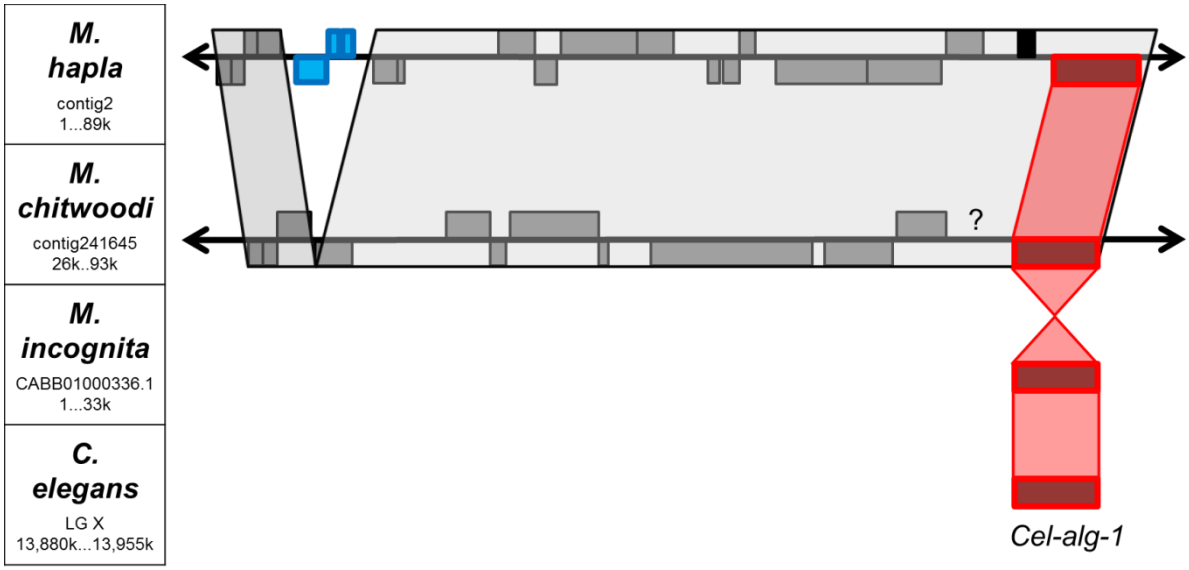


Figure 9. Conserved contig between *M. hapla* and *M. chitwoodi* on which *Cel-alg-1* was aligned. Though *M. chitwoodi* showed sequence similarity to *M. hapla*, it was without one gene annotation, implying a missing annotation of *M. chitwoodi* genome.

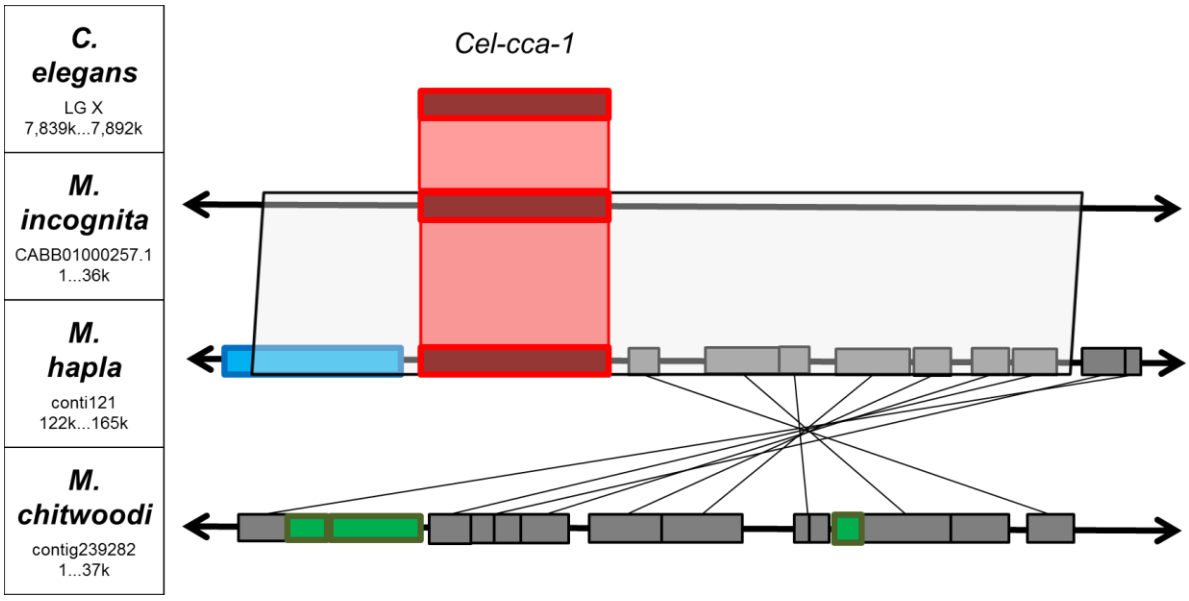


Figure 10. *M. hapla* contig on which *Cel-cca-1* was aligned. This *M. hapla* contig was more conserved with *M. incognita* whereas it was partially and reversely conserved with *M. chitwoodi*. *cca-1* was deleted on the *M. chitwoodi* contig, but it contained three gene insertions indicated by green box.

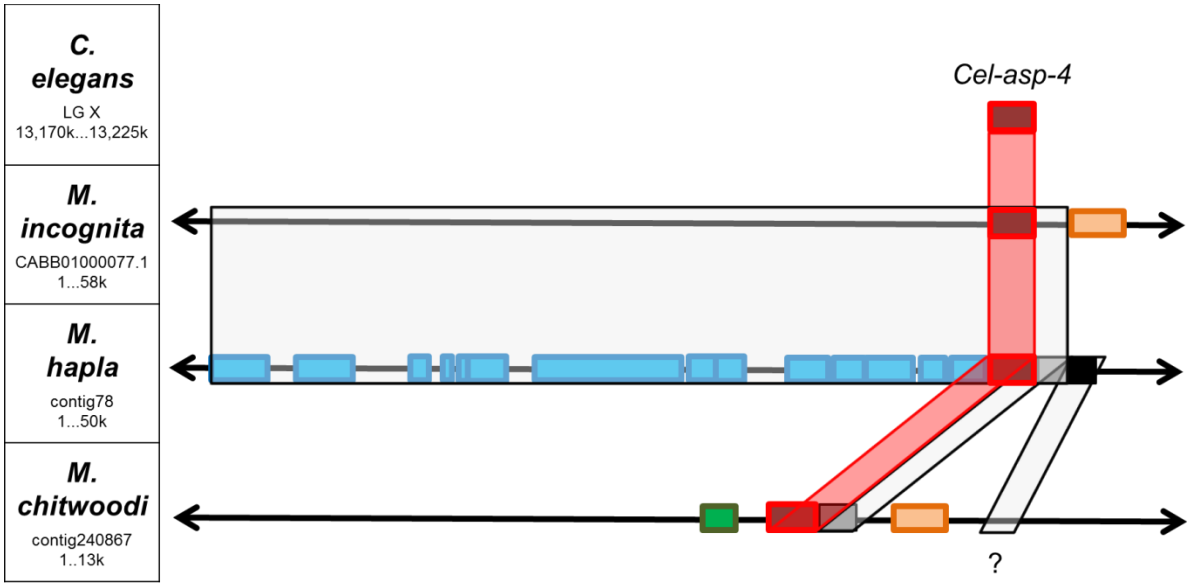


Figure 11. *M. hapla* contig on which *Cel-asp-4* was aligned. This contig was more conserved with *M. incognita* whereas it was partially conserved with *M. chitwoodi*. One region of *M. chitwoodi* with high sequence similarity to *M. hapla* lacked gene model, indicating a possible missing annotation of *M. chitwoodi*.

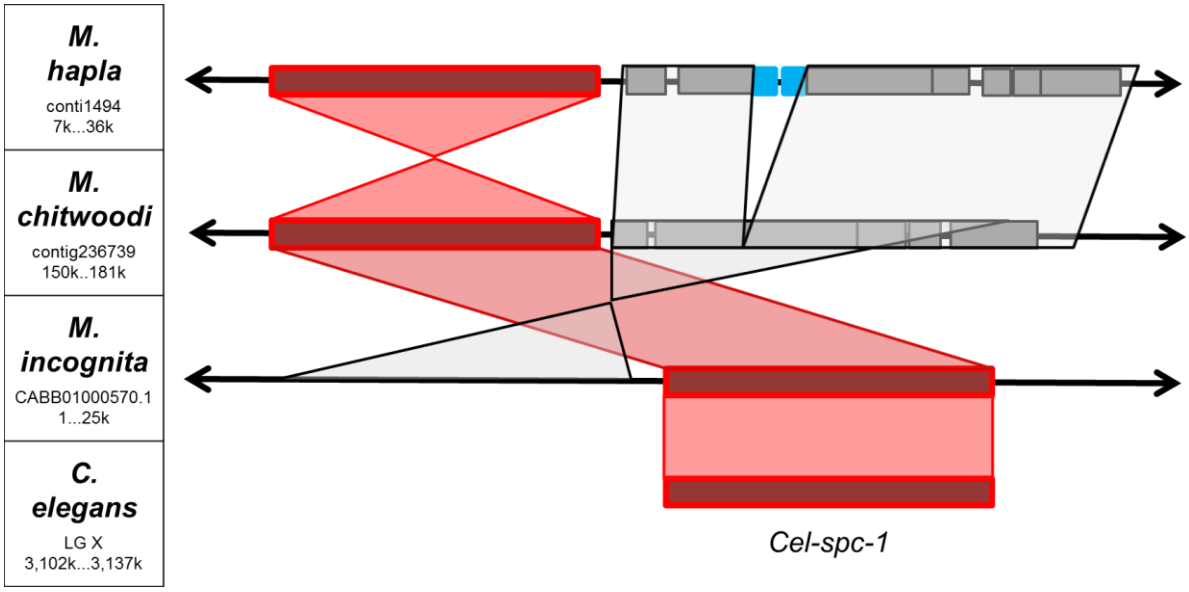


Figure 12. Contigs between *M. hapla* and *M. chitwoodi* showing conserved regions except for a two-gene insertion and one gene reversion in the *M. hapla* contig. *Cel-spc-1* is involved in the formation of body wall, locomotion and larval development. Also correct localization of SPC-1 is aided by UNC-70.

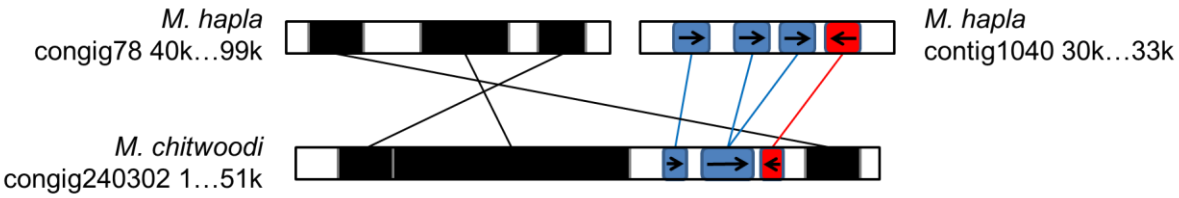


Figure 13-1 (A). An *M. hapla* contig containing collagen gene and another *M. hapla* contig together forming the conserved region with the *M. chitwoodi* contig. The gene models in right arm of this region were depicted in Figure 13-1 (B).

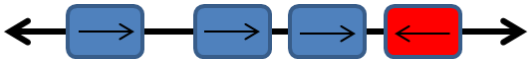
M. hapla	*M. chitwoodi*
Contig1040:22,500…35,000	Contig240302:40,000…48,500
gene11, gene12, gene13, gene14	g14319, g14320, g14321
	

Figure 13-1 (B). Gene model comparison between *M. hapla* and *M. chitwoodi*. Collagen gene was indicated by red box. Two *M. hapla* genes in the middle were concatenated to one gene in *M. chitwoodi* (Figure 13-1). The function of this fused gene was predicted to be CUTiclin-Like (cutl-6) by tblastx.

```
Ce MFEEKLLVGIASLASTIAILTCVVVIPGLYSTINEMHDEVLDGVKMFRADTDAAWVEMLDVQVMVSPPSQPKENPFNSVFRQKRRSTFSGLPAWCQCEPTKPKCP
   M E K+++GIA  +S +AI+  +VV+P LYS IN+++  V DGV+ FR +TD+AW +++++Q+ V+PPS+P+ NPF S++RQKR     GLPA+C C+P + K
Mh MQETKVVIGIACFSSLLAIMATLVVMPQLYSQINDLNLRVRDGVQAFRVNTDSAWNDLMELQISVTPPSKPRSNPFQSLYRQKR-----GLPAYCICQPLEIKGA
   MQETKVVIGIACFSSLLAIMATLVVMPQLYSQINDLNLRVRDGVQAFRVNTDSAWNDLMELQ++VTP SKPRSNPFQSLYRQKR      LP YCICQPLEI  A
Mi MQETKVVIGIACFSSLLAIMATLVVMPQLYSQINDLNLRVRDGVQAFRVNTDSAWNDLMELQVAVTPQSKPRSNPFQSLYRQKR-----SLPDYCICQPLEINCA
   MQETKVVIGIACFSSLLAIMATLVVMPQLYSQINDLNLRVRDGVQAFRVNTDSAWNDLMELQ+AVTP SK RSNPFQSLYRQKR      LP YCICQPLEINC
Mc MQETKVVIGIACFSSLLAIMATLVVMPQLYSQINDLNLRVRDGVQAFRVNTDSAWNDLMELQIAVTPPSKARSNPFQSLYRQKR-----QLPAYCICQPLEINCK

Ce RGPPGPPGHPGQRGIPGIPGRNGQDNYNTIRAPACPPRNQDCIKCPAGPPGPSGTCGQVGRPGPDGRPGQPGRRGNDGRPGQPGPQGNAGQPGRDGNPGQPGHPGKDGRRGHGS
    GPPGPPG PGQ G PG PG  GQ       AP CP   Q C +CPAG PG  G  G  G+PG  GRPG PG+    G PG  GPQG  G  G+ G PGQPG PGK+G
Mh PGPPGPPGPPGQPGHPGQPGHVGQPGSPGQPAPPCPLPMQACQRCPAGAPGTPGKQGPAGQPGQPGRPGPPGKGTGAGPPGPAGPQGPPGPAGKHGGPGQPGQPGKNGVSSPTI
   PGPPGPPGPPGQPGHPGQPGHVGQPGSPGQPAPPCPLP QACQRCPAGAPGTPGKQGPAGQPGQPGRPG PGK +GAGPPGPAGPQGPPGPAGKHGGPGQPGQPGKNGVS PTI
Mi PGPPGPPGPPGQPGHPGQPGHVGQPGSPGQPAPPCPLPQQACQRCPAGAPGTPGKQGPAGQPGQPGRPGAPGKSSGAGPPGPAGPQGPPGPAGKHGGPGQPGQPGKNGVSHPTI
   PGP GPPGPPGQPGHPGQPGHVGQPG PGQ APPCPLPQQACQRCPAGAPGTPGKQGP+GQPGQPGRPG PG     GPPGPAGPQGPPGPAGKHGGPGQPGQPGKN  S PTI
Mc PGPQGPPGPPGQPGHPGQPGHVGQPGQPGQQAPPCPLPQQACQRCPAGAPGTPGKQGPSGQPGQPGRPGPPGLKYFKGPPGPAGPQGPPGPAGKHGGPGQPGQPGKNAYSQPTI

Ce PGAPGRAGQPGRQGAP---GNPGRPGERGPSGPCGPAGRSGQPGNRGSDGHPGAPGNPGLQGSDAAYCACPTRSVMFLKRH
   PG  G +G PG+ G P   G PG+PG  GP GP GPAG SG+P   G+ G PG  G PG  G DA YC CP RS +   +
Mh PGPKGPSGSPGQPGKPGPAGEPGKPGPEGPPGPIGPAGPSGKPGAPGQPGPHGPPGQPGQDAQYCPCPPRSSVLKAKKRASMQETKVVIGI
   PGPKGPSGSPGQPGKPGPAG  GK GPEGPPGP+GPAGPSGKPGAPGQPGPHGPPGQPGQDAQYCPCPPRS   + +   R  M+  K +I
Mi PGPKGPSGSPGQPGKPGPAGVAGKTGPEGPPGPVGPAGPSGKPGAPGQPGPHGPPGQPGQDAQYCPCPPRSLCSRQRNALRKQMKRKKAMI
   PG +GPSG PGQPG+PGP G + K+GPEGPPGPVGPAGPSGKPGAPGQPGPHGPPGQPGQDAQYCPCP RS  S  +   R  M+  K +I
Mc PGKQGPSGQPGQPGRPGPPGKSEKSGPEGPPGPVGPAGPSGKPGAPGQPGPHGPPGQPGQDAQYCPCPQRSSVLKAKKRASMQETKVVIGI
```

Figure 13-2. Gene encoding collagen positioned within the synteny of *M. hapla* and *M. chitwoodi*; this positioning of the gene indicates that it is orthologous between the two species. The amino acid sequences were aligned with *M. incognita* collagen (lemmi5; GenBank ID, AF006727.1) and *C. elegans* collagen (*Col-149* on LG. V). Collagen is normally characterized by glycine positioned in every third amino acid (colored in yellow). The sequences from the RKN species shared higher similarity, compared to that of *C. elegans*.

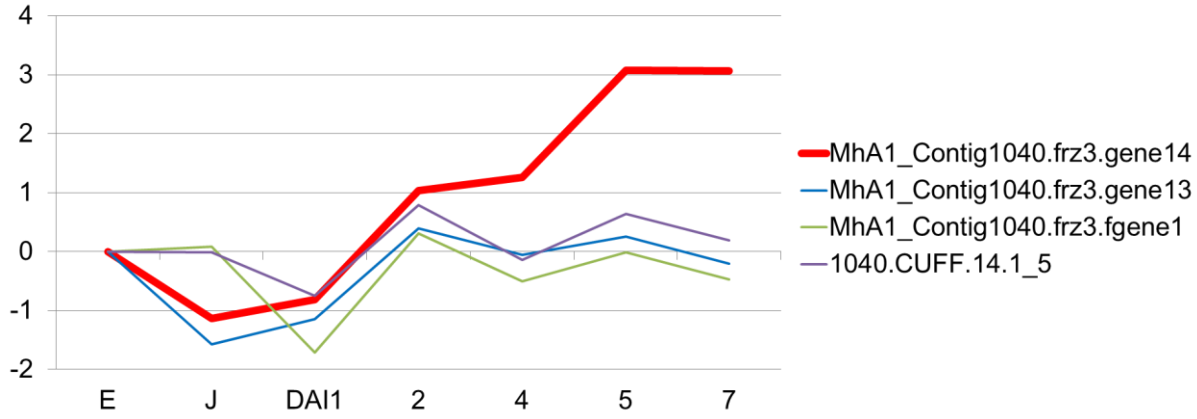


Figure 13-3. Gradual increase in the expression of *M. hapla* collagen gene from J2 to 4 days after infection (DAI) as observed in the longitudinal experiment. (A, B, C: time points at night; D, E, F: time points during the day)

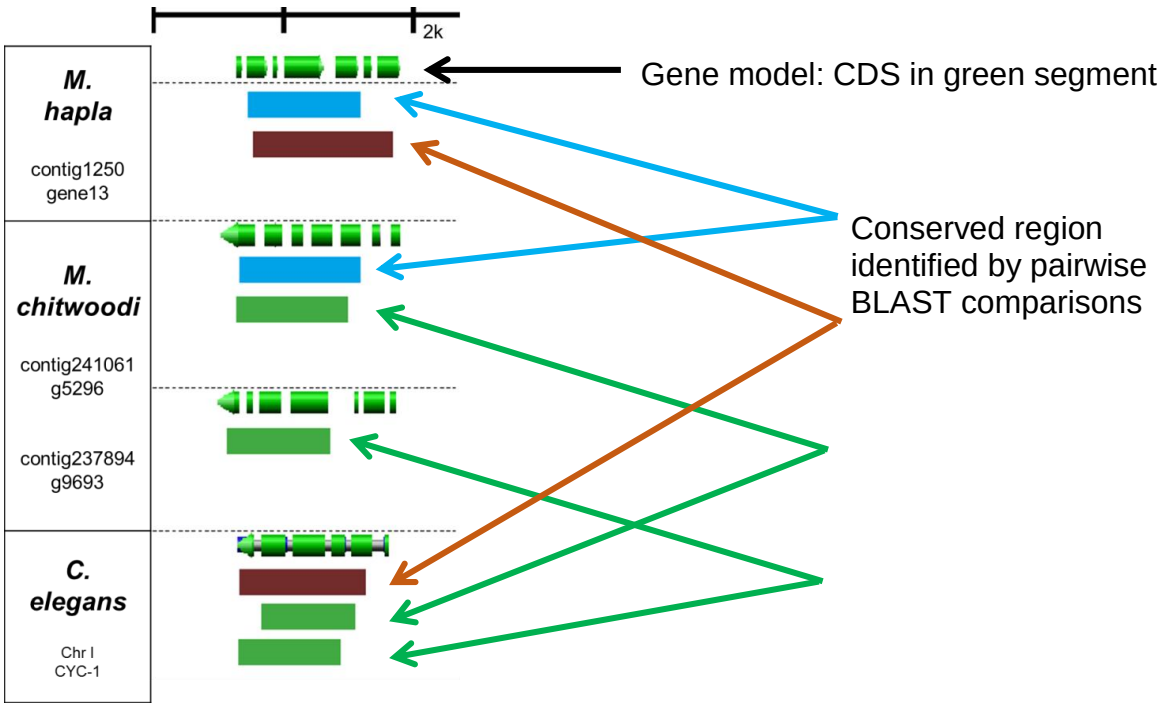


Figure 14. Sequence alignments of protein, KOG3052 (cytochrome c1), with one gene of *M. hapla*, two genes of *M. chitwoodi* and one gene of *C. elegans*.

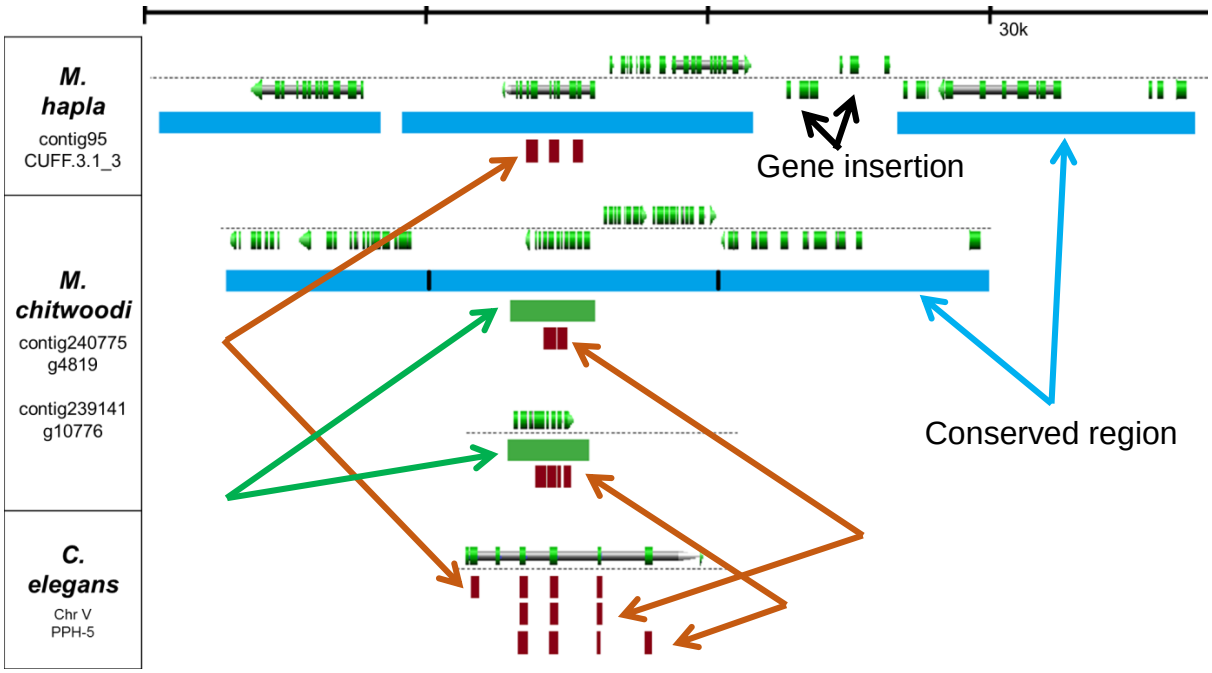


Figure 15. Sequence alignments of protein, KOG0376 (serine-threonine phosphatase 2A, catalytic subunit), with one gene of *M. hapla*, two genes of *M. chitwoodi*, and one gene of *C. elegans*. This region is conserved between *M. hapla* and *M. chitwoodi* except for two gene insertions in *M. hapla*.

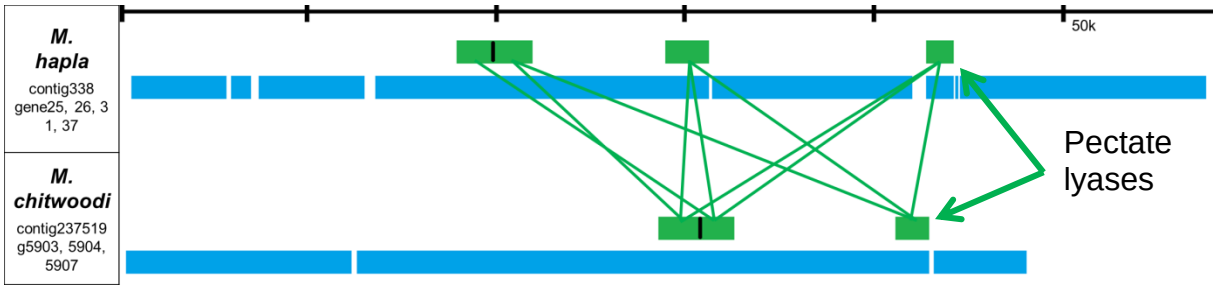


Figure 16-1 (A). Location of pectate lyases within the region conserved between *M. hapla* and *M. chitwoodi.*

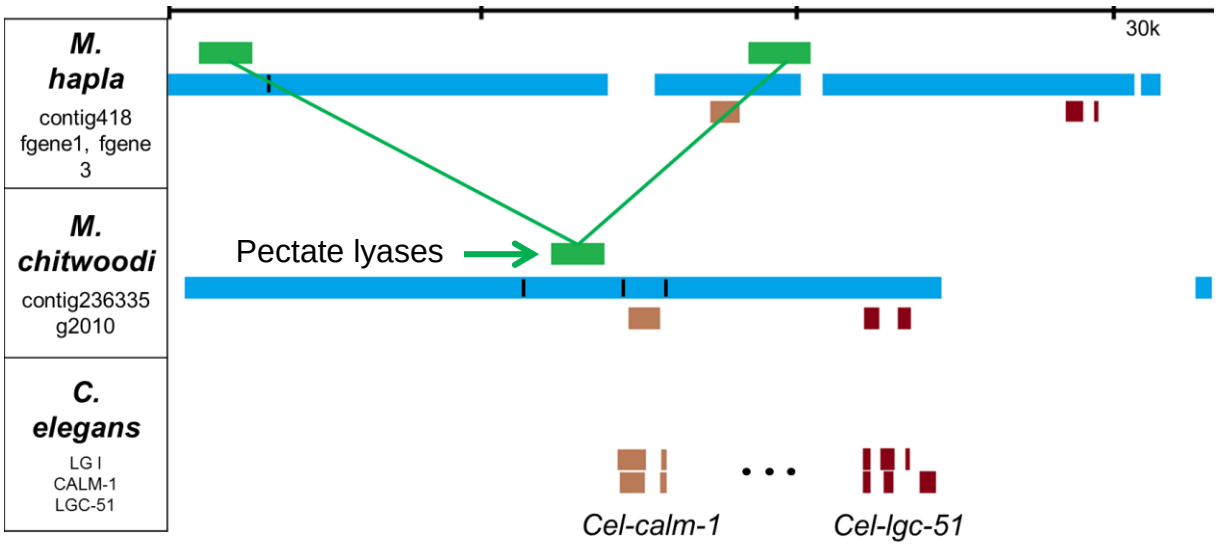


Figure 16-1 (B). The pectate lyases within region conserved between *M. hapla* and *M. chitwoodi*. The syntenic region of *M. hapla* contig 418 and *M. chitwoodi* contig 236335 was also aligned by *calm-1* (calcium and integrin binding protein) and *lgc-51* (ligand-gated ion channel), two of which were 7,225 Kb apart on LG. I of *C. elegans*. Also, this synteny of *M. hapla* and *M. chitwoodi* was predicted to contain the genes of LBP/BPI/CETP family (lipopolysaccharide-binding protein/bactericidal permeability-increasing protein/ cholesteryl ester transfer protein), which is known to function in the innate immune response (Bingle and Craven 2004).

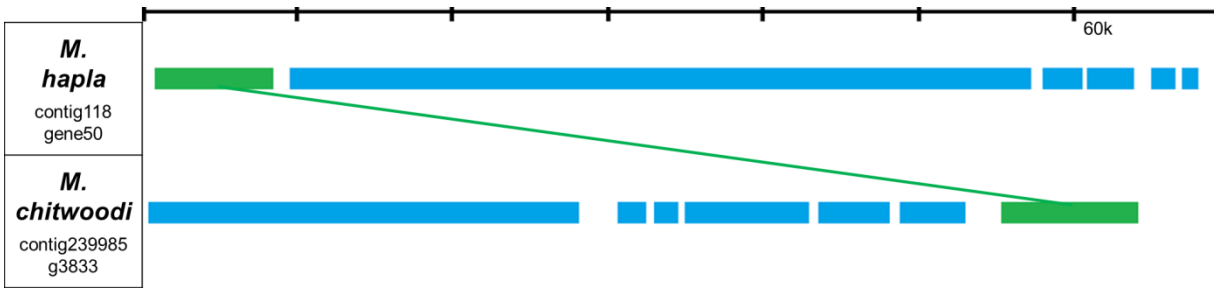


Figure 16-1 (C). The pectate lyases within region conserved between *M. hapla* and *M. chitwoodi*.

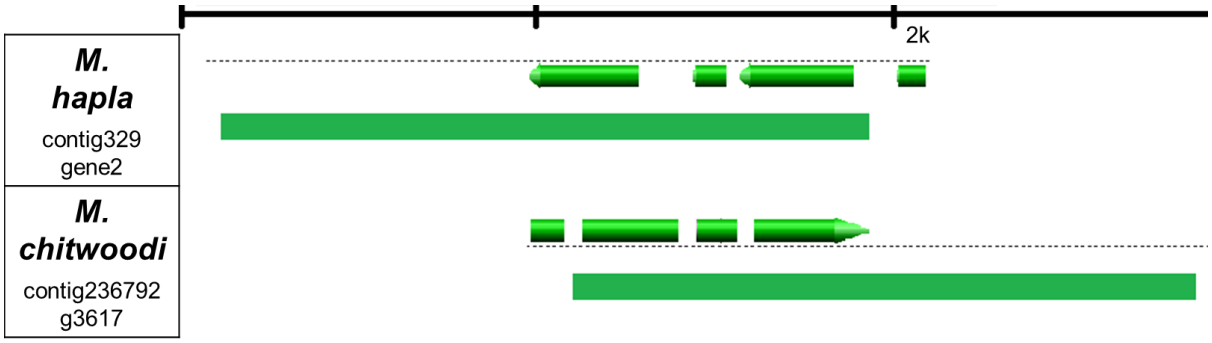


Figure 16-1 (D). Alignment of pectate lyases of *M. hapla* and *M. chitwoodi*.

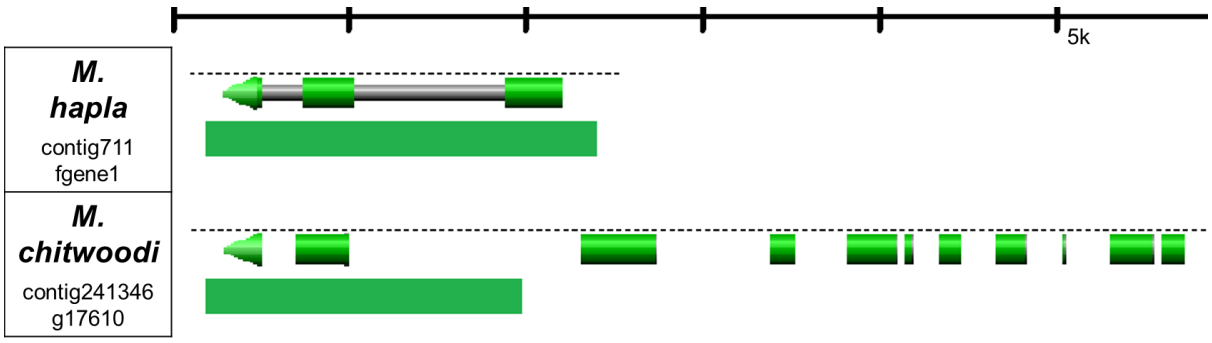


Figure 16-1 (E). Alignment of pectate lyases of *M. hapla* and *M. chitwoodi*.

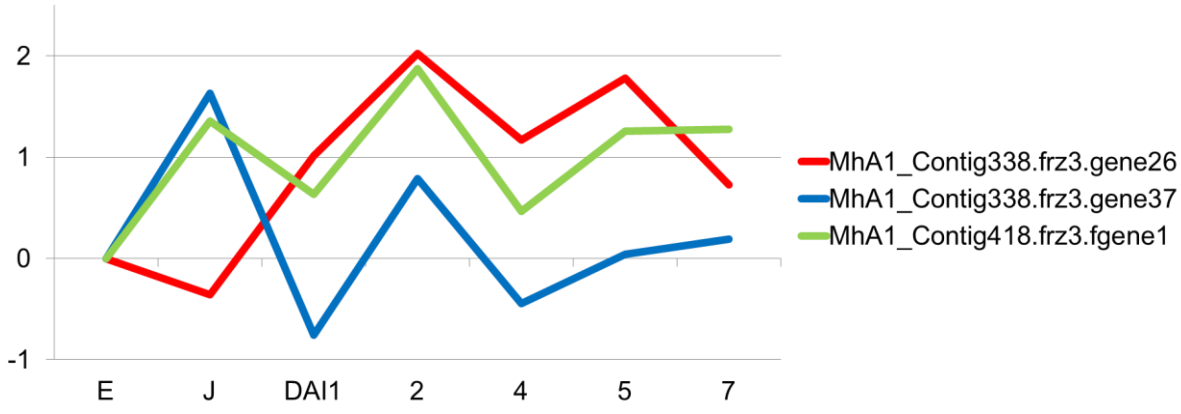


Figure 16-2. Differential expression pattern of three *M. hapla* pectate lyases, implying the distinct functions of the paralogous genes.

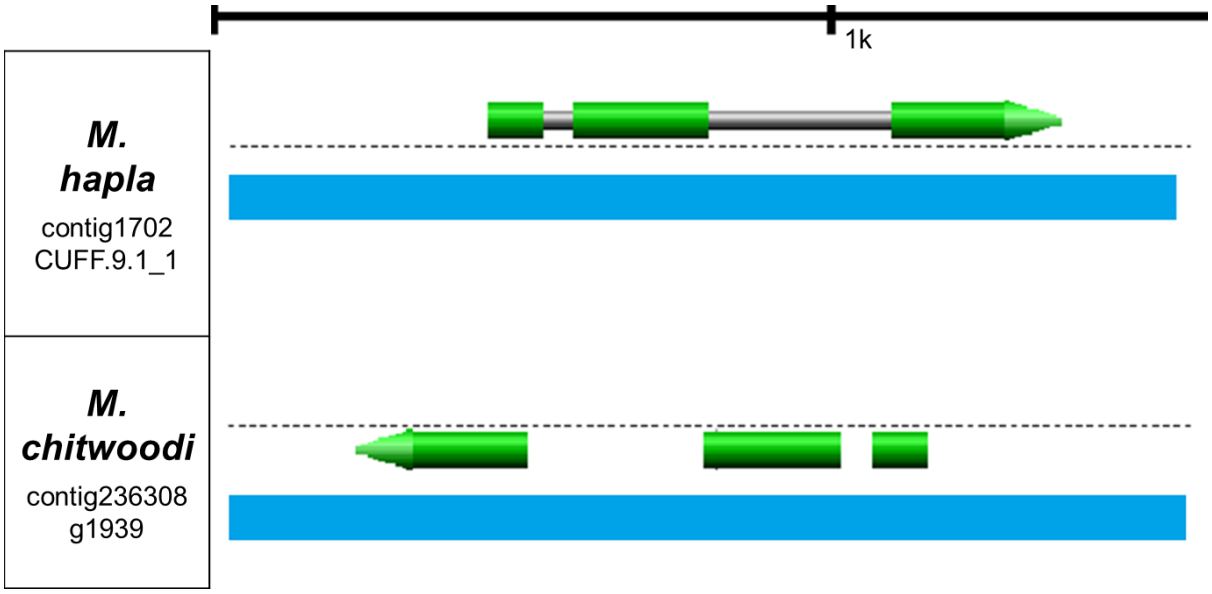


Figure 17-1. Alignment of chorismate mutase sequences of *M. hapla* and *M. chitwoodi*.

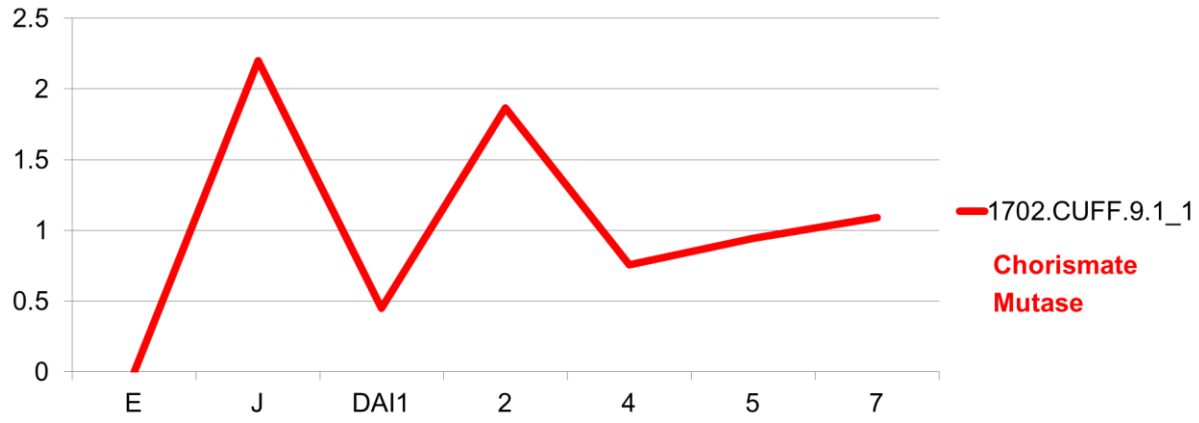


Figure 17-2. Large increase in the expression of *M. hapla* chorismate mutase from egg stage to J2.

CHAPTER 5

CONCLUSION

Summary. The different chapters embodied in this dissertation presented *de novo* assembly of a whole genome, differential transcriptome analysis, and comparative genomics, as assessed by exploiting the bioinformatics tools with the aim of elucidating the RKN-plant interactions. The results presented herein should form the basis of future studies in further understanding of such interactions as well as should aid in devising better strategies for combat against the pathogenicity of RKNs.

In the second chapter, the optimization of *de novo* assembly of a whole *M. chitwoodi* genome using a variant simple grid search was demonstrated. By comparing different assembler software, with combinations of key parameters, it was shown that the contiguity of the assembled genome was quite sensitive to specifications. The optimized assembly was further probed by mapping the core eukaryotic proteins conserved across a wide range of taxa, detecting on average two orghologs in the *M. chitwoodi* genome and revealing the possibility of genome duplication. Furthermore, as *M. chitwoodi* and *M. hapla* appear to occupy the same niche, they presumably share similar complements. To this end, I was motivated to compare the RKN genomes for verification, as documented in the last chapter. As implied above, obtaining a reference genome is the first step in figuring out not only the entire gene expression pattern of an organism but also for tracking the evolutionary history of different organisms.

Compared to other RKN species, *M. hapla* has been relatively well studied over many years since its genome was first assembled in 2008 (Opperman et al. 2008). For example, horizontal gene transfer on *M. hapla* genome was evidenced through multi-loci phylogenetic analysis (Scholl et al. 2003). Moreover, the C-terminally Encoded Peptide (CEP)-like mimicry was demonstrated to be encoded by *M. hapla*. Furthermore, the genome annotation of *M. hapla* was recently updated (Guo et al. 2014). Based on these established findings, in the third chapter, I investigated transcriptional expression patterns in RKN and host plant throughout their parasitic cycles and along their diurnal rhythm, respectively. Our analysis scheme was designed in such a way that it could still divulge the inherent biology. For example, the developmental stages were tied together to represent either control or treatment groups, namely pre-feeding site formation (egg, J2, and 1 DAI stages) vs. post-feeding site formation (2, 4, 5, and 7 DAI stages) for testing the differential expression in *M. hapla,* and infected root (1, 2, 4, 5, and 7 DAI stages) vs. uninfected root (1, 2, 4, 5, and 7 DAI stages) for testing the differential expression in *M. truncatula*. The DEGs identified in these tests were further binned into one of the expression pattern groups that were generated. The underlying concept of the different pattern groups was discretization of the continuous expression levels into one of the three directions, significant increase (↗), significant decrease (↘), and non-significance (—), between any two time points, by which insignificant noisy expressions could be dampened. More importantly, expressions with non-significant fold changes between consecutive time points but significant fold changes between non-consecutive time points allowed the pattern groups to include all the possible expression directions even though the actual gene expression would show either gradual

increases or decreases. Finally, the DEGs labeled with a specific expression pattern group were further compared with the results from diurnal experiments. In the diurnal data analysis, I evaluated the night vs. day expressions at whole, central, or time points. Collectively, all *M. hapla* genes differentially expressed in both the longitudinal and diurnal experiments exhibited (gradual) increase from that in egg to that in juveniles a week after inoculation; activation was especially observed at central night. Antithetically, a majority of *M. truncatula* DEGs in the two experiments showed (gradual) decrease upon initiation up to a week after infection. Interestingly, both RKN and the plant showed oscillations in the expression having convex or concave curves. Alhough some *M. hapla* DEGs were interpreted based on functional predictions by InterProScan or BLAST, many genes were recognized as uncharacterized hypothetical proteins, or "pioneers". Thus, it might need more thorough biological examination to characterize their precise functions in parasitism.

To extend the findings described above, the genomes of two RKNs, *M. hapla* and *M. chitwoodi*, were compared along with other nematode species, *M. incognita* and *C. elegans*. The featured syntenies presented in the fourth chapter were exemplars, which were typically observed while navigating the multiple genome alignments on MAUVE and CoGe software. In most cases, it was noticed that the sequences of *M. hapla* contig were highly conserved with that of *M. chitwoodi* contig, though sometimes they were in reversed orientation. Also, the synteny was broken by several gene insertions or deletions. This conserved region was often partially conserved with *M. incognita* contig. This implied that there were more synteny breaks on speciation to *M. incognita*. Yet, it could be observed that the contigs of all the three *Meloidogyne* species were conserved each other at least partially. On the other hand,

more than two neighboring genes in the same order were hardly identified between RKNs and *C. elegans*. Rather, only a single gene of *C. elegans* was likely to be mapped onto the contigs, with regions that were still conserved with the RKN species. This was expected because these observations were consistent with genetic distances in the phylogenetic tree. Moreover, while browsing the synteny, it was found that two or three separate nearby genes of *M. hapla* were aligned together to one *C. elegans* gene in the left, middle, or right part. When these *M. hapla* genes were searched for their expressions in longitudinal and diurnal experiments, they showed either similar or opposite expression patterns. Here, several possible interpretations could be postulated as follows: (1) those *M. hapla* orthologs that were annotated as different genes should actually be a single gene, (2) they might be iso-forms derived by alternative splicing, and (3) other genes inserted in between the *M. hapla* orthologs could explain the compact genome of *M. hapla*, which was achieved through evolution. As described, it could be informative to incorporate the transcriptome profiles of *M. hapla* with the scrutiny of syntenic regions. This could enhance or even rectify the existing annotations of the *M. hapla* genome. Alternatively, if there exists a missed annotation on *M. chitwoodi* genome, it could be amended by comparing the orthologs within the conserved region between the two RKNs because they are expected to share similar gene model inherited from the common ancestral synteny. Moreover, we could assume the circumstance under which the functions of *M. chitwoodi* genes should be predicted for a specific stage of lifecycle. If the *M. chitwoodi* gene of interest is positioned within the synteny, its function might be inferred from the orthologous *M. hapla* gene, which accompanies spatio-temporal transcriptome profiles. Moreover, it was confirmed in the

synteny between *M. hapla* and *M. chitwoodi* that pectate lyases probably acquired by horizontal gene transfer from bacteria or fungi were expanded through gene duplications on the same contig, several times. Overall, assessing the commonalities in the results between seemingly different studies of *M. hapla* and *M. chitwoodi*, it was proposed that complementation of the results could reinforce the available information of the RKN in a counterpart species. As a final remark, I would mention that the biology of *M. hapla* and *M. chitwoodi*, including their evolutionary adaptations for plant parasitism, sympatric lineages of RKN, and parasitic gene expressions at specific times, is reflected in their genomes.

Web-Resources. Owing to a cascade of enormous genomic data, readily available online database have increasingly been in demand. Particularly, with the overflowing numbers resulting from complicated experimental designs and overwhelming statistical data analysis, it is essential to extract and integrate crucial information only according to the genomic features of interest. Motivated by this, we constructed databases encompassing all the research results presented in the previous chapters. The databases are available at the following URLs: https://brcwebportal.cos.ncsu.edu/haplatome/1-0_chitwoodi.php for *M. chitwoodi*, https://brcwebportal.cos.ncsu.edu/haplatome/1-0_hapla.php for *M. hapla*, and https://brcwebportal.cos.ncsu.edu/haplatome/1-0_medicago.php for *M. truncatula*.

For example, if users would like to search for how many genes encoding chorismate mutase are predicted to exist in the *M. chitwoodi* genome, where they are positioned, and what their sequences are, inserting a keyword "chorismate mutase" in the search box would query the *M. chitwoodi* database and display a succinct table of the compiled results (Figure 1). Similarly, by searching for the same keyword in the *M. hapla* database, which contains

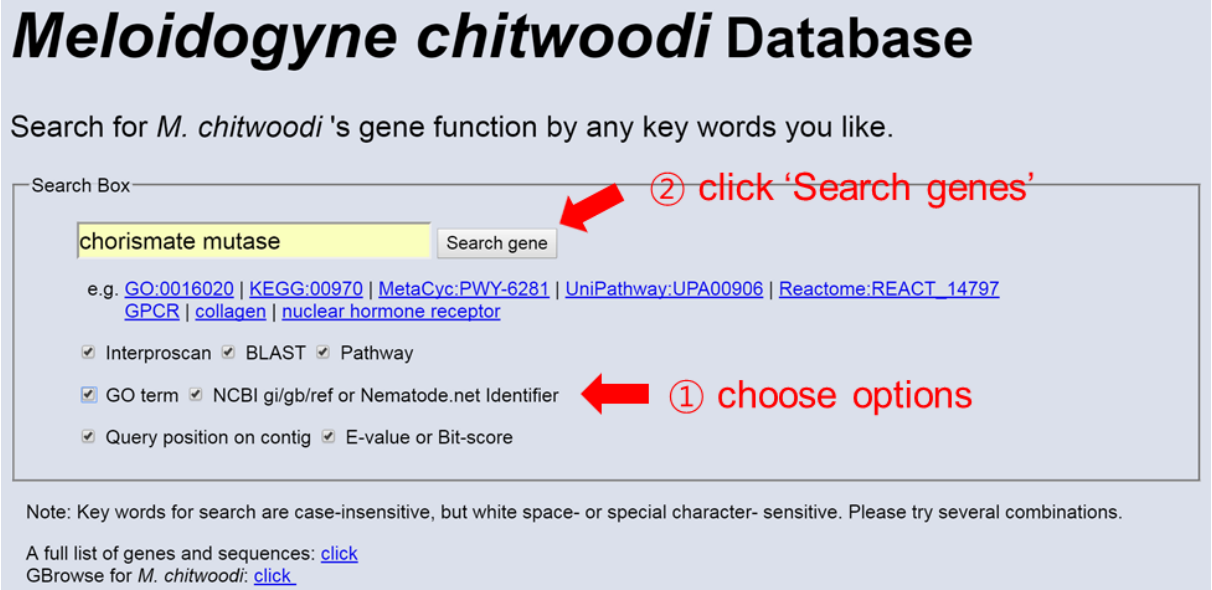
transcriptome expression levels throughout the developmental and diurnal stages, the temporal and spatial functions of *M. chitwoodi* genes could be inferred from cross-species comparisons (Figure 2). Also, based on the extracted DNA sequences of the same gene from *M. hapla* and *M. chitwoodi*, their amino acids sequences could be easily compared *via* the NCBI tBLASTx tool.

Furthermore, in the *M. truncatula* database, the *A. thaliana* genes could be queried to identify candidate orthologs of *M. truncatula* genes. For example, if AT5G54680 and AT3G47640 known to contain basic helix-loop-helix domain are searched as keywords, then it shows four and one *M. truncatula* gene orthologs, respectively. Of these resulting genes, Medtr2g089000 was identified to be a DEG in tests comparing root vs. shoot of infected plants, root vs. shoot of uninfected plants, and night vs. day conditions (Figure 3). As another example, AT2G22540, an AGAMOUS-LIKE22 gene, was found to correspond to eight candidate orthologs of *M. truncatula*. This extended applicability for cross-species comparisons based on differential transcriptome expressions under different spatial and temporal conditions could give deeper insights into the gene functions.

Future Directions. As long as even one more piece of information could be added, annotation work would not be considered completed. Rather, research should be ongoing for accurate and comprehensive understanding of gene models. Based on our extensive studies accumulated over years in our lab and those of our collaborating groups, exhaustive annotation work on the *M. chitwoodi* genome will have to be performed. This will further improve our web databases by equipping them with inter-connected information among the

RKN species. Also, our database will be upgraded to provide more accessibility, effective navigations, and insightful visualization in the near future.

Moreover, accelerated by recent acquisition of assembled genome of *M. chitwoodi*, comparative genomics of RKN species will spur discovery of novel aspects of plant–pathogen parasitism, which could enhance crop resistance to PPNs and achieve creative methods for nematode control.

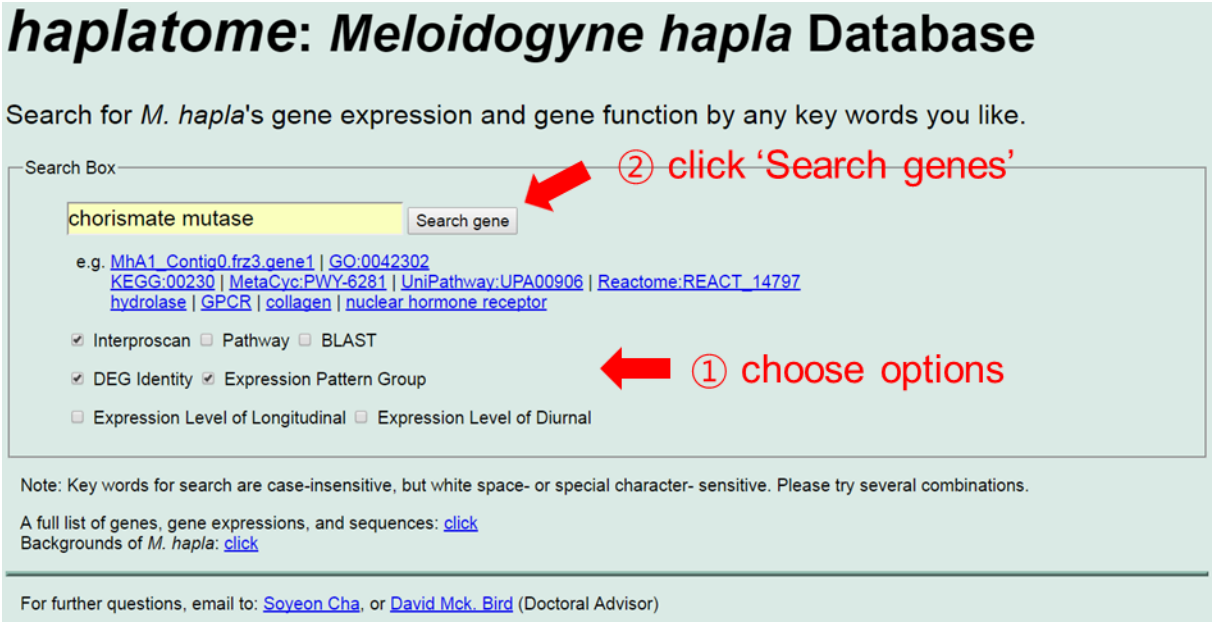


Seq	CONTIG_ID	IPRS_DB	MEMBERS_DSCR	ENTRY_DSCR	BLAST_DSCR	PATHWAY	GO_TERMS	GENE_ID	QPOS_START	QPOS_END	IPRS_EVAL	BLAST_SCORE
seq	a_236308	ProSiteProfiles	Chorismate mutase domain profile.	Chorismate mutase	putative nuclear encoded protein Method: similarity and extension	KEGG:00400+5.4.99.5 | MetaCyc:PWY-3461 | MetaCyc:PWY-3462 | MetaCyc:PWY-6120 | MetaCyc:PWY-6627 | MetaCyc:PWY-7626 | UniPathway:UPA00120	GO:0046417	MC00308	1607	1389	9.239	eval 6e-49 bitscore 145
seq	a_236308	SUPERFAMILY	.	Chorismate mutase, type II	putative nuclear encoded protein Method: similarity and extension	KEGG:00400+5.4.99.5 | MetaCyc:PWY-3461 | MetaCyc:PWY-3462 | MetaCyc:PWY-6120 | MetaCyc:PWY-6627 | MetaCyc:PWY-7626 | UniPathway:UPA00120	GO:0046417	MC00308	1607	1389	2.93E-12	eval 6e-49 bitscore 145
seq	a_238000	Gene3D	.	Endoribonuclease L-PSP/chorismate mutase-like	putative nuclear encoded protein Method: similarity and extension	.	.	MC00503	1359	1030	7.00E-09	eval 1e-75 bitscore 181
seq	a_238000	SUPERFAMILY	.	Endoribonuclease L-PSP/chorismate mutase-like	putative nuclear encoded protein Method: similarity and extension	.	.	MC00503	1359	1030	7.23E-08	eval 1e-75 bitscore 181
seq	a_238000	SUPERFAMILY	.	Endoribonuclease L-PSP/chorismate mutase-like	putative nuclear encoded protein Method: similarity and extension	.	.	MC00503	985	671	1.26E-09	eval 1e-75 bitscore 120
seq	a_238000	Gene3D	.	Endoribonuclease L-PSP/chorismate mutase-like	putative nuclear encoded protein Method: similarity and extension	.	.	MC00503	985	671	9.80E-11	eval 1e-75 bitscore 120

③ click 'seq'

>a_238000 | 16605...16874, target MC05198, putative nuclear encoded protein Method: similarity and extension, eval 8e-14, bitscore 73.2
GTTCGCTATCGTGACCGTGTAACAATACTCAGAGGGAATCATGAGTCTCGTCAAATCACACAAGTTTATGGATTTTATGATGAATGCTTGCGT
>a_238000 | 41843...41688, target MC00546, putative nuclear encoded protein Method: similarity and extension, eval 7e-14, bitscore 72.8
TTGTAGTGCAATCAAGGCAAGCCCCTTCCATTTAAAATTTTTCCCATCTGTGACATTGTTAGCTAATTCTCTGACCAAACGTATGAAAGAACC
>a_238000 | 41695...41522, target MC00546, putative nuclear encoded protein Method: similarity and extension, eval 2e-10, bitscore 62.0
AACCTCTCTTCTCAATTTCCTCAATAAATGAACATCTTTTGGCATAAGCGTAACACGTCCAGTATGGAGAGCAAGCAAATTAGCACCCTCAA

Figure 1. An example for navigation of *M. chitwoodi* database. On the first page, ① choose options, ② put the keyword of interest, and click "search genes". The results will be shown on the following page: contig on which the searched gene exists, InterProScan predictions (databases and gene function descriptions), BLAST search descriptions, KEGG pathways, GO terms, gene ID of nematode.net, gene positions, and BLAST scores. To see the corresponding sequences, ③ click "seq" in the first column.



Seq	GENE_ID	LONG1_DEG	LONG2_DEGcount	DI1_DEG	DI2_DEG	DI3_DEG	SUM_DEGcount	G1	IPRS_DB	MEMBERS_DSCR	ENTRY_DSCR	PATHWAY
seq	1702.CUFF.9.1_1	0	2	0	0	0	2	u_____	ProSiteProfiles , SUPERFAMILY	Chorismate mutase domain profile.	Chorismate mutase , Chorismate mutase , type II	KEGG:00400+5.4.99.5 | MetaCyc:PWY-3461 | MetaCyc:PWY-3462 | MetaCyc:PWY-6120 | MetaCyc:PWY-6627 | MetaCyc:PWY-7626 | UniPathway:UPA00120
seq	MhA1_Contig740.frz3.gene23	1	2	0	0	0	3	_uooo_	ProSiteProfiles , SUPERFAMILY	Prokaryotic membrane lipoprotein lipid attachment site profile.	Chorismate mutase , type II	KEGG:00400+5.4.99.5 | MetaCyc:PWY-3461 | MetaCyc:PWY-3462 | MetaCyc:PWY-6120 | MetaCyc:PWY-6627 | MetaCyc:PWY-7626 | UniPathway:UPA00120

>MhA1_Contig740.frz3.gene23 | Mh:0740:MhA1_Contig740 | CDS | 62045...62159 | +
ATGAAAATTAAAGCTTCTCCTAACTTTTATTTATTTTTATTTCTTGCTTCAATTATTATATCTTGTGTGAAAGGCGATGAGGATAAAGAA
>MhA1_Contig740.frz3.gene23 | Mh:0740:MhA1_Contig740 | CDS | 62403...62620 | +
AAAACCTCGAGAATCTTATCAATTTGGTCGATGCTCGTTTGGATGTCATTGAAAAAGTGACAATGTATAAGTACCTGGAAGGATTACC

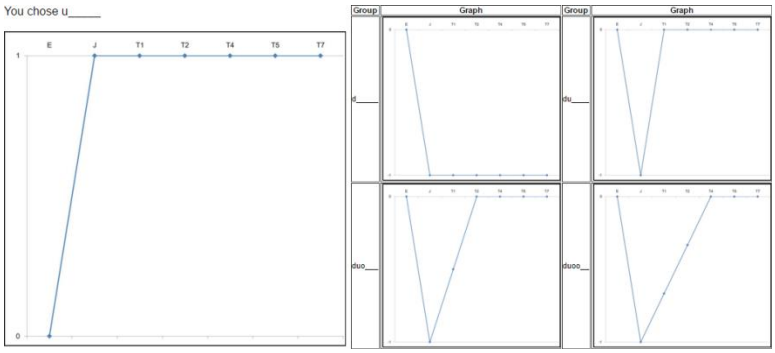
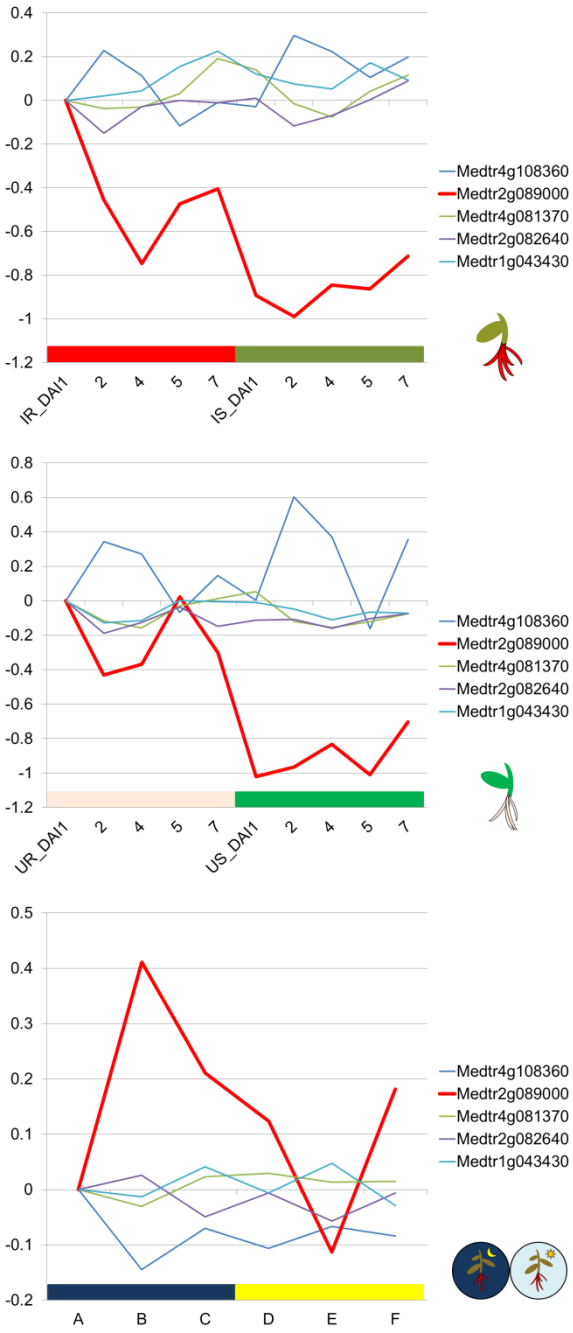


Figure 2. An example for navigation of *M. hapla* database. Steps from ① to ③ are similar to those described in Figure 1. In *M. hapla* database, more features are added, including the time of being identified as DEG in differential analysis of longitudinal and diurnal experiments. The expression group to which a DEG is classified is displayed as graphs (④).



Infected plants

Root vs. Shoot

Medtr2g089000

down-regulated in shoot

Uninfected plants

Root vs. Shoot

Medtr2g089000

down-regulated in shoot

Infected plants

Night vs. Day

Medtr2g089000

down-regulated during a day

Figure 3. Reads Per Kilobase of transcript per Million mapped reads (RPKM) of orthologous candidates of *M. truncatula* and *A. thaliana*. AT5G54680 was found to be likely orthologous to Medtr1g043430, Medtr2g089000, Medtr2g082640, and Medtr4g081370 whereas AT3G47640 might possibly be related to Medtr4g108360. Of these, Medtr2g089000 was uniquely identified as a DEG in our longitudinal and diurnal study.

REFERENCES

Abad P et al. Nat Biotechnol. 2008 26(8):909.

Anders S, Pyl PT, Huber W. 2015 Bioinformatics. 31(2):166-9.

Bar-Joseph Z, Gitter A, Simon I. 2012. Nat Rev Genet. 18;13(8):552-64.

Bartlem DG, Jones MGK, Hammes UZ. 2014 J Exp Bot. 65(7):1789-1798.

Benjamini Y, Hochberg Y. 1995. J R Stat Soc. 57:289–300.

Bhattarai KK, Xie QG, Mantelin S, Bishnoi U, Girke T, Navarre DA, Kaloshian I. 2008 Mol Plant Microbe Interact. 21(9):1205-14.

Bingle CD, Craven CJ. 2004 Trends Immunol. 25(2):53-5.

Bird AF. 1968 Int J Parasitol. 13:343-348.

Bird DM, Jones JT, Opperman CH, Kikuchi T, Danchin E. 2013 Parasitology, pp.1-14. <10.1017/S0031182013002163>. <hal-01137177>.

Bird DM, Kaloshian I. 2002 Physiological and Molecular Plant Pathology. 62:115–123.

Bird DM, Koltai H. 2000 J Plant Growth Regul. 19(2):183-194.

Bird DM, Williamson VM, Abad P, McCarter J, Danchin EG, Castagnone-Sereno P, Opperman CH. 2009. Annu Rev Phytopathol. 47:333-51.

Bird DM. 1996 J Parasitol. 82(6):881-8.

Bird DM. 2004 Curr Opin Plant Biol. 7(4):372-6.

Blaxter M, Koutsovoulos G. 2015 Parasitology. 142 Suppl 1:S26-39.

Blaxter M. 1998 Science. 282(5396):2041-6.

Bobay BG, DiGennaro P, Scholl E, Imin N, Djordjevic MA, Bird DM. 2013 FEBS Lett. 587(24):3979-85.

Brenner ED, Lambert KN, Kaloshian I, Williamson VM. 1998 Plant Physiol. 118(1):237-47.

Brown CR, Yang CP, Mojtahedi H, Santo GS, Masuelli R. 1996 Theor Appl Genet. 92(5):572-6.

Burow MD, Simpson CE, Paterson AH, Starr JL. 1996 Molecular Breeding. 2: 369-319.

Byrd DW, Nusbaum CJ, Barker KR. 1996 Plant Dis. Rep. 50:954–957.

C. elegans Sequencing Consortium. 1998 Science. 282(5396):2012-8.

Caillaud MC, Dubreuil G, Quentin M, Perfus-Barbeoch L, Lecomte P, de Almeida Engler J, Abad P, Rosso MN, Favery B. 2008 J Plant Physiol. 165(1):104-13.

Castagnone-Sereno P. 2006 Heredity (Edinb). 96(4):282-9.

Cha S, Bird DM. 2016 Bioinformation. 12(2): 36-40

Chen C, Liu S, Liu Q, Niu J, Liu P, Zhao J, Jian H. 2015 PLoS One. 10(4):e0122256.

Chikhi R & Medvedev P. 2014 Bioinformatics. 30(1):31.

Chitwood DJ. 2002 Annu Rev Phytopathol. 40:221-49.

Chitwood DJ. 2003 Pest Manag Sci. 59:748–753.

Conesa A et al. Bioinformatics 2005 21(18):3674.

Dalzell JJ, McVeigh P, Warnock ND, Mitreva M, Bird DM, Abad P, Fleming CC, Day TA, Mousley A, Marks NJ, Maule AG. 2011. PLoS Negl Trop Dis. 5(6):e1176.

Darling AC, Mau B, Blattner FR, Perna NT. 2004 Genome Res. 14(7):1394-403.

Decraemer W, Hunt D. 2006 Structure and classification. Pp. 3–32 in R. N. Perry and M. Moens, eds. Plant nematology. Wallingford, UK: CAB International.

Denancé N, Sánchez-Vallet A, Goffner D, Molina A. 2013 Front Plant Sci. 4:155.

Elling AA. Phytopathology. 2013. 103(11):1092-1102.

Ferris H. 2010 J Nematol. 42(1):63-7.

Freckman DW. 1988 Agric Ecosys Envir. 24:195-217.

Gao B, Allen R, Maier T, Davis EL, Baum TJ, Hussey RS. 2003. Mol Plant Microbe Interact. 16(8):720-6.

Gao X, Starr J, Gobel C, Engelberth J, Feussner I, Tumlinson J, Kolomiets M. 2008 Mol Plant Microbe Interact. 21(1):98-109.

Gems D. 2000 Biogerontology 1:289–307.

Gheysen G, Fenoll C. 2002 Annu Rev Phytopathol. 40:191-219.

Giebel J. 1974 J Nematol. 6(4):175-84.

Glazebrook J. 2005 Annu Rev Phytopathol. 43:205-27.

Goverse A, Rouppe van der Voort J, Roupe van der Voort C, Kavelaars A, Smant G, et al. 1999 Mol Plant-Microbe Interact. 12:872–81.

Greenham K, McClung CR. 2015. Nat Rev Genet. 16(10):598-610.

Guiliano DB, Hall N, Jones SJ, Clark LN, Corton CH, Barrell BG, Blaxter ML. 2002 Genome Biol. 3(10):RESEARCH0057.

Guiran G. 1979 Revue Nénzatol. 2(2):223-231.

Gundy SDV. 1965 Annual Review of Phytopathology. 3:43–68.

Guo Y et al. Worm 2014 3:29158.

Haegeman A, Jones JT, Danchin EG. 2011 Mol Plant Microbe Interact. 24(8):879-87.

Hewezi T, Baum TJ. 2013 Mol Plant Microbe Interact. 26(1):9-16.

Ho JY, Weide R, Ma HM, van Wordragen MF, Lambert KN, Koornneef M, Zabel P, Williamson VM. 1992 Plant J. 2(6):971-82.

Hoff KJ & Stanke M, Nucleic Acids Res. 2013 41(Web Server issue):123.

Huang G, Dong R, Allen R, Davis EL, Baum TJ, Hussey RS. 2006 Mol Plant-Microbe Interact. 19:463–70.

Hugot JP, Baujard P, Morand S. 2001 Nematology. 3:199–208.

Hussey RS, Davis EL, Baum TJ. 2002. Braz. J. Plant Physiol. 14:183–194.

Hussey RS, Mims CW. 1990 Protoplasma. 156:9 18.

Imin N, Mohd-Radzman NA, Ogilvie HA, Djordjevic MA. 2013 J Exp Bot. 64(17):5395-409.

Jaouannet M, Magliano M, Arguel MJ, Gourgues M, Evangelisti E, Abad P, Rosso MN. 2013 Mol Plant Microbe Interact. 26(1):97-105.

Jasmer DP, Goverse A, Smant G. 2003 Annu Rev Phytopathol. 41:245-70.

Jones JDG and Dangl JL, 2006 Nature. 444:323-329

Jones JT, Furlanetto C, Bakker E, Banks B, Blok V, Chen Q, Phillips M, Prior A. 2003. Mol Plant Pathol. 1;4(1):43-50.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. 2014 Bioinformatics. 30(9):1236-40.

Karajeh M. R. 2008. J Plant Protection Res. 48(2):182-187.

Klein-Lankhorst R, Rietveld P, Machiels B, Verkerk R, Weide R, Gebhardt C, Koornneef M, Zabel P. 1991 Theor Appl Genet. 81(5):661-7.

Klumpp S, Hermesmeier J, Selke D, Baumeister R, Kellner R, Krieglstein J. 2002. J Cereb Blood Flow Metab. 22(12):1420-4.

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A. 2015. Nucleic Acids Res. 43(Database issue):D1113-6.

Kumari C, Dutta TK, Banakar P1, Rao U. 2016 Sci Rep. 6:22846. doi: 10.1038/srep22846.

Kunkel BN, Brooks DM. 2002 Curr Opin Plant Biol. 5(4):325-31.

Kyndt T, Denil S, Haegeman A, Trooskens G, Bauters L, Van Criekinge W, De Meyer T, Gheysen G. 2012 New Phytol. 196(3):887-900.

Kyndt T, Fernandez D, Gheysen G. 2014 Annu Rev Phytopathol. 52:135-53.

Lacey LA, Georgis R. 2012 J Nematol. 44(2):218-25.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G. 2011 Nucleic Acids Res. 39(Database issue):D28-31.

Li X, Schuler MA, Berenbaum MR. 2002. Nature. 419(6908):712-5.

Lohar DP, Schaff JE, Laskey JG, Kieber JJ, Bilyeu KD, Bird DM. 2004 Plant J. 38(2):203-14.

Lu SW, Chen S, Wang J, Yu H, Chronis D, Mitchum MG, Wang X. 2009 Mol Plant-Microbe Interact. 22:1128–42.

Lu ZX, Sosinski B, Reighard GL, Baird WV, Abbott AG. 1998 Genome. 41(2): 199-207.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S,

Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. 2012 Gigascience. 1(1):18.

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M. 2008 Plant Physiol. 148(4):1772-81.

Mai WF, Abawi GS. 1987 Annual Review of Phytopathology. 25:317-338.

Martin J et al. Nucleic Acids Res. 2015 Jan;43(Database issue):D698-706.

Mayer WE, Schuster LN, Bartelmes G, Dieterich C, Sommer RJ. 2011 BMC Evol Biol. 11:13.

McCord PH. 2012 J Nematol. 44(4):387-90.

Messeguer R, Ganal M, de Vicente MC, Young ND, Bolkan H, Tanksley SD. 1991 Theor Appl Genet. 82(5):529-36.

Mitchum MG, Hussey RS, Baum TJ, Wang X, Elling AA, Wubben M, Davis EL. 2013. New Phytol. 199(4):879-94.

Nahar K, Kyndt T, De Vleesschauwer D, Höfte M, Gheysen G. 2011 Plant Physiol. 157(1):305-16.

Nahar K, Kyndt T, Hause B, Höfte M, Gheysen G. 2013 Mol Plant Microbe Interact. 26(1):106-15.

Nguyen PV, Bellafiore S, Petitot AS, Haidar R, Bak A, Abed A, Gantet P, Mezzalira I, de Almeida Engler J, Fernandez D. 2014 Rice (N Y). 7(1):23.

Niebel A, Gheysen G, Van Montagu M. 1994 Parasitol Today. 10(11):424-30.

Niu J, Liu P, Liu Q, Chen C, Guo Q, Yin J, Yang G, Jian H. 2016 Sci Rep. 6:19443.

Ogilvie HA, Imin N, Djordjevic MA. 2014 BMC Genomics 15:870.

Olsen AN, Skriver K. 2003 Trends Plant Sci. 8:55–57.

Opperman CH et al. Proc Natl Acad Sci USA. 2008 105(39):14802.

Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, Cromer J, Diener S, Gajan J, Graham S, Houfek TD, Liu Q, Mitros T, Schaff J, Schaffer R, Scholl E, Sosinski BR, Thomas VP, Windham E. 2008. Proc Natl Acad Sci U S A. 30;105(39):14802-7.

Parkinson J, Mitreva M, Hall N, Blaxter M, McCarter JP. 2003 Trends Parasitol. 19(7):283-6.

Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH, McCarter JP, Blaxter ML. 2004. Nat Genet. 36(12):1259-67.

Parra G et al. Bioinformatics 2007 23(9):1061.

Patel N, Hamamouch N, Li C, Hewezi T, Hussey RS, Baum TJ, Mitchum MG, Davis EL. 2010. J Exp Bot. 61(1):235-48.

Perry RN, Moens M. 2011. In: Genomics and Molecular Genetics of Plant-nematode Interactions. Jones JT, Gheysen G, Fenoll C, editor. Dordrecht, The Netherlands: Springer; Introduction to plant-parasitic nematodes: modes of parasitism; pp. 3–20.

Perry RN. 1989 Parasitol Today. 5(12):377-83.

Pinot F, Beisson F. 2011. FEBS J. 278(2):195-205.

Postma WJ, Slootweg EJ, Rehman S, Finkers-Tomczak A, Tytgat TO, van Gelderen K, Lozano-Torres JL, Roosien J, Pomp R, van Schaik C, Bakker J, Goverse A, Smant G. 2012 Plant Physiol. 160(2):944-54.

Poulin R, Morand S. 2000 Q Rev Biol. 75(3):277-93.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009 Plant Cell. 21(12):3718-31.

Qin L, Kudla U, Roze EH, Goverse A, Popeijus H, Nieuwland J, Overmars H, Jones JT, Schots A, Smant G, Bakker J, Helder J. 2004 Nature. 427(6969):30.

Quentin M, Abad P, Favery B. 2013 Front Plant Sci. 4:53.

Quevillon E et al. Nucleic Acids Res. 2005 33(Web Server issue):116.

Rehman S, Gupta VK, Goyal AK. 2016 BMC Microbiol. 16: 48.

Rehman S, Postma W, Tytgat T, Prins P, Qin L, Overmars H, Vossen J, Spiridon LN, Petrescu AJ, Goverse A, Bakker J, Smant G. 2009 Mol Plant Microbe Interact. 22(3):330-40.

Replogle A, Wang J, Bleckmann A, Hussey RS, Baum TJ, Sawa S, Davis EL, Wang X, Simon R, Mitchum MG. 2011 Plant J. 65:430–40.

Robinson MD, McCarthy DJ, Smyth GK. 2010. Bioinformatics. 26(1):139-40.

Sauerbrunn N, Schlaich NL. 2004. Planta. 218(4):552-61.

Scholl EH, Bird DM. 2005 Mol Phylogenet Evol. 36(3):536-45.

Scholl EH, Bird DM. 2011 BMC Biol. 9:9.

Scholl EH, Thorne JL, McCarter JP, Bird DM. 2003 Genome Biol. 4(6):R39. Epub.

Shannon AJ, Browne JA, Boyd J, Fitzpatrick DA, Burnell AM. 2005 J Exp Biol. 208(Pt 12):2433-45.

Slaymaker DH, Navarre DA, Clark D, del Pozo O, Martin GB, Klessig DF. 2002. Proc Natl Acad Sci U S A. 99(18):11640-5.

Smant G, Stokkermans JP, Yan Y, de Boer JM, Baum TJ, Wang X, Hussey RS, Gommers FJ, Henrissat B, Davis EL, Helder J, Schots A, Bakker J. 1998 Proc Natl Acad Sci U S A. 95(9):4906-11.

Smit A et al. RepeatMasker at http://repeatmasker.org.

Smith PG. 1944 Proc. Am. Soc. Hort. Sci. 44: 413-416.

Sommer RJ, Streit A. 2011 Annu Rev Genet. 45:1-20.

Stanke M et al. Nucleic Acids Res. 2004 32(Web Server issue):309.

Stein LD et al. Genome Res. 2002 12(10):1599-610.

Sun Y1, Liu G, Li Z, Chen Y, Liu Y, Liu B, Su Z. 2013. Immunology. 138(4):370-81.

Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter M, Lunt DH. 2015 bioRxiv, 034884.

Thomma BP, Nürnberger T, Joosten MH. 2011. Plant Cell. 23(1):4-15.

Thurau T, Ye W, Menkhaus J, Knecht K, Tang G, Cai D. 2010 Sugar Tech. 12(3-4):229-237.

Torsten Seemann. Velvet Advisor at http://dna.med.monash.edu.au/~torsten/velvet_advisor/, http://www.vicbioinformatics.com/velvetk.pl.

Torto-Alalibo T, Collmer CW, Lindeberg M, Bird D, Collmer A, Tyler BM. 2009. BMC Microbiol. 9 Suppl 1:S3.

Trapnell C, Pachter L, Salzberg SL. 2009 Bioinformatics. 25(9):1105-11.

van der Linden AM, Beverly M, Kadener S, Rodriguez J, Wasserman S, Rosbash M, Sengupta P. 2010. PLoS Biol. 8(10):e1000503.

Wang X, Allen R, Ding X, Goellner M, Maier T, de Boer JM, Baum TJ, Hussey RS, Davis EL. 2001 Mol Plant Microbe Interact. 14(4):536-44.

Wang W, Barnaby JY, Tada Y, Li H, Tör M, Caldelari D, Lee DU, Fu XD, Dong X. 2011 Nature. 470(7332):110-4.

Wasserstein RL, Lazara NA. 2016 Am Stat. DOI:10.1080/00031305.2016.1154108.

Weerasinghe RR, Bird DM, Allen NS. 2005 Proc Natl Acad Sci U S A. 102(8):3147-52.

Williamson VM, Gleason CA. 2003 Current Opinion in Plant Biology. 6:1–7.

Williamson VM, Hussey RS. 1996 Plant Cell. 8(10):1735-45.

Williamson VM, Kumar A. 2006 Trends Genet. 22(7):396-403.

Williamson. 1999 Currenft Opinion in Plant Biology. 2:327–331.

Wubben MJ, Jin J, Baum TJ. 2008 Mol Plant Microbe Interact. 21(4):424-32.

Zerbino DR & Birney E, Genome Res. 2008 18(5):821.

Zhang ZQ. 2013 Zootaxa. 3703:1-82.

APPENDICES

Appendix A

Group	InterProScan Functions
D1∩D2∩D3	(-) PAN/Apple domain, Antifreeze protein, Nematode polyprotein allergen, Laminin-type EGF, Intermediate filament protein, Collagen triple helix repeat, Myosin, Nematode cuticle **collagen**, Nuclease-related domain, **Thioredoxin**-like fold, **Immunoglobulin**, Alpha-2-macroglobulin, Cysteine-rich Golgi apparatus protein, Immunoglobulin-like fold, E3 ubiquitin ligase, Protein kinase, Clathrin, WD40 repeat, Protein kinase, **Low-density lipoprotein receptor** (+) Nuclear hormone receptor, Peptidase
(D2∩D3)\D	(+) **Glutathione S-transferase**, C-type lectin domain
(D1∩D2)\D	(-) **Low-density lipoprotein (LDL) receptor**, Cys/Met metabolism pyridoxal phosphate-dependent enzyme, Dihydroprymidine dehydrogenase, **Adrenodoxin reductase**, Malic enzyme, cuticle **collagen**, Protein kinase, ATP-citrate lyase, FERM domain, domain of beta-TrCP, Sec23/Sec24 (+) Phospholipase, **Glutathione S-transferase**, tRNA endonuclease
(D1∩D3)\D	(-) Myosin head, PAN/Apple domain, Carboxypeptidase, Immunoglobulin-like fold, Kinesin motor, Cadherin, (+) Ankyrin repeat
D2\D	(+) Cysteine-rich secretory protein family, SGS domain, **Actinin-type actin-binding domain**, Caspase-like domain

Table 1-1. InterProScan Functions of *M. hapla* DEGs identified in diurnal tests 1, 2, and 3 ((-), down-regulation; (+), up-regulation).

Group	InterProScan Functions
D1∩D2∩D3	(-) Male sterility protein, **Pectinesterase**, Subtilase family, Male sterility protein, Carbohydrate, **Toll - interleukin 1 resistance**, Cytochrome P450, Glycosyl hydrolase, GTP cyclohydrolase I, oxidases, Serine/Threonine protein kinases, ABC-transporter, Multicopper oxidase, Disease resistance protein, **Leucine-rich repeat**, Tyrosine protein kinases, **Stress up-regulated Nod 19**, Exo70 exocyst complex, Transferase family, **Salt stress response**, UDP-glucosyl transferase, Serine/Threonine protein kinases, Protein kinase, maltase-glucoamylase, **Plant heme peroxidase**, Plant CLC chloride channel signature, **Lipoxygenase**, Beta-L-arabinofuranosidase, Pollen allergen, **Pectate lyase**, Bulb-type lectin, Heat shock factor (HSF), D-arabinono-1,4-lactone oxidase, **Plant lipid transfer protein** / seed storage protein / trypsin-alpha **amylase inhibitor**, Plant phospholipid transfer protein signature, **Protease inhibitor**/seed storage/LTP family, **Chitinases**, globulin family, Voltage-dependent anion channel, BURP domain, Copper amine oxidase, FAE1/Type III polyketide synthase-like protein, 3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase, **Proline rich extensin**, Cysteine-rich secretory protein, Lipase/Acylhydrolase, basic helix-loop-helix (bHLH) domain, **Ferric reductase**, Phosphofructokinase, apical meristem protein, Plant invertase, **pectin methylesterase inhibitor**, Formin, Sodium/calcium exchanger protein, POT family, **WRKY** domain, ACT domain, GRAS, Chalcone and stilbene synthases, cysteine protease signature, subtilase, Sulfate permease, Dirigent-like protein, dioxygenase, SAM-dependent O-methyltransferase, Macrophage migration inhibitory factor (MIF), Sugar transport proteins, Acyl-ACP thioesterase, Sodium/hydrogen exchanger family, Acetyltransferase, Exo70 exocyst complex subunit, Phytocyanin, BURP domain profile, PDDEXK-like family, TIGR01615, **Polysaccharide lyase**, Galactose mutarotase, Haemolysin-III related, Plastocyanin-like domain, **Aldo/keto reductase** family, Domain of unknown function (DUF4228), HEAT repeat profile, Phosphoenolpyruvate carboxykinase, Pathogenesis-related protein Bet v I family, GRAS domain, Weak chloroplast movement under blue light, metalloprotease, Matrixins cysteine switch, PAP_fibrillin, MatE, GTP cyclohydrolase I, **Soybean trypsin inhibitor (Kunitz) protease inhibitors**, Phenylalanine and histidine ammonia-lyases, Aromatic amino acid , phenylalanine ammonia-lyase, Cys/Met metabolism, Fructose-1-6-bisphosphatase, Xyloglucan fucosyltransferase, Soluble glutathione S-transferase, Wall-associated receptor kinase, dioxygenase, Kinesin-associated protein (KAP)Armadillo/beta-catenin-like repeats, Retinal pigment epithelial

	membrane protein, Ureide permease, Heat shock hsp70 proteins, VHS domain, **Lipolytic enzymes**, Chalcone-flavanone isomerase, Metallothionein, Beta-L-arabinofuranosidase, Armadillo/beta-catenin-like repeats, **Glutathione S-transferase**, ABC transporter, 1,3-beta-glucan synthase subunit, **Ethylene responsive element**, **Gibberellin regulated protein**, Staygreen protein, Hydroxymethylglutaryl-coenzyme A reductase, Asp/Glu/Hydantoin racemase, Ubiquitination Targets, Remorin, C-terminal region, Asparagine synthase, Foie gras liver health family 1, Cellulose synthase, Amidase, Intron-binding protein aquarius N-terminus, Late nodulin protein, Acyl-Esterase, Biotinyl/lipoyl, Staphylococcal nuclease, S-locus glycoprotein, Pyridoxal-dependent decarboxylase, Copine, **KNOX1**, Filament-like plant protein, Plant CLC chloride channel signature, Pyridoxal-phosphate dependent enzyme, Nodulin-like, Telomerase activating protein, HOX, Spondin, Exostosin, Fasciclin, Carbon-nitrogen hydrolase, Phosphatidylinositol phosphate kinase (PIPK), Aldehyde dehydrogenases, **14-3-3 protein**, FAD binding domain, Flavodoxin-like domain profile, **Oxidoreductase**, NAD-binding domain, Pumilio, Cathepsin propeptide inhibitor, Urease, CRAL/TRIO, Phosphate-induced protein, Rieske [2Fe-2S] iron-sulfur (+) basic helix-loop-helix domain, Homeobox domain, Oxoglutarate, dioxygenase, UDP-glucosyltransferase, **Thioredoxin**, **Gibberellin regulated protein**, NADPH-dependent FMN reductase, S-adenosyl-L-methionine-dependent methyltransferase, hydroxyacid dehydrogenase, **Glycolipid transfer protein**, Cysteine-rich secretory protein, Cytochrome P450, Haem peroxidase
(D2∩D3)\D	(-) Germin, UDP-glucosyltransferase, **lectin/glucanase**, Serine/threonine-protein kinase, **Wall-associated receptor kinase**, **galacturonan-binding domain**, Terpenoid cyclases, Terpene synthase, Leucine-rich repeat, Bet v I type allergen, Cobalamin biosynthesis, plant lipid transfer protein, Hydrophobic seed protein, **Auxin responsive protein**, Bulb-type lectin domain, Alginate lyase, triphosphate hydrolase, **Gibberellin regulated protein**, Oxoglutarate, Non-haem dioxygenase, **Proteinase inhibitor**, **Ferredoxin reductase**, **lipase/esterase**, Homeodomain, **Haem peroxidase**, Cytochrome P450, **Auxin efflux carrier**, Pectinesterase inhibitor, antimicrobial extrusion protein, Glycoside hydrolase, **Endo-1,3(4)-beta-glucanase**, **Phospholipase**, Pyridoxal phosphate-dependent transferase, **Pectin lyase**, FMN reductase, Cytochrome P450, Glycosyl transferase, O-methyltransferase, Inorganic pyrophosphatase, Membrane attack

	complex/perforin (MACPF) domain, Proteolipid membrane potential modulator (+) Ubiquitin domain, Sieve element occlusion, Tetraspanin, Glycosyl hydrolase, Cupredoxin, Phytocyanin, meristem regulator, Ankyrin repeat, Acyl-CoA N-acyltransferase, Serine-threonine/tyrosine-protein kinase, lectin/glucanase, Glycoside hydrolase, Glycosyl transferase, Oligopeptide transporter, Sulfotransferase, triphosphate hydrolase
(D1∩D2)\D	(-) Proton-dependent oligopeptide transporter, Toll/interleukin-1 receptor, Bulb-type lectin, Protein kinase, Serine-threonine/tyrosine-protein kinase, Leucine-rich repeat, nucleoside triphosphate hydrolase, **Phospholipid-translocating P-type ATPase**, **oxidase**, basic helix-loop-helix (bHLH) domain, Tetratricopeptide-like helical domain, Homeobox, plant lipid transfer protein, Cyclic nucleotide-binding domain, Pyruvate kinase, Proton-dependent oligopeptide transporter, lectin/glucanase domain, **Nodulin-like**, **Wall-associated receptor kinase**, **glucanase domain**, glutamyltranspeptidase, Ankyrin repeat, Phosphatidylinositol 3-/4-kinase, Protein phosphatase (+) Nucleophile aminohydrolases, Glycosyl transferase, Ribosomal protein, Histone, Photosystem I PsaJ, basic helix-loop-helix (bHLH) domain, Protein chlororespiratory reduction, Auxin efflux carrier
D2\D	(-) Myb domain, Transferase, **Ethylene responsive element binding protein**, Protein kinase, **actin-binding protein**, sugar transporter, **Toll - interleukin 1 - resistance**, threonine/tyrosine-protein kinase, Cysteine-rich secretory protein, Alpha/beta hydrolase, Leucine-rich repeat, Histidine phosphatase, Phosphoglycerate/bisphosphoglycerate mutase, Legume lectin domain, NB-ARC, nucleoside triphosphate hydrolase, Cytochrome P450, Glycoside hydrolase, S-adenosyl-L-methionine-dependent methyltransferase, **Pectin lyase**, **Xylanase inhibitor**, **Chitin-binding**, Phosphate-induced protein, plant lipid transfer protein, basic helix-loop-helix (bHLH) domain, **Auxin responsive protein**, Phospholipase, Oxoglutarate, Non-haem dioxygenase, UDP-glucuronosyl/UDP-glucosyltransferase, Cyclic nucleotide-binding, glucanase, Ubiquitin domain, plant lipid transfer protein, **Phospholipase**, Rhamnogalacturonan lyase (+) Major intrinsic protein, Xanthine, permease, Retinoblastoma-associated protein, plant lipid transfer protein/seed storage helical

	domain, Rhamnogalacturonate lyase, Carboxypeptidase, basic helix-loop-helix (bHLH) domain, **Xylanase inhibitor**, plant mobile domain, **Phloem protein**

Table 1-2. InterProScan Functions of *M. truncatula* DEGs identified in diurnal tests 1, 2, and 3.

Groups	InterProScan Functions
UOOOOO	ShKT domain
UOOOO_	Histidine phosphatase, serine carboxypeptidase, UDP-glucuronosyl/UDP-glucosyltransferase
UOOO__	Cysteine peptidase
UOO___	Aspartic peptidase
UO____	Glycoside hydrolase, **C-type lectin**, Trypsin Inhibitor
U_____	**C-type lectin**, **Concanavalin A-like lectin/glucanase**, **Chondroitin AC/alginate lyase**, **Alginate lyase**, **Pectin lyase fold/virulence factor**, **Pectin lyase**, Fibronectin, **Immunoglobulin**, **Pectate lyase**, **Glycoside hydrolase**, Catalase immune-responsive domain, **Cellulose-binding family II/chitobiase**, **Carbohydrate-binding domain**, **Glycoside hydrolase**, Peptidase, subtilisin, Galactose-binding domain, Proteinase inhibitor, propeptide, Proprotein convertase, subtilisin, Cysteine peptidase, papain, Metallopeptidase, Short-chain dehydrogenase/reductase, ShKT domain, Aspartic, astacin, Metallopeptidase, NAD(P)-binding domain, Short-chain dehydrogenase/reductase SDR, Alpha/Beta hydrolase fold, Peptidase S10, Serine carboxypeptidase, Peptidase S10, Serine carboxypeptidase, Active site, Protein kinase, Phospholipase D/Transphosphatidylase, Carboxypeptidase, Glycoside hydrolase, Pectin lyase fold, Pectate lyase, Immunoglobulin, Glycoside hydrolase, Chitinase, Alpha carbonic anhydrase, Carbonic anhydrase, Alpha-class, Conserved site, Tyrosinase copper-binding domain, Uncharacterised domain, di-copper center, transmembrane domain, Cation-transporting P-type ATPase, cytoplasmic domain, Aromatic amino acid hydroxylase, Protein-tyrosine phosphatase-like, Galactosyltransferase, Beta-1, 4-galactosyltransferase, UDP-glucuronosyl/UDP-glucosyltransferase, Glycosyl transferase, **Chorismate mutase**, NHL repeat, TolB, Low-density lipoprotein (LDL) receptor class A repeat, Nuclear hormone receptor, Serum albumin, N-terminal, ShKT domain, RlpA-like double-psi beta-barrel domain, Major facilitator superfamily domain, ENTH domain, Dbl homology (DH) domain, PH, Calcium-binding site EF-hand domain, CAP domain, Cysteine-rich secretory protein, Allergen V5/Tpx-1-related, Ets domain, RmlC, Acireductone dioxygenase, Trypsin inhibitor, Follistatin, Pleckstrin, Transthyretin-like, **FMRFamide-related peptide**, Thioredoxin, Calsequestrin, Surfeit locus 4, Beta-catenin-interacting ICAT, Sodium:neurotransmitter symporter, Na+ channel, Amiloride-sensitive, Proteolipid membrane potential modulator, *Sulfolobus islandicus* filamentous virus, Orf14, **Cytochrom**, **Serpentine type 7TM GPCR chemoreceptor Srsx**, **GPCR**, **Rhodopsin**, 7TM, DUF148
_UOOOO	ELO family
UOOO	**Cytochrome** P450, **Nematode cuticle collagen**, Ribokinase-like, Carbohydrate kinase PfkB, Chorismate mutase, CAP domain, Aspartic peptidase domain, Thrombospondin type-1 (TSP1) repeat, Peptidase M12B, ADAM/reprolysin, Metallopeptidase

_UOO__	**Nematode cuticle collagen, C-type lectin**
_UO___	Sulfatase, Alkaline-phosphatase, Sulfatase, Alkaline phosphatase
_U____	Glycoside hydrolase, Tyrosinase copper-binding domain, Aspartic peptidase, **Pectin lyase** fold/virulence factor, **Pectate lyase**, Alkaline-phosphatase, Sulfatase
UDOO	(Uncharacterized)
UDOOO_	**Pectate lyase**
UDOO__	(Uncharacterized)
DUOOOO	Nematode cuticle **collagen**, papain, **Cysteine peptidase**, Thioredoxin
DUOOO_	Nematode cuticle collagen, Nucleoside triphosphate hydrolase, Helicase
DUOO__	Ribonuclease H-like domain, Piwi domain, PAZ domain, T antigen
DUO___	(Uncharacterized)
DU____	**C-type lectin**, NUDIX hydrolase
D_____	Zona pellucida domain, Glycosyl transferase, Family 31, T antigen, Ori-binding, Annexin, C3HC4 RING-type, Nuclear hormone receptor, Ubiquitin-conjugating enzyme, Ubiquitin specific protease domain, Ubiquitin carboxyl-terminal hydrolase, WD40 repeat, UDP-glucuronosyl/UDP-glucosyltransferase, von Willebrand factor, Zona pellucida domain, **Antifreeze protein**, K Homology domain, STAR protein, Tyrosinase copper-binding domain, Thymidine kinase, carboxypeptidase, Serine proteases, Proteasome, Protein kinase, Serine/threonine-protein kinase, PLAT/LH2 domain, Ras-like guanine nucleotide exchange factor, Metallo-dependent phosphatase, Serine/threonine-specific protein phosphatase/bis(5-nucleosyl)-tetraphosphatase, Serine proteases, trypsin, Peptidase, Serine carboxypeptidase, Nuclease-related domain, Carbohydrate/puine kinase, NUDIX hydrolase, papain C-terminal, Alpha/Beta hydrolase, Peptidase, Phospholipase, Phosphomannose isomerase, Acyl transferase/acyl hydrolase/lysophospholipase, Aldehyde dehydrogenase, C-terminal, Glycoside hydrolase, Aldehyde dehydrogenase, Glycoside hydrolase superfamily, Aldolase, Amidase, Pyridoxal phosphate-dependent transferase, Heme peroxidase, Aromatic amino acid hydroxylase, Metallopeptidase, Astacin, Catalase immune-responsive domain, Tyrosinase copper-binding domain, Metallopeptidase, Nucleophile aminohydrolases, N-terminal, Gamma-glutamyltranspeptidase, Concanavalin A-like lectin/glucanase, Tyrosinase, oxidoreductase, Fatty acid desaturase, Flavin monooxygenase, Glycosyl hydrolase, Intein, Ataxin-1/HBP1 module (AXH), Hedgehog protein, Integrase, Aspartic peptidase, Kinesin, Pleckstrin, Citron, Mini-chromosome maintenance, Malate/L-lactate dehydrogenase, Metallopeptidase, Ribosomal protein, Short-chain dehydrogenase/reductase, Carboxylesterase, Lipase EstA/Esterase EstB, SLC26A/SulP transporter domain, Thioredoxin, Glutathione S-transferase, S-adenosyl-L-methionine-dependent methyltransferase, Nucleotide-diphospho-sugar transferase, Neurotransmitter-gated ion-channel ligand-binding domain, Phospholipid/glycerol acyltransferase, CobB/CobQ glutamine amidotransferase,

	Methyltransferase FkbM, Nucleotide-diphospho-sugar transferases, Glycosyltransferase, Galactosyltransferase, Major facilitator, Potassium channel domain, Na+ channel, Amiloride-sensitive, ShKT domain, Calponin, Stomatin, TauD/TfdA, BRCT domain, PAS domain, PAZ domain, PDZ domain, SH2 domain, HAD-like domain, DM13 domain, DnaJ domain, EB domain, Patched, F-box domain, MADF domain, SKP1/BTB/POZ domain, BTB/POZ domain, **Chitin binding domain**, HSP20-like chaperone, Chromo domain, Homeobox domain, Thionin, Pleckstrin homology domain, PH domain-like, Anillin homology domain, PapD-like, **Major sperm protein (MSP) domain**, Laminin, Lectin/glucanase, Galectin, C-type lectin, EGF-like calcium-binding domain, Aspartate/asparagine hydroxylation site, Ferredoxin, Follistatin/Osteonectin EGF domain, SPARC/Testican, **Nematode cuticle collagen**, Kinetochore protein, Serum albumin, Lipid-binding serum glycoprotein, permeability-increasing protein, Pancreatic trypsin inhibitor Kunitz domain, Transthyretin-like, Mad3/Bub1 homology region 1, **Moulting cycle MLT-10-like protein**, Innexin, Calycin, HSP20-like chaperone, Immunoglobulin, Kunitz inhibitor, Cold-shock protein, Drought induced 19 protein, Chondroitin proteoglycan 4, Histone, Annexin repeat, Plectin repeat, BIR repeat, Thrombospondin repeat, Lustrin, PAN/Apple domain, Thrombospondin, Armadillo, Importin, Pumilio, Low-density lipoprotein (LDL), Leucine rich repeat, Myc-type, Basic helix-loop-helix (bHLH) domain, Basic-leucine zipper domain, Ets domain, Winged helix-turn-helix DNA-binding domain, Sterile alpha motif/pointed domain, Pointed domain, Cytoplasmic polyadenylation element-binding protein, 14-3-3 protein, T05H10.3, DUF236

Table 2-1. InterProScan functions of *M. hapla* differentially expressed genes (DEGs) identified in the longitudinal study approach 2. In DEG analysis across seven different time points (egg-J2-1 DAI-2 DAI-4 DAI-5 DAI-7 DAI), DEGs were classified into one of the pattern groups as denoted in the table.

Groups	InterProScan Functions
__D_	TB2/DP1, HVA22, **Cytochrome**, Signal transduction histidine kinase, phosphotransfer (Hpt), **Ethylene responsive element binding protein**, Ankyrin repeats
_D__	Ribulose bisphosphate carboxylase
D___	Ubiquinone/plastoquinone **oxidoreductase**, Mitochondrial ATP synthase, Alpha-carbonic anhydrases, **Auxin response factor**, Pseudobarrel, **Cytochrome**, Cytochrome oxidase, Oxidoreductase, Cytochrome C oxidase transmembrane, ATP synthase, F1/V1/A1 complex, Cytochrome biogenesis protein CcbS, Epsin N-terminal homology (ENTH), Phosphoinositide-binding clathrin adaptor, **Ethylene responsive element binding protein** APETALA2 AP2/ERF, Oxoglutarate/iron-dependent dioxygenase, Isopenicillin N synthase, Non-heme dioxygenase in morphine synthesis, Glycosyl hydrolases, Glycoside hydrolase, Galactose mutarotase, Myc-type, Basic helix-loop-helix (bHLH) domain, Heme-copper oxidase, Cytochrome c oxidase subunit III, Leucine-rich repeat, Major facilitator superfamily, Multi antimicrobial extrusion protein, **Auxin efflux carrier**, Mlo, Nicotianamine synthase, Plant heme **peroxidase**, Photosystem I P700 chlorophyll a apoprotein psaA and psaB, Polyketide cyclase/dehydrase and **lipid transport**, Proton-dependent oligopeptide transporter, Proline rich **extensin**, Serine-threonine/tyrosine-protein kinase Concanavalin A-like lectin/glucanase, Leucine-rich repeat, Malectin-like carbohydrate, DUF2647, Proton-conducting membrane transporter, DH-ubiquinone oxidoreductase, PXA, Ribosomal protein, SecY translocase, SecY/SEC61, Serine carboxypeptidases, Sulfotransferase domain, Voltage-dependent anion channel, B-box-type zinc finger, Zinc finger Dof-type, Endonuclease/exonuclease/phosphatase
D_D_	P-loop containing nucleoside triphosphate hydrolase, ATPase, F1/V1/A1 complex, Ribulose bisphosphate carboxylase, Photosystem I PsaA/PsaB, Ubiquitin-conjugating enzyme, Multi antimicrobial extrusion protein
__DO	Oleosin, EF-hand calcium-binding, **Sugar efflux transporter** for intercellular exchange
_DO_	Plant heme **peroxidase**, **Cytochrome**, Aminotransferase, Large T-antigen origin-binding domain (OBD), Pyridoxal phosphate-dependent transferase, **Ethylene responsive element binding protein**, AP2/ERF, FBox and BRCT domain, Leucine Rich Repeat, FAD linked oxidase, CO dehydrogenase flavoprotein, Berberine, Dehydrin, Aldehyde dehydrogenases glutamic acid active site Aldehyde dehydrogenase family, Gnk2-, Salt stress response/antifungal, **Chitinases**, globulin, Glycosyl hydrolases, Glycoside hydrolase, Proton-dependent oligopeptide transporter, Aconitase/3-isopropylmalate dehydratase, Male sterility protein, **Fatty acyl-CoA reductase**, Homeodomain-likeSANT/Myb domain, ABC transporter, P-loop

	containing nucleoside triphosphate hydrolase, DUF1602, Pseudobarrel, **Protease inhibitor**/seed storage/LTP family, Plant **lipid transfer** protein/trypsin-alpha amylase inhibitor, CTP synthase, Glutamine amidotransferase, **Pectinesterase**, Plant invertase/pectin methylesterase inhibitor, **Pectin lyase**, Virulence factor, Pectinesterase inhibitor, Piwi domain, DUF241, Solute carrier protein, Ring finger domain, Protein of unknown function (DUF688), FBD, Universal stress protein family, Pyridoxal-phosphate dependent enzyme
DO__	Isoprenoid synthase, Zinc/iron permease, Serine/threonine-protein kinase, Tyrosine kinase, PAN/Apple domain, Bulb-type lectin, S-locus glycoprotein, Concanavalin A-like lectin/glucanase, Legume lectin, Proton-dependent oligopeptide transporter, P-loop containing nucleoside triphosphate hydrolase, Macrophage migration inhibitory factor, Tautomerase/MIF, ATPases associated with a variety of cellular activities, Disease resistance protein, AAA+ ATPase, P-loop containing nucleoside triphosphate hydrolase, NB-ARC, Toll/interleukin-1 receptor, Leucine-rich repeat, Cyclin-dependent kinase, Regulatory subunit, S-adenosyl-L-methionine-dependent methyltransferase, DUF1442DH, Ubiquinone/plastoquinone **oxidoreductase**, **Peroxidase**, Plant heme peroxidase, Serine/Threonine protein kinases, Tyrosine kinase, SGNH hydrolase-type esterase, UDP-glucuronosyl/UDP-glucosyltransferase, **H+-transporting ATPase** (proton pump), Cation transporter/ATPase, Non-heme dioxygenase, Oxoglutarate/iron-dependent dioxygenase, Isopenicillin, Squalene-hopene cyclase, Terpenoid cyclases/protein prenyltransferase, LysM, Fusaric acid resistance protein-like, Periplasmic binding protein, Leucine Rich Repeat, NB-ARC domain, Disease resistance protein signature, Toll/interleukin-1 receptor, P-loop containing nucleoside triphosphate hydrolase, Glucose/ribitol dehydrogenase, **Dehydrogenases/reductases**, Cation efflux protein transmembrane, DUF1602, **Ferric reductase**, **Ferredoxin reductase**, UDP-glucuronosyl/UDP-glucosyltransferase, **Cytochrome**, ABC transporter transmembrane, Transcription factor GRAS, Disease resistance protein, NB-ARC domain, Toll/interleukin-1 receptor–resistance, P-loop containing nucleoside triphosphate hydrolase, Serine-threonine/tyrosine-protein kinase Bulb-type lectin, APPLE/PAN domain, S-locus glycoprotein, DUF3403, Polyketide cyclase/dehydrase and **lipid transport**, Basic helix-loop-helix (bHLH) domain profile, Transcription factor GRAS, **Sugar (and other) transporter**, Pyruvate kinase
_DOO	Male sterility protein, Thiolase, polyketide synthase, Ketoacyl-CoA synthase, Proline rich extensin, Alpha/beta hydrolase, ABC-transporter extracellular, Plant PDR ABC transporter, Acyl-ACP thioesterase, **Phospholipid/glycerol acyltransferase**, Subtilase family, Peptidase inhibitor, T-complex protein, Glycerol-3-phosphate dehydrogenase, **Cytochrome**, UPF0160, NB-ARC domain, Ankyrin repeats, DUF4219, Plant heme **peroxidase**, **WRKY**,

	Chloramphenicol acetyltransferase, Fructose-bisphosphate aldolase, Phosphogluconate dehydrogenase, Hydroxyisobutyrate dehydrogenase, Sugar-metabolising enzyme, Soybean trypsin inhibitor (Kunitz), **protease inhibitors**, Plant **lipoxygenase**, **Ethylene responsive element binding protein** APETALA2 and EREBPsAP2/ERF, **Nodulin**, Myc-type, Basic helix-loop-helix (bHLH), **H+-transporting ATPase** (proton pump), Leucine-rich repeat, NB-ARC, Disease resistance protein, Toll-interleukin 1-resistance, P-loop containing nucleoside triphosphate hydrolase, AAA+ ATPase, PAN/Apple domain, D-mannose binding lectin, S-locus glycoprotein, Bulb-type mannose-specific lectin, Homeobox domain, Cysteine-rich TM module stress tolerance, Sulfate permease, STAS, SLC26A/SulP transporter, Calmodulin binding protein, Serine-threonine/tyrosine-protein kinase, Bowman-Birk serine protease inhibitors, Glycosyl Glycoside hydrolase, Xyloglucan endo-transglycosylase, Concanavalin A-like lectin/glucanase domain, Translation elongation factor, S-adenosyl-L-methionine-dependent methyltransferase, G-box binding protein multifunctional mosaic region, Zinc finger RING-type, Protease inhibitor/seed storage/**lipid transfer** protein/trypsin-alpha amylase inhibitor/Par allergen, LEM3 (ligand-effect modulator 3)/CDC50, Tify, CO/COL/TOC1, Glyoxalase/Bleomycin resistance protein/Dioxygenase, Dihydroxybiphenyl dioxygenase, Chalcone and stilbene synthases, Thiolase, Serine-threonine/tyrosine-protein kinase, Salt stress response/antifungal, Gnk2
DOO_	Stress up-regulated Nod, ABC transporter transmembrane region, Plant heme **peroxidase**, likeOxoglutarate/iron-dependent dioxygenase, non-heme dioxygenase in morphine synthesis, Glycosyl hydrolases, Glycoside hydrolase Globulin, 2A0109, Phosphate permease, Phosphate:**H+ symporter, Sugar (and other) transporter**, Cation efflux
DOOO	**Ferric reductase**, **Cytochrome**
DOU_	Stress up-regulated Nod, ABC transporter transmembrane region, Plant heme **peroxidase**, likeOxoglutarate/iron-dependent dioxygenase, non-haem dioxygenase in morphine synthesis, Glycosyl hydrolases, Glycoside hydrolase Globulin, 2A0109, Phosphate permease , phosphate:**H+ symporter, Sugar (and other) transporter**, Cation efflux
DOUO	Bifunctional inhibitor/plant **lipid transfer** protein/seed storage
D_U_	UDP-glucuronosyl/UDP-glucosyltransferase, S-adenosyl-L-methionine-dependent methyltransferase, **Ethylene responsive element binding protein**, APETALA2 and EREBPs, AP2/ERF, Heme **peroxidase**, Peroxidases, Myc-type, Basic helix-loop-helix (bHLH) domain, UDP-glucuronosyl/UDP-glucosyltransferase
D_UO	Homeodomain-like, SANT/Myb domain, SANT domain

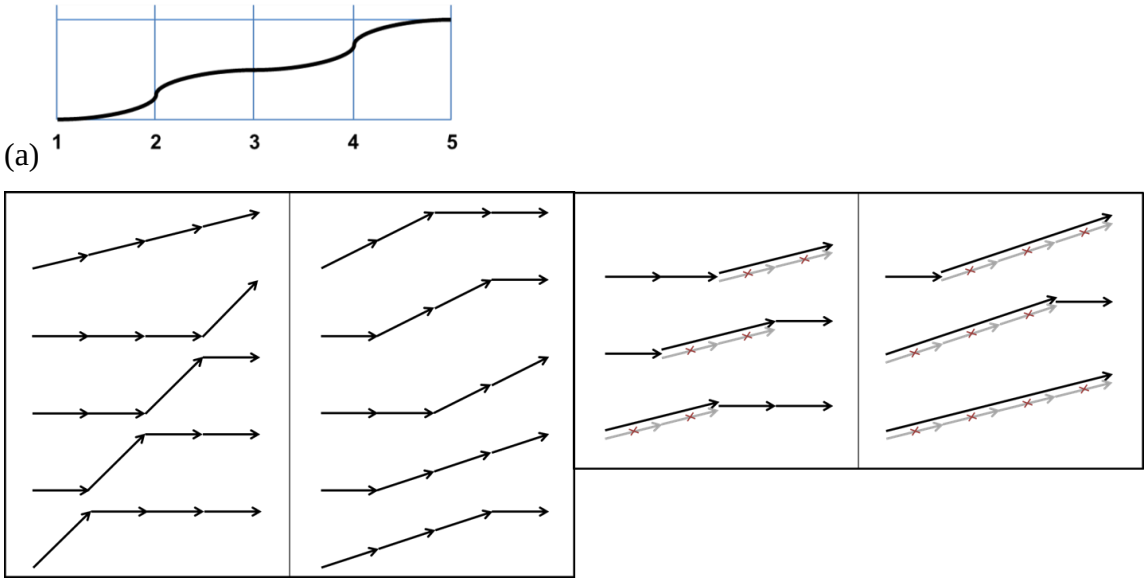
DUO_	Transcription factor GRAS, DUF588, myb_SHAQKYF, Sulfotransferase, Retinal pigment epithelial membrane protein, Carotenoid oxygenase, Polyketide cyclase/dehydrase and **lipid transport**, **Cellulase**, Glycosyl hydrolase, Glycoside hydrolase, Phosphatase, **Xylanase inhibitor**
DUOO	Molybdate transporter, Gnk2, Salt stress response/antifungal, Photosystem I reaction, Carboxyl transferase, Acetyl-coenzyme, Acetyl-CoA carboxylase
___U	Ribosomal protein, DH-ubiquinone/plastoquinone **oxidoreductase**
__U_	Carbonic anhydrase, Receptor for ubiquitination targets F-box domain, Phloem protein, Plant **lipid transfer** protein/seed storage protein/trypsin-alpha amylase inhibitor, Hydrophobic seed protein, **Trypsin and protease inhibitor**, Soybean trypsin inhibitor (Kunitz), Protease inhibitors, Endonuclease/exonuclease/phosphatase, **Cellulose** synthase
_U__	**Thioredoxin**
U___	Aminotransferase, Pyridoxal phosphate-dependent transferase, Oxoglutarate/iron-dependent dioxygenase, Isopenicillin, Non-heme dioxygenase in morphine synthesis, Hydroxysteroid dehydrogenase/isomerase, Epimerase/dehydratase, PDDEXK, Nucleophile aminohydrolases, Bowman-Birk type **proteinase inhibitor**, Bowman-Birk serine protease inhibitor, CCT motif, **Cytochrome**, **Ferric reductase**, EF-hand calcium-binding, D-isomer specific 2-hydroxyacid dehydrogenase, D binding domain, Alcohol dehydrogenase, GroES-like domain, PAN/Apple domain, Bulb-type lectin, Serine/Threonine/tyrosine protein kinases, Bulb-type mannose-specific lectin, S-locus glycoprotein domain, EGF-like domain, DUF4408, Ferritin-like diiron, **Gibberellin regulated protein**, Globin, Plant hemoglobins, IQ motif, EF-hand binding site,BAG domain, Lanthionine synthetase, Leucine rich repeat, Soluble **glutathione S-transferase**, Concanavalin A-like lectin/glucanase, **Thioredoxin**, Lateral organ boundaries, Multi antimicrobial extrusion protein, Myb domain, Homeodomain, Early **nodulin** 93 ENOD protein, PDDEXK, TIGR01615, Proton-dependent oligopeptide transporter, AGC-kinase, DUF1677, DUF760, FAR1 DNA-binding domain, MULE transposase, plant mutator transposase, SWIM zinc finger, Tify, CO/COL/TOC1, Chloramphenicol acetyltransferase, Universal stress protein family, Rossmann-like alpha/beta/alpha sandwich fold, Tetratricopeptide, Rhodanese, P-loop containing nucleoside triphosphate hydrolase
_UD_	(Uncharacterized)
U_D_	Heat shock proteins, Glutamate dehydrogenase, Lateral organ boundaries DUF260, Late embryogenesis abundant (LEA) LEA-25/LEA-D113, Phosphatidylethanolamine-binding protein, **Thioredoxin** phorbol-ester/DAG, SANT/Myb domain, Homeodomain, basic helix-loop-helix (bHLH) domain, F-

	box domain, Ferritin, Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase, Glyoxalase/fosfomycin resistance, DUF3339, LysM domain, Cold acclimation protein WCOR413, **Protease inhibitor**/seed storage/Plant **phospholipid transfer protein**/trypsin-alpha amylase inhibitor/Par allergen, Linker histone, Dehydrins, Trp-Asp (WD) repeats, cystathionine beta-synthase, Endonuclease/Exonuclease/phosphatase, Concanavalin A-like lectin/glucanas/Glycosyl hydrolases/Glycoside hydrolase, Xyloglucan endo-transglycosylase, **Expansin**, **Cellulose-binding-like domain**, Pollen allergen, **endoglucanase**, Zinc finger, RING/FYVE/PHD-type (The RING-variant domain is a C4HC3 zinc-finger like motif found in a number of cellular and viral proteins. Some of these proteins have been shown both *in vivo* and *in vitro* to have ubiquitin E3 ligase activity), PPM-type phosphatase, Thiamine pyrophosphate enzyme, Thiamine pyrophosphate enzyme, EF-hand calcium-binding domain, Seed maturation protein, Late embryogenesis abundant protein, LEA-25/LEA-D113, Dormancy/**auxin associated protein**, TPL-binding domain in **jasmonate signalling**, Ferritin-like diiron, NB-ARC domain, Late embryogenesis abundant protein, **Aldo/keto reductase**, Sodium/hydrogen exchanger family, **Cation/H+ exchanger**, UDP-glycosyltransferases signatureUDP-glucoronosyl and UDP-glucosyl transferase, **Sugar (and other) transporter**, Serine/threonine phosphatases, MULE transposase
UD__	**Chitinases**, Glycosyl hydrolases, Glycoside hydrolase, Globulin, **Pectate lyase**, **Pectin lyase** fold/virulence factor
UDD_	**Dehydrogenases/reductases**, Glucose/ribitol dehydrogenase, Cupin
U_DO	Soybean trypsin inhibitor, Kunitz, **Protease inhibitors**, Chloramphenicol acetyltransferase, Sieve element occlusion, Phosphate acyltransferases, **Phospholipid/glycerol acyltransferase**
UDO_	Chloramphenicol acetyltransferase, **Phospholipid/glycerol acyltransferase**, Myc-type, Basic helix-loop-helix (bHLH) domain, BAG, Heat shock proteins, **Ethylene responsive element binding protein**, TspO/MBR, Soluble **glutathione S-transferase**, Armadillo, Thaumatin, Pathogenesis-related protein signature, Heavy-metal-associated domain, UPF0114, Sodium/hydrogen exchanger, **Cation/H+ exchanger**, Peptidase, Reticulon, Sieve element occlusion, Glycosyl transferase family 8, Kinesin, Calponin, Hemopexin, Soybean trypsin inhibitor (Kunitz), **Protease inhibitors**, Cupin, Aluminium activated malate transporter, Potato inhibitor, DUF4782, Proteolipid membrane potential modulator, EXS, SPX, **Cytochrome**, UDP-glucoronosyl and UDP-glucosyl transferase, Myb domain,SANT/Myb domain,Homeodomain, SAM-dependent O-methyltransferase, DUF1262, Trp-Asp (WD) repeats, Epimerase/dehydratase family, GDP-mannose 4,6 dehydratase, Proton-dependent oligopeptide transporter, Terpenoid cyclases/protein prenyltransferase, Terpene synthase, Aldehyde

	dehydrogenases, **Sugar (and other) transporter**, Small hydrophilic plant seed protein, F-box domain, Receptor for ubiquitination targets, DUF760, Cleavage site for pathogenic type III effector avirulence factor Avr, Sodium/hydrogen exchanger family, Prokaryotic membrane **lipoprotein lipid attachment** site profile, Phosphofructokinase, IQ calmodulin, AMP-dependent synthetase/ligase, Glycosyl hydrolases, Xyloglucan endo-transglycosylase, GDP-mannose 4,6 dehydratase, CBS domain, P-loop containing nucleoside triphosphate hydrolase, Glycosyl hydrolase, Seipin, Glycoprotein membrane precursor, GPI-anchored lipoprotein, AT5G19250, DUF241, Glycosyl transferase, Nucleotide-diphospho-sugar transferases, Glucose/ribitol dehydrogenase, **Dehydrogenases/reductases**, Late embryogenesis abundant (LEA), Cupin, **Aldo/keto reductase**, GDSL-like Lipase/Acylhydrolase, Serine/Threonine protein kinases, Lipase, DUF223, Papain family cysteine protease, thiol (cysteine) proteases, asparagine, Cathepsin propeptide inhibitor domain, LMBR1 membrane protein
UDOO	Plant **phospholipid transfer** protein, Amino acid kinase, Aspartate/glutamate/uridylate kinase, Aldehyde dehydrogenase, DUF3475, Aminotransferase, Pyridoxal phosphate-dependent transferase, Male sterility protein, Glutamine amidotransferase, Raffinose synthase or seed imbibition protein Sip1, Glycosyl hydrolases, Glycoside hydrolase, Transferase, **Xylanase inhibitor**, *Arabidopsis* broad-spectrum mildew resistance protein RPW8, NB-ARC domain, P-loop containing nucleoside triphosphate hydrolase, Powdery mildew resistance protein, Leucine-rich repeat, Partial alpha/beta-hydrolase lipase, Trehalose-phosphatase, Glycosyltransferase, glycanase, NMD3, NmrA, Non-heme dioxygenase, Carbonic anhydrase, PDDEXK, Myo-inositol-1-phosphate synthase, **Cytochrome**, Proline rich **extensin** signature, Alpha/beta hydrolase
__UO	**Gibberellin regulated protein**, Trp-Asp (WD) repeats, DUF599
_UO_	Transcription factor, SBP-box, **Lipoxygenase**, Lipase, Glycosyl hydrolases, Xyloglucan endo-transglycosylase, Concanavalin A-like lectin/glucanase domain, Glycoside hydrolase, Proline rich **extensin**, Plant **peroxidase** signature, Chlorophyll A-B binding protein, Cysteine-rich secretory protein, Allergen V5/Tpx-1/SCP/Tpx-1/Ag5/PR-1/Sc7 extracellular domains, Phloem protein, GDA1/CD39 (nucleoside phosphatase), Glyceraldehyde 3-phosphate dehydrogenase, Haloacid dehalogenase, Nitrophenylphosphatase, Fructose-bisphosphate aldolase, Photosystem I PsaG/PsaK, S-adenosyl-L-methionine-dependent methyltransferase, Glycosyl hydrolase, Glycoside hydrolase, **Ethylene responsive element**, AP2/ERF, **Sugar (and other) transporter**, 2A0109, Phosphate permease, myb_SHAQKYF, Terpene synthase, Phosphorylase, membrane lipoprotein lipid attachment site, Plant disease resistance response protein, **Cytochrome**, Glycosyl hydrolase, Glycoside hydrolase, Pyridoxal 5′-phosphate synthase, Transcription factor GRAS,

	PdxS/SNZ, Phosphopantetheine, Acyl carrier protein, Chloramphenicol acetyltransferase, Photosystem I PsaD, Organ specific protein DUF2775, glutamyl reductase, quinate 5-dehydrogenase, Protein phosphatase, **Expansin**/pollen allergen, **endoglucanase** (RlpA), No apical meristem, Phytocyanin, Cupredoxin
UO__	Alcohol dehydrogenase, Ferritin, Lactate/malate dehydrogenase, **Legume lectin**, Glycosyl hydrolases, **Pectate lyase**, Pathogenesis-related protein, Thaumatin, **Plant peroxidase**, 3′-5′ exonuclease, EF-hand calcium-binding domain, **Sugar transport proteins**, **Glutaredoxin**, **Thioredoxin**
UOD_	RCD1-SRO-TAF4 (RST) plant domain, RmlC-like cupin, Cysteine oxygenase/2-aminoethanethiol dioxygenase, AMP-dependent synthetase/ligase, **Universal stress protein**, Rhodanese
_UOO	FBD, Leucine-rich repeat, Ribosomal protein S15, Rpb1, Ribosomal protein S8, Ribosomal protein S3, Tetratricopeptide repeat, S4 RNA-binding domain, Phosphatase, Homeodomain
UOO_	**Pectin lyase** fold/virulence factor, Glycoside hydrolase, Oligopeptide transporter

Table 2-2. InterProScan Functions of *M. truncatula* differentially expressed genes (DEGs) identified in the longitudinal study approach 2. In DEG analysis across five different time points (1DAI-2DAI-4DAI-5DAI-7DAI) within infected root, DEGs were classified into one of the pattern groups as denoted in the table.



(a)

(b) (c)

Figure 1. Example showing the limitation of ignoring significant changes in the expression levels between non-consecutive time points. This puts an emphasis on gradual changes when gene expression patterns are classified (arrow heading to upper right: positive fold changes between two time points; arrow in grey with red cross: non-differential expression between consecutive time-points). (a) Supposed actual gene expression levels, which are gradually increasing along the five time points. (b) Some of the possible results when pairwise exact test was conducted only between the consecutive time-points. Any significant change between consecutive time-points could be identified. (c) Cases that could possibly be ignored by consecutive time-point comparisons. If the changes between consecutive time-points are not significant, but the sum of the gradual increases across more than two time points is significant, they could be missed though they have to be captured.

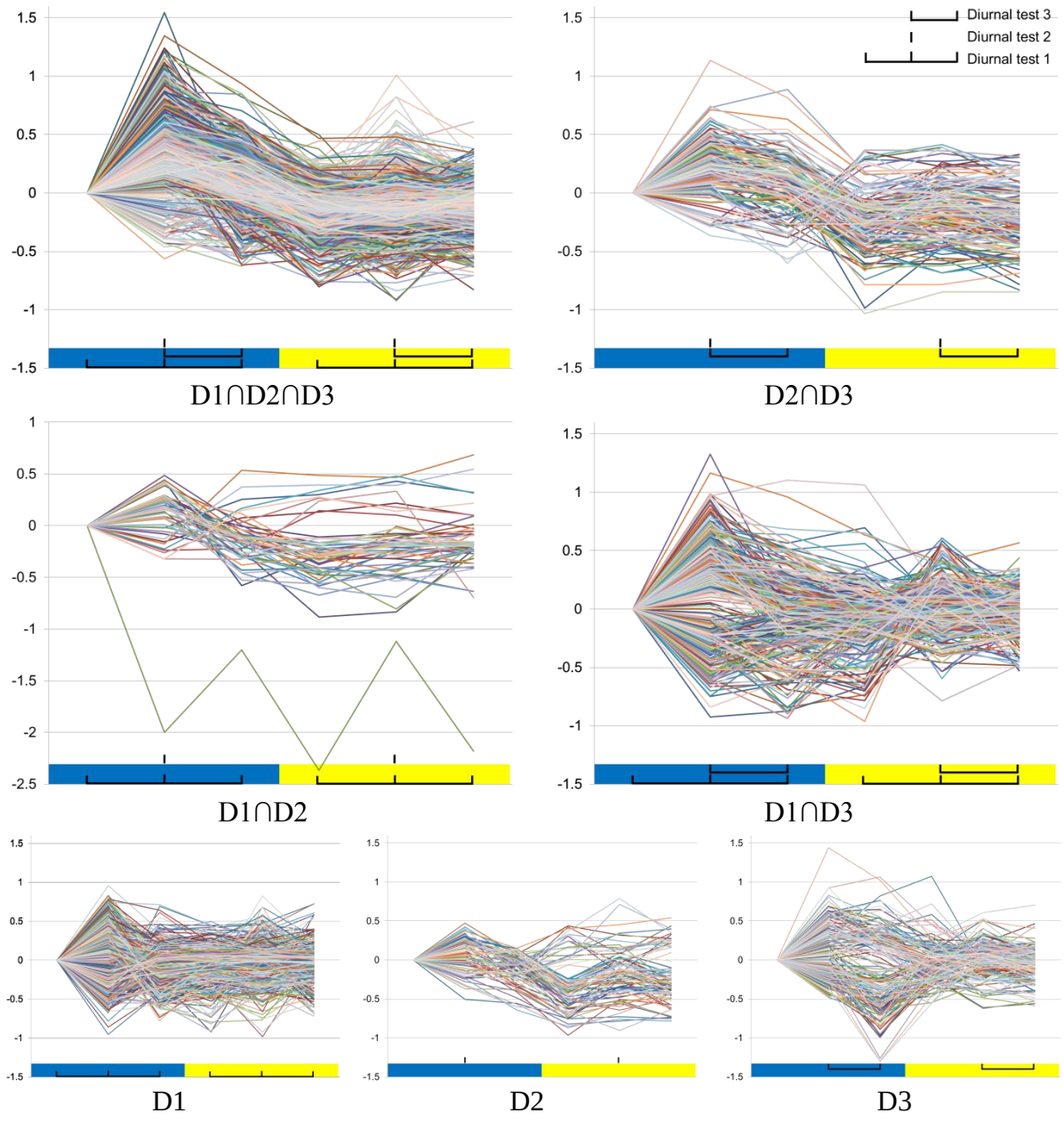
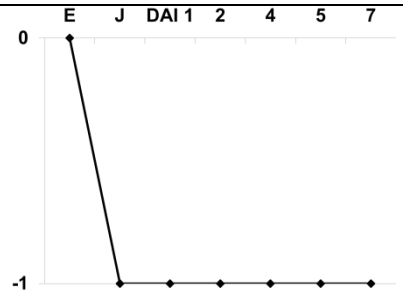
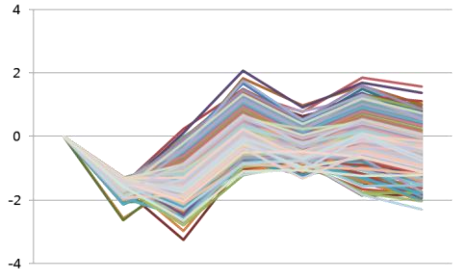
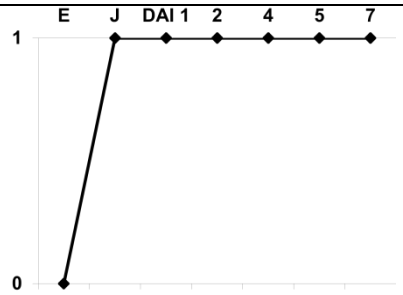
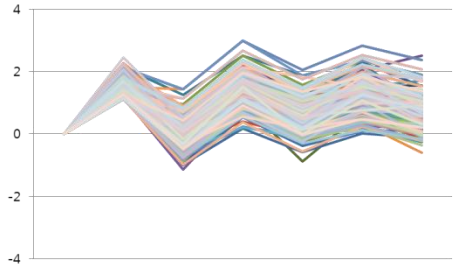
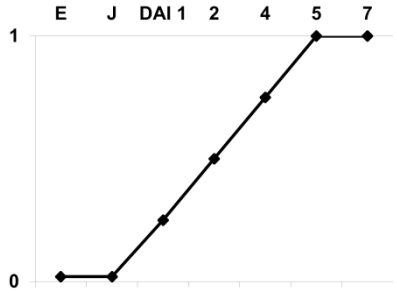
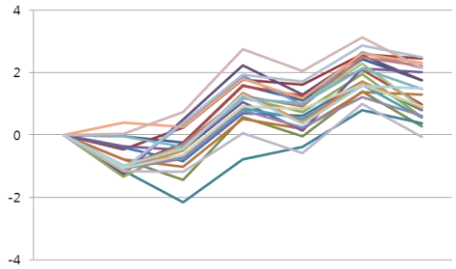
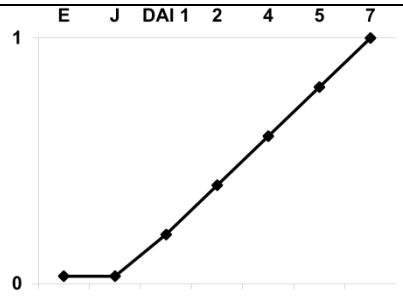
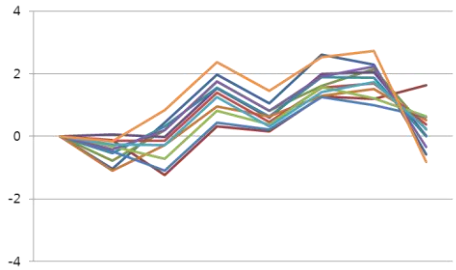
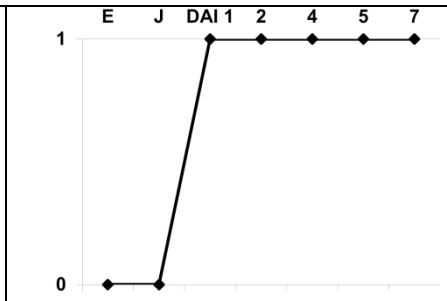
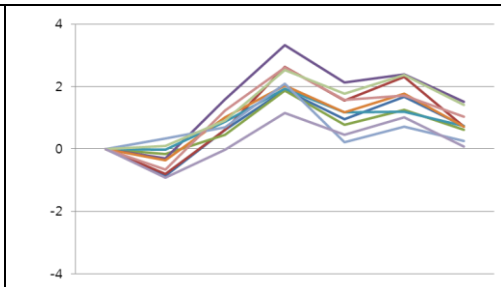
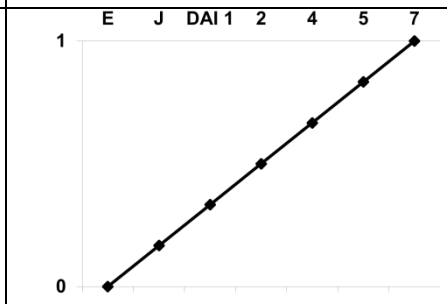
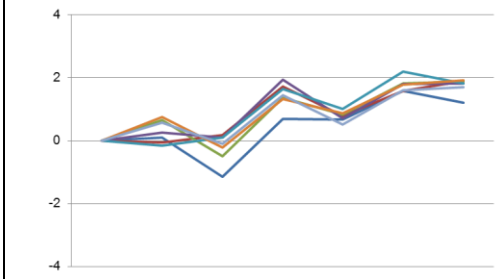
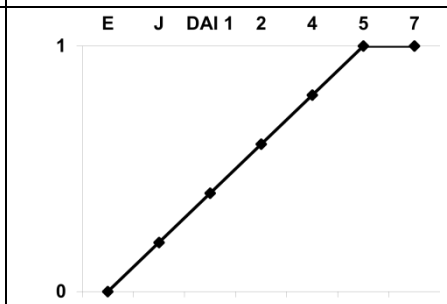
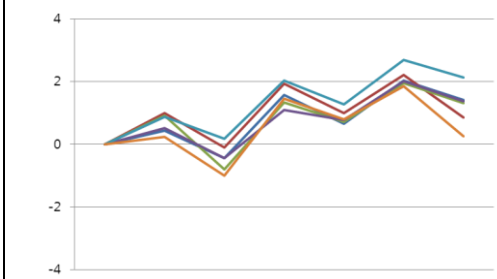
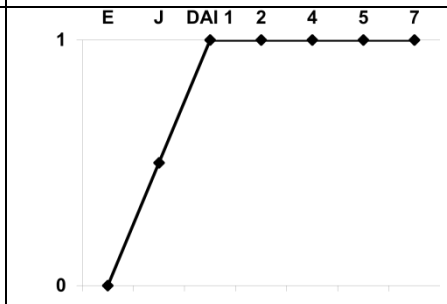
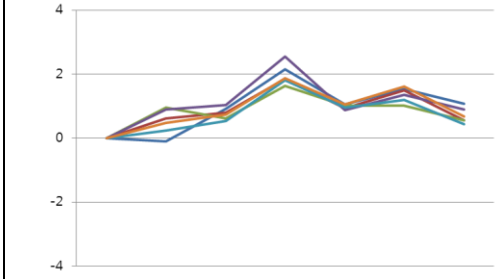
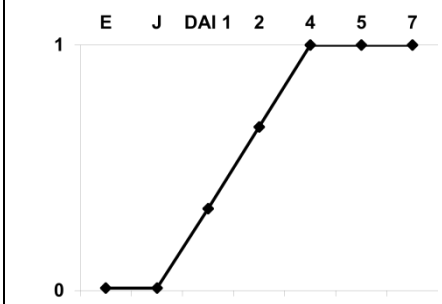
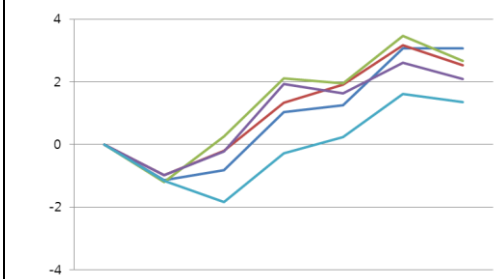
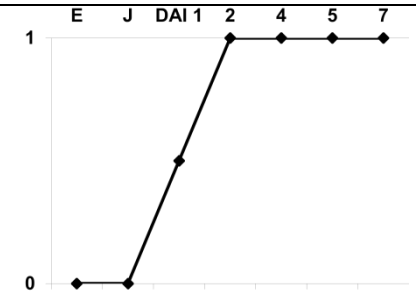
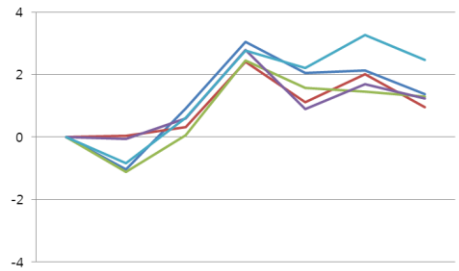
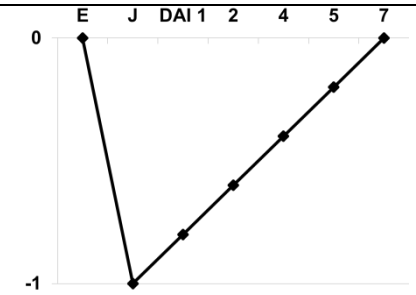
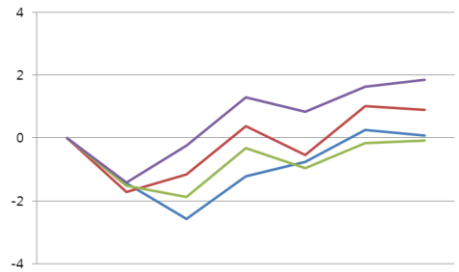
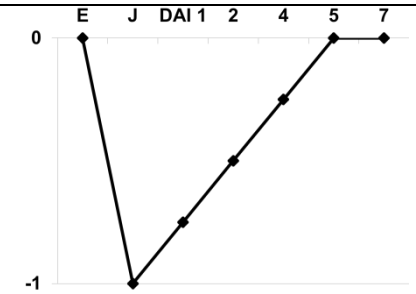
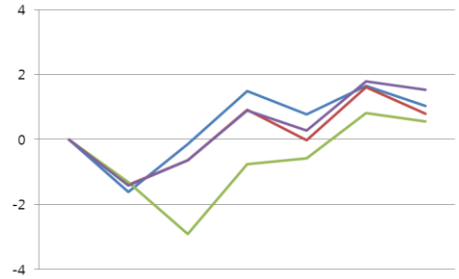
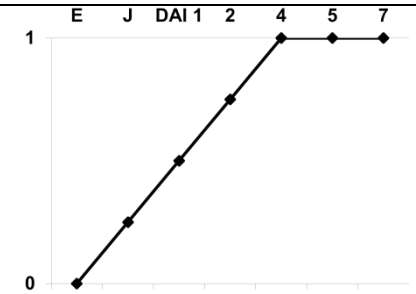
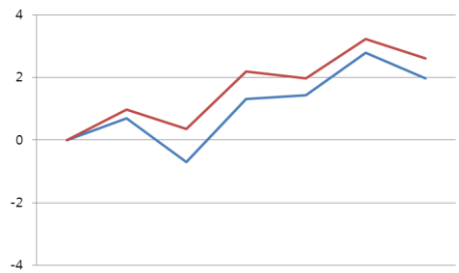
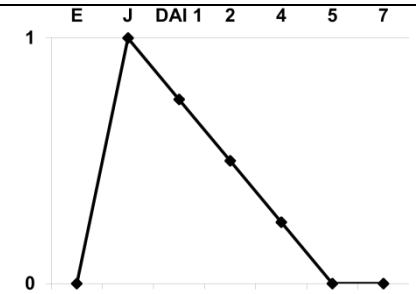
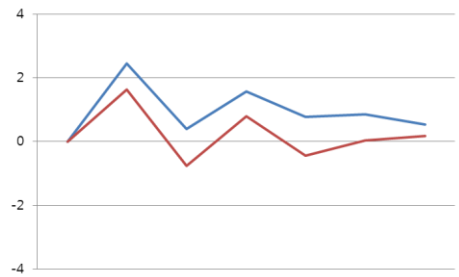
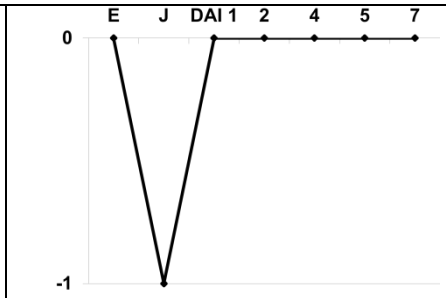
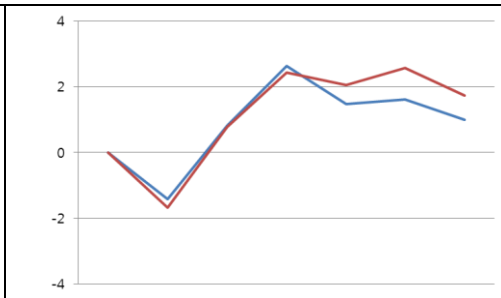
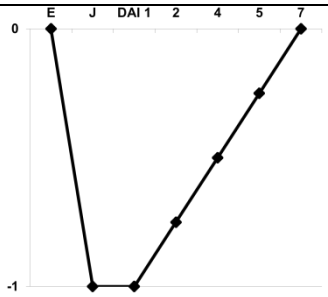
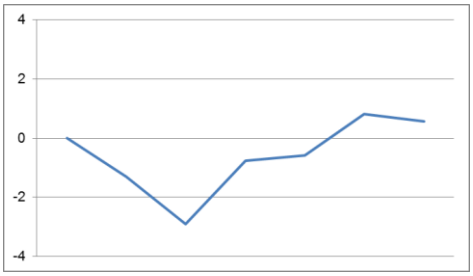
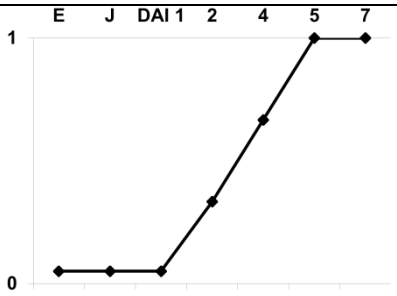
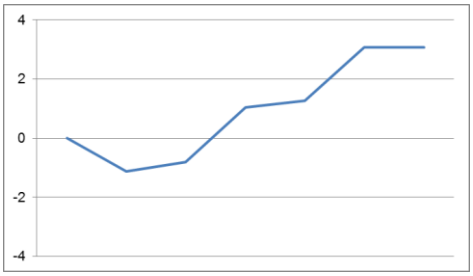
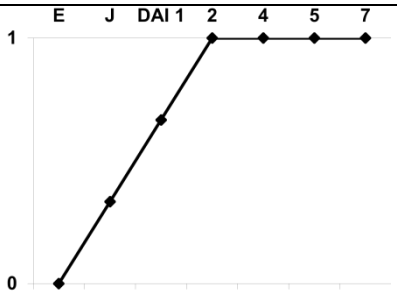
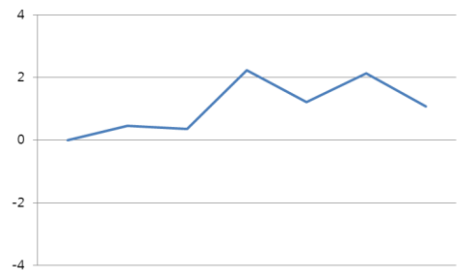
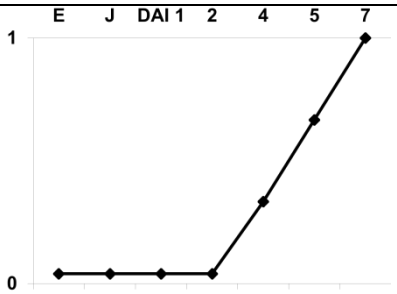
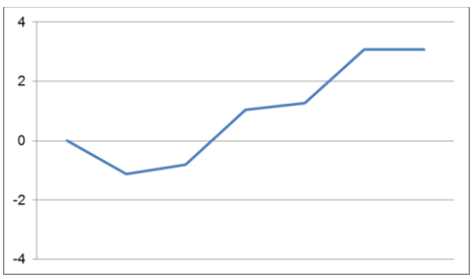


Figure 2. Expression profiles of *M. truncatula* DEGs in each test or in overlapping set of tests.

Groups		Gene Expression Profiles
D_____		
U_____		
UOOO		
_UOOOO		

_U____		
UOOOOO		
UOOOO_		
UO____		
_UOO__		

_UO___		
DUOOOO		
DUOOO_		
UOOO__		
UDOOO_		

DU____		
D_UOOO		
__UOO_		
UOO___		
___UOO		

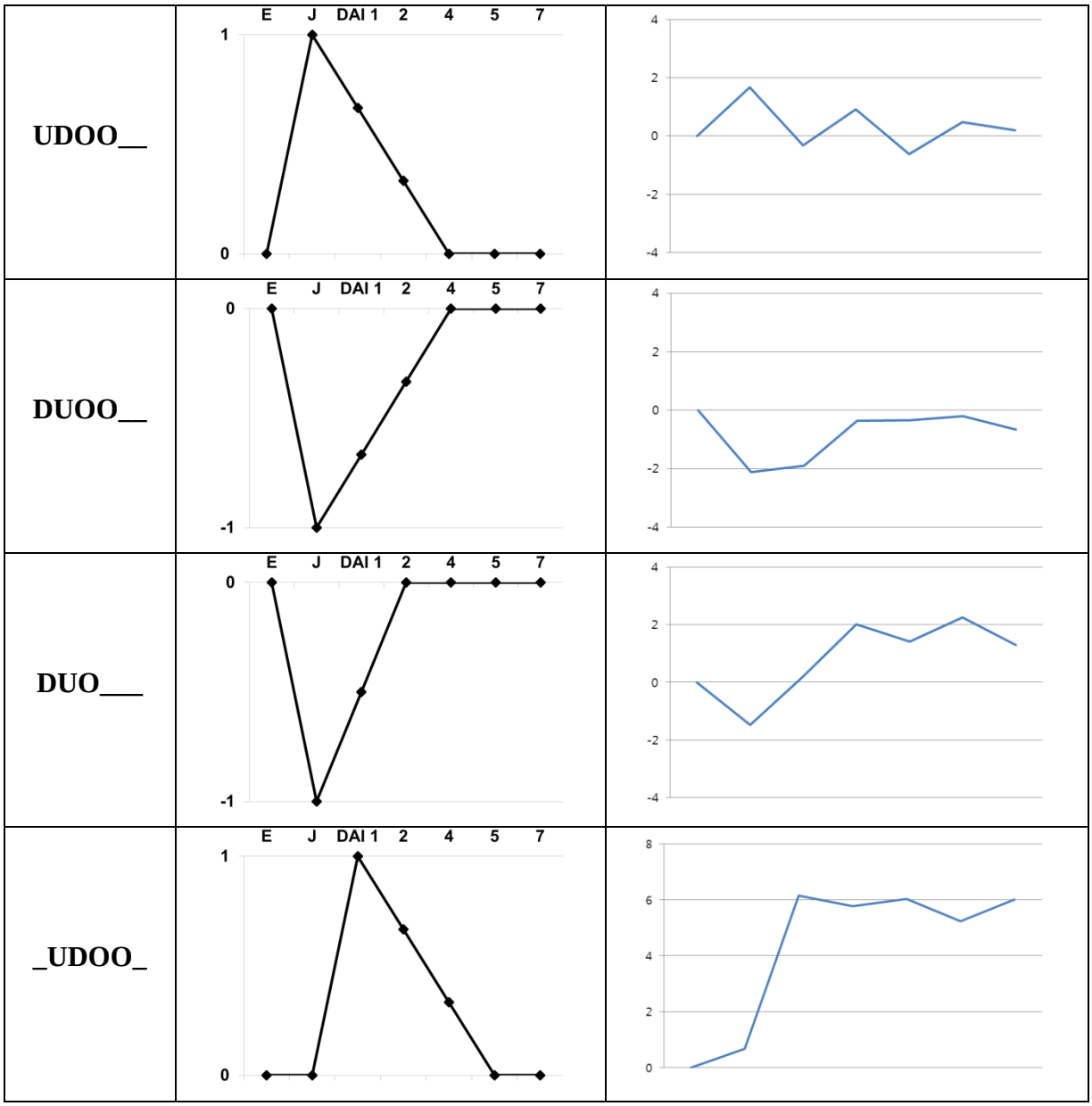
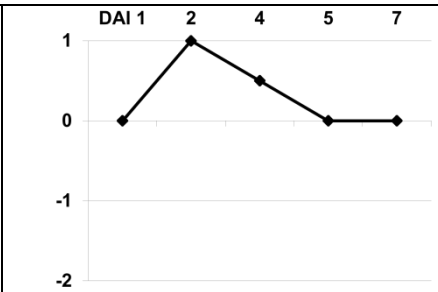
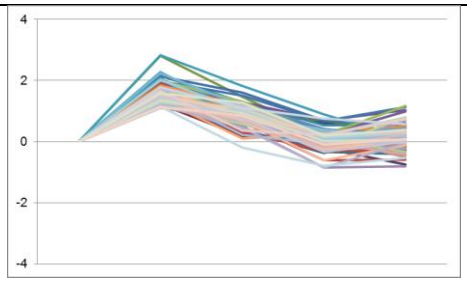
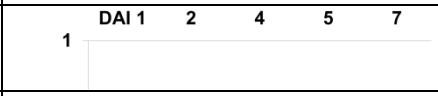
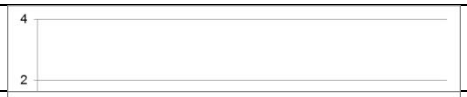
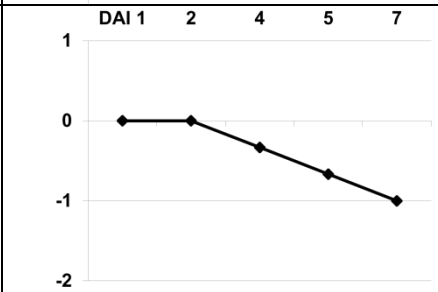
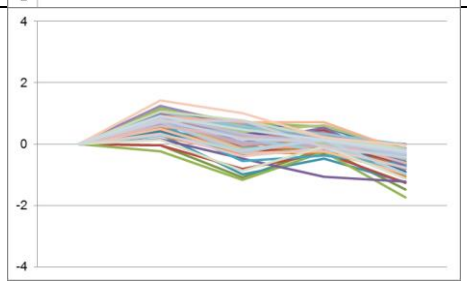
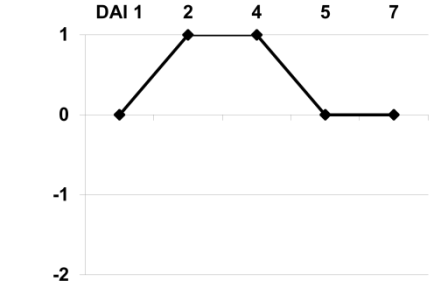
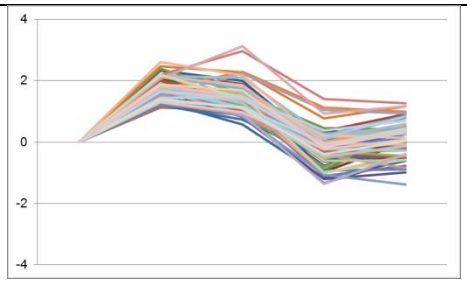
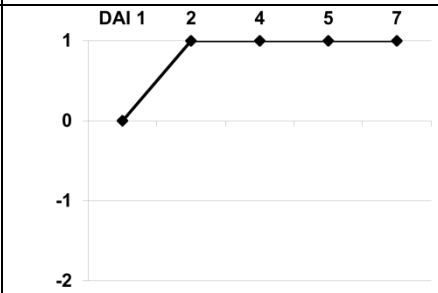
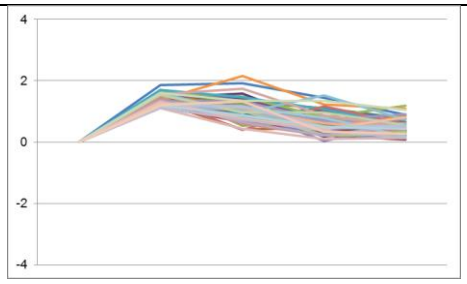
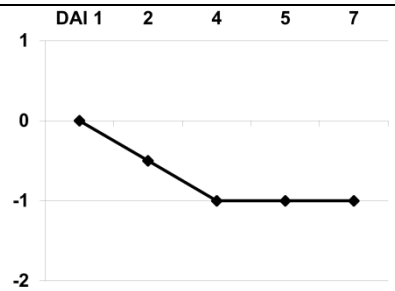
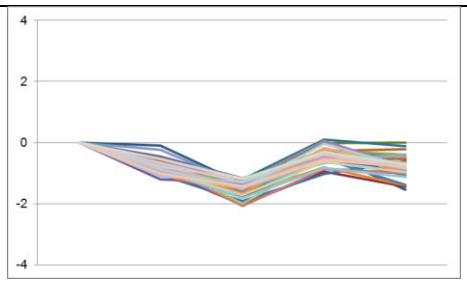
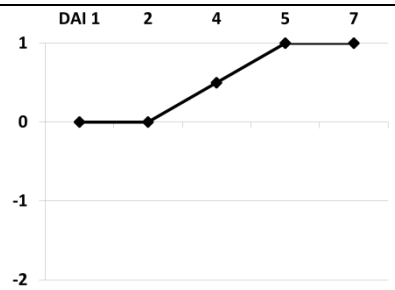
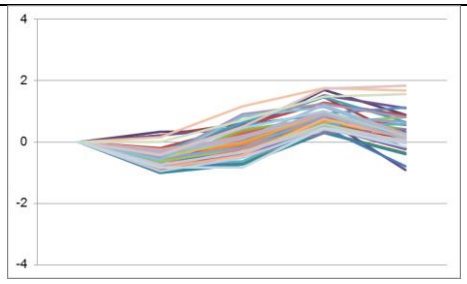
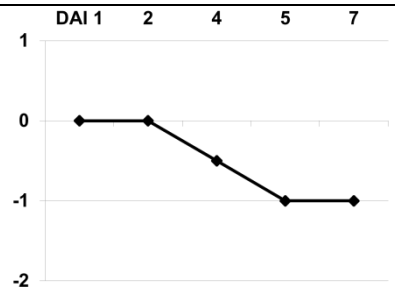
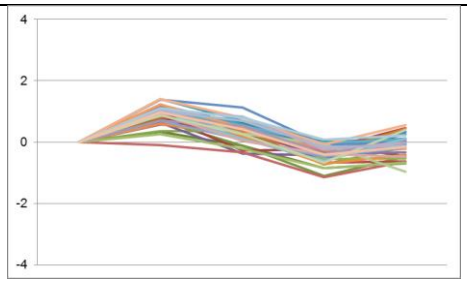
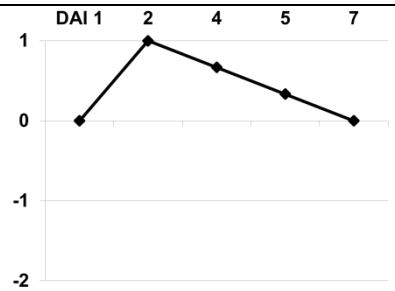
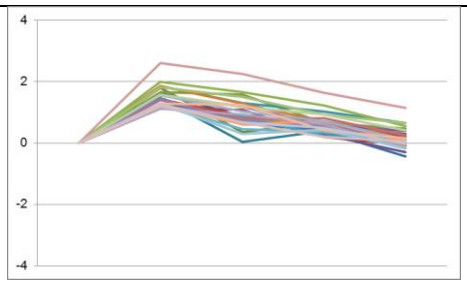
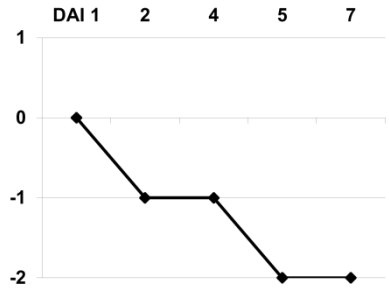
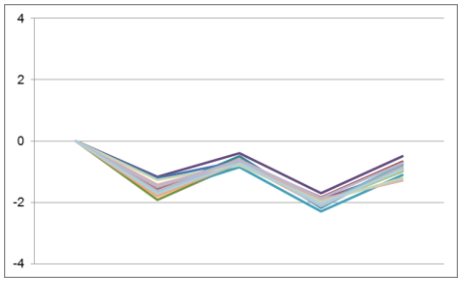
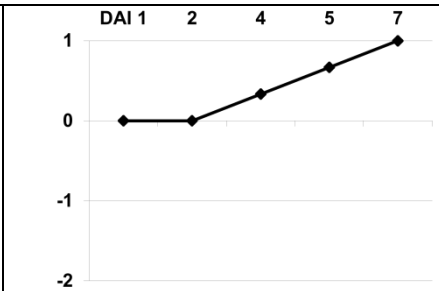
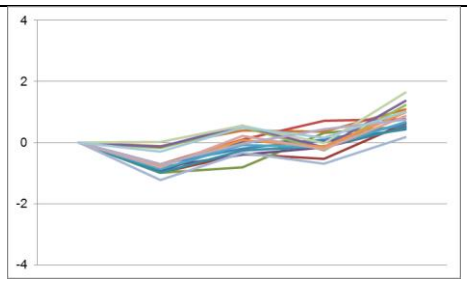
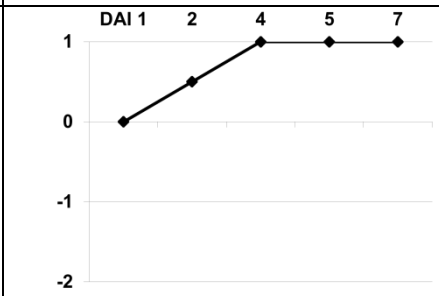
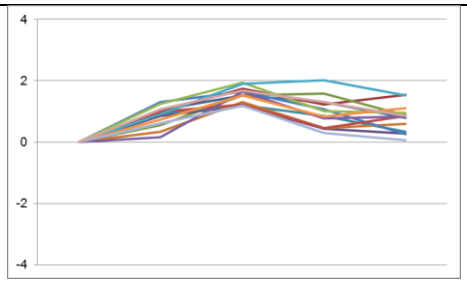
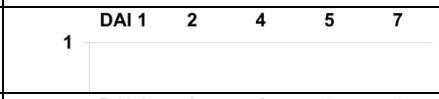
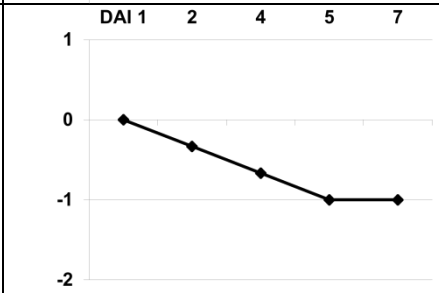
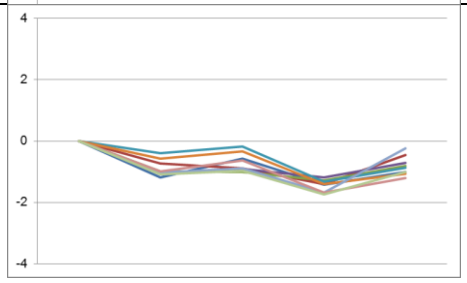
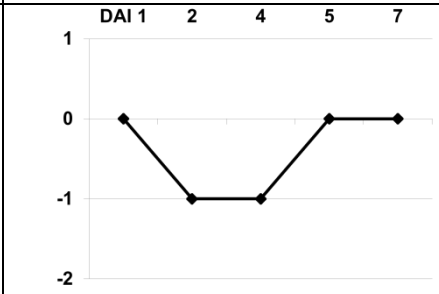
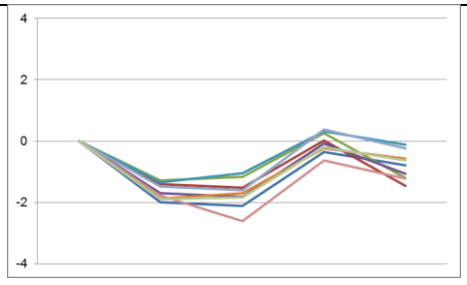
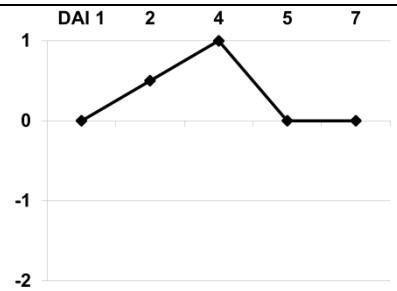
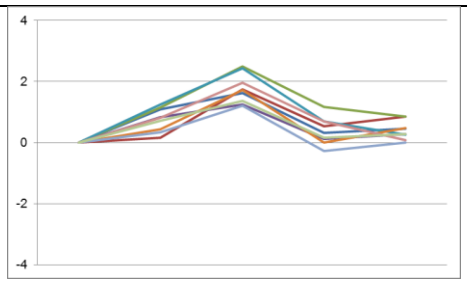
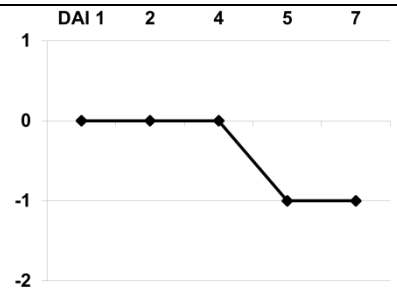
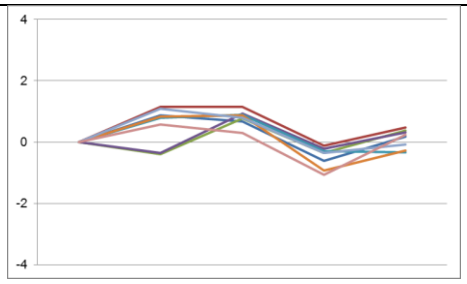
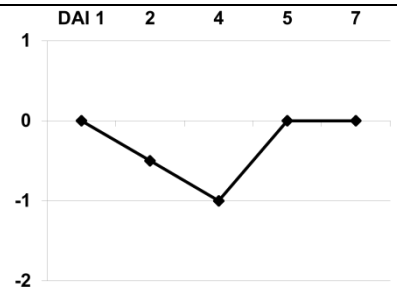
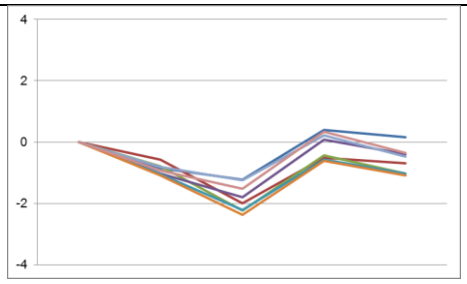
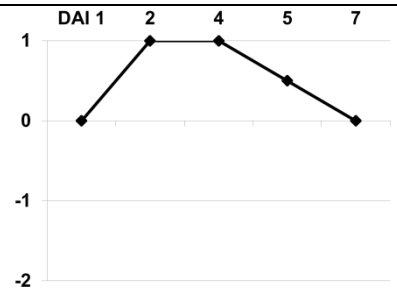
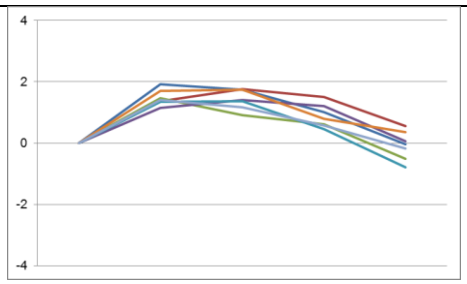
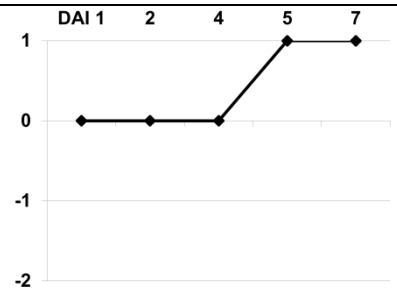
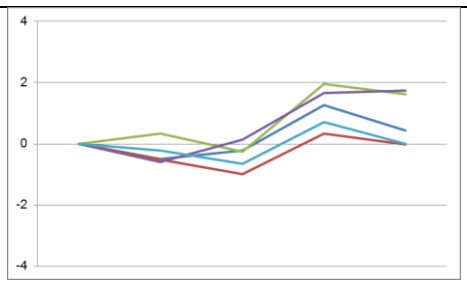


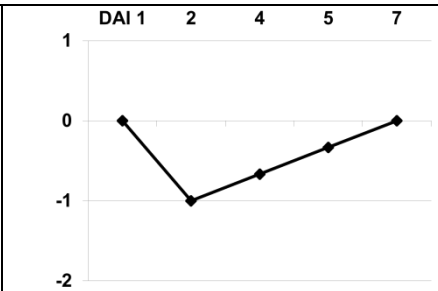
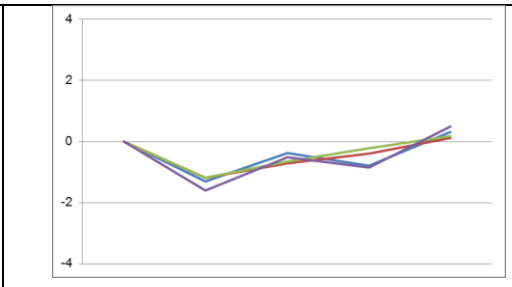
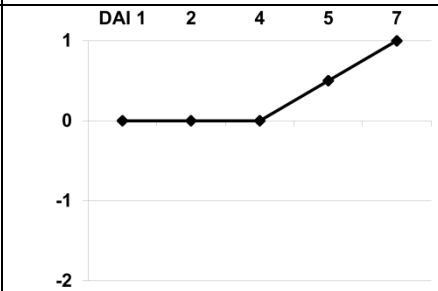
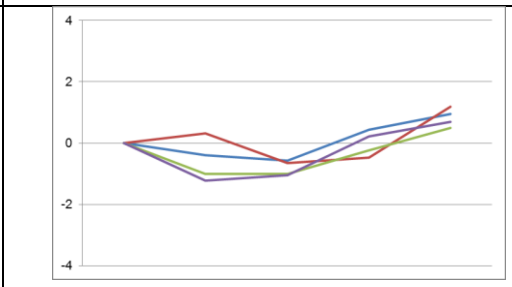
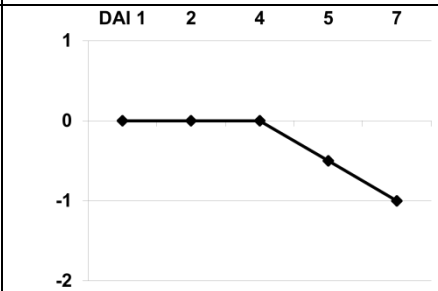
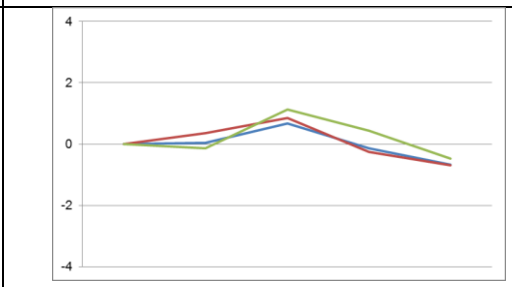
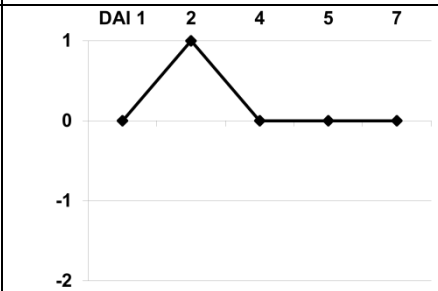
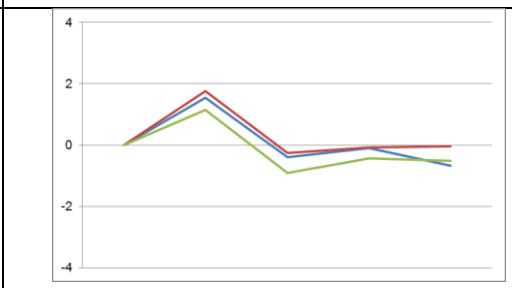
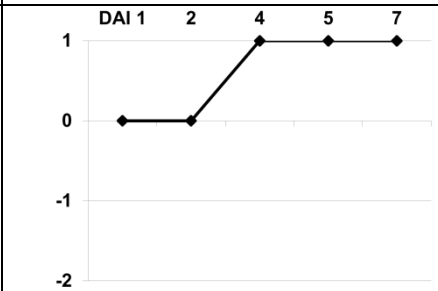
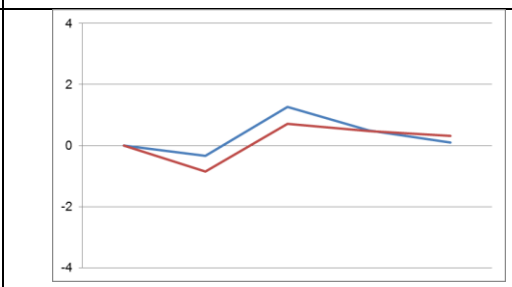
Figure 3-1. Differentially expressed genes (DEGs) identified in approach 2. Of the 2,131 DEGs identified for *M. hapla*, 989 were sorted into 20 groups (Column1: symbols represent our expression group patterns; Column2: graphical representation of column1; Column3: RPKM expression profiles classified into the same pattern group).

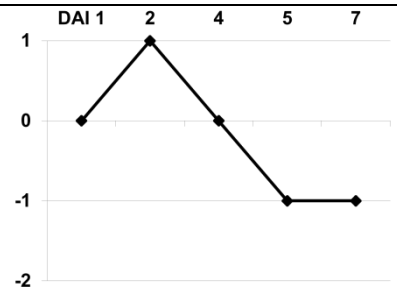
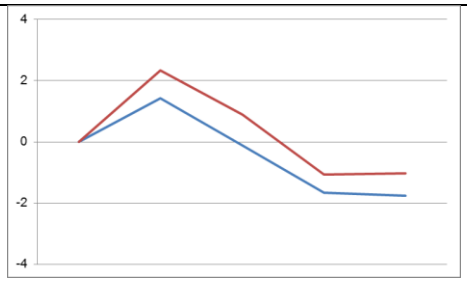
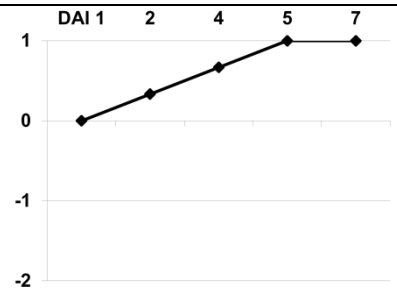
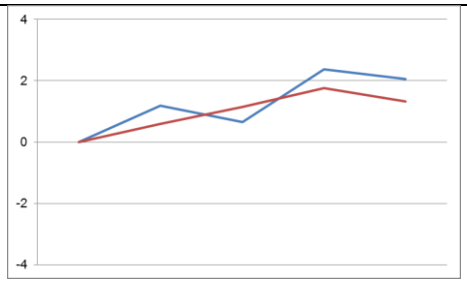
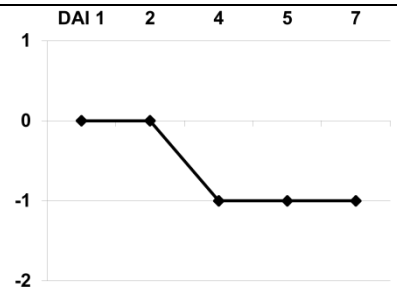
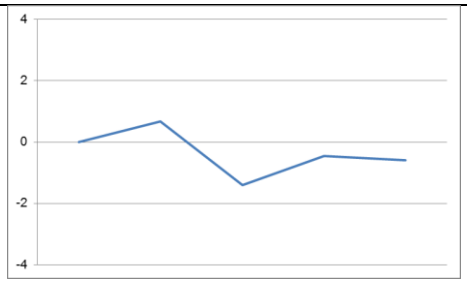
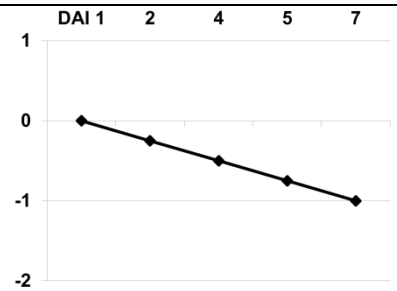
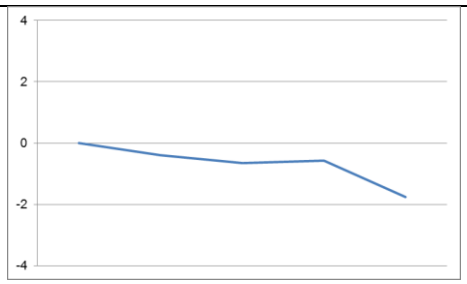
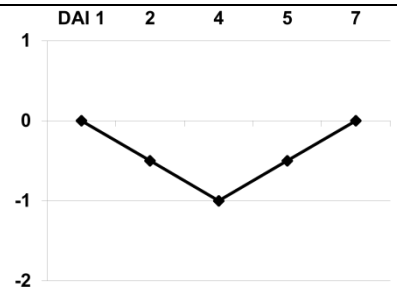
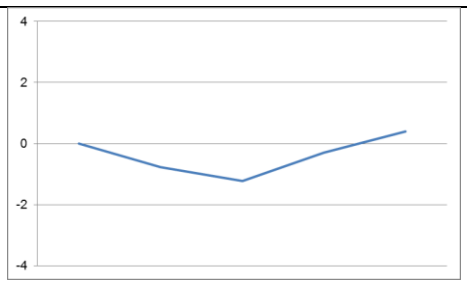
Groups		Gene Expression Profiles
UDO_		
D___		
_DOO		
U_D_		
U___		

DO__		
_UO_		
_DO_		
UDOO		
D_D_		

_UOO		
UO__		
DUO_		
DOO_		
D_U_		

UOD_		
__D_		
DOU_		
U_DO		
__U_		

DUOO		
__UO		
__DO		
UD__		
_U__		

UDD_		
UOO_		
_D__		
DOOO		
DOUO		

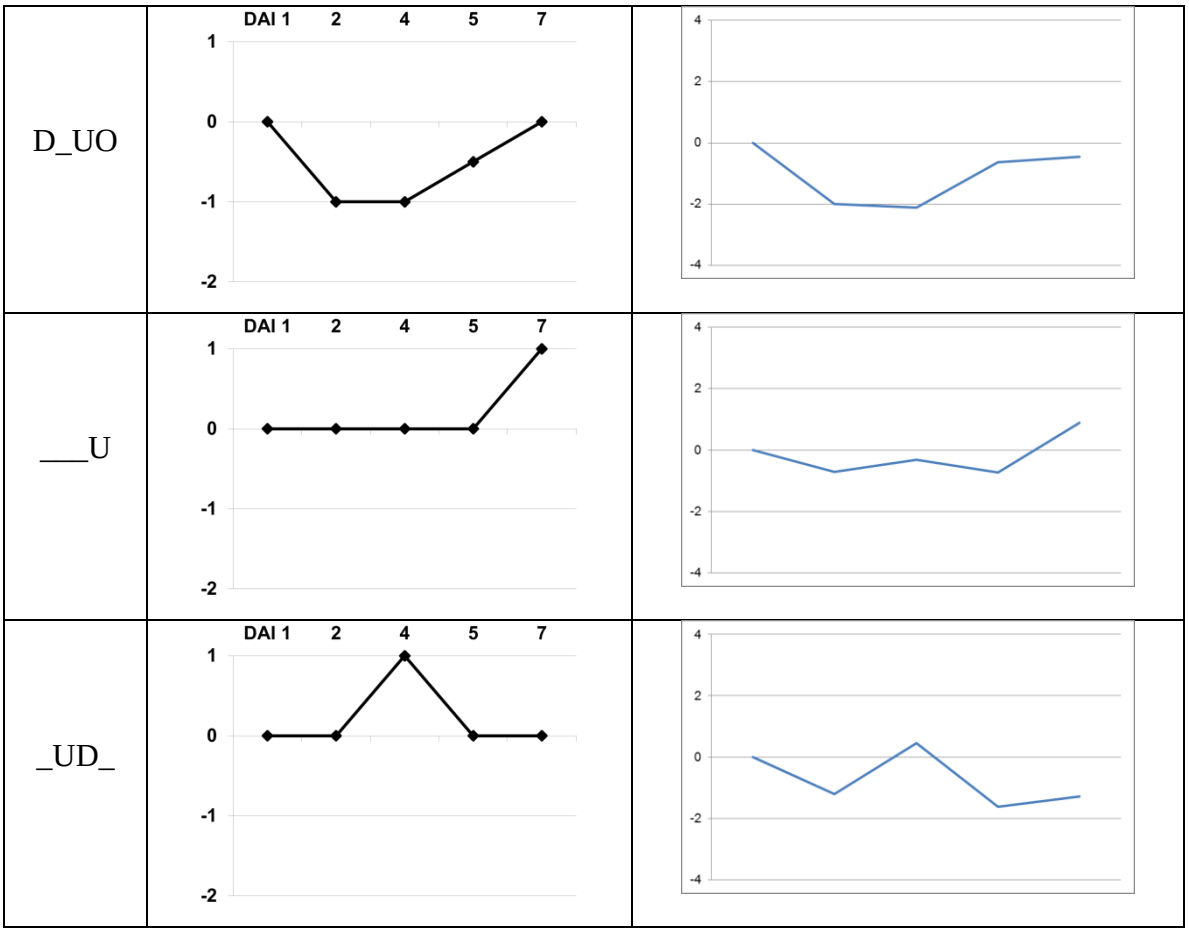


Figure 3-2. Of all the possible 153 groups for *M. truncatula*, 699 DEGs in infected roots identified over time-courses were classified into 33 groups (Column1: symbols represent our expression group patterns; Column2: graphical representation of column1; Column3: RPKM expression profiles of genes classified into the same pattern group).

Appendix B

M. hapla Contig952	Predicted Function	*M. chitwoodi* Contig241648
frz3.gene12		g18771
frz3.gene13 (9)	protein kinase, **GCK-3**	g18772 (11)
frz3.gene14 (9)	**SNX-6**	g18773 (9)
frz3.gene15		g18774
frz3.gene16	WD40 repeat	g18775
frz3.gene17 (2)	hypothetical protein	g18776 (18)
frz3.gene18 (5)	hypothetical protein	
frz3.gene19 (12)	protein kinase, **PAK-1**	
frz3.gene20 (24)	Rpb1, **AMA-1**	g18777 (27)
CUFF.12.1_5	WD40-repeat	g18778
frz3.gene23		g18779
frz3.gene24		g18780
frz3.gene25		g18781
frz3.gene26	Fork head domain	

Table 1. Genes within the synteny between *M. hapla* contig952 and *M. chitwoodi* contig 241648. This syntenic contig was predicted to possess 13 genes for *M. hapla* and 11 genes for *M. chitwoodi*, with the three *M. hapla* genes corresponding to one *M. chitwoodi* gene. The number of CDS is indicated by the numbers in parenthesis.

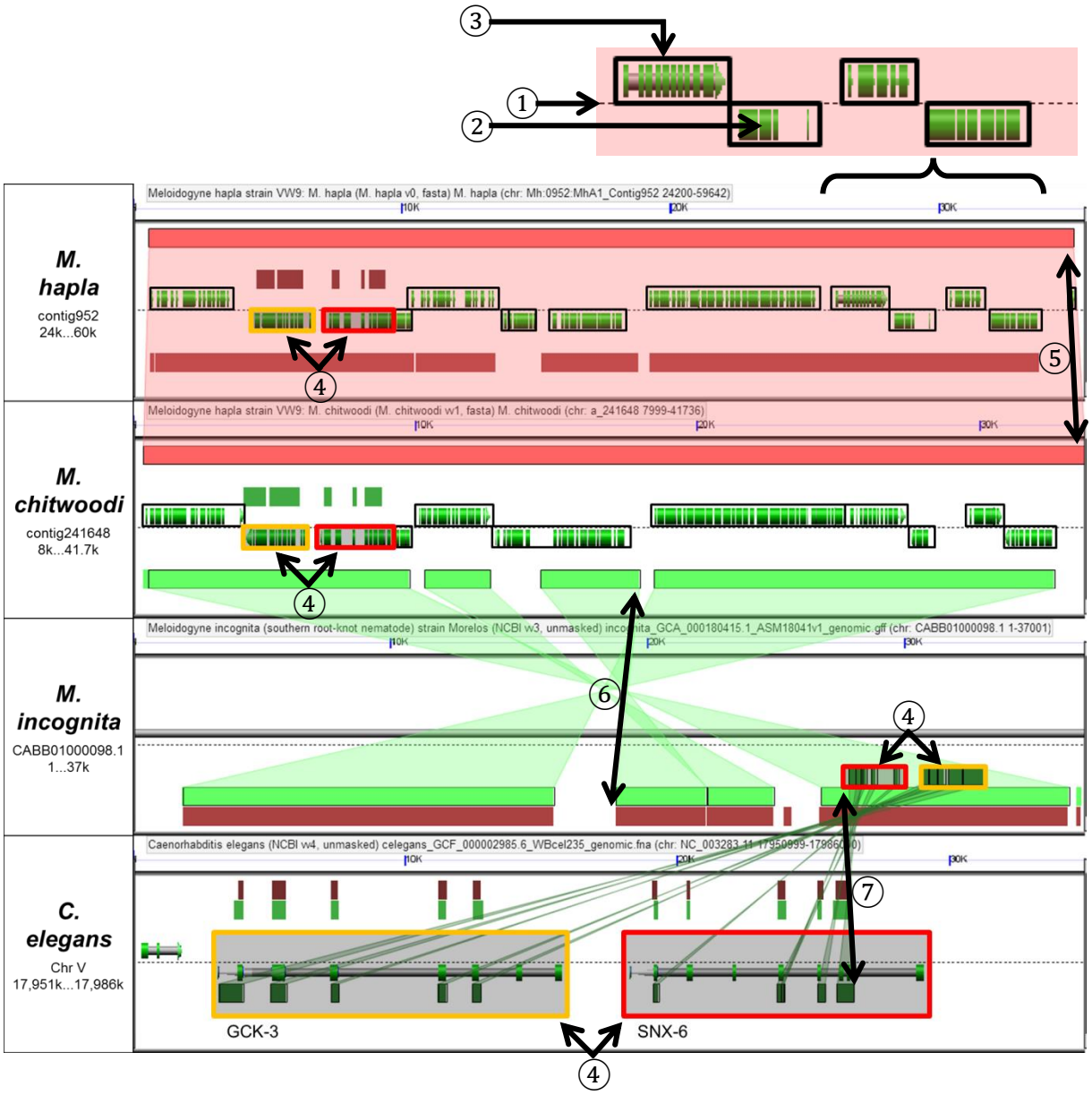


Figure 1 (A). Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-gck-3* and *Cel-snx-6* were aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008), a web-based tool that compares local chromosomal regions of multiple organisms. In each horizontal figure which represents a user-specified contig region of one organism, the dashed line (①) in the middle divides the strand into two orientations, the positive strand above the line and the negative strand below the line. Around the dividing line, each CDS is

represented by green segments (②) which are further tied by black outline (③) to indicate one gene partition. The gene conserved among RKN species and aligned with the *C. elegans* gene is outlined in red (④). Above or below the gene models, the regions identified by all the pairwise sequence comparisons of BLAST as conserved between two species are drawn with thick color-filled lines (⑤) which are further connected through translucent dimensions. For example, color box with deep pink (⑥) corresponds to the pairwise comparison between *M. hapla* and *M. chitwoodi*, whereas the light green (⑥) or deep green box (⑦) is for *M. hapla* and *M. incognita* or *M. incognita* and *C. elegans*, respectively. These color boxes all together implicate synteny. The figure legend attributes apply for all the figures from Figure 1 to 12.

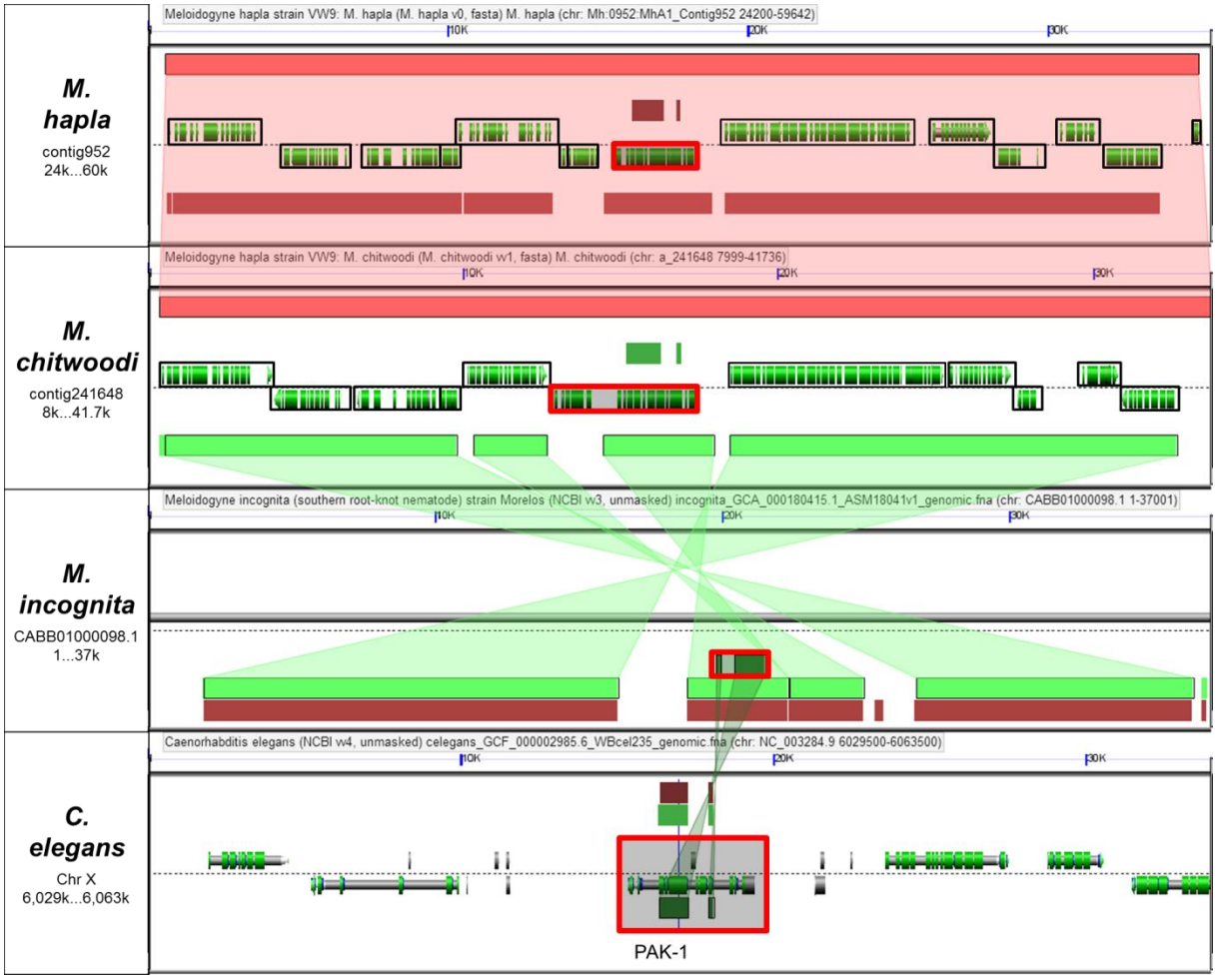


Figure 1 (B). Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-pak-1* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-pak-1* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green or deep green was for *M. chitwoodi* and *M. incognita* or *M. incognita* and *C. elegans*, respectively. These color boxes all together implicated synteny.

Figure 1 (C). Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-ama-1* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-ama-1* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green or deep green was for *M. chitwoodi* and *M. incognita* or *M. incognita* and *C. elegans*, respectively. These color boxes all together implicated synteny.

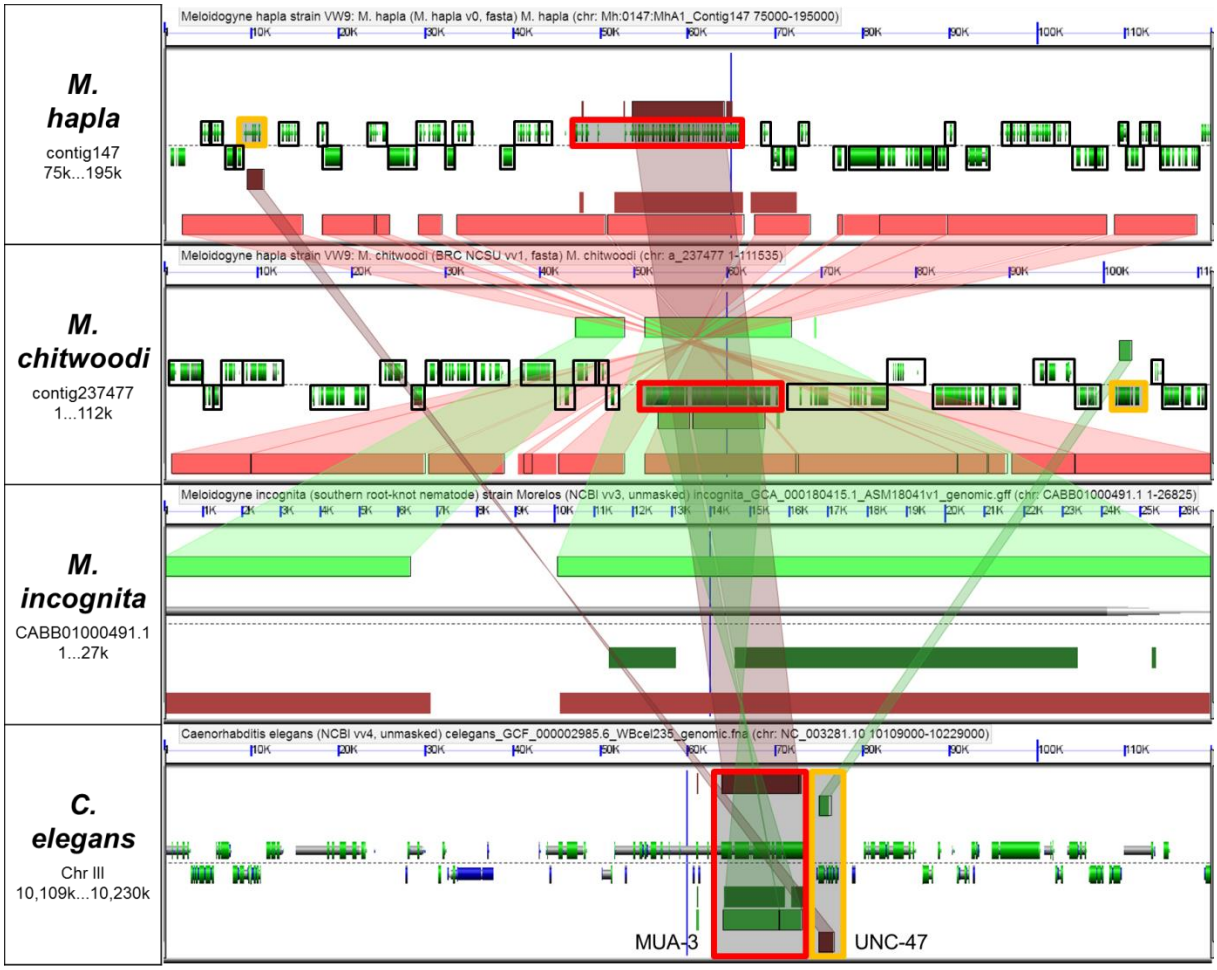


Figure 2. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-mua-3* and *Cel-unc-47* were aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The genes conserved among RKN species and aligned with *Cel-mua-3* and *Cel-unc-47* were outlined in red and yellow, respectively. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green, deep green, or dark brown was for *M. chitwoodi* and *M. incognita*, *M. incognita* and *C. elegans*, or *M. hapla* and *C. elegans*, respectively. These color boxes all together implicated synteny.

Figure 3 (A). Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-unc-70* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-unc-70* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the deep green or dark brown was for *M. chitwoodi* and *C. elegans*, or *M. hapla* and *C. elegans*, respectively. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

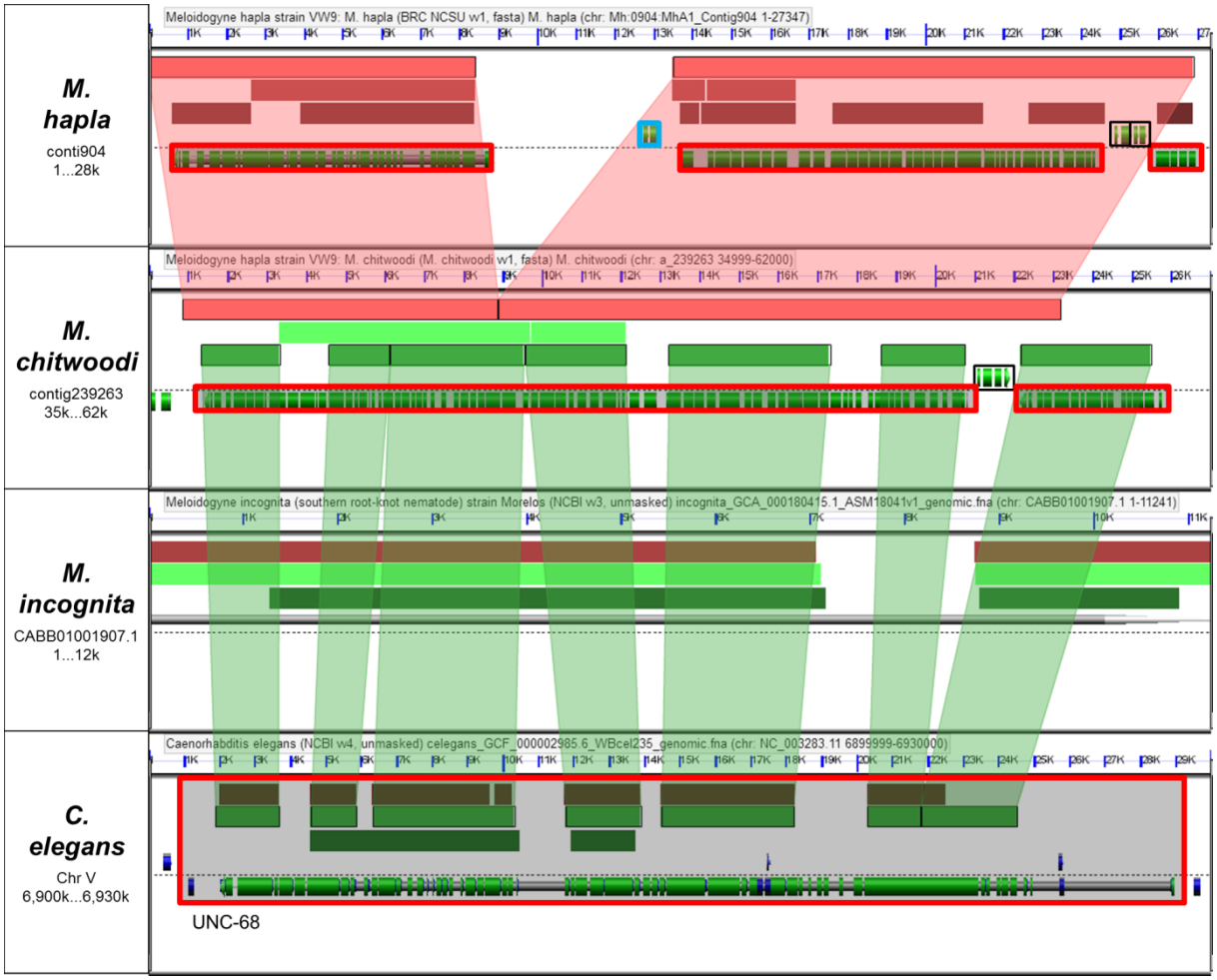


Figure 3 (B). Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-unc-68* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-unc-68* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the deep green was for *M. chitwoodi* and *C. elegans*. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

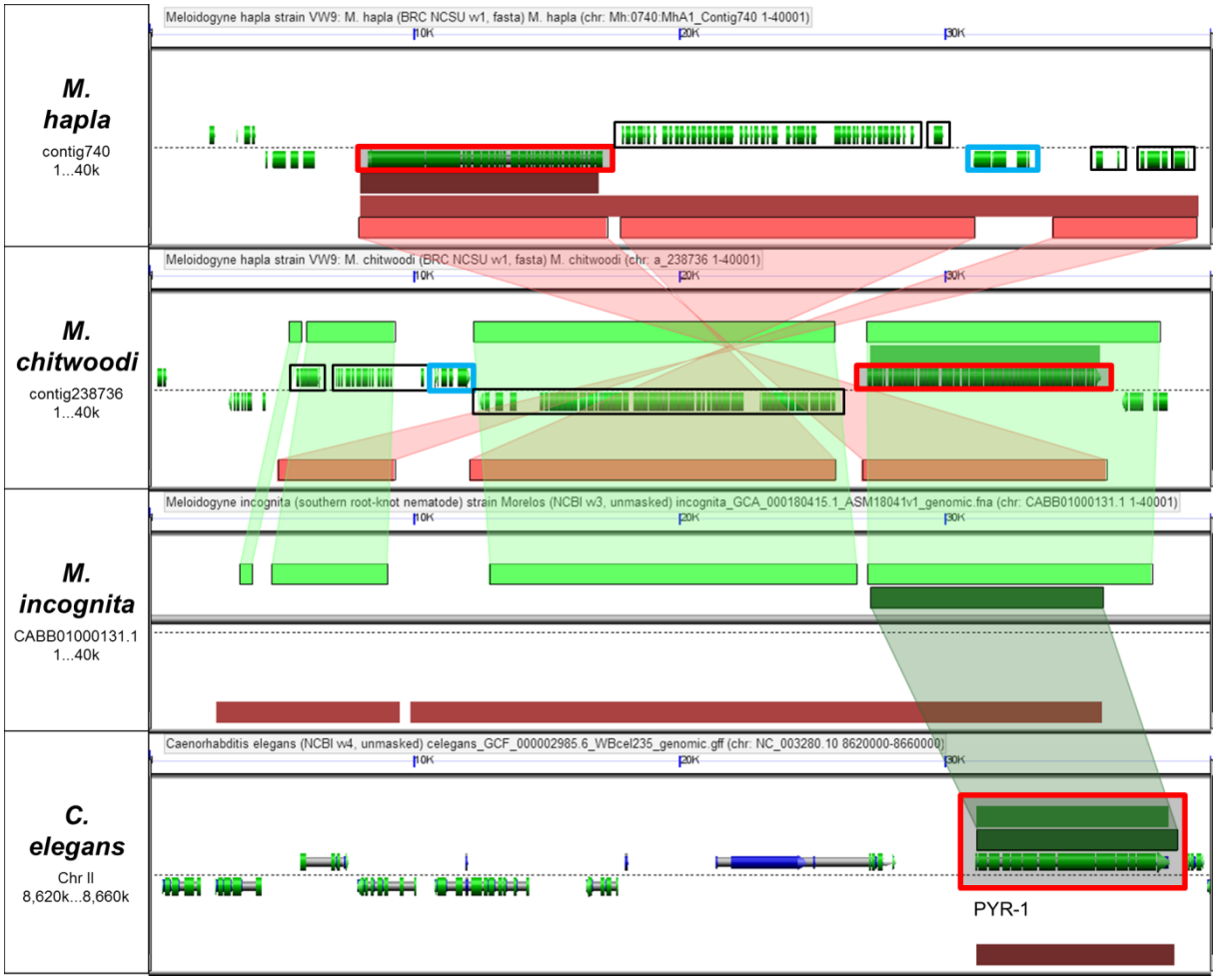


Figure 4 (A). Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-pyr-1* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-pyr-1* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green or deep green was for *M. chitwoodi* and *M. incognita*, or *M. incognita* and *C. elegans*, respectively. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

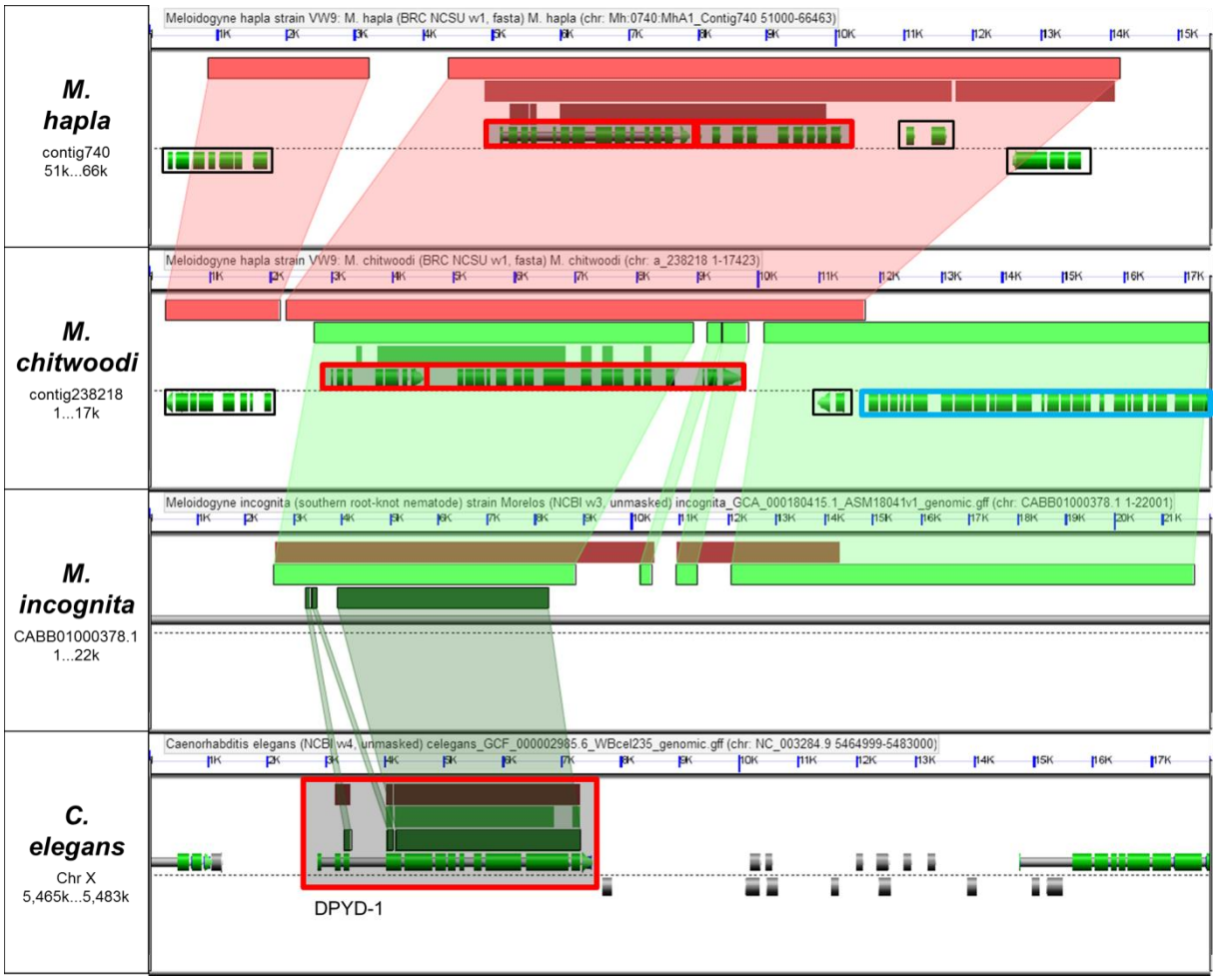


Figure 4 (B). Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-dpyd-1* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-dpyd-1* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green or deep green was for *M. chitwoodi* and *M. incognita*, or *M. incognita* and *C. elegans*, respectively. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

Figure 5. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-myo-1* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-myo-1* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green or deep green was for *M. chitwoodi* and *M. incognita*, or *M. incognita* and *C. elegans*, respectively. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

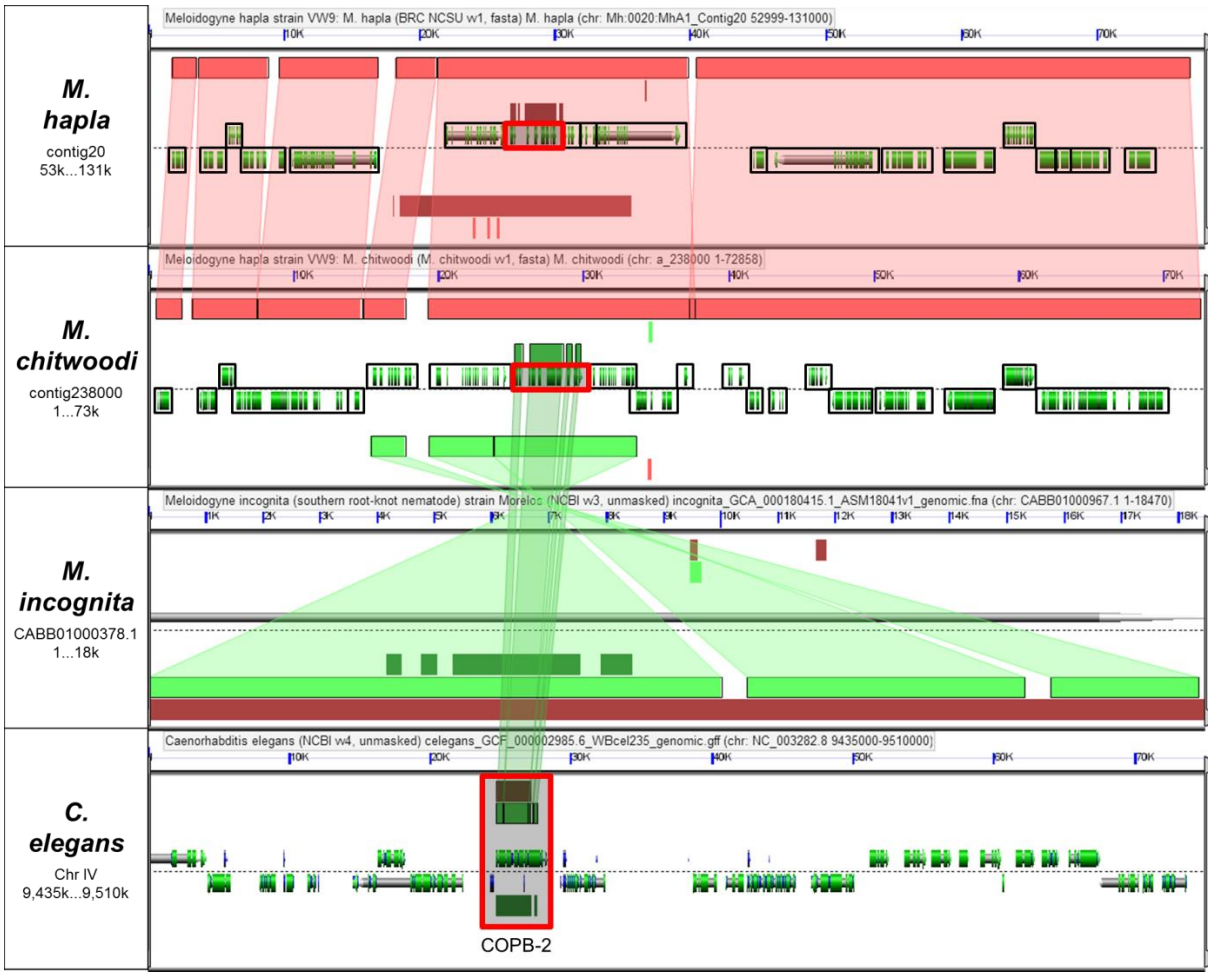


Figure 6. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-copb-2* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-copb-2* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green or deep green was for *M. chitwoodi* and *M. incognita*, or *M. chitwoodi* and *C. elegans*, respectively.

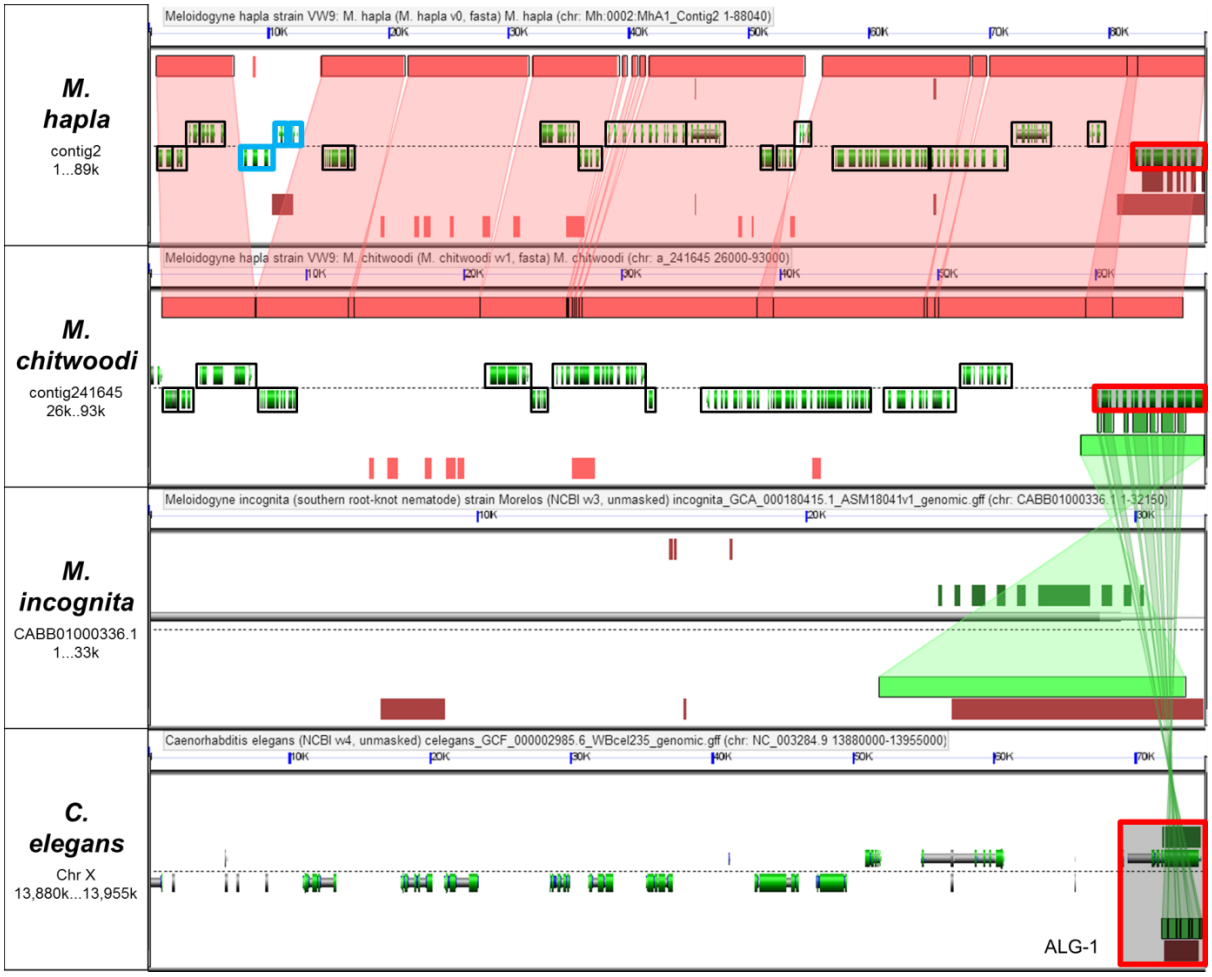


Figure 7. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The contig on which *Cel-alg-1* was aligned was syntenic between *M. hapla* and *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-alg-1* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green or deep green was for *M. chitwoodi* and *M. incognita*, or *M. chitwoodi* and *C. elegans*, respectively. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

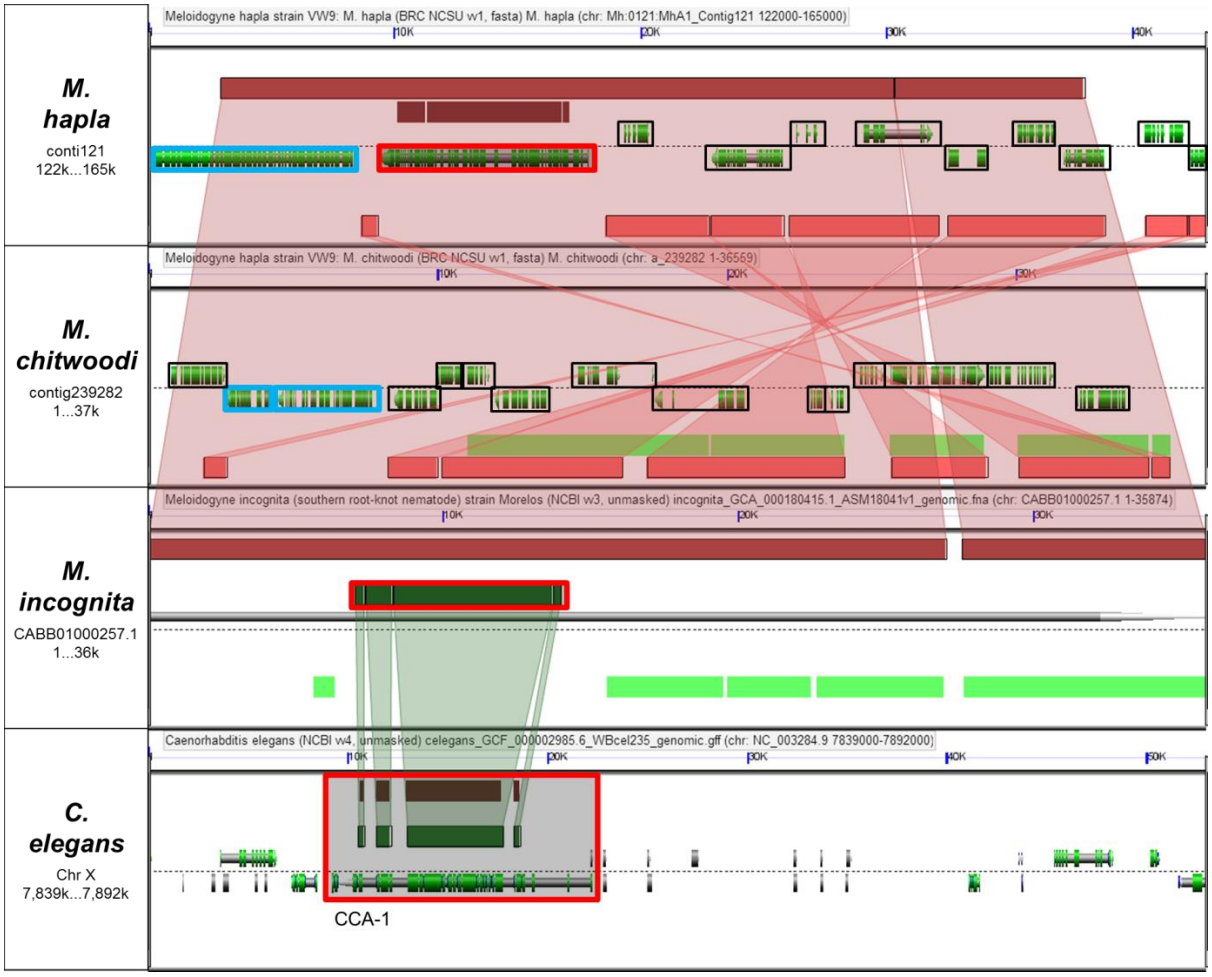


Figure 8. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom) on which *Cel-cca-1* was aligned. The *M. hapla* contig was more conserved with *M. incognita* whereas it was partially and reversely conserved with *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-cca-1* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the deep green or dark brown was for *M. incognita* and *C. elegans*, or *M. hapla* and *M. incognita*, respectively. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

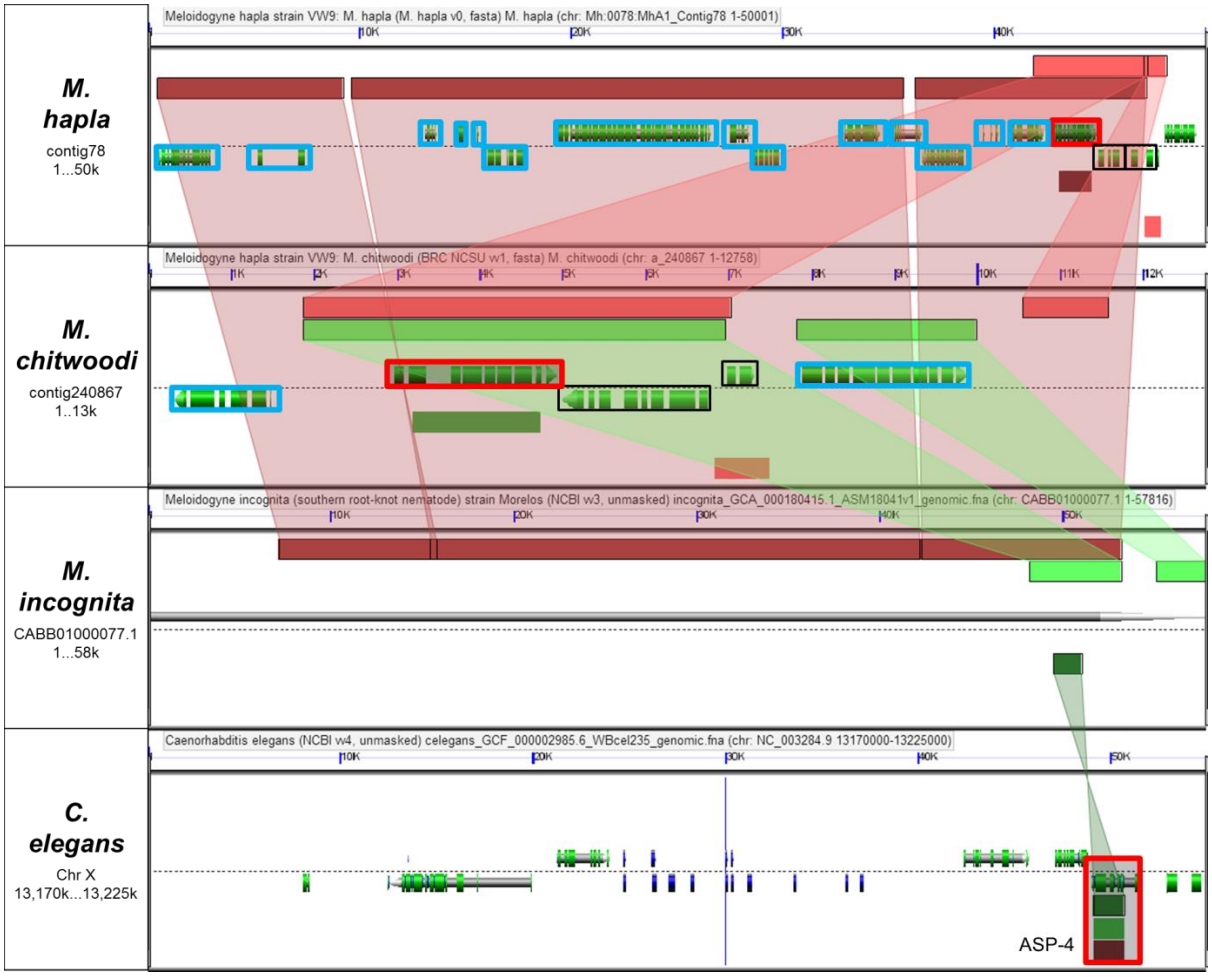


Figure 9. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom) on which *Cel-asp-4* was aligned. The *M. hapla* contig was more conserved with *M. incognita* whereas it was partially conserved with *M. chitwoodi*. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-asp-4* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the dark brown, light green or deep green was for *M. hapla* and *M. incognita*, and *M. chitwoodi* and *M. incognita*, or *M. incognita* and *C. elegans*, respectively. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

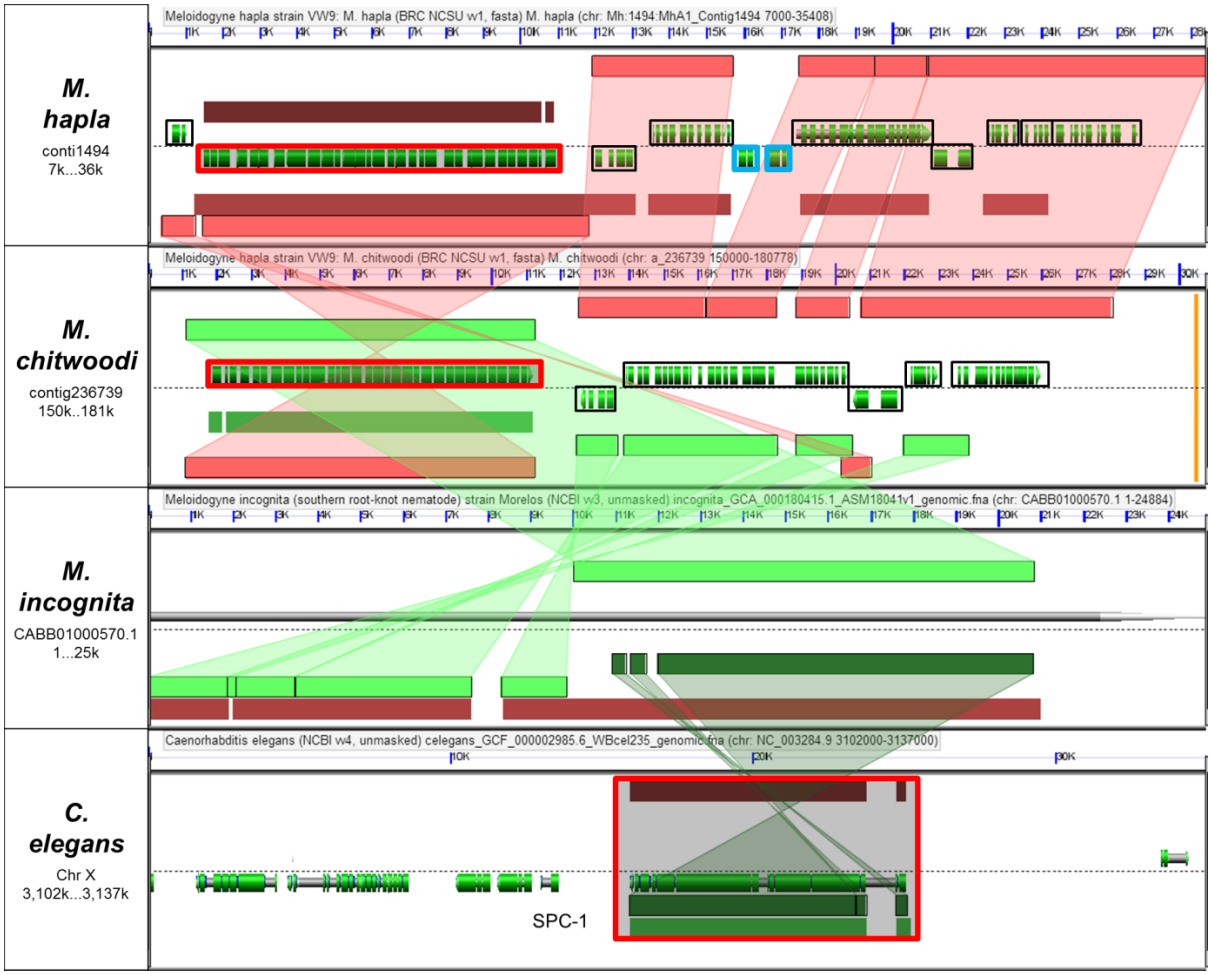


Figure 10. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom) on which *Cel-spc-1* was aligned. The region was conserved between *M. hapla* and *M. chitwoodi*, the only exceptions being the two gene insertions and one gene reversion on the *M. hapla* contig. The graphical visualization was performed using CoGe (Lyons et al. 2008). The gene conserved among RKN species and aligned with *Cel-spc-1* was outlined in red. The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the light green or deep green was for *M. chitwoodi* and *M. incognita*, or *M. incognita* and *C. elegans*, respectively. These color boxes all together implicated synteny. Gene insertion was outlined in blue box.

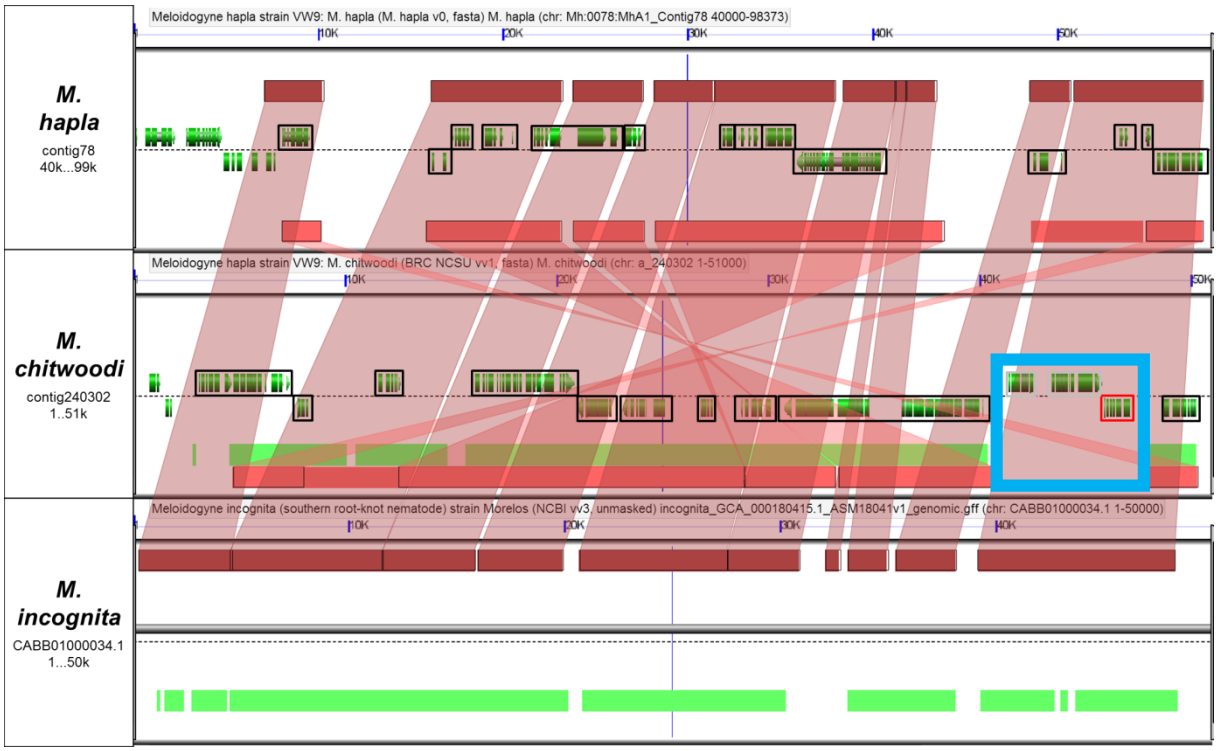


Figure 11. Syntenic region of *M. hapla*, *M. chitwoodi*, *M. incognita*, and *C. elegans* (top to bottom). The *M. hapla* contig was reversely conserved with *M. chitwoodi*. The region of M. chitwoodi outlined in blue box was syntenic with the *M. hapla* contig 1040. The graphical visualization was performed using CoGe (Lyons et al. 2008). The regions identified by all the pairwise sequence comparisons of BLAST as conserved between *M. hapla* and *M. chitwoodi* were drawn with pink color-filled lines whereas the dark brown was for *M. hapla* and *M. incognita*. These color boxes all together implicated synteny.

Figure 12. Alignment of translated sequences of cytochrome c1 (KOG3052) among different species, including *M. hapla*, *M. chitwoodi*, *Drosophila melanogaster*, *Arabidopsis thaliana*, human, *Schizosaccharomyces pombe*, and *yeast*.

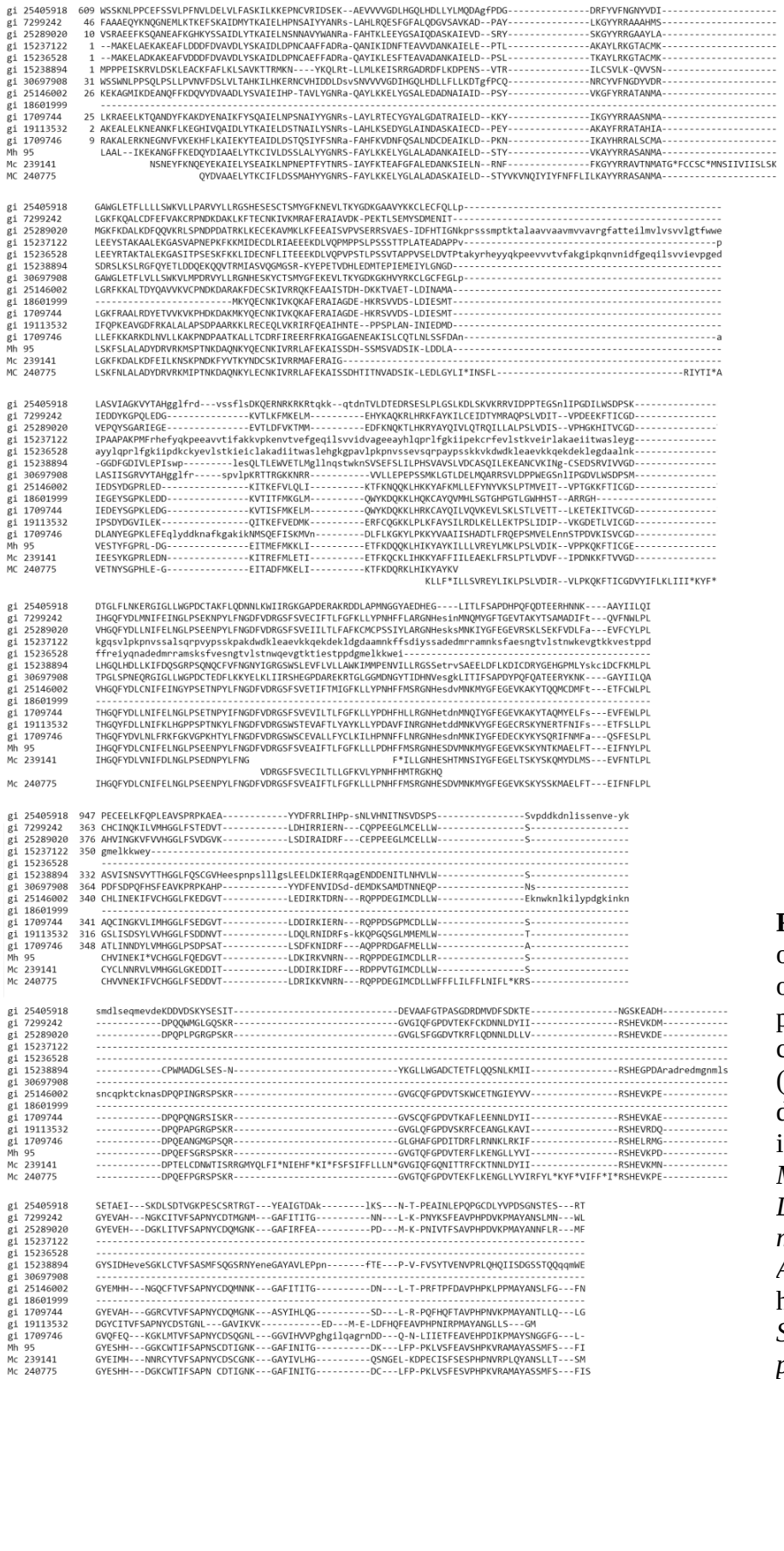


Figure 13. Alignment of translated sequences of serine-threonine phosphatase 2A, catalytic subunit (KOG0376) among different species, including *M. hapla*, *M. chitwoodi*, *Drosophila melanogaster*, *Arabidopsis thaliana*, human, *Schizosaccharomyces pombe*, and *yeast*.

```
Mh FPDVFKNVIILTNKVLDFARPVALFEFANNHSMDHPEKEDAFLQRIDDKAGEIGLDKEFARLFFIDQINASKVVQASKILL
Mc FSEIFKKVLILTNQVLDFSRPVALFEFANNHSIDHPDKEDAFIQRIEAKANEIGLDKEFARLYFIDQIKASKIVQVNNLNL
Ma          IILVDNVLDFSRPVALFEYANNHPIDHPDKEAAFLERIKATAVEIGLDTEFARLFFADQINASKVVQSA
Gp        IVGVANKRLMLAKDVALYKYINNNSIDDFEREKVVLQNVLAQANSAGISDNYGEPFFQDQMDANKVIQVKRFRL

Mh KLFRLAAYFNLWNKNGMPNEQVVDIHTTEFQTNMGKVTMNLAEALLPIVKYRDS
Mc KFFI*KAYFAVWNKSGLPNEPVIDIHTTEFQTKMGNVTMSLAEALLPFVKYRDN
Ma SAYFALWNRTGLPNETVVDIHTTEFQSNMGKVTMDLEKALLPIVKYRD

Mh KIDVSSEFKRDKAFGVALNHLCSNNPKYK*
Mc QIDICEEYKRNEGFKLAISHLCSNNP*
Ma KIDISCEFNRDKAFVFALRHLCSNRIFYN*
```

Figure 14. Alignment of translated sequences of chorismate mutase among M. hapla, *M. chitwoodi*, *M. arenaria* and *G. pallida*.