Abstract

SCHOLL, ELIZABETH HOFFMANN. Molecular Evolution and Horizontal Gene Transfer in *Meloidogyne* spp. (Co-Advisors: David McK. Bird and Jeffrey L. Thorne)

Most phylogenetic analyses of plant-parasitic nematodes have been based on a small number of character sets and have provided little insight into the evolution of parasitism within the species analyzed. A strategy was designed to recover the most robust phylogeny for five *Meloidogyne* species (*M. arenaria*, *M. chitwoodi*, *M. hapla*, *M. incognita*, and *M. javanica*), three closely related taxa (*Heterodera glycines*, *Globodera pallida* and *G. rostochiensis*) and the more distant taxon, *Caenorhabditis elegans*. This multiple-gene approach is based on sampling more than 80,000 tylenchid sequences present in public databases. Forty-seven genes were identified which unambiguously could be assigned as orthologues, and an alignment performed, so that all 47 genes could be concatenated to create one multi-gene alignment. Results from Bayesian analyses place *M. incognita* and *M. javanica* as sister taxa, with *M. arenaria* basal to these. Placement of the other taxa is congruent with previous studies.

A method for a high-throughput genome screen for horizontally acquired genes is further presented, illustrated using expressed sequence tag (EST) data from *M. incognita*, *M. javanica* and *M. arenaria*. Applying a phylogenetic filter to a series of homology searches revealed previously postulated horizontally transferred genes and six new candidates. Computational and experimental methods verified the horizontal gene transfer candidates as bona fide nematode genes. Phylogenetic analysis implicated rhizobial ancestors as donors of horizontally acquired genes in *Meloidogyne*. Analysis of these horizontally transferred gene candidates suggests a link between horizontally transferred genes in *Meloidogyne* and parasitism.

MOLECULAR EVOLUTION AND

HORIZONTAL GENE TRANSFER

IN *MELOIDOGYNE* SPP.

by

Elizabeth Hoffmann Scholl

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

BIOMATHEMATICS

Raleigh

2004

APPROVED BY:

Co- Chair of Advisory Commitee	Co- Chair of Advisory Commitee

Dedication

To the three people whose love will always be with me:

Mom, Dad and Reinhard.

I wouldn't have made it without each of you, and I know you're all proud of me.

Thank you.

Personal History

The story starts one evening in the middle of a southern California heat wave when, much to Penny and Bob's surprise, Caleb Robert Scholl arrived as a girl. Realizing that Caleb Robert was no longer an ideal name for their third child, Penny decided the girl should be called Betsy. Bob disagreed. He is reported to have said "When my daughter becomes the first female President of the United States, she will have a *real* name!" They both got their way – the girl was named Elizabeth Hoffmann Scholl, but the world was to know her as Betsy. Betsy's family left California when she was two and returned to their home in Setauket, New York. It is the North Shore of Long Island that Betsy calls "home".

After exploring her educational and extracurricular options in High School, she applied for early acceptance at Cornell University. She was accepted. Depending on which friend you asked, it was a sure thing she was going to study theater… or hotel and restaurant management… or science. But being Betsy, she did the unexpected – she went to school to study Agricultural Economics.

As a freshman, she had trouble staying awake during Economics 101. That was when she knew a change of major was in store.

Her brother Adam, veteran of a few of colleges himself, suggested she try Statistics. Not quite knowing what Statistics was, other than that it had to do with math, Betsy took his advice, and thus started on her true academic path. It was during her senior year at Cornell that Professor Gary Churchill recommended a Population Genetics class. This was when Betsy discovered she could combine math and genetics into a single field of study, setting her up for another 10 years of schooling.

Staying true to her New Yorker self, Betsy went back to Long Island to attend SUNY Stony Brook for graduate school. Being assured she could combine math and biology with no problem, she went into the Applied Mathematics and Statistics (AMS) Program.

She also got a job designing databases for the RHIC project at Brookhaven National Lab. It was in those years that she learned AMS students cannot get academic credit for Genetics classes. She also learned how to tie the correct knot to keep the dinghy from coming loose from the back of the sailboat. She took a Master's degree and, for the first time in her life, voluntarily moved out of New York and down to North Carolina. She started her time at NC State modeling idiotypic networks in the human immune system and working full-time as a webmaster for the North Carolina State Center for Health Statistics.

It was during this time that there were many comings and goings. On New Years Eve, 1998, her father left her world. Three months after that, Reinhard entered it.

In the spring of 2000 Betsy learned how to file a K-1 visa petition to ask the US Government for permission to marry Reinhard Höpperger. Two days after Betsy's 29th birthday, Reinhard left the paradise of Bad Ischl, Austria to make her the most thankful woman alive. They were married soon after (within 90 days, as mandated by the INS).

Comings and goings were happening in more than her personal life. Also in 2000 Betsy's academic advisor decided to leave North Carolina. It was then that Drs. Jeffrey Thorne and David Bird agreed to take her under their wings.

For the next three years Betsy learned a lot about evolution and bioinformatics, picked up a bit of biology, and was repeatedly informed that Wisconsin cheese is better than New York cheddar. Betsy has also learned to love her research so much that she needed to be told to stop working and start writing. And from that comes this dissertation.

It's not the Presidency of the United States, but she's sure her Dad would still be proud to see the name Elizabeth Hoffmann Scholl on these pages.

Acknowledgements

I would not have been able to do any of this without my co-advisors, Jeffrey Thorne and David Bird. The two of you took me on and let me start over. With the right combination of patience, care, and pushing you have seen me through to the end. Beyond learning in the academic sense, I have learned to love what I do. I will always be grateful. To the other members of my committee, Brian Wiegmann and Charles Smith, thank you for staying with me through my switch in advisors and in projects. I know I ended up doing something very different than what I started with, and I appreciate your understanding that life works out that way sometimes. I also want to thank Brian for the weekly discussion group. I learned as much there as anywhere and the interaction with a larger group helped expand my horizons.

Also thanks to the NSF, the agency that funded this research through grant DBI-0077503.

Nothing can compare to the intellectual and friendly interactions I've had in my time here. To the Thorne Group, Stéphane Aris-Brosou, Doug Robinson,Tae-Kun Seo and Jiaye Yu, thank you for sharing ideas, reading manuscripts, helping me over rough spots and explaining things I didn't understand. Special thanks to Stéphane who has lately been a mentor to me as well. And a big thanks to Doug for helping me keep what little sanity I had left, even from afar, and helping me appreciate the benefits of losing the rest of my mind. I hope you do take over the world some day; you'd only make it a better place. Also at the BRC I need to give a special thanks to Debbie Hibbard. Your door was always open, as was your heart. Our conversations let me exercise other parts of my brain

and more importantly my soul. I hope you find whatever it is you want in this world. You deserve it all and then some.

To everyone in the Center for the Biology of Nematode Parasitism, thank you for your kindness, the office space, and for explaining what you were doing in the lab whenever I asked. I have great respect for your work, and am glad to have been able to witness some of it. Thank you for making me a part of the lab, even though what I do is foreign to you. What I did would not have been possible without the help of Mark Burke, who maintains the most amazing computer system to which I've ever had access. Your knowledge of Bioinformatics tools is nothing short of incredible. Thanks for making all of it available to me. Thanks especially to Jenn Schaff for being a good friend and answering a *lot* of questions, and to Reenah Schaffer, just for being unique.

There are people who are not at NC State who have also been wonderful to work with. I have to thank Keith Davies for getting me all worked up about research that isn't even in this dissertation. You may say you don't know anything, but you know how to get excited about new discoveries, and how to share that excitement with others. Jim McCarter has always been more than happy to read manuscripts and send useful advice and comments back quickly. Jim and the crew at WashU are also the people who keep nematode.net up and running – thanks for the access to the data. And the University of Tokyo is lucky to have Hirohisa Kishino, whose hospitality and guidance while I was in Japan is very appreciated. Anyone who has a chance to learn from you is being given a true gift. Thank you.

I also would not have made it this far if not for the love of my family. To my Dad, who taught me so much about life and what's important, I miss you and I hope you can

see what I've accomplished. Mom, you're the strongest woman I've ever met, and when things get tough, I think of your strength and it helps me through. You have not just my love, but also my admiration. And most importantly to Reinhard, who left his country to come live in a strange land so I could finish my degree and marry my love, and who never gave up on me even on the days when I felt like giving up on myself – I thank you endlessly.

I am grateful for the support of the rest of my family as well. My surrogate parents, Aunt Judy and Uncle Sid, saw me through some of the toughest times in my life, and without them I wouldn't be where I am today. And my brothers, Adam and Matt, and their wives, Lisa and Alyssa, have given me the most precious gifts of all – Abby, Julia and Emma. As they grow so fast, I can see each one of them is smarter than me in their own way. Especially Abby, who explained so well to Julia that when Aunt Betsy goes back to school she does not go to Kindergarten. (Aunt Betsy is already in the first grade.) My mother-in-law, Margaretha, has always told me how "stolz" she is and how "brav" I am. And to Bert and Phil - you may have been gone for a while now, but it is the humor I inherited from you that gets me through every day. You'll always live in my smiles.

Finally thank you to the people of Sumatra for their wonderful coffee. Without them my eyes would have had trouble opening in the morning and staying open throughout the day. Keith, Mick, Charlie and Ronnie, thanks for your little blues band. It's brought me friends, it's brought me the soundtrack of my life, and it's brought me to the front row and beyond.

A dissertation cannot be completed without a lot of help and support. To those I forgot to mention by name, know I haven't forgotten you in my heart.

Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Review

Introduction

Nematodes are the most abundant and speciose metazoans, comprising an estimated 80% - 90% of all animals (Boucher and Lambshead, 1994). They occupy diverse ecological niches and employ a variety of lifestyle strategies, from the free-living bacteriovores to obligate parasites of most plants and animals. The impact of the parasites is felt through such widespread problems as human disease, debilitation of livestock and crop damage. Plant-parasitic nematodes are responsible for an estimated $100 billion in annual crop damage worldwide (Koenning et al., 1999). The root-knot nematode (*Meloidogyne* spp.) is a member of the family Heteroderidae, which are responsible for the largest proportion of crop loss by the plant-parasites. Despite their obvious economic importance much still needs to be learned about the biology of parasitism in *Meloidogyne* and the responses elicited in their hosts.

In order to understand the origin of parasitism, it is essential to start with a general evolutionary framework. Without understanding how the orders of nematodes are inter-related, we cannot understand when, where, or how often parasitism evolved in the history of the phylum. Likewise, resolving a species phylogeny can lead to new insights regarding commonalities and differences in host specificity and host response and enhance our ability to identify targets for control.

I present here a brief review of the history of nematode phylogenetics. The difficulties facing the proper placement of the phylum Nematoda into the higher-level metazoan tree will be discussed, as will the relationships established between the nematode orders based on morphology and how these differ from molecular analyses.

I also consider the discovery of putative horizontally acquired genes in Heteroderidae, and what the implications of these acquisitions are relative to the evolution of parasitism. I will examine some of the traditional methods for detecting horizontally transferred genes and discuss the issues with these methods, especially related to eukaryotic involvement in those transfers.

Placement of the Phylum Nematoda

In the early 19th century, the animal kingdom was divided on the basis of shared body plan, or morphology, into bifurcating groups called phyla (Cuvier 1817, as cited in Adoutte et al., 1999). The sorting of the metazoa into their different phyla had been generally easy to accomplish. However, almost 200 years after this first attempt at classification, the precise relationship between the phyla is still not fully resolved. Considerable disagreement has centered on the proper placement of the unsegmented round worms, or phylum Nematoda.

Traditional classifications of metazoan phylogeny were organized to reflect a general increase in body complexity as evolution progressed. This placed extant simple animals near the base of the tree. The most basal division was between sponges, which have loose cell organization and no nerve cells, and all other animals. The more complex metazoa were then further divided into two groups, based on the number of germ layers exhibited by the taxa within those groups. The Cnidarians, or diploblasts, have two germ layers; the ectoderm, which is external, and the internal endoderm. The Triploblasts, on the other hand, contain a third layer, the mesoderm, which lies between the other two and gives rise to the development of muscles and internal organs. The Cnidarians and the

Triploblasts primarily follow demarcations of radial and bilateral symmetry, respectively. Therefore the Triploblasts are commonly referred to as Bilateria.

The Triploblasts were further divided into the acoelomates, comprised primarily of the flatworms and characterized by a mesoderm filled with loosely organized cells, and the coelomates, which have a hollow mesoderm. However, there exist organisms that fit into neither category. Termed the pseudocoelomates, these animals lack a true coelom, yet still displayed some internal cavities. These pseudocoelomates, which included the Nematoda, are sometimes designated as a separate super-phylum and placed in an evolutionary intermediate position between the acoelomates and those Bilateria with a true coelom (Marshall and Williams, 1974). This traditional understanding of metazoan phylogeny was fully established by 1940 and the hypothesis of progressive increase in body complexity over evolutionary time is still taught today (Adoutte et al., 1999).

It wasn't until 1988 that molecular phylogenetic techniques were applied to metazoan phylogeny (Field et al., 1988). The results, based on partial 18S rRNA sequences from 20 taxa distributed across 10 phyla, largely agreed with the morphology-based phylogenies previously published. However, some incongruencies existed with regards to the coelomates. For centuries, arthropods and annelids shared membership in the Articulata, a clade of animals with segmented body plans. Molecular results dissolved this clade and the annelids were placed with the mollusks. Reanalysis of this newly-defined relationship has been further confirmed in a more detailed examination of spiralian metazoan morphology (Eernisse, 1992).

The placement of the pseudocoelomates within the Bilateria was similarly redefined. Traditionally the pseudocoelomates could not be firmly established within one or the

other of the Bilateria super-phyla due to the ambiguity of observed morphological characteristics. Although a large proportion of the pseudocoelomates are nematodes, it has been suggested that the acoelomate condition is ancestral in nematodes (Winnepenninckx et al., 1995). If this were the case, the pseudocoelomate super-phyla would be monophyletic. However, another analysis of an 18S rRNA dataset suggested the pseudocoelomates were polyphyletic, falling into three separate clades (Winnepenninckx et al., 1995). The Priapulida became associated with the arthropods, the Gastrotrichia and Platyhelminthes formed their own clade, and the Nematoda were no longer paired with the flatworms in a single super-taxa, but were placed as a separate lineage basal to the Bilateria (Winnepenninckx et al., 1995). An analysis of cytochrome c and globin amino acid sequences resulted in a similar placement of the Nematoda as a basal metazoan lineage (Vanfleteren et al., 1994). Despite the congruence of two genes, the authors of both analyses (Winnepenninckx et al., 1995; Vanfletern et al., 1994) cautioned that this placement of Nematoda could be an artifact of accelerated evolution in the model nematode, *C. elegans*, resulting in a phenomenon of phylogenetic inference called "long branch attraction", in which taxa with proportionally higher evolutionary rates are shifted to a basal position closer to the outgroup used in the analysis (Felsenstein, 1978).

Due to the implications of long branch attraction, Aguinaldo et al. (1997) reexamined the 18S rDNA sequence data in an attempt to determine the proper placement of Nematoda. When both rapidly and slowly evolving nematode genes were included in the phylogenetic analysis, the nematodes were placed basally in the phylogeny, originating before the divergence of protostomes and deuterostomes. This result agreed with the

traditional view of metazoan evolution. However, when only the 18S rDNA sequence from *Trichinella spiralis* was included, a gene believed to have a rate of evolution closer to that of the other 18S sequences in the study, the nematodes were placed higher in the phylogeny. No longer considered closely related to the flat worms, they rather became a sister taxon to the arthropods. The resulting clade has been dubbed "Ecdysozoa", as it contains animals that undergo ecdysis, or moulting (Aguinaldo et al., 1997). Further analyses have not clarified which of the two hypotheses, the "Ecdysoza hypotheis" or the traditional coelomata hypothesis, is correct. Analysis of Hox (de Rosa et al., 1999) and β-Thymosin (Manuel et al., 2000) sequence data support the ecdysozoa hypothesis while an analysis of another sample of genes from a larger variety of organisms (fungi, plants, invertebrates and vertebrates) supports the more traditional coeleomata hypothesis (Blair et al., 2002). To date, there is no consensus as to the proper placement of Nematoda in the larger framework of animal evolution, and the subject remains divisive.

Resolving the Relationships Within the Phylum Nematoda

Determining the proper relationship of the different orders of nematodes within the Nematoda has been as challenging as resolving the proper placement of the phylum Nematoda within the animal kingdom. Whilst many schemes of nematode phylogeny have been proposed (as reviewed by Coomans, 2000), the most widely accepted hypothesis had been based on a split between the terrestrial nematodes (Secernentea) and the marine nematodes (Adenophorea), which form two distinct monophyletic clades. Beyond this fundamental split the resolution of the evolutionary relationships of the different orders of nematodes had been much more ambiguous. Much of this ambiguity

was due to the use of morphological characters for taxonomic classification. Nematodes have exhibited a distinctive lack of distinguishable homologous characters across the phylum as a whole, likely due to a combination of morphological conservation in some features and accelerated evolution in others. Homoplasy due to convergent evolution or rapid diversification is also thought to play a role in nematode morphology, further confounding the issue (Bürglin et al., 1998; Blaxter et al., 1998). Therefore the segregation of the orders within Secernentea was initially based on lifestyle and divided into nine clades, with five containing exclusively animal-parasitic taxa.

In 1998, the first widespread molecular phylogenetic analysis of Nematoda was performed using small subunit (SSU) sequences from 53 taxa, sampled from parasitic as well as free-living groups (Blaxter et al., 1998). The results revolutionized the way nematode evolution is understood. The Adenophorea were shown to be paraphyletic, including the Secernentea ancestral taxa. In both of these groupings, the parasites were scattered throughout multiple clades and, with one exception, paired with free-living nematodes. The Secernentea were divided into three clades, rather than the traditional nine. One, labeled "Clade III", contains only animal-parasites. The other two (Clade IV and Clade V) contain taxa whose lifestyles include free-living nematodes, insect-parasites, invertebrate-parasites and vertebrate-parasites (Blaxter et al., 1998). Perhaps most important was the suggestion that parasitism arose on multiple independent occasions; at least three times for the plant-parasites and at least four times for animal-parasites. Because only 53 taxa were used in the analysis, this number may increase as more data becomes available (Dorris et al., 1999).

The Heteroderidae, the plant-parasites that are the focus of this research, fall into Clade IV. They are closely related to fungivores and invertebrate-parasites as well as free-living bacteriovores.

The Genus *Meloidogyne*

Meloidogyne is a large genus containing more than 60 species (Eisenbeck and Traintaphyllou, 1991) of obligate plant-parasites encompassing a very broad host-range (Sasser, 1980). As such, there is a widespread interest in establishing the evolutionary history within the genus as well as within the family Heteroderidae. Having a robust phylogeny for these plant pathogens can aid in understanding more about the evolution of parasitism within the genus and the differences between the individual species that might account for variations in host specificity and host resistance.

Meloidogyne phylogenetics has been an ongoing study stretching back decades (Dickson et al., 1971; Dalmasso and Bergé, 1978; Esbenshade and Traintaphyllou, 1987; Powers et al., 1986 and Powers and Sandall, 1988) but it was not until fairly recently that analysis on the molecular level was possible (Hugall et al., 1999; De Ley et al., 2002). Even these molecular analyses, based on a small number of genetic loci, are considered preliminary, with further analysis suggested (De Ley et al., 2002). A brief history of *Meloidogyne* phylogenetics, especially with respect to those *Meloidogyne* species with large agricultural impact (*M. arenaria*, *M. chitwoodi*, *M. hapla*, *M. incognita* and *M. javanica*) is presented in the introduction of Chapter 2 of this dissertation.

The Role of Horizontal Gene Transfer in the Evolution of Parasitism

With the availability of a reliable evolutionary framework, it will be easier to formulate more biologically plausible hypotheses about the evolution of parasitism in root-knot nematodes and other tylenchids. Presumably, the free-living lifestyle is an ancestral state, with symbiosis being a viable option only after the emergence of the host species. For vascular plants this is estimated at approximately 400 million years ago. As a derived trait, the evolution to symbiosis requires acquisition of genes that allow the organism to contact and enter the host and establish a source of nutrition, as well as overcome host resistance. These traits are common for all symbiotic relationships, whether mutualistic, commensualistic, or parasitic. Thus a link between parasitism and beneficial symbioses can be established through these basic requirements. Indeed, strong similarities have been found between genomes of plant mutualists and plant pathogens (Paulsen et al., 2002). The tree of life places all three types of symbionts in close association in many clades, suggesting independent and repeated acquisition of the symbiotic lifestyle (Ochman and Moran, 2001).

The relationship between non-parasitic and parasitic symbiosis can be seen within a single organism. *Escherichia coli*, a common microbe found in the digestive system of many animals, can be transformed into an enteropathogen through acquisition of a type III secretion system cassette, called LEE (McDaniel et al., 1995). Type III secretion systems are also partially responsible for virulence in *Yersinia*, *Erwinia*, *Pseudomonas*, *Salmonella* and other bacterial species (Van Gijsegem et al., 1993). *Salmonella* contains two such cassettes, or pathogenicity islands. Both pathogenicity islands are required for the development of the parasitic lifestyle and both are likely to have been acquired by

horizontal gene transfer (Groisman and Ochman, 1993). The transfer of such pathogenicity islands by horizontal gene transfer (HGT) has been well established as one of the key events in the emergence of pathogenicity in prokaryotic species (as reviewed in Ochman and Moran, 2001).

Whilst HGT involving eukaryotes is much more controversial than HGT between prokaryotes (Salzberg et al., 2001; Stanhope et al., 2001; Brinkman et al., 2002), there are indications of horizontally acquired genes in Heteroderidae (Smant et al., 1998; Lambert et al., 1999; Veronico et al., 2001; Jaubert et al., 2002). Most have been identified on the basis of biochemical and immunological criteria. In particular, the discovery of a putatively horizontally acquired β-1,4-endoglucanase gene allowing endogenous production of a cellulase in cyst nematodes (Smant et al., 1998) prompts the question whether horizontal gene transfer may have played a role in adaptation to the parasitic lifestyle.

Thus far, those genes implicated in HGT in nematodes appear to be bacterial in origin. As they were identified serendipitously, based mainly on biological observations, it is possible that other horizontally acquired genes are present in the genome but have not yet been identified. If more transfer candidates exist, bacteria would continue to be the most likely donor source of genetic material.

Soil-dwelling nematodes share an ecological niche with many different species of bacteria. Indeed, free-living nematodes are, for the most part, bacteriovores. Thus the intimate interaction between nematode species and bacteria is well established. Assuming that an ancestral nematode species (not necessarily extant) incorporated and integrated bacterial genetic material into its own genome, it would perhaps explain how nematodes

were able to evolve the full range of attributes needed to develop and maintain a parasitic lifestyle. It could also provide at least a partial explanation for the multiple independent origins of parasitism in Nematoda.

Detecting Horizontal Gene Transfer

Acquisition of a gene (or set of genes) by HGT has three basic requirements: 1. genetic material from a donor must be available to the recipient (i.e. physical proximity); 2. the recipient cell or organism must be in a competent state; and 3. the genetic material must be integrated into the genome and become functional (Mazodier and Davies, 1991).

The most commonly implemented methods for searching for horizontally acquired genes in a given genome involve looking for anomalies in G+C content or codon usage in a given gene or open reading frame (ORF). As such genome characteristics vary amongst different organisms, yet are generally consistent within a given species, statistically significant divergences from the average G+C content (Médigue et al., 1991; Lawrence and Ochman, 1997) or codon adaptation index (Sharp and Li, 1987) in a given gene can signal possible xenologous origin, or horizontal acquisition. Using G+C content and synonymous codon usage metrics, Lawrence and Ochman (1998) analyzed the *E. coli* genome and concluded that approximately 17.6% of the ORFs were acquired via HGT, a number similar to a previous estimate (Médigue et al., 1991). Lawrence and Ochman (1997) also examined the rate of amelioration, the process by which an acquired gene undergoes mutational pressures to homogenize G+C content and codon bias to more properly reflect the genome of its new host organism. They used a measure of rate of ameiloration in order to estimate the time of introgression of genes in *E. coli* by

comparing it to the *Salmonella enterica* genome. However, further comparison between *E. coli* and *Salmonella typhi* genomes suggests that G+C content and codon bias do not provide ample evidence for horizontally acquired genes (Koski et al., 2001). The authors determined that unusual base composition and/or codon bias may give indications of putative horizontally transferred genes, but are likely to lead to overestimates of this phenomenon in a genome. Further, because HGT is more likely to happen between closely related organisms (such as between different strains of *E. coli*, as mentioned above), some horizontally acquired genes may look like "native" genes, and therefore will not be detected. Even should G+C content or codon usage be detectably different between donor and recipient organisms, amelioration can occur at a rate which is sufficient to mask this signal. This may be especially true for transfers between eukaryotes and prokaryotes as presumably structural differences, such as the addition or loss of introns and promoter sites, would be required to make the xenologous gene functional within its new genome.

Another method for verifying a HGT event is phylogenetic analysis. A xenologous origin could explain incongruence between a phylogeny for a given gene and the expected relationships between the set of species containing that gene. Arguments for HGT due to phylogenetic incongruence have notably been given for transposable P-elements in *Drosophila* (Clark and Kidwell, 1997; Haring et al., 2000) and in the human genome (International Human Genome Sequencing Consortium, 2001). The discovery of human genes with unexpected phylogenies lead to claims of vast horizontal gene transfer between vertebrates and bacteria. Two refutations of these claims (Salzberg et al., 2001; Stanhope et al., 2001) used similar methods to show that the degree of HGT in the human

genome had been dramatically overstated. In the first (Salzberg et al., 2001), a larger sample of organisms was used to search for homology between the putatively horizontally acquired human genes and other eukaryotes. A correlation between the number of organisms included in the comparisons and a reduction of HGT candidates in the human genome was found. The investigation eliminated 182 of the human HGT candidates, leaving only 41 genes with possible bacterial origin. Also noted was that the phylogenies for these remaining 41 HGT candidates could be explained through multiple independent gene loss in certain lineages. This number might also be further reduced as more eukaryotic species are sequenced. Similarly, the second group (Stanhope et al., 2001) used publicly available expressed sequence tag (EST) sequences to determine that most of the putative horizontally acquired genes are present in simple eukaryotes. The authors concluded that these genes are not likely to have been horizontally acquired; rather the presence of the genes in scattered bacterial genomes and the human genome and lack of sequence information in invertebrates is likely due to a bias originating from the selection of organisms with sequencing projects. These two analyses highlight a drawback of using phylogenetic analysis and homology searches in an attempt to prove horizontal gene transfer, namely that just because a gene cannot be found in a database does not mean the gene does not exist in an organism. Even those genomes which are "completely sequenced" are, for the most part, only close to completion with annotation ongoing.

Although there are tools and methods which can be used to identify genes that may have been horizontally acquired, caution must be taken regarding such claims as these

tools cannot give conclusive results, but rather only provide anecdotal evidence for xenology.

Conclusion

With the recent availability of a large number of tylenchid sequences (McCarter et al., 2000; 2003), a more robust phylogenetic analysis can be completed to resolve the proper relationships between the *Meloidogyne*, and within the Heteroderidae. While there are no complete genomes for root-knot nematodes, the fully annotated *C. elegans* genome sequence (The *C. elegans* Sequencing Consortium, 1998) can serve as an essential resource for identifying orthologous genes in the tylenchid datasets. The second chapter of this dissertation presents a strategy for mining the Heteroderidae sequence data for orthologues. This strategy is similar to the construction of clusters of orthologous groups (COGs) (Tatusov et al., 1997) deposited with the National Center for Biotechnology Information (NCBI: http://www.ncbi.nlm.nih.gov). As further presented in chapter 2, the COGs constructed from the nematode datasets were also used to resolve a *Meloidogyne* phylogeny that appears to be more robust than those previously published.

The third chapter demonstrates the use of a similar dataset (available *Meloidogyne* sequences and the completed *C. elegans* genome) as well as the complete *Drosophila melanogaster* genome and available bacterial sequences to identify HGT candidates in *Meloidogyne*. By using the fully sequenced *C. elegans* and *D. melanogaster* genomes in a directed search for horizontal gene transfer events that may have occurred since the divergence of the *Meloidogyne* and *C. elegans* ancestors, errors in HGT identification due

to insufficient sequencing in these taxa can potentially be reduced. The final chapter contains a brief review of my findings and some suggestions for future research.

References

Adoutte A., Balavoine G., Lartillot N., de Rosa R. 1999. Animal evolution: The end of the intermediate taxa? *Trends in Genetics* 15:104-108.

Aguinaldo A. M. A., Turbeville J. M., Linford L. S., Rivera M. C., Garey J. R., Raff R. A., Lake L. A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489-493.

Blaxter, M. L., DeLey P., Garey J., Liu L. X., Scheldeman P., Vierstraete A., Vanfletern J., Mackey L. Y., Dorris M., Frisse L. M., Vida J. T., Thomas W. K. 1998. A molecular evolutionary framework for the phylum Nematoda. *Nature* 392:71–75.

Blair J. E., Ikeo K., Gojobori T., Hedges S. B. 2002. The evolutionary position of nematodes. *BMC Evol. Biol.* 2:7.

Boucher G., Lambshead P. J. D. 1995. Ecological biodiversity of marine nematodes in samples from temperate, tropical, and deep-sea regions. *Conserv. Biol.* 9:1594-1604.

Brinkman F. S. L., Blanchard J. L., Cherkasov A., Av-Gay Y., Brunham R. C., Fernandez R. C., Finlay B. B., Otto S. P., Ouellette B. F. F, Keeling P. J., Rose A. M., Hancock R. E. W., Jones S. J. M. 2002. Evidence that plant-like genes in *Chlamydia* species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the choloroplast. *Genome Res.* 12:1159-1167.

Bürglin T. R., Lobos E., Blaxter M. L. 1998. *Caenorhabditis elegans* as a model for parasitic nematodes. *Int J. Parasitol.* 28:395-411.

Clark J. B., Kidwell M. G. 1997. A phylogenetic perspective on P transposable element evolution in *Drosophila*. *Proc Natl Acad Sci* 94:11428-11433.

Coomans, A. 2000. Nematode systematics: past, present and future. *Nematology* 2:3-7

Dalmasso A., Bergé J. G. 1978. Molecular polymorphism and phylogenetic relationship in some *Meloidogyne* spp.: Application to the taxonomy of *Meloidogyne*. *J. Nematol.* 10:323–332.

De Ley I. T., De Ley P., Vierstraete A., Karssen G., Moens M., Vanfleteren J. 2002. Phylogenetic analyses of *Meloidogyne* small subunit rDNA. *J. Nematol.* 34:319–327.

de Rosa R., Grenier J. K., Andreeva T., Cook C. E., Adoutte A., Akam M., Carroll S. B., Balavoine G. 1999. Hox genes in brachiopods and priapulids and protostome evolution. *Nature* 399:772-776.

Dickson D. W., Huisingh D., Sasser J. N. 1971. Dehydrogenases, acid and alkaline phosphatases, and esterases for chemotaxonomy of selected *Meloidogyne*, *Ditylenchus*, *Heterodera*, and *Aphelenchus* spp. *J. Nematol.* 3:1–16.

Dorris M., De Ley P., Blaxter M. L. 1999. Molecular analysis of nematode diversity and the evolution of parasitism. *Parasitol. Today* 15:188–193.

Eernisse D. J., Albert J. S., Anderson F. E. 1992. Annelida and Arthopoda are not sister taxa: A phylogenetic analysis of spiralian metazoan morphology. *Syst. Biol.* 41:305-330.

Eisenbeck J. D., Triantaphyllou H. H. 1991. Root-knot nematodes: *Meloidogyne* species and races. In: Nickle, W. R. (Ed.), Manual of Agricultural Nematology. Marcel Dekker Inc., New York, pp. 191–274

Esbenshade P. R., Triantaphyllou A. C. 1987. Enzymatic relationships and evolution in the genus *Meloidogyne* (Nematoda: Tylenchida). *J. Nematol.* 19:8–18.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401-410.

Field K. G., Olsen G. J., Lane D. J., Giovannoni S. J., Ghiselin M. T., Raff E. C., Pace N. R., Raff R. A. 1988. Molecular phylogeny of the animal kingdom. *Science* 239:748-753.

Groisman E. A., Ochman H. 1993. Cognate gene clusters govern invasion of host epithelial cells by *Salmonella typhimurium* and *Shigella flexneri. EMBO J*. 12:3779-3787

Haring E., Hagemann S., Pinsker W. 2000. Ancient and recent horizontal invasions of drosophilids by P elements. *J. Mol. Evol.* 51:577-586.

Hugall A., Stanton J., Mortiz C. 1999. Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic *Meloidogyne*. *Mol. Biol. Evol.* 16:157–164.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 407:508-513.

Jaubert S., Laffaire J.-B., Abad P., Rosso M.-N. 2002. A polygalacturonase of animal origin isolated from the root-knot nematode *Meloidogyne incognita*. *FEBS Letts.* 522:109-112.

Koenning S. R., Overstreet C., Noling J. W., Donald P. A., Becker J. O., Fortnum B. A. 1999. Survey of crop losses in response to phytoparasitic nematodes in the United States for 1994. *J. Nematol.* 31:587-618.

Koski L. B., Morton R. A., Golding G. B. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Molec. Biol. Evol.* 18: 404-412.

Lambert K. N., Allen K. D., Sussex I. M. 1999. Cloning and characterization of an esophageal-gland-specific chorismate mutase from the phytoparasitic nematode *Meloidogyne javanica*. *Molec. Plant Microbe Interacts*. 12:328-36

Lawrence J. G., Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44:383-397.

Lawrence J. G., Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci.* 95:9413-9417.

Manuel M., Kruse M., Muller W. E., LeParco Y. 2000. The comparison of beta-thymosin homologues among metazoan supports an arthropod-nematode clade. *J. Mol. Evol.* 51:378-381.

Marshall A. J., Williams W. D. 1974. Textbook of Zoology Invertebrates (7^{th} edition). The Macmillian Press Ltd. London and Basingstoke, UK, pp 17-18.

Mazodier P., Davies J. 1991. Gene transfer between distantly related bacteria. *Annu. Rev. Genet.* 25:147-171.

McCarter J., Abad P., Jones J. T., Bird D. McK. 2000. Rapid gene discovery in plant parasitic nematodes via Expressed Sequence Tags. *Nematology* 2:719–731.

McCarter J. P., Mitreva M. D., Martin J., Dante M., Wylie T., Rao U., Pape D., Bowers Y., Theising B., Murphy C., Kloek A. P., Chiapelli B., Clifton S.W., Bird D. McK., Waterston R. 2003. Analysis and functional classification of transcripts from the root-knot nematode *Meloidogyne incognita*. *Genome Biol.* 4:R26.1–R26.19.

McDaniel T. K., Jarvis K. G., Donnenberg M. S., Kaper J. B. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc. Natl. Acad. Sci.* 92:1664-1668.

Médigue C., Rouxel T., Vigier P., Henaut A., Danchin A. 1991. Evidence for horizontal gene-transfer in *Escherichia coli* speciation. *J. Mol. Biol.* 222:851-856.

Ochman H., Moran N. A. 2001. Genes lost and genes found: Evolution of bacterial pathogensis and symbiosis. *Science* 292:1096-1098.

Paulsen I. T., Seshadri R., Nelson K. E., Eisen J. A., Heidelberg J. F., Read T. D., Dodson R. J., Umayam L., Brinkac L. M., Beanan M. J., et al. 2002. The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proc Natl Acad Sci* 99:13148-13153.

Powers T. O., Platzer E. G., Hyman B. C. 1986. Species-specific restriction site polymorphism in root-knot nematode mitochondrial DNA. *J. Nematol*. 18:288–293.

Powers T. O., Sandall L. J. 1988. Estimation of genetic divergence in *Meloidogyne* mitochondrial DNA. *J. Nematol.* 20:505–511.

Salzberg S. L., White O., Peterson J., Eisen J. A. 2001. Microbial genes in the human genome: larteral transfer or gene loss? *Science* 292:1903-1906.

Sasser J. N. 1980 Root-knot nematodes: A global menace to crop production. *Plant Dis.* 64:36–41.

Sharp P. M., Li W.-H: 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nuc. Acids Res.* 15:1281-1295.

Smant G., Stokkermans J. P. W. G., Yan Y., De Boer J. M., Baum T. J., Wang X., Hussey R. S., Gommers F. J., Henrissat B., Davis E. L., Helder J., Schots A., Bakker J. 1998. Endogenous cellulases in animals: Isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proc. Natl. Acad. Sci.* 95:4906-4911.

Stanhope M. J., Lupas A., Italia M. J. , Koretke K. K., Volker C., Brown J. R. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 11:940-944.

Tatusov R. L., Natale D. A., Garkavtsev I. V., Tatusova T. A., Shankavaram U. T., Rao B. S., Kiryutin B., Galperin M. Y., Fedorova N. D., Koonin E. V. 1997. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. 29:22–28.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012-2018.

Vanfleteren J. R., Van De Peer Y., Blaxter M. L., Tweedie S. A. R., Trotman C., Lu L., Van Hauwaert M.-L., Moens L. 1994. Molecular genealogy of some nematode taxa as based on Cytochrome c and Globin amino acid sequences. *Mol. Phyl. Evol.* 3:92-101.

Van Gijsegm F., Genin S., Boucher C. 1993. Conservation of secretion pathways for pathogenicity determinants of plant and animal bacteria. *Trends Microbiol.* 1:175-180.

Veronico P., Jones J., Di Vito M., De Giorgi C. 2001. Horizontal transfer of a bacterial gene involved in polyglutamate biosynthesis to the plant-parasitic nematode *Meloidogyne artiellia*. *FEBS Letts.* 508:470-474.

Winnepenninckx B., Backeljau T., Mackey L. Y., Brooks J. M., De Wachter R., Kumar S., Garey J. R. 1995. 18S rRNA data indicate that Aschelminthes are polyphlyetic in origin and consist of at least three distinct clades. *Mol. Biol. Evol*. 12:1132-1137.

Chapter 2

Resolving evolutionary relationships in *Meloidogyne* through multiple gene analysis

Elizabeth H. Scholl,[1,2] and David McK. Bird[1,2]*

[1]Center for the Biology of Nematode Parasitism and [2]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA.

*Corresponding author:

David Bird
Center for the Biology of Nematode Parasitism
Box 7253, NC State University
Raleigh, NC 27695-7253
Phone: +1 (919) 515-6813
Fax: +1 (919) 515-9500
e-mail: david_bird@ncsu.edu

Keywords: Bayes, COG, cyst nematode, orthologue, phylogeny, root-knot nematode.

Abstract

Most phylogenetic analyses of plant-parasitic nematodes have been based on a small number of character sets. Such analyses report gene-trees which may not reflect the true species tree. We designed a strategy to suppress such incongruencies, and performed an analysis to recover the most robust phylogeny for five *Meloidogyne* species (*M. arenaria*, *M. chitwoodi*, *M. hapla*, *M. incognita*, and *M. javanica*), three closely related taxa (*Heterodera glycines*, *Globodera pallida* and *G. rostochiensis*) and the more distant taxon, *C. elegans*. Our multiple-gene approach is based on sampling more than 80,000 tylenchid sequences present in public databases. Reciprocal, pairwise BLAST analyses were performed for EST cluster consensus sequences in each of the nine species to identify orthologous genes. To minimize ascertainment bias, gene ontology was deliberately ignored. We identified 47 genes which could unambiguously be assigned as orthologues, and performed an alignment, so that all 47 genes could be concatenated to create one multi-gene alignment. Bayesian inference was employed for phylogenetic reconstruction. Our results place *M. incognita* and *M. javanica* as sister taxa, with *M. arenaria* basal to these. Placement of *M. hapla* and *M. chitwoodi* are congruent with previous studies, as are relationships with the other taxa examined.

Introduction

Meloidogyne is a large genus in the order Tylenchida containing more than 60 species (Eisenback and Triantaphyllou, 1991). Of these, four (*M. incognita*, *M. arenaria, M. javanica* and *M. hapla*) are major agricultural pests worldwide. Other root-knot nematodes (RKN), such as *M. chitwoodi*, are important pests regionally. As a genus, *Meloidogyne* has a very broad host-range (Sasser, 1980) but individual species may have a more restricted host preference. For any given host, responses to invasion by different RKN vary qualitatively and quantitatively. For example, *M. incognita* infection of *Lotus japonicus* yields large galls whereas *M. hapla* infection produces small galls which typically sprout lateral roots (Lohar and Bird, 2003). Similarly, the presence of a resistance gene can effectively block infection of an otherwise susceptible host (Bird and Kaloshian, 2003). Because this diversity has obvious practical consequences in agriculture, there has been a long history of research aimed at identifying RKN species and also a long-standing interest in establishing evolutionary relationships within the genus and between closely-related genera such as the cyst nematodes (*Heterodera* and *Globodera*).

The traditional approach to identifying individual *Meloidogyne* species is based on differentiation of morphological characteristics (reviewed by Eisenback and Triantaphyllou, 1991), which also formed the initial basis for evolutionary studies. A strength of this approach is that morphology reflects the integrated outcome of the expression of many genes. Unfortunately it is not known what those genes are, nor what contribution each gene makes to the organism as a whole. Furthermore, attempts to discretize otherwise continuous morphological characters makes identification and

classification of those differences subjective. Cytogenetic criteria have been independently examined for evolutionary and diagnostic purposes (Triantaphyllou, 1966; 1985), but these chromosomal features merely represent an alternative morphological coding.

Subsequent attempts to more directly survey genetic characters have employed DNA hybridization techniques (Xue et al., 1992; Castagnone-Sereno et al., 1993) and isozyme analysis (Dickson et al.,1971; Dalmasso and Bergé, 1978; Esbenshade and Triantaphyllou, 1987). Other early attempts to establish *Meloidogyne* phylogenies were based on mitochondrial genes (Powers et al., 1986; Powers and Sandall, 1988), but these results showed that mitochondria seem to be hypervariable, with the resultant tree reflecting only the relationship of the genes sampled for the analysis rather than the underlying species relationships. Recently, the high variability found in nuclear ribosomal internal transcribed spacers has been exploited for phylogenetic inference between closely related *Meloidogyne* species (Hugall et al., 1999), and the portion of this locus encoding the highly conserved 18S rRNA subunit (SSU) was exploited to examine more diverged relationships within *Meloidogyne*, and between *Meloidogyne* and other tylenchid genera (De Ley et al., 2002).

A limitation of using single, individual loci is that locus history within a collection of species will not always accurately reflect the relationship of the species as a whole (Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991; Doyle, 1992; Dorris et al., 1999). Issues such as gene duplication, lineage sorting and especially horizontal gene transfer, which seems prevalent in Tylenchida (Scholl et al., 2003), may contribute to varying degrees of discordance between a gene tree and a species tree. Furthermore, analysis of

sequences with high numbers of tandem repeats, such as ribosomal genes, creates a risk of choosing genes that have undergone gene conversion or concerted evolution, and are therefore not actually true orthologues (Slowinski and Page, 1999). One solution to increase the reliability of capturing the correct underlying phylogenetic signal from gene sequences is to sample a larger gene set; any biases presented by a single gene with a history not reflecting that of the species will hopefully be offset with the larger selection of genes that do better reflect that proper relationship (Takahata, 1989). Ideally, entire genomes would be compared, but complete genomes remain unavailable for any tylenchids. However, large numbers of tylenchid sequences have been generated as Expressed Sequence Tags (ESTs) (McCarter et al., 2000; Bird et al., 2002; McCarter et al., 2003), and we have used these sequences for a multi-gene approach to resolve evolutionary relationships between this important group of plant-parasitic nematodes. The extensively annotated and completed *C. elegans* genome sequence (The *C. elegans* Sequencing Consortium, 1998) serves as an essential reference.

Rather than select genes based on some preconceived notion of what might be the "best" Tylenchida candidates, we designed an approach that was blind to gene function, based on three minimal requirements. First, the gene must be expressed and so potentially be present in a cDNA library. In fact, because the EST data were obtained from a relatively small number of clones (compared to the organism's total gene complement), a practical consequence was that each gene should be sufficiently expressed to have been identified as an EST. Second, the gene must have an identifiable orthologue in *C. elegans*. The presence of a gene across diverse nematode orders presumably reflects some evolutionary constraint (Blaxter et al., 1998); such slowly evolving genes are prime

candidates for studying evolutionary relationships (Smith, 1989; Mindell et al., 1990; Hillis and Dixon, 1991; Kumazawa and Nishida, 1993). Hence this requirement will tend to eliminate genes which may have undergone accelerated evolution during some specific aspect of the ecology of the particular nematode species, and thus lead to a gene tree that might be discordant with the general species tree. Further, it may also eliminate genes acquired by horizontal transfer (Scholl et al., 2003), which clearly would bias the species trees. Genes meeting these two criteria were compared to identify orthologues of each in each species. Such groups of orthologues present in multiple species are termed "clusters of orthologous groups of proteins" or COGs (Tatusov et al., 1997) and are identified based on an iterative series of reciprocal, pairwise comparisons for each gene in each species, leading to a network defining the COG (Figure 2.1). A closed network (Figure 2.1A) is suggestive of true orthologues, whereas deviations (Figure 2.1B) flag ambiguity, such as might arise from paralogues. COGs thus constructed were subjected to a third minimal criterion that each selected gene be present in the available dataset either of three cyst nematodes plus *C. elegans*, or three *Meloidogyne* plus *C. elegans,* or five species total.

Bootstrapping (Felsenstein, 1985) has been typically employed to determine support for phylogenetic hypotheses determined by maximum likelihood or protein distance methods. However, advances in computing power permit the use of more complex statistical approaches, particularly by means of Bayesian analysis, to more realistically implement evolutionary models and more accurately measure statistical uncertainty. We also employ maximum likelihood and distance methods to support the results from our Bayesian analysis.

The analysis we present recovers the most robust phylogeny for nine nematode species (*Meloidogyne arenaria*, *M. chitwoodi*, *M. hapla*, *M. incognita*, *M. javanica, Heterodera glycines*, *Globodera pallida, G. rostochiensis* and *C. elegans*). We used a set of 47 orthologous genes initially identified from genome-wide EST analysis, and employed a powerful and more realistic statistical model for recovering the most phylogenetically relevant information from a concatenation of the alignments of these sequences.

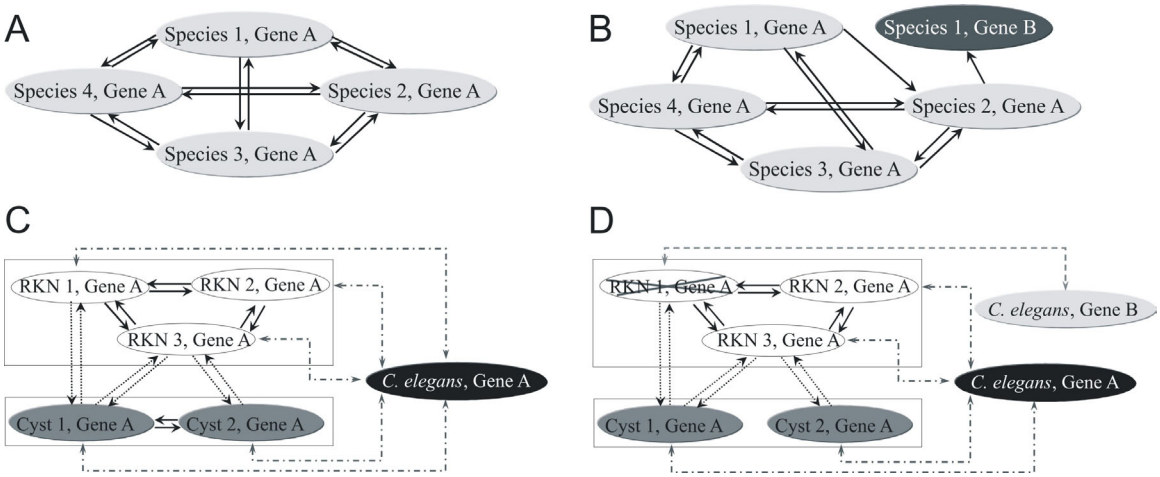

Figure 2.1: Schemata conceptualizing design and construction of clusters of orthologous groups (COGs). The tail of each arrow represents a query sequence for a BLAST analysis, and the head defines the best match within the target species. Double-headed arrows summarize reciprocal best matches. A) A hypothetical, four-species COG depicting the relationship of gene A in each species to its orthologue in each of the other three species. The network is closed and complete because all potential pairwise matches are present and concordant. This represents one orthologous group. B) A more typical COG network exhibiting reciprocal match disagreement. The orthologues of gene A from Species 1 are identifiable in Species 3 and 4, but the relationship of this gene between Species 1 and Species 2 is ambiguous, owing to a best match of gene A in Species 2 to gene B in Species 1. Genes A and B in Species 1 may be paralogous. C) Linking COGs. Solid arrows depict relationships within an initial COG, and dotted arrows reflect links between pre-assembled COGs. Dot-dash arrows reflect the incorporation of *C. elegans* orthologues. Because all genes have the same reciprocal best match to *C. elegans*, the cluster is closed and complete. D) Some analyses as in (C) result in a gene in one species having a best match to a gene in *C. elegans* (dashed line) different from the rest of the COG; such genes (RKN 1, Gene A in this example) were stricken from the COG. RKN = *Meloidogyne* spp.; Cyst = *Heterodera* and *Globodera* spp.

Materials and Methods

Available data

All sequences were obtained from resources publicly available on April 30, 2003. The full *Caenorhabditis elegans* genome and the full set of predicted proteins (WormPep) were downloaded as FASTA files from the National Center for Biological Information (NCBI) GenBank database (Wheeler et al., 2002). *Globodera pallida* ESTs (1,878 entries) were similarly obtained. *Meloidogyne hapla* (13,869) and *M. chitwoodi* (3,456) ESTs were obtained from the nematode.net ftp server (McCarter et al., 2003). These ESTs had been subjected to a rigorous, automated quality filtering pipeline prior to being made publicly available, but may retain some errors. The *G. pallida* ESTs had been processed through an equivalent quality pipeline. *Meloidogyne arenaria* (3,321), *M. incognita* (5,662), *M. javanica* (5,574), *Heterodera glycines* (4,307) and *G. rostochiensis* (5,039) data were kindly provided as cluster consensus sequences by the Plant Parasitic Nematode sequencing project (Bird et al., 2002; www.nematode.net). In addition to the quality pipeline, these sequences were further processed through the semi-automatic NemaGene pipeline (McCarter et al., 2003), leading to a very low error rate. *C. elegans* genome sequence is effectively error-free (The *C. elegans* Sequencing Consortium, 1998).

EST clustering

As a means to suppress potential errors, *M. chitwoodi* and *M. hapla* EST sequences were clustered into consensus sequences using the Incremental Clustering Algorithm

(ICA) as implemented by the n2tool program from the ICA-tools suite (Parsons, 1995), with the subsequence similarity score threshold set to 20; both clone orientations were compared. Each cluster was confirmed by manual examination of the alignments. The smaller *G. pallida* EST dataset was clustered by pairwise BLASTN analysis (Altschul et al., 1990). Because our cluster consensus sequences (along with those from nematode.net) can be considered to define genes, we refer to them as such for the subsequent analyses.

Identifying orthologues

Genes were gathered into clusters of orthologous groups (COGs) with a procedure that is essentially the same as described by Tatusov et al. (1997). To minimize ascertainment bias, gene ontology was deliberately ignored during gene selection and COG assembly. Gene-for-gene, pairwise comparisons were made between different species using the Hardware Accelerated, Tera-BLAST algorithm (Time-Logic, Crystal Bay, NV) executed on 3 TimeLogic DeCypher nodes. As the initial filter, single FASTA files containing the genes for each parasitic species were used as queries for a six-phase translated BLASTX comparison to the *C. elegans* WormPep database. Parasite genes that failed to produce a match to *C. elegans* at an e-value of less than $1.0e^{-20}$ were eliminated. Remaining sequences were used in additional 6-phase to 6-phase Tera-TBLASTX searches with each of the genes from the chosen parasitic species, and the top match for each gene in each species for each search was recorded. Matches were considered significant for e-values of less than $1.0e^{-15}$.

We commenced COG assembly within the genus *Meloidogyne* by comparing the reciprocal best hits between all pairs of *Meloidogyne* species. An initial pairwise grouping was arbitrarily selected (*M. incognita* and *M. arenaria*), and orthologues between these two species were established. The nascent COG was extended to a third species (*M. javanica*) by examining reciprocal best hits for the COG members established in the initial two species. The procedure was continued for each *Meloidogyne* species, such that multi-dimensional networks of orthologues were developed that incorporated each of the BLAST results (Figure 2.1A). Once all species were exhausted based on the initial pairing of *M. incognita* and *M. arenaria*, a new "initial" pair was chosen (in this case *M. incognita* and *M. javanica*) and the procedure repeated. This step ensured that genes which might be present in the sequence databases of these (and other) species, but absent from the *M. arenaria* dataset, could also be grouped into a COG. Similar procedures were continued until all pairs of orthologues between *Meloidogyne* species were incorporated into a networked grouping. Such networks potentially range in size from two genes (identified in one pair of species by reciprocal best hits) to five genes (sequence for a given homologue available for all *Meloidogyne* species tested). Once completed for *Meloidogyne*, the same procedure was implemented for the three cyst nematode species.

Meloidogyne COGs and cyst nematode COGs were linked by examining pairwise reciprocal best matches between the *Meloidogyne* genes within one COG and their cyst nematode orthologues to form a new COG spanning the family Heteroderidae. To add *C. elegans* genes to this COG, each gene in a Heteroderidae COG was used in a 6-phase to 6-phase Tera-BLASTX search against the *C. elegans* nucleotide database (Figure 2.1C).

Any gene in a group of parasitic nematodes that did not have the same best match to *C. elegans* as the majority of the other genes within that same group was removed from the COG (Figure 2.1D). An advantage of this step is that parasite genes, which might be sufficiently related to genes in a particular COG to be added to that COG, but which in reality are not true orthologues, should be resolved as unique genes in the fully sequenced *C. elegans* genome. Examples would include members of gene families.

As a tool to flag errors in the original EST clustering and also to further identify paralogous genes which might spuriously join the COG, we identified situations in which more than one gene from a given species was represented in a COG (Figure 2.1B). The multiple sequences from that species were aligned to the *C. elegans* nucleotide sequence using CLUSTAL-W (Thompson et al., 1994). When the alignment revealed distinct parasite EST clusters which clearly defined a single gene (i.e., had not correctly been assigned a single cluster) those EST clusters were concatenated to reflect the true relationship of the ESTs to a gene, and to facilitate restoration of the COG network (Figure 2.1A). In cases where the alignment did not reveal EST clustering anomalies, the multiple genes from a single species were deemed to be putative paralogues. Such COGs were removed from subsequent analyses.

Phylogenetic analysis

Based on deduction of the correct reading frames for the EST clusters from Tera-BLASTX alignments to *C. elegans* WormPep, individual genes (one COG at a time) were conceptually translated and aligned using MegAlign (DNASTAR, Inc., Madison WI) to implement the CLUSTAL-W algorithm (Thompson et al., 1994) with default

parameter values. Thus reflecting a codon alignment, original nucleotide sequences were recovered. Adjustments to each alignment were made by hand and individual sequence alignments were concatenated to create one multi-gene alignment.

Bayesian inference and Markov chain Monte Carlo (MCMC) techniques were employed for phylogenetic reconstruction using version 3.0b4 of the MrBayes software (Huelsenbeck and Ronquist, 2001). A general time reversible model was selected, allowing for gamma-distributed, among-site rate variation with four categories. Four chains were run for 100,000 iterations, sampling every 20 iterations for a total of 5,000 samples. The first 1,000 iterations were sufficient to reach stationarity and were thus discarded as burn-in. Three additional runs to confirm convergence of parameters were performed for 50,000 iterations each sampling every 50 steps with the first 1,000 iterations discarded as burn-in.

A topology consistent with the results of the Bayesian analysis was recovered by maximum likelihood analysis using the HKY85 + Γ model with eight categories and stepwise addition for tree searching as implemented in the PAML v3.13a package of programs (Yang, 1997). In addition, all sequences in the alignment were conceptually translated and PHYLIP (Felsenstein, 1993) was employed to calculate protein distances with the Jones-Taylor-Thornton rate matrix (Jones et al., 1992). Tree topologies were inferred from the protein distances using neighbor joining. Non-parametric bootstrapping with 10,000 replicates provided an estimate of overall support for the topology. Gaps were treated as missing data in all analyses.

Results

EST clustering

Clustering generated 1,283 consensus sequences from the *M. hapla* EST set, and 839 cluster consensuses for *M. chitwoodi*. As is typically observed (McCarter et al., 2003), the distribution of the cluster sizes (Figure 2.2) approximately followed that expected for a Poisson process. However, an attempt to cluster the smaller, *G. pallida* EST set using ICA-tools revealed deviation from the anticipated distribution. Rather than use these clusters, we performed a manual clustering based on BLAST queries of each sequence against the database of all *G. pallida* sequences (i.e., a self BLAST), yielding 989 clusters.

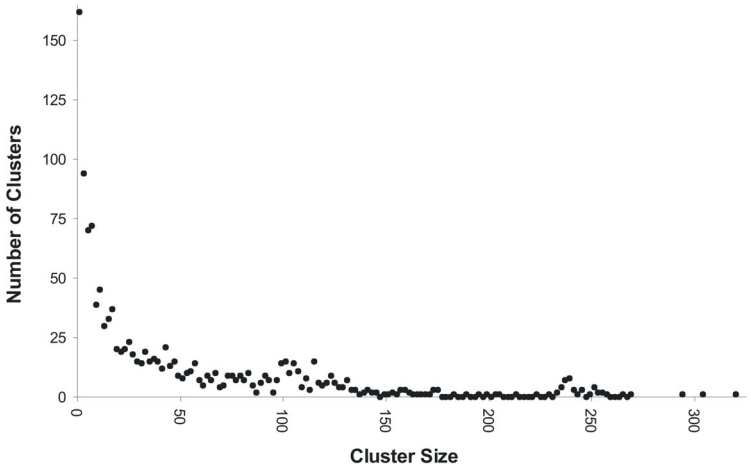

Figure 2.2: Distribution of ESTs by cluster size for *M. hapla*.

COG assembly

Seventy-four clusters of orthologous groups containing genes from at least three plant-parasitic nematodes plus *C. elegans* were deduced (Table 2.1). In four instances (in

three COGs) the best BLAST match of an individual gene to *C. elegans* contradicted the remainder of the COG, and so those genes were disqualified from their COG (Figure 2.1D). Based on the best match to *C. elegans*, it was further apparent that some of the COGs actually defined the same gene; manual resolution of these redundancies reduced the number of COGs to 58 (Table 2.1). Three additional COGs were eliminated because we considered the global alignment to be too short to be phylogenetically informative.

Table 2.1: COG enrichment during the clustering process.

	COG number[a]	Putative Paralogues[b]
Step 1[c]	74	31
Step 2[d]	58	15
Step 3[e]	56	15
Step 4[f]	55	14
Step 5[g]	47	0

[a]The number of clusters of orthologous genes (COGs) remaining following sequential steps to eliminate spurious genes.
[b]Number of COGs containing putative paralogues at each step of the COG enrichment.
[c]Step 1: Initial COG number as determined by pair-wise BLAST results.
[d]Step 2: COG number following merging different COGs based on identical best matches to *C. elegans*.
[e]Step 3: COG number following correction of EST clustering anomalies.
[f]Step 4: COG number following removal of members with poor global alignment.
[g]Step 5: Final number of COGs after removal of sets harboring putative paralogues.

One consequence of our COGing strategy is that non-overlapping, parasite EST clusters which actually define a single gene would be flagged as putative paralogues (Figure 2.1B). Indeed, alignment with *C. elegans* revealed six such instances. Correct

concatenation of those EST clusters reflects the true relationship to their *C. elegans* orthologue reduced the total number of COGs with putative paralogues from 14 to eight. Because it is not apparent which, if either, of the paralogues in such COGs truly is an orthologue, these eight COGs were disregarded. The remaining 47 COGs (Table 2.1) were deconstructed by species, and the sequences concatenated into a multi-gene alignment of 22,920 base pairs. An average of six species was represented for each gene, with an average aligned sequence length of 511 base pairs per gene. The G+C content, which averaged 45%, and ranged from 40% in *M. chitwoodi* to 52% in *G. rostochiensis* showed no strong species bias. Ontology was determined for each of the 47 genes *post hoc* (Figure 2.3), and the WormPep identification numbers, which are searchable at WormBase (Harris et al. 2003; www.wormbase.org), serve as a convenient way to identify each COG (Appendix A).

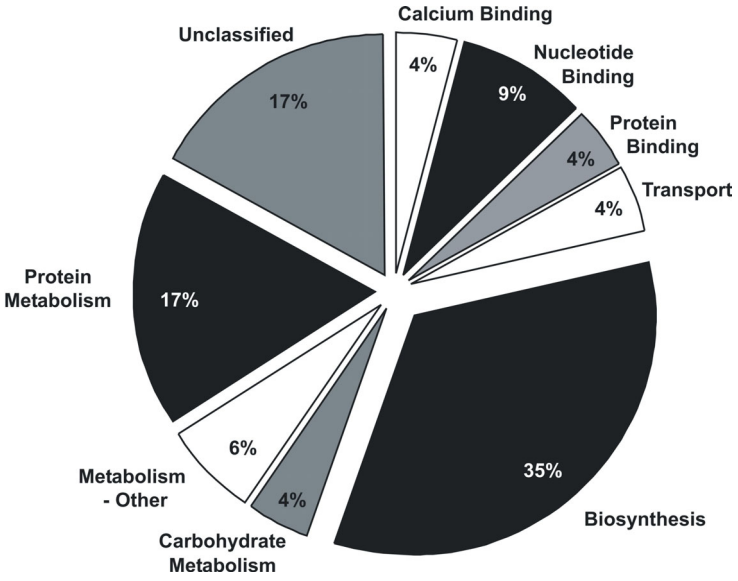

Figure 2.3: Summary of ontology classifications of the 47 *C. elegans* genes defining the COGs established for this study, as curated by WormBase release WS100, May 10, 2003; http://www.wormbase.org (Harris et al., 2003).

Phylogenetic analysis

Bayesian analysis recovered only four topologies with a nonzero approximated posterior probability (Figure 2.4). The topology which places *M. incognita* and *M. javanica* as sister taxa (Figure 2.4A) has a posterior probability of approximately 97.4%. In other words, 97.4% of the samples from the MCMC simulations support this conclusion. Further, the topologies provide 99.3% bipartition support grouping *M. incognita* and *M. javanica* as sister taxa. The next best topology (Figure 2.4B), with a posterior probability of approximately 1.89%, retains the relationships between *M. incognita* and *M. javanica*, and *M. arenaria*, differing from the first topology only in the inferred relationship between *M. hapla* and *M. chitwoodi*. The third topology (Figure 2.4C) groups *M. incognita* as most closely related to *M. arenaria*. The final topology (Figure 2.4D), which places *M. javanica* and *M. arenaria* together, as did the phylogeny reported from an analysis of 18S rRNA data (De Ley et al., 2002), has a posterior probability of only about 0.32%. Convergence was confirmed by the fact that each independent run of the Bayesian analysis resulted in topologies and parameters with similar posterior estimates, regardless of the randomly chosen initial state for the Markov chain.

Based on the posterior probabilities from the Bayesian analyses, we find strong support for the first topology (Figure 2.4A). In addition, we undertook maximum likelihood and protein distance analyses of our data. Both approaches produced phylogenies congruent with our Bayesian findings. Bootstrapping, conducted as part of the protein distance analysis, indicates high support for the relationships. Figure 2.5

summarizes the phylogeny we recovered, and indicates branch lengths, and posterior probabilities and bootstrap support for each bifurcation in the tree.

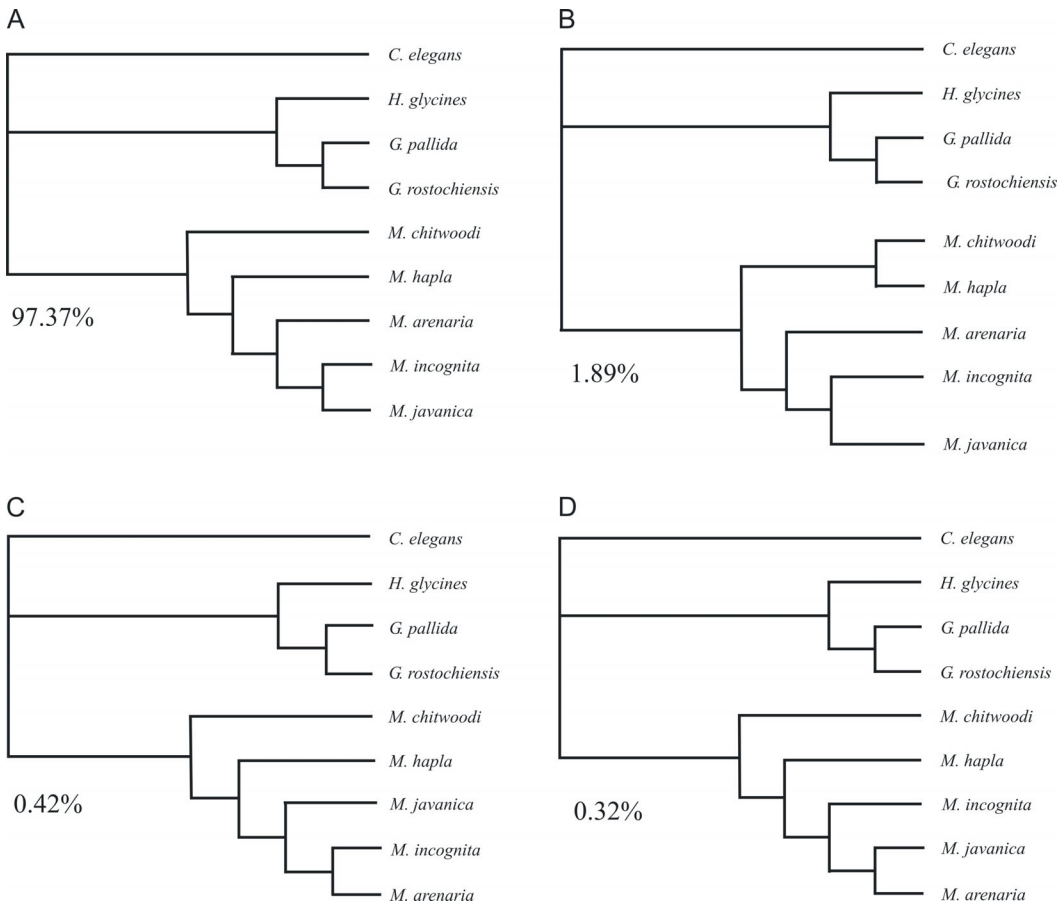

Figure 2.4: The four topologies with non-zero posterior probability (A–D) reported by the Bayesian analysis of 47 concatenated genes assembled from individual gene alignments of COG members. The posterior probability for each topology is indicated in %.

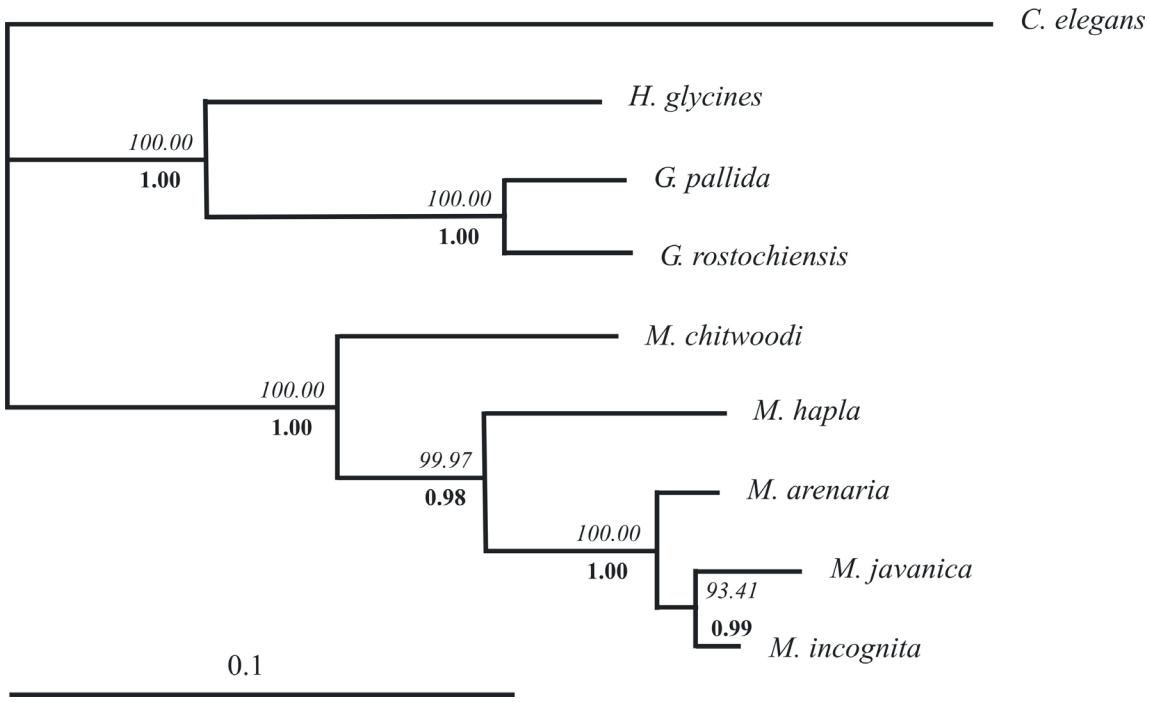

Figure 2.5: Phylogenetic reconstruction of tyenchid species, using *C. elegans* as an outgroup. Branch lengths, determined from Bayesian analysis, indicate number of nucleotide substitutions per site along the length of the branch. Scale bar reports estimated number of nucleotide substitutions per site. Bootstrap support for 10,000 replicates in protein distance analysis is indicated as italicized numerals, and posterior probability of clade support from Bayesian analysis is indicated in bold.

Discussion

The goal of many phylogenetic analyses is to establish the evolutionary relationship between a chosen set of species. However species trees can be obscured by reporting relationships specific to the chosen trait(s). We have employed a strategy to simultaneously examine many genes that are unlikely to have experienced differential evolutionary pressures. Through this strategy, we have potentially suppressed any sampling bias that may result in the inferred history of chosen gene to improperly reflect the evolutionary history of the species (Rokas et al., 2003).

Traits widely conserved across the phylum are less likely to have experienced accelerated evolution, and thus are more likely to successfully recover the true underlying species relationships. A convenient way to sample such traits is through direct sampling of gene sequences. We required that each tylenchid gene we sampled also be present in *C. elegans*. Tylenchida and Rhabditidae are in different major clades of the Nematoda (Blaxter et al., 1998). Therefore these genes are more likely to be evolutionarily conserved. Further, by selecting groups of genes based solely on sequence availability rather than some preconceived notion of which genes are more likely to resolve the phylogeny, we minimize biasing the results towards a specific relationship between genes, although as discussed above, certain loci (such as rRNA genes) present specific problems (such as gene conversion). That is not to say that any individual gene widely distributed across the phylum might not demonstrate a history which does not reflect the true relationship between the species, but the additional use of multiple genes will tend to counteract any spurious phylogenetic signal from such genes (Rokas et al., 2003). An additional requirement of gene-based phylogenetic analysis is that truly orthologous

genes be compared between species, and we employed stringent criteria to identify networks of orthologues across species (COGs).

Although we constructed our COGs with no *a priori* knowledge of gene function, it is instructive to examine the identity of the genes defining each COG. With only one exception (*F40E10.6*, specifying WormPep protein CE31508), all the genes encode proteins with identifiable biochemical function. Even *F40E10.6* clearly is a protein-coding gene, and although its RNAi phenotype is wild type (Kamath et al., 2003), it is broadly expressed in various tissues throughout the worms' lifecycle. Included in the list are genes originally defined using classical genetics, such as *daf-21* (HSP90), *dpy-14* (S-adenosylhomocysteine hydrolase), *egl-21* (Zn carboxypeptidase), and *let-70* (ubiquitin-conjugating enzyme), as well as genes inferred from sequence data, such as ribosomal subunit proteins (including *rpl-10*, *rpl-12* and *rps-16*), enzymes (such as the *F48E8.5* protein phosphatase 2A subunit) and structural proteins (such as the actin-binding, *K06A4.3* product). Collectively, the COGs defined here reflect a cross-section of what are mainly core metabolism genes (Figure 2.3). As such, they are likely to be slowly evolving genes. Taken together, our methods of gene selection in conjunction with the use of multiple genes for phylogenetic reconstruction is a powerful strategy to yield a tree reflecting the actual species relationships.

Resolution of the correct phylogeny for the five species of *Meloidogyne* allows for further dissection of biological traits found within the individual species. For example, there are at least two hypotheses which attempt to explain the triploid chromosome number observed in *M. arenaria*. The first argues that the increase in chromosome number is a recently acquired trait occurring as a result of a hybridization of two other

Meloidogyne species, leading to the genesis of *M. arenaria* (Triantaphyllou, 1985; Hugall et al., 1999). Another postulates that the triploid chromosome number is an ancestral trait and chromosome loss and/or fusion has occurred in more recently evolved species (Triantaphyllou, 1985). If *M. arenaria* is a hybrid of two other *Meloidogyne* species, our phylogeny suggests that this hybridization involves neither *M. incognita* nor *M. javanica*, as neither are ancestral to *M. arenaria*. Alternatively, the idea of triploid chromosome number being an ancestral trait and chromosome fusion creating the hypotriploid species, such as *M. incognita* and *M. javanica*, is consistent with our phylogenetic framework. Similarly, our phylogeny can help address the phenomenon of chromosome clustering during an unusually long prophase in *M. incognita*, which makes it difficult to discern distinct chromosomes in the oocytes (Triantaphyllou, 1985). The most parsimonious explanation of this trait suggests it was recently derived, rather than having existed in an ancestral population and subsequently lost in *M. arenaria* and *M. javanica*. A fully resolved phylogeny provides us with the framework needed to develop sound hypotheses for various characteristics of the different *Meloidogyne* species in a biologically relevant context. Issues as diverse as pathogen-binding to *Meloidogyne* (Davies et al., 2001) and host-choice (Bird and Kaloshian, 2003) are most fruitfully examined in light of a robust phylogeny.

A goal of the morphological approach to understanding evolution was to ultimately sample the entire genome by use of physical attributes. This multi-gene analysis is a first step towards accomplishing genome-wide coverage, but at a finer resolution than the morphological characteristics were able to provide. We expect that in the near future

complete genomes will be available for direct sampling, permitting the morphologists' goals to finally be realized.

Acknowledgements

The authors thank Drs. B. Adams, S. Aris-Brosou, K. G. Davies, J. P. McCarter and J. L. Thorne for their constructive criticism and advice. This research was supported by NSF Plant Genome award DBI0077503.

References

Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Bird D. McK., Clifton S. W., Kepler T., Kieber J. J., Thorne J., Opperman C. H. 2002. Genomic dissection of a nematode-plant interaction: A tool to study plant biology. *Plant Physiol.* 129:394-395.

Bird D. McK., Kaloshian I. 2003. Are roots special?: Nematodes have their say. *Physiol. Mol. Plant P.* 62:115–123.

Blaxter M. L., DeLey P., Garey J., Liu L. X., Scheldeman P., Vierstraete A., Vanfletern J., Mackey L. Y., Dorris M., Frisse L. M., Vida J. T., Thomas W. K. 1998. A molecular evolutionary framework for the phylum Nematoda. *Nature* 392:71-75.

Castagnone-Sereno P., Piotte C., Uijthof J., Abad P., Wajnberg E., Vanlerberghe-Masutti F., Bongiovanni M., Dalmasso A. 1993. Phylogenetic relationships between amphimictic and parthenogenetic nematodes of the genus *Meloidogyne* as inferred from repetitive DNA analysis. *Heredity* 70:195-204.

Dalmasso A., Bergé J. G. 1978. Molecular polymorphism and phylogenetic relationship in some *Meloidogyne* spp.: Application to the taxonomy of *Meloidogyne*. *J. Nematol.* 10:323-332.

Davies K. G., Fargette M., Balla G., Daud A. I., Duponnois R., Gowen S. R., Mateille T., Phillips M. S., Sawadogo A., Trivino C., Vouyoukalou E., Trudgill D. L. 2001.

Cuticle heterogeneity as exhibited by *Pasteuria* spore attachment is not linked to the phylogeny of parthenogenetic root-knot nematodes (*Meloidogyne* spp.). *Parasitology* 122:111-120.

DeLey I. T., DeLey P., Vierstraete A., Karssen G., Moens M., Vanfleteren J. 2002. Phylogenetic analyses of *Meloidogyne* small subunit rDNA. *J. Nematol.* 34:319-327.

Dickson D. W., Huisingh D., Sasser J. N. 1971. Dehydrogenases, acid and alkaline phosphatases, and esterases for chemotaxonomy of selected *Meloidogyne*, *Ditylenchus*, *Heterodera*, and *Aphelenchus* spp. *J. Nematol.* 3:1-16.

Dorris M., DeLey P., Blaxter M. L. 1999. Molecular analysis of nematode diversity and the evolution of parasitism. *Parasitol. Today* 15:188-193.

Doyle J. J. 1992. Gene trees and species tree: Molecular systematics as one-character taxonomy. *Syst. Bot.* 17:144-163.

Eisenbeck J. D., Triantaphyllou H. H. 1991. Root-knot nematodes: *Meloidogyne* species and races. In: Nickle, W. R. (Ed.), Manual of Agricultural Nematology. Marcel Dekker Inc., New York, pp. 191-274

Esbenshade P. R., Triantaphyllou A. C. 1987. Enzymatic relationships and evolution in the genus *Meloidogyne* (Nematoda: Tylenchida). *J. Nematol.* 19:8-18.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783-791.

Felsenstein J. 1993. PHYLIP (Phylogeny Inference Package) version 3.6a2. Department of Genetics, University of Washington, Seattle, WA

Harris T. W., Lee R., Schwarz E., Bradnam K., Lawson D. , Chen W., Blasier D., Kenny E., Cunningham F., Kishore R., Chan J., Muller H.-M., Petcherski A., Thorisson G., Day A., Bieri T., Rogers A., Chen C.-K., Spieth J., Sternberg P., Durbin R., Stein L. D. 2003. WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.* 31:133-137.

Hillis D. M., Dixon M. T. 1991. Ribosonal DNA: Molecular evolution and phylogenetic inference. *Quart. Rev. Biol.* 66:411-453.

Huelsenbeck J. P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754-755.

Hugall A., Stanton J., Mortiz C. 1999. Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic *Meloidogyne*. *Mol. Biol. Evol.* 16: 157-164.

Jones D. T., Taylor W. R., Thornton J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.* 8:275-282.

Kamath R. S., Fraser A. G., Dong Y., Poulin G., Durbin R., Gotta M., Kanapin A., Le Bot N., Moreno S., Sohrmann M., Welchman D. P., Zipperlen P., Ahringer J. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421:231-237.

Kumazawa Y., Nishida M. 1993. Sequence evolution of mitochondrial tRNA genes and deep-branch animal phylogenetics. *J. Mol. Evol.* 37:380-398.

Larget B., Simon D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16:750-759.

Lohar D. P., Bird D. McK. 2003. *Lotus japonicus*: A new model to study root-parasitic nematodes. *Plant Cell. Physiol.* 44:1176-1184.

McCarter J., Abad P., Jones J. T., Bird D. McK. 2000. Rapid gene discovery in plant parasitic nematodes via Expressed Sequence Tags. *Nematology* 2:719-731.

McCarter J. P., Mitreva M. D., Martin J., Dante M., Wylie T., Rao U., Pape D., Bowers Y., Theising B., Murphy C., Kloek A. P., Chiapelli B., Clifton S. W., Bird D. McK., Waterston R., 2003. Analysis and functional classification of transcripts from the root-knot nematode *Meloidogyne incognita*. *Genome Biol.* 4:R26.1-R26.19.

Mindell D. P., Honeycutt R. L. 1990. Ribosomal RNA in vertebrates: Evolution and phylogenetic applications. *Annu. Rev. Ecol. Syst.* 21:541-566.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568-583.

Parsons J. D. 1995. Improved tools for DNA comparison and clustering. *Bioinformatics* 11:603-613

Powers T. O., Platzer E. G., Hyman B. C. 1986. Species-specific restriction site polymorphism in root-knot nematode mitochondrial DNA. *J. Nematol.*18:288-293.

Powers T. O., Sandall L. J. 1988. Estimation of genetic divergence in *Meloidogyne* mitochondrial DNA. *J. Nematol.* 20:505-511.

Rokas A., Williams B. L., King N., Carroll S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798-804.

Sasser J. N. 1980 Root-knot nematodes: A global menace to crop production. *Plant Dis.* 64:36-41.

Scholl E. H., Thorne J. L., McCarter J. P., Bird D. McK. 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. *Genome Biol.* 4:R39.

Slowinski J. B., Page R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Syst. Bio.* 48:814-825.

Smith A. B. 1989. RNA sequence data in phylogenetic reconstruction: Testing the limits of its resolution. *Cladistics* 5:321-344.

Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957-966.

Tatusov R. L., Natale D. A., Garkavtsev I. V., Tatusova T. A., Shankavaram U. T., Rao B. S., Kiryutin B., Galperin M. Y., Fedorova N. D., Koonin E. V. 1997. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29:22-28.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012-2018.

Thompson J. D., Higgins D. G., Gibson T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.

Triantaphyllou, A.C. 1966. Polyploidy and reproductive patterns in the root-knot nematode *Meloidogyne hapla*. *J. Morphol.* 118:403-414.

Triantaphyllou, A.C. 1985. Cytogenetics, cytotaxonomy, and phylogeny of root-knot nematodes. In: J.N. Sasser and C.C. Carter, (Eds.), An advanced treatise on *Meloidogyne*, vol. I. Biology and control. North Carolina State University Graphics, Raleigh, NC, pp. 113-126.

Wheeler D. L., Church D. M., Lash A. E., Leipe D. D., Madden T. L., Pontius J. U., Schuler G. D., Schriml L. M., Tatusova T. A., Wagner L., Rapp B. A. 2002. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.* 30:13-16.

Wu, C. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* 127:429-435.

Xue B., Baillie D., Beckenbach K., Webster J. 1992. DNA hybridization probes for studying the affinities of three *Meloidogyne* populations. *Fun. App. Nematol.* 15:35-41.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comp. App. Biosci.* 13:555-556.

Yang Z., Goldman N., Friday A. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11:316-324.

Appendix A

WormPep identification numbers for the *C. elegans* genes that are members of the 47 COGs used in this study, as reported from WormBase (http://wormbase.org):

CE00664, CE01030, CE01225, CE01236, CE01253, CE01543, CE02249, CE02255, CE02618, CE02883, CE03025, CE03278, CE03482, CE03567, CE04561, CE04821, CE05441, CE05598, CE05860, CE06090, CE06107, CE06577, CE06652, CE07033, CE07370, CE07537, CE09156, CE09650, CE09654, CE09682, CE09901, CE10884, CE11052, CE12898, CE12918, CE14956, CE16260, CE16954, CE17986, CE20218, CE20547, CE27706, CE30997, CE31508, CE31508, CE31867, CE32364.

Chapter 3

Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach

Elizabeth H. Scholl[*†], Jeffrey L. Thorne[†], James P. McCarter[‡§] and David McK. Bird[*†°]

[*]Center for the Biology of Nematode Parasitism, Box 7253, North Carolina State University, Raleigh, NC 27695, USA. [†]Bioinformatics Research Center, Box 7566, North Carolina State University, Raleigh, NC 27695, USA. [‡]Genome Sequencing Center, Department of Genetics, Box 8501, Washington University School of Medicine, St. Louis, MO 63108, USA. [§]Divergence Inc., 893 North Warson Road, St. Louis, MO 63141, USA.

[°]Corresponding author

David Bird
Center for the Biology of Nematode Parasitism
Box 7253, NC State University
Raleigh, NC 27695-7253
Phone: +1 (919) 515-6813
Fax: +1 (919) 515-9500
e-mail: david_bird@ncsu.edu

Published in Genome Biology 2003, 4:R39

Abstract

Background

Published accounts of horizontally acquired genes in plant-parasitic nematodes have not been the result of a specific search for gene transfer *per se*, but rather have emerged from characterization of individual genes. We present a method for a high-throughput genome screen for horizontally acquired genes, illustrated using expressed sequence tag (EST) data from three species of root-knot nematode, *Meloidogyne* species.

Results

Our approach identified the previously postulated horizontally transferred genes and revealed six new candidates. Screening was partially dependent on sequence quality, with more candidates identified from clustered sequences than from raw EST data. Computational and experimental methods verified the horizontal gene transfer candidates as bona fide nematode genes. Phylogenetic analysis implicated rhizobial ancestors as donors of horizontally acquired genes in *Meloidogyne*.

Conclusions

High-throughput genomic screening is an effective way to identify horizontal gene transfer candidates. Transferred genes that have undergone amelioration of nucleotide composition and codon bias have been identified using this approach. Analysis of these horizontally transferred gene candidates suggests a link between horizontally transferred genes in *Meloidogyne* and parasitism.

Introduction

Nematodes are the most abundant and speciose metazoans, and account for up to 80% of the kingdom's members (Boucher and Lambshead, 1994). Not surprisingly, nematodes have evolved to occupy diverse ecological niches. Like the well-studied *Caenorhabditis elegans*, most are free-living and graze on microbes or detritus, and as such, have no obvious direct impact on humans. Others, however, are adapted as parasites and are responsible for such widespread problems as human disease, debilitation of livestock and crop damage. Plant-parasitic forms are responsible for an estimated $100 billion in annual crop damage worldwide (Koenning et al.,1999). The most damaging family (the Heteroderidae) includes the root-knot (*Meloidogyne* spp.) and the cyst (*Globodera* and *Heterodera* spp.) nematodes. Root-knot nematodes penetrate plant hosts and migrate between the cells in roots, where they induce formation of large multinucleate cells called 'giant cells'. Galls form around the giant cells, and the roots become distorted, often leading to compromised root function and retardation of plant growth (Bird and Koltai, 2000).

It is not clear which genetic differences between the plant-parasitic and non-parasitic forms may be responsible for conferring parasitic ability. On the basis of phylogenetic analysis (Blaxter et al., 1998) it appears that plant-parasitism arose independently at least three times over the course of nematode evolution. Consequently, one cannot be assured that any gene or set of genes that aid in the parasitic lifestyle in one nematode species will also exist in another. Conceptually, several mechanisms affecting evolution to parasitism can be envisioned. These include adaptation of pre-existing genes to encode new functions; changes in genes regulating metabolic or developmental pathways; gene

duplication; gene loss; and acquisition of genes from other species (horizontal gene transfer, HGT). HGT has become a widely accepted mechanism of rapid evolution and diversification in prokaryotic populations (Jain et al., 1999; Lawrence, 1999; Ochman et al., 2000). Recent genome analyses of primitive eukaryotes, such as the sea squirt (*Ciona intestinalis*) (Dehal et al., 2002) and single-celled parasitic diplomonads (Andersson et al., 2003), implicate HGT events in early eukaryotic evolution. In contrast, the extent of horizontal transfer involving higher eukaryotes has been controversial, with many cases of hypothesized horizontally transferred genes (Stephens et al., 1998; Wolf et al., 1999; Lange et al., 2000; Royo et al., 2000; International Human Genome Sequencing Consortium, 2001) having been refuted by later studies (Stanhope et al., 2001; Brinkman et al., 2002).

On the basis of biochemical and immunological criteria, genes have been identified in *Globodera rostochiensis* and *Heterodera glycines* that allow these nematodes to endogenously produce enzymes that can degrade cellulose and pectin, the two major components of plant cell walls. A possible ancient bacterial origin of these genes has been theorized (Smant et al., 1998; Yan et al., 1998; Popeijus et al., 2000). A bacterial origin for a number of root-knot nematode (RKN) genes also has been proposed, although their possible role in parasitism is less clear. Some, such as a gene encoding chorismate mutase (Lambert et al., 1999), were likewise identified on the basis of biochemical properties, whereas others, including a polygalacturonase gene (Jaubert et al., 2002), were identified from expressed sequence tag (EST) datasets, the latter from our data (McCarter et al., 2000) using a keyword search. Veronico et al. (2001) isolated a presumed polyglutamate synthetase gene with bacterial homology by sequencing

neighboring regions of the *M. artiellia* chitin synthetase locus. We wished to determine whether other RKN genes might have been acquired by horizontal gene transfer, particularly as such genes might potentially be related to parasitism.

Claims of HGT have frequently pivoted on incongruencies between a particular gene tree and the assumed underlying species tree. Acquisition of new sequence data has often revealed that genes believed to be absent in a species were merely missing in the database rather than missing from the genome (Stanhope et al., 2001). Obviously, because full genomes are not available for all plant and animal species, we are not able to make definitive statements about the presence or absence of a particular gene in every organism. However, with the completed *C. elegans* genome available as a reference 'model' nematode, it is now possible to examine the emerging genetic resources for *Meloidogyne* comprehensively, to begin to address the question of evolution of parasitism and, in particular, a possible role for HGT.

A similarity to a bacterial protein sequence is the simplest criterion for considering a nematode protein, and thus the gene that encodes it, as a possible HGT candidate. For that candidate truly to define an HGT event, its presence must be incongruent with nematode phylogeny (Figure 3.1). Nevertheless, the presence of a gene in one nematode species (such as *Meloidogyne*) but its absence in another (such as *C. elegans*) might merely reflect a gene loss in the latter lineage. In addition to *C. elegans*, several other invertebrate genomes have been completely sequenced, and at the time of this study, the best characterized of those was *Drosophila melanogaster*. Consequently, we chose this resource as a tool to identify genes which may be present in nematodes, but which are absent in *C. elegans*. A bacteria-like gene present in *Meloidogyne* and *Drosophila*, but

absent in *C. elegans*, is unlikely to have experienced HGT, but may rather reflect a gene loss in the *C. elegans* lineage. We therefore developed a 'phylogenetic filter' based on these relationships to rapidly reveal *Meloidogyne* HGT candidates identified by sequence similarity to bacterial proteins. The intent of this filter is to efficiently eliminate spurious HGT candidates.

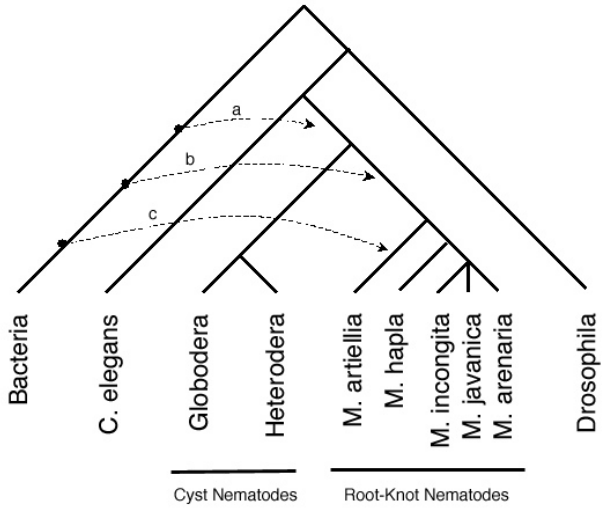

Figure 3.1: Schematic species tree indicating relationships between *Drosophila*, *C. elegans* and plant-parasitic nematodes in the Heteroderidae. The locations of three possible horizontal events that would pass through our initial phylogenetic filter by dotted lines. Transfer 'a' occurs after divergence of the to *C. elegans* and Heteroderidae, transfer 'b' after divergence nematodes and cyst nematodes, and transfer 'c' to the lineage specific *Meloidogyne* species. Adapted from Tandingan-De Ley (2002).

Surprisingly, the relationship between the invertebrate phyla Nematoda and Arthropoda (which includes *Drosophila*) is controversial. The traditional view is that arthropods are more closely related to annelids than to nematodes, but some recent molecular phylogenies place nematodes and arthropods together in a high-level taxon named Ecdysozoa, which does not include annelids (Aguinaldo et al., 1997; Mallat and

Winchell, 2002). Other molecular studies give conflicting results (Blair et al., 2002; Hedges 2002). Regardless of the evolutionary relationship between Nematoda and Arthropoda, *C. elegans* and *Drosophila* remain useful and valid models for our analyses, and the relationships shown in Figure 3.1 are consistent with both hypotheses.

Genes that were transferred from bacteria to nematodes would pass through our phylogenetic filter if the transfer event occurred subsequent to the divergence of the *C. elegans* and *Meloidogyne* lineages (Figure 3.1). Should a gene appear to be present in other closely related plant-parasites, such as the cyst nematodes, the transfer event probably affected a common ancestor of the two families of parasitic nematodes (event 'a' in Figure 3.1). Alternatively, the transfer event may be more recent, such as to the progenitor of the *Meloidogyne* lineage since its divergence from the cyst nematodes (event 'b' in Figure 3.1), or in a lineage leading to a single *Meloidogyne* species (event 'c').

Although bacteria-like *Meloidogyne* genes that are not present in *C. elegans* and *Drosophila* comprise a preliminary pool of candidates, multiple gene loss may be responsible for the presence/absence pattern revealed by the filter. To test this more thoroughly, we established a screen to compare the now small pool of preliminary candidates with all other sequences in the public databases. The most parsimonious explanation to be drawn from candidates with no significant matches to any metazoan genes is that they arose by horizontal gene transfer from a non-metazoan pool, as opposed to multiple independent gene losses in the metazoan lineages. Candidates thus identified were subsequently validated through phylogenetic analysis of relationships between the most similar matches from our screening processes

We describe here a comprehensive two-step search for HGT candidates in *M. incognita*, *M. javanica* and *M. hapla* using EST data (McCarter et al., 2000; 2002; http://www.nematode.net). Genome-to-genome comparisons were made to discover patterns of presence and absence that would indicate horizontally acquired genes. Second, kingdom-wide comparisons further reduced the candidate pool; these genes were then examined from an evolutionary standpoint. Twelve *Meloidogyne* candidates were discovered and their potential role in plant pathogenicity is discussed.

Materials and methods

Available data

Sequences were obtained from the Parasitic Nematode Sequencing Project (PNSP) (http://www.nematode.net) including clustered *Meloidogyne* ESTs built with the NemaGene approach (McCarter et al., 2003). We analyzed 1,799 *M. incognita* (WMi) sequences and 3,119 *M. javanica* (WMj) sequences from these PNSP clusters. Additional raw sequences were extracted from the July 31, 2002 NCBI GenBank dbEST build with the Entrez Search and Retrieval System (Table 3.1) (Wheeler et al., 2002). *Meloidogyne incognita* and *M. javanica* datasets from NCBI (NMi and NMj respectively) contain the individual ESTs generated by the PNSP, and from which the clusters for the WMi and WMj datasets were generated. In addition, the NMi and NMj datasets included some sequences from sources other than the PNSP. *M. hapla* sequences (NMh) were also retrieved from NCBI. Entrez was used to extract all available nuclear sequences for *D. melanogaster*, *C. elegans* and bacterial sequences from the GenBank non-redundant (nr) database (May 1, 2002 build).

Table 3.1: Efficiency of each step of screening *Meloidogyne* datasets for HGT candidates

Name	Original	1st Screen	2nd Screen	Final Candidates
WMi	1,799	16 (0.889%)	12 (0.667%)	12 (0.667%)
WMj	3,119	11 (0.353%)	7 (0.224%)	7 (0.224%)
NMi	12,841	99 (0.771%)	27 (0.210%)	5 (0.038%)
NMj	5,630	54 (0.959%)	16 (0.284%)	6 (0.107%)
NMh	6,514	4 (0.061%)	0 (0.00%)	0 (0.00%)

Clustered ESTs (W) were from the Parasitic Nematode Sequencing Project at Washington University. Raw ESTs (N) were extracted from NCBI's GenBank. Mi, *Meloidogyne incognita*; Mj, *M. javanica*; Mh, *M. hapla.* 'Original number' gives the size of the initial dataset. For both screens, matches were declared when e-values were less than $1.0e^{-10}$. The percentage of the original number of sequences remaining after each screen is listed in parentheses. 'Final candidates' reflects total number of candidates after removal of redundancy.

Candidate search algorithm

Analyses of the WMi and WMj data were performed via a local installation of WU-BLAST 2.0 (http://blast.wustl.edu). Each sequence in WMi and WMj was extracted into individual FASTA format files using Perl scripts and submitted for three six-phase translated WU-BLASTX searches, once each against the *C. elegans*, *Drosophila* and bacterial protein databases. WU-BLASTX parameters were E = 10, W = 3, T = 12. E-values were extracted for the best match for each query sequence in each of the three searches.

Meloidogyne sequences from NCBI were analyzed using the Tera-BLAST Hardware Accelerated BLAST algorithm (TimeLogic, Crystal Bay, NV). Single FASTA files were submitted for three six-phase translated Tera-TBLASTX queries against six-phase translated *C. elegans* and *Drosophila* genomic databases. Tera-TBLASTX parameters were Open Penalty = 8, Extend Penalty = 2, Word Size = 4, Query Increment = 3 and Neighborhood Threshold = 18. Perl scripts were employed to parse the query name and associated best e-value from each of the nine analyses (three each for NMi, NMj and NMh).

As a first round of phylogenetic filtering, automated comparison of e-values for each sequence allowed us to eliminate sequences with a best match to either *C. elegans* or *Drosophila* from further analysis. The remaining sequences, those with a best match to bacteria of order $1.0e^{-10}$ or better, provided a preliminary pool of candidates for each dataset. A BLASTX search was carried out for each candidate against the nr database, using the above parameters. The results from this second filter were examined and any

Table 3.2: List of horizontal gene transfer candidates from *M. incognita*

	Best Bacterial Match		Best Eukaryotic Match*	
	Name	e-value, %identity	Name	e-value, %identity
β-1,4-endoglucanases				
MI00537	*Bacillus* sp. KSM-N252	($2.7e^{-24}$, 40%)	*Orpinomyces joyonii*	($5.6e^{-10}$, 32%)
MI01011	*Pseudomonas fluorescens*	($2.5e^{-75}$, 47%)	*Orpinomyces joyonii*	($9.4e^{-41}$, 36%)
MI01381	*Streptomyces coelicolor*	($6.9e^{-13}$, 31%)	*Orpinomyces joyonii*	(0.013, 27%)
MI01842	*Pseudomonas fluorescens*	($1.2e^{-35}$, 44%)	NONE	
Pectinases				
MI00252	*Ralstonia solanacearum*	($8.8e^{-61}$, 50%)	*Arabidopsis thaliana*	($5.1e^{-7}$, 40%)
MI00592	*Streptomyces coelicolor*	($3.9e^{-12}$,31%)	*Fusarium solani*	($1.9e^{-7}$, 33%)
Rhizobia Matches				
NodL	*Rhizobium leguminosarum*	($8e^{-54}$, 58%)	*Saccharomyces cerevisiae*	($5e^{-38}$, 46%)
Glutamine synthetase	*Mesorhizobium loti*	($9e^{-45}$, 56%)	*Blumeria graminis*	($2e^{-15}$, 33%)
L-threonine aldolase	*Brucella melitensis*	($1e^{-23}$, 48%)	*Leishmania major*	(0.096, 25%)
Unknown function	*Sinorhizobium meliloti*	($9e^{-45}$, 51%)	*Caenorhabditis elegans*	(3.9, 26%)
Unknown Function				
MI01406	*Amycolatopsis mediterranei*	($4.9e^{-28}$, 53%)	*Arabidopsis thaliana*	($2.5e^{-4}$, 33%)
MI00267	*Amycolatopsis mediterranei*	($3.0e^{-28}$, 58%)	*Aspergillus fumigatus*	($5.4e^{-6}$, 32%)

The best bacterial and eukaryotic matches are listed with their e-values from a BLASTX search and percent identity as reported by BLAST.

*Best match to any eukaryote other than a plant-parasitic nematode.

sequence with a significant match to a metazoan other than a closely related plant-parasitic nematode was removed from further analysis. An e-value of $1.0e^{-10}$ was the threshold used to declare a match. The remaining sequences provided our final set of candidates for horizontally transferred genes (Tables 3.1, 3.2).

Codon usage analysis

The protein alignment of the *M. incognita* and *R. leguminosarum nodL* sequences was trimmed such that only identical amino acids remained, and the sequences back-translated, retaining the correct codon usage. Ten thousand pairs of simulated sequences were generated by independently permuting the homologous codon pairs in the actual data. In other words, the probability that the i^{th} codon in the first simulated sequence was assigned the i^{th} codon from the actual *M. incognita* sequence and the i^{th} codon in the second simulated sequence was assigned the i^{th} codon from the actual *R. leguminosarum* sequence was set to 0.5 and the probability that the i^{th} codon assignments in the simulated sequences were reversed was also set to 0.5. Codon adaptation indices were computed for each simulated sequence using the EMBOSS suite of sequence analysis tools (Rice et al., 2000).

Phylogenetic analysis of candidates

For each candidate, the protein sequences for the top 15 matches with an e-value of $1.0e^{-10}$ or less were extracted from the BLASTX search against the nr database. If there were not 15 matches with an e-value meeting this criterion, all sequences with e-values lower than $1.0e^{-10}$ were selected. Alignments of these sequences with the translated

candidate sequence were constructed with CLUSTAL-X (Thompson et al., 1997); improvements to the CLUSTAL-X alignments were performed manually. Sequences from the same species with more than 95% identity after alignment were considered possible paralogs and deemed redundant information for this analysis. Only one sequence from each of these sets was used in further analysis. Poorly aligned sequences were also discarded.

Distances between aligned proteins were estimated with the Dayhoff amino-acid replacement model (Dayhoff, 1978). Tree topologies were then inferred from these distances via neighbor-joining (Saitou et al., 1987) and 1,000 non-parametric bootstrap replicates were used to estimate clade support (Felsenstein, 1985). Maximum likelihood analysis produced topologies consistent with the neighbor-joining analysis. All phylogenetic reconstructions were performed with the PHYLIP and PAML software packages (Felsenstein, 1993; Yang, 1997).

Additional analyses of the putative *nodL* gene were conducted with Version 3.0b4 of the MrBayes software (Huelsenbeck et al., 2001). For these analyses, the Jones-Taylor-Thornton model of amino-acid replacement (Jones et al., 1992) was adopted and variation of replacement rates among sites was incorporated by a discretized gamma distribution with four rate categories (Yang, 1994). Each Markov chain Monte Carlo analysis used four heated chains and employed a burn-in period of 10,000 cycles, followed by 990,000 additional cycles. Convergence of the Markov chain was diagnosed by performing two different runs from different initial parameter states. Prior distributions for all parameters were the default distributions incorporated in the MrBayes software.

Results and discussion

Genome-to-genome comparisons act as a phylogenetic filter in candidate searching

Given the large number of sequences to examine and the expectation that most were not horizontally acquired, we developed a phylogenetic filter based on genome-to-genome sequence comparisons. Further, because the available data included raw ESTs from the National Center for Biotechnology Information (NCBI) GenBank (dbEST) as well as clustered ESTs from the Parasitic Nematode Sequencing Project (McCarter et al., 2000; 2002; 2003), for which the data can be presumed to be significantly more reliable, we wished to compare the efficiency of reducing each dataset with this filter. *Meloidogyne* sequences from NCBI dbEST (*M. incognita*, *M. javanica* and *M. hapla* sequences, named NMi, NMj, and NMh respectively) were translated in six frames and individually compared to conceptual six-phase translations of the *C. elegans* and *Drosophila* genomes as well as all available bacterial sequences. This first filter, which makes no assumptions about gene annotation in the target genomes, and which employed the relatively error-prone raw ESTs, reduced the pool of HGT candidates by eliminating more than 99% of the original ESTs for all three species tested (Table 3.1). Using clustered ESTs (*M. incognita* and *M. javanica* sequences, named WMi and WMj) as queries to the worm, fly and bacterial protein databases (which are based on gene annotation) produced a similar degree of reduction (Table 3.1). Importantly, genes previously predicted to be the result of HGT events were identified by, and passed through, the phylogenetic filter.

The main objective of the phylogenetic filter was to reduce the computational load necessary to screen HGT candidates against all metazoan proteins. A second filter, consisting of a BLAST analysis against the GenBank nonredundant (nr) protein database, served to eliminate genes that may have been independently lost in the *C. elegans* and *Drosophila* lineages, but are still representative of a more ancient animal gene (Table 3.1). This filter eliminated four candidates from the WMi data set. Examination of these showed a putative copper homeostatis protein and a protein of unknown function, both with significant matches to *Homo sapiens* (e-values of $1.10e^{-23}$ and $4.20e^{-18}$ respectively), one aldehyde dehydrogenase with a significant match to *Mus musculus* ($2.70e^{-26}$) and one asparaginyl-tRNA synthetase. Three of the four had best matches to bacteria (Table 3.3). Interestingly, manual inspection revealed that all four sequences did have significant matches to *C. elegans*, but passed through our initial phylogenetic filter because the bacterial matches were stronger than those for *C. elegans* or *Drosophila*. The 12 final candidates in WMi had no significant match to *C. elegans* or *Drosophila* in the preliminary screen. The best eukaryotic matches to these candidates from the BLAST search against nr are shown in Table 3.2. The second filter generated similar enrichment in WMj, reducing the number of candidates from eleven to seven.

The fact that more candidates from the raw datasets were eliminated during second-round filtering (for example, from 99 to 27 in NMi) reflects the redundancy in the datasets. If multiple EST sequences representing a single gene pass through the first filter, each of those EST sequences will be in the preliminary candidate pool. The second filter is likely to simultaneously remove more than one of these homologous sequences if it removes any at all. Therefore, searching with raw EST sequences is likely to result in a

larger absolute decrease in the candidate number than will searching with clustered EST sequences.

Table 3.3: Sequences from WMi that passed the preliminary screen but were removed from candidate pool after second screen

NemaGene ID	Putative Function	Bacteria	*Drosophila*	*C. elegans*	Other
MI01839	Copper homeostatis protein	**3.90e^{-31}**	1.70e^{-20}	4.50e^{-22}	1.10e^{-23} *(Homo Sapiens)*
MI00665	Aldehyde dehydrogenase	4.30e^{-22}	1.40e^{-10}	4.30e^{-18}	**2.70e^{-26}** (*Mus musculus)*
MI01016	Asparaginyl-tRNA synthetase	**2.30e^{-46}**	1.80e^{-35}	4.30e^{-29}	4.50e^{-39} (*Arabidopsis thaliana)*
MI00754	Hypothetical protein	**3.70e^{-61}**	9.80e^{-01}	8.10e^{-17}	4.20e^{-18} (*Homo Sapiens*)

The best match in the preliminary screen was to bacteria. Significant matches to other eukaryotes (including *C. elegans* and *Drosophila*) exist for each sequence. E-value for overall best match is listed in **bold**.

The number of final candidates listed in Table 3.1 are candidate HGT genes after clustering. A smaller number of candidates was discovered from the raw EST datasets compared with the clustered sequences, which suggests that our method of HGT candidate searching is partially dependent on sequence quality. The lower number of final candidates obtained using raw EST data is principally due to filtering of areas of low complexity and tandem repeats, and uncertainty of similarity matching for shorter sequences during BLAST searches. Similarly, the size of the dataset influences the number of final candidates obtained. Thus, the absence of candidates in *M. hapla* is likely to be due to a combination of the small number of unique ESTs analyzed (because of redundancy in the data), and possibly overall quality of the raw ESTs, rather than to a

lack of laterally acquired genes in the genome. Despite the lowered efficiency of candidate discovery when using the lower-quality, raw EST sequences, this tool was able to recover five candidates from the NMi dataset, compared to the 12 candidates identified from the higher-quality clustered sequences in the WMi dataset. The fact that candidates were discovered across disparate sequence-quality conditions not only provides additional validation of our methods, but also suggests a high degree of flexibility and robustness in the tool.

Identification of previously hypothesized HGT candidates

The literature reports seven genes postulated to have been horizontally acquired by *M. incognita*, *M. hapla* or *M. javanica* during evolution of plant-parasitic nematodes (Smant et al., 1998; Yan et al., 1998; Lambert et al., 1999; Popeijus et al., 2000; Jaubert et al., 2002); our search algorithm revealed six of these genes. The notable exception is *Mj-CM*, which is postulated to encode chorismate mutase in *M. javanica* (Lambert et al., 1999). To examine why this gene was not identified by our filtering process, we used both *Mj-CM* sequences found in GenBank (AF095949, AF095950) in a series of BLASTX queries. No significant matches were found in the *Drosophila*, *C. elegans* or bacterial databases, nor in the *Meloidogyne* datasets used in this study. Recent BLAST searches at nematode.net (http://www.nematode.net) against all *Meloidogyne* ESTs, including sequences not available when our analyses were first conducted confirm that the chorismate mutase gene is absent from WMi and WMj, although a single, significant match to an *M. arenaria* chorismate mutase EST was revealed. Another RKN gene also postulated to have been acquired by HGT, and which encodes polyglutamate synthetase,

was previously identified in *M. artiellia* (Veronico et al., 2001). Significantly, hybridization data showed that this particular gene is absent from both the *M. javanica* and *G. rostochiensis* genomes (Veronico et al., 2001). We speculate that acquisition of this gene by *M. artiellia* is a recent HGT event (event 'c', Figure 3.1), and thus it is truly absent from the *Meloidogyne* genomes from which our datasets were derived. In other words, failure to 'discover' this gene was not a failure of our screening process, but is likely to be a correct reflection of the biology.

The most extensively studied HGT candidates are four genes encoding β-1,4-endoglucanase, initially identified in the cyst nematodes *G. rostochiensis* and *H. glycines* (Smant et al., 1998; Yan et al., 1998). These four genes (NemaGene Contig IDs MI00537, MI01011, MI01381 and MI01842) (http://www.nematode.net) appear to define two sets of paralogs formed before divergence of the cyst and root-knot nematodes. As noted (Smant et al., 1998; Yan et al., 1998), β-1,4-endoglucanases presumably equip these nematodes with the ability to endogenously degrade the most abundant component of cell walls, namely cellulose. Similarly, the second most abundant component of cell walls (pectin) is the assumed target of nematode-encoded pectate lyase and exo-polygalacturonase, both functions also postulated to have been acquired by HGT. The pectate lyase gene (MI00592) was identified in *G. rostochiensis* and *H. glycines* (Popeijus et al., 2000) and the exo-polygalacturonase (MI00252) was identified in our *M. incognita* data (Jaubert et al, 2002; McCarter et al., 2000). Because of the obvious role of nematode genes that allow endogenous production of cell-wall degrading enzymes in attacking a plant host, it has been hypothesized that their acquisition by HGT may have been key steps in the evolution of plant-parasitic nematodes from ancestral free-living

forms (Bird and Koltai, 2000). In that model, an intermediate, symbiotic association of a soil-dwelling (but free-living) nematode with a soil bacterium possessing these enzymes is postulated before the HGT event. It was suggested (Bird and Koltai, 2000) that acquisition of these new functions (either by symbiosis or HGT) permitted previously free-living nematodes to expand their range into a new ecological niche (the plant) as a prelude to speciation into parasitic forms.

Also revealed by our tool were six new HGT candidates, including homologs for glutamine synthetase, L-threonine aldolase and *nodL*, and three to which function could not be unequivocally ascribed.

Rhizobial origin of Meloidogyne genes

Of the six newly identified HGT candidates, four have highest similarity to genes in the nitrogen-fixing soil bacteria that nodulate plant roots and which are collectively termed rhizobia. *Meloidogyne* and rhizobia are sympatric (that is, they share an ecological niche in the soil (Bird and Koltai, 2000), and arguably in the plant too (Koltai et al., 2001), satisfying the minimal requirement for an HGT to occur, namely physical proximity. Interestingly, models of bacterial evolution suggest HGT as a mechanism of adaptation into either symbiosis or parasitism (Ochman and Moran, 2001). This is specifically thought to be the case for divergent species of rhizobia, such as the symbiont *Sinorhizobium meliloti* and the pathogen *Rhizobium radiobacter* (formerly known as *Agrobacterium tumefaciens*), where differential selection and gene maintenance is likely to be responsible for different lifestyle strategies (Wood et al., 2001).

Two of the *Meloidogyne* genes revealed by our filters, which encode an L-threonine aldolase gene (MI01644) and a deduced protein of unknown function (MI00109), exhibit striking amino-acid identity to rhizobial proteins (48% and 51% respectively), but a complete absence of meaningful homology with any eukaryotic sequence (Table 3.2). Consequently, these genes are strong candidates for having entered nematodes via HGT, presumably from a rhizobial ancestor.

The deduced product of a third *M. incognita* gene (MI00426) has striking sequence similarity to glutamine synthetase (GS). Glutamine synthetases fall into two structurally and functionally distinct classes. GSI, which to date appears restricted to prokaryotes (Ludwig, 1980), is involved in ammonium assimilation as part of the nitrogen-fixation pathway in rhizobia (Turner and Young, 2000). The ability to be reversibly adenylylated at Tyr397 of the active site is a characteristic of GSI. The second class, GSII, is found in all eukaryotes and a small number of prokaryotes, and appears to be involved in purine synthesis (Turner and Young, 2000). Unlike GSI, GSII is not adenylylated (and lacks the conserved tyrosine). On the basis of both amino-acid sequence similarity (Table 3.2) and a Pfam (Bateman et al., 2002) HMM search (e-value $4.3e^{-24}$), it is clear that the RKN glutamine synthetase is a GSI homolog, implying a prokaryotic origin. Strikingly, the nematode protein has greatest similarity (56% amino-acid identity) to GSI from the rhizobial bacterium, *Mesorhizobium loti*, including conservation of Tyr397. The best match to a eukaryotic glutamine synthetase (GSII) is substantially lower (Table 3.2), strongly implicating the RKN gene as a robust candidate for an HGT event.

The fourth rhizobial-like HGT candidate (MI01045) identified by our filter has 58% amino-acid identity ($8.8e^{-54}$) to NodL from *Rhizobium leguminosarum* (Table 3.2). This

protein encodes an N-acetyltransferase previously thought to be present only in rhizobia (Downie and Young, 2001), where it functions in the biosynthesis of Nod factor. Nod factors are a rhizobial species-specific family of lipo-chito-oligosaccharides which function in signal exchange between the bacterium and its symbiotic partner plant (Van Rhijn and Vanderleyden, 1995). The first visible signs of nodule formation (root-hair deformation) as part of the symbiotic pathway are triggered by Nod factors (Göttfert, 1993), and although the specific mechanisms of Nod factor function remain unknown, it is clear that it has a central role in initiation of cell division and possibly also nodule differentiation in the root (Ditt et al., 2001). For most rhizobia, the product of *nodD* acts as a transcriptional activator and induces expression of a set of *nod* genes. Experimental evidence (Göttfert, 1993) shows that lack of either *nodABC* or *nodD* in rhizobia results in a Nod$^-$ phenotype (that is, a strain unable to initiate nodule formation on the host plant). By contrast, *R. radiobacter*, which forms a parasitic relationship with plants by producing a crown gall rather than nodules, lacks these genes, and appears to possess only *nodL*, *nodX* and *nodN*, suggesting these three *nod* genes are sufficient to affect root growth and are involved in a parasitic lifestyle rather than being specific to symbiosis (Wood et al., 2001).

To examine further the relationship between putative *nodL* candidates found in *M. incognita* and *M. javanica* with the cognate genes in rhizobia, we undertook a phylogenetic analysis and found that the two nematode genes fall squarely within the rhizobial *nodL* clade (Figure 3.2). This analysis further grouped other sequences with significant similarity to the deduced *Meloidogyne* NodL protein. Not surprisingly, these enzymes clustered according to specific enzymatic function of the different classes of

acetyltransferase. Significantly, the solitary significant match of the *Meloidogyne* NodL sequences to a eukaryote is to a yeast serine-acetyltransferase, an enzyme clearly separated from the RKN by function as well as in our phylogeny (Figure 3.2). Bayesian analysis of the amino-acid alignment confirms this grouping. The posterior probability of the two *Meloidogyne* sequences being most closely related to the one other eukaryotic sequence, from *Saccharomyces cerevisiae*, is estimated to be 0. Instead, a group consisting of the two *Meloidogyne* sequences along with the *Sinorhizobium meliloti* and *Rhizobium leguminosarum* sequences is estimated to have a posterior probability of 1.0. For the clade consisting solely of the four rhizobial and the two *Meloidogyne* sequences, the posterior probability is estimated to be about 0.657, and almost all of the remaining posterior probability is accounted for by adding the *Streptomyces coelicolor* sequence to this clade of rhizobial and *Meloidogyne* sequences.

Using polymerase chain reaction (PCR) primers designed from the *Meloidogyne* sequence we have attempted to amplify *nodL* from a range of nematode species. For each of the *Meloidogyne* species tested (including *M. hapla*), we have been able to confirm the presence of the gene. However, similar experiments do not yield amplification products from the cyst nematodes we tested. Although other interpretations can be made, these results are consistent with *nodL* being acquired by an 'event b' HGT (see Figure 3.1).

Meloidogyne nodL truly is a nematode gene

A question that arises in analyzing eukaryotic sequences with strong matches to bacterial proteins, especially when the match is unique, is whether the gene in question truly was isolated from a eukaryote, or whether it represents a prokaryotic contaminant

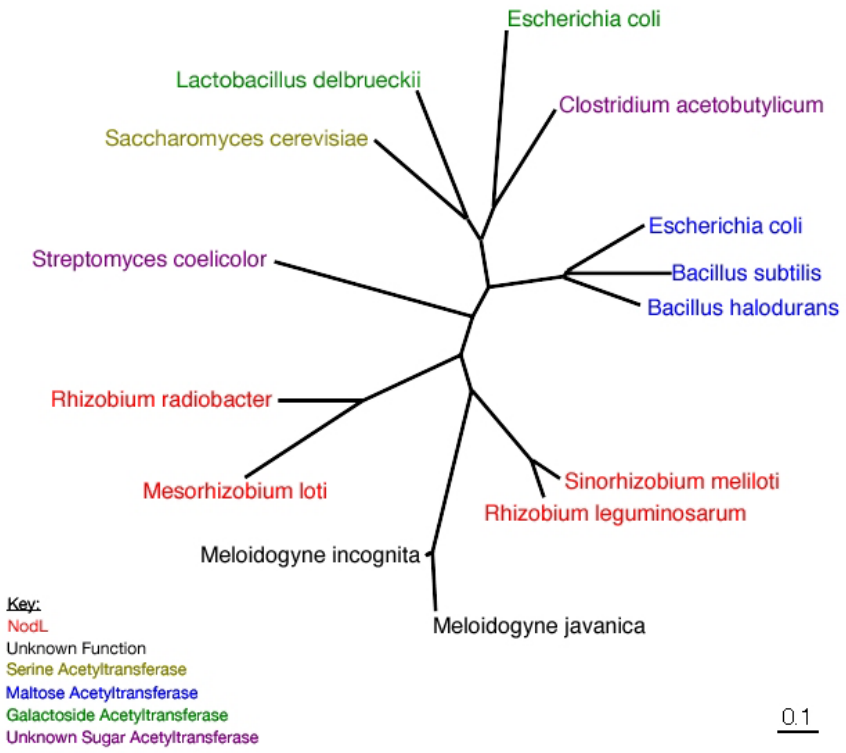

Figure 3.2: Cladogram of NodL-like proteins. The unrooted tree is generated by protein-distance and neighbor-joining methods and shows relationships of the deduced, putative *Meloidogyne* NodL proteins with similar enzymes, color-coded according to known function. Numbers indicate percent support from 1,000 non-parametric bootstrap replicates. The scale bar represents 0.1 amino-acid replacements per site across the length of a given branch.

(any nucleic acid matches of ESTs to prokaryotes,which probably would be contaminants, were removed before database submission (McCarter et al., 2003)). Claims of nematode genes having been acquired by HGT (Popeijus et al., 2000; Smant et al., 1998; Yan et al., 1998 ; Veronico et al., 2001) have addressed this issue in a number of ways. To provide experimental evidence that the *Meloidogyne nodL* sequences represent nematode loci, we cloned and sequenced a full-length transcript from *M. incognita* (*Mi-NodL*). Identification of the SL1 trans-splice leader at the 5' end of the message (Ray et al., 1994), and a poly(A) tail at the 3' end, confirmed that this is a bona fide nematode

gene (Figure 3.3). Analysis of genomic *Mi-NodL* sequences revealed an intron (Figure 3.3), further reinforcing the notion that this gene is integrated within the *M. incognita* genome.

In cases of a recent HGT, it has been suggested that the nucleotide composition of the transferred gene might reflect that of the donor species rather than the recipient species (Lawrence and Ochman, 1997). To establish a baseline nucleotide composition of *M. incognita* transcripts, we calculated the average G+C content for our entire *M. incognita* (WMi) sequence dataset, obtaining a value of 34.3%. By contrast, the average G+C content of rhizobial species ranges from 57 to 65% (Capela et al., 2001). Consistent with the average for *M. incognita*, the G+C content of *Mi-NodL* is 36%. This value is strikingly different for the *nodL* genes in *Rhizobium leguminosarum* (57% G+C) and *Mesorhizobium loti* (68% G+C). We similarly examined the G+C content of all 12 HGT candidates, and found the values to be consistently representative of *Meloidogyne*. Another way to consider nucleotide composition is through codon usage. In particular, we considered how similar the *Meloidogyne* codon usage is to that of a 'typical' rhizobial protein by using the codon adaptation index (CAI) (Sharp and Li, 1987). From an *R. leguminosarum* codon-usage table, we calculated the CAI for those amino acids precisely conserved between *Mi-NodL* and the rhizobial NodL protein to be 0.621 and 0.703 respectively. To evaluate the null hypothesis that the expected codon usage between the two *nodL* genes is identical, the difference in CAI values was adopted as a test statistic. The observed value of this test statistic was 0.082 and its null distribution was approximated by simulating 10,000 datasets as described in Materials and methods. Because the absolute value of the test statistic calculated from the simulated datasets

exceeded 0.082 only 62 of 10,000 times, we reject the null hypothesis of identical expected codon usage in the *M. incognita* and *R. leguminosarum nod*L genes and conclude that codon usage in these genes is significantly different between the species. Collectively, comparison of the nematode and rhizobial *nodL* genes suggests that each is adapted for function in the organism in which it resides, and despite the high degree of similarity between the amino-acid sequences of these genes, the DNA sequences are strikingly different.

From the Lawrence and Ochman model (1997), in which differences in G+C and codon bias are diagnostic for HGT events, it might be argued that our findings on the base composition of bacterial and nematode sequences are inconsistent with HGT. However, analyses in which synteny and phylogenetic information were also considered suggest that codon bias and G+C content are poor indicators of HGT (Koski et al., 2001). A role for 'amelioration', whereby structural characteristics of the foreign gene are eventually homogenized to resemble those of the recipient species, has been assumed, but the rate was postulated to be the same as the rate of random, forward mutation (Lawrence and Ochman, 1997). In addition to alterations in codon usage (as reflected in G+C content), for a bacterial gene to function efficiently in a nematode presumably requires acquisition of regulatory elements (including a promoter) and structural elements (including a poly(A) tail and, optionally, a trans-spliced leader). Other elements (such as introns) might also be acquired. It is possible that a careful phylogenetic analysis comparing rates of evolution of *Meloidogyne* genes acquired by HGT with those present in the more ancient nematode lineage, might shed light on the rate of amelioration of gene structure following inter-kingdom HGT.

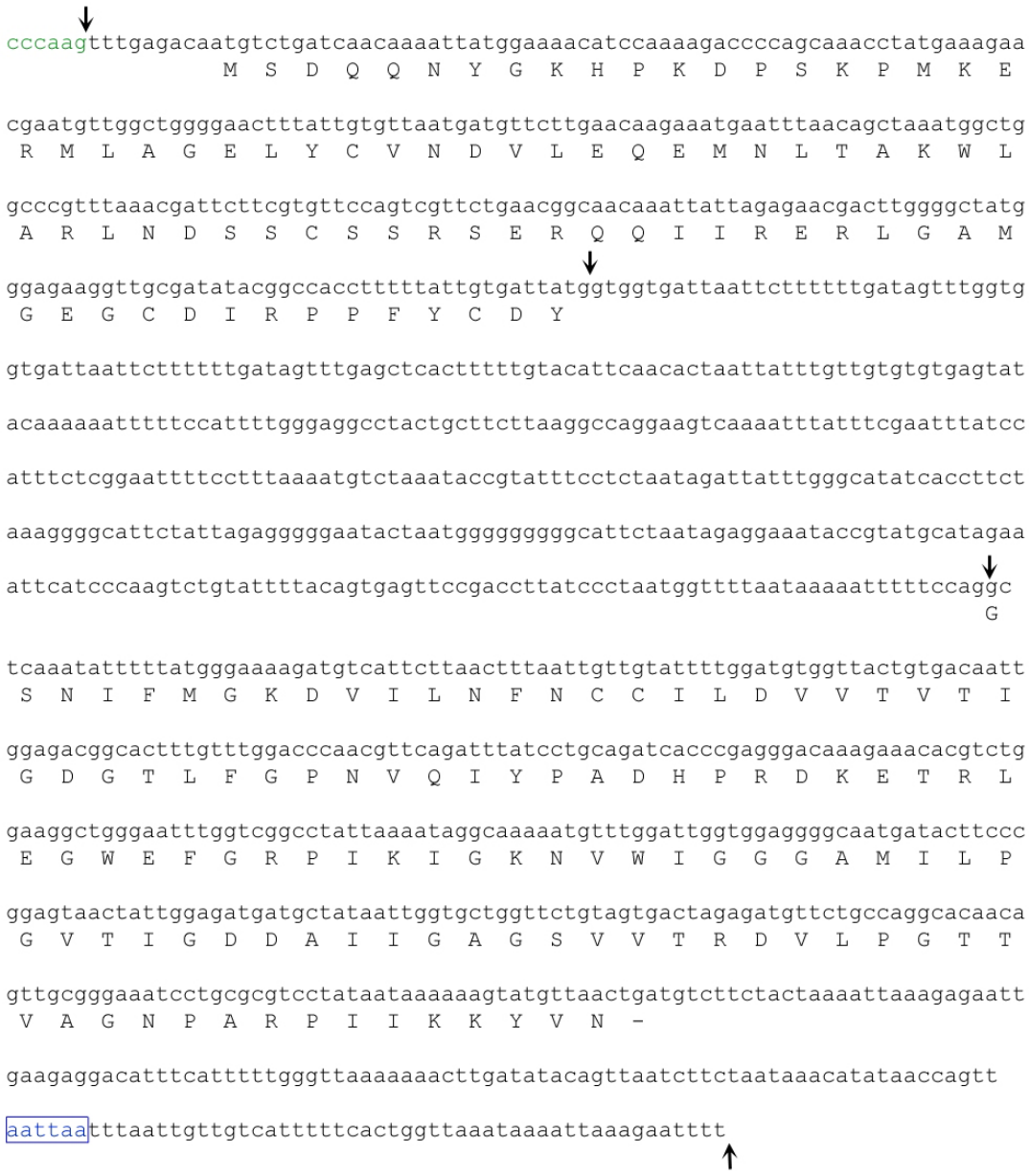

Figure 3.3: Structure of *Meloidogyne incognita NodL* and its deduced translation product. Features of the genomic sequence were established by comparison with that of a full-length cDNA clone, and are indicated by arrows in the following order: addition site of SL-1 *trans*-splice leader; beginning of intron; end of intron; and site of poly(A) tail.

Patterns of HGT from rhizobia

In the absence of an assembled genome sequence for *Meloidogyne*, it is not yet possible to examine conserved, genome-wide gene order of HGT candidates between nematodes and the hypothesized bacterial donor. Nevertheless, because the origin of many of the nematode HGT candidates appeared to be rhizobial, we wished to investigate the organization of the bacterial homologs. Unlike many prokaryotes, in which the genome resides largely on a single, circular chromosome, with varying numbers of small episomes, rhizobial genomes are typically organized in a manner conceptually more like eukaryotes. *Sinorhizobium meliloti*, for example, has three large, single-copy plasmids (Capela et al., 2001), and the primary *Mesorhizobium loti* chromosome is linear. Rhizobia have the ability to transfer genes horizontally to other bacteria, and *M. loti* carries a 'symbiosis island' which spans approximately 9% of its genome and has been shown to have a role in rhizobial evolution via HGT (Sullivan and Ronson, 1998). This symbiosis island contains certain genes involved in nodulation and nitrogen-fixation functions, but none of these is a homolog of the nematode HGT candidates we have identified. However, four of these genes do map to the same *M. loti* linear chromosome (Figure 3.4), including *nodL* and the glutamine synthetase gene, both of which are involved in nodulation/nitrogen fixation in rhizobia. Together with the L-threonine aldolase homolog candidate, these three genes are found within 257 kb of each other, a distance that represents only 3.65% of the *M. loti* chromosome, which is less than half the size of the symbiosis island. The fourth candidate, of unknown function, lies approximately 149 kb from the opposite side of the symbiosis island from the other three (Figure 3.4). Interestingly, examination of the colinearity and gene arrangements between *S. meliloti*,

R. radiobacter and *M. loti* indicates that the location of the genes in *M. loti* probably represents a more primitive state (Wood et al., 2001) and is therefore more likely to reflect the proximity of these genes in rhizobial ancestral species. Although it cannot be known if these genes were acquired in a single transfer event between a rhizobial ancestor and an ancestor to *Meloidogyne*, remnants of the HGT event (other than the already identified genes) may remain, and candidates are currently being mapped into the *M. incognita* genome to examine possible synteny with *M. loti*. BLAST analysis of the genes in the intervening span of chromosome indicates only three significant matches to the *M. incognita* (WMi) data set, all with significant matches to *C. elegans*, that is, they are not HGT candidates.

Figure 3.4: Schematic map (not to scale) of four genes on the *Mesorhizobium loti* linear chromosome with putative homologs in *M. incognita*, encoding NodL, L-threonine aldolase, glutamine synthetase and an unknown function. Also indicated is the 612 kb transferable *M. loti* symbiosis island.

Conclusions

We have demonstrated that a high-throughput bioinformatics approach based on EST sequences is an efficient and effective way to identify possible HGT candidates in plant-parasitic nematodes. Previous reports of horizontally acquired genes have been based mainly on biochemical or immunological criteria. Using an informatics approach, we rediscovered previously identified candidates (thus validating our method), and were able to identify new candidates for HGT. Strikingly, a common theme underpinning the HGT candidates is their apparent direct relationship to the parasitic lifestyle of *Meloidogyne* (Bird and Koltai, 2000). Also striking was our finding that phylogenetically, rhizobia appear to be the predominant group of 'donor' bacteria. This is significant for two reasons. First, root-knot nematodes and rhizobia occupy similar niches in the soil and in roots, and thus the opportunity for HGT may be omnipresent. Second, both organisms establish intimate developmental interactions with host plants, and mounting evidence suggests that the mechanisms for these interactions are also shared (Koltai et al., 2001). It seems a reasonable hypothesis that the origin of parasitism in *Meloidogyne* may have been facilitated by acquisition of genetic material from soil bacteria through horizontal transfer. Indeed, such events may have represented key steps in speciation of plant-parasitic nematodes

Acknowledgements

We thank M. Burke for his technological support and advice, H. Kishino for his helpful comments and insights, and M. Dante and J. Martin for NemaGene clusters. This research was supported by NSF grant DBI-0077503 to D.B. J.P.M. was supported by a Helen Hay Whitney/Merck Postdoctoral Fellowship. E.H.S. and J.L.T. were supported by NSF grant INT-990934, and J.L.T. was further supported by BIRD of Japan Science and Technology Corporation. D.B. and J.M. are equity holders of Divergence Inc.; none of this research was funded by Divergence Inc.

References

Aguinaldo A. M., Turbeville J. M., Linford L. S., Rivera M. C., Garey J. R., Raff R. A., Lake J. A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489-493.

Andersson J. O., Sjögren Å. M., Davis L. A. M., Embley T. M., Roger A. J. 2003. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr Biol* 13:94-104.

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S. R., Griffiths-Jones S., Howe K. L., Marshall M., Sonnhammer E. L. L. 2002. The Pfam protein families database. *Nucleic Acids Res* 30:276-280.

Bird D. McK., Koltai H. 2000. Plant parasitic nematodes: habitats, hormones, and horizontally-acquired genes. *J Plant Growth Regul* 19:183-194.

Blair J. E., Ikeo K., Gojobori T., Hedges S. B. 2002. The evolutionary position of nematodes. *BMC Evol Biol* 2:7.

Blaxter M. L., DeLey P., Garey J., Liu L. X., Scheldeman P., Vierstraete A., Vanfletern J., Mackey L. Y., Dorris M., Frisse L. M., Vida J. T., Thomas W. K. 1998. A molecular evolutionary framework for the phylum Nematoda. *Nature* 392:71-75.

Boucher G., Lambshead P. J. D. 1994. Ecological biodiversity of marine nematodes in samples from temperate, tropical, and deep-sea regions. *Conserv Biol* 9:1594-1604.

Brinkman F. S. L., Blanchard J. L., Cherkasov A., Av-Gay Y., Brunham R. C., Fernandez R. C., Finlay B. B., Otto S. P., Ouellette B. F. F., Keeling P. J., *et al*. 2002. Evidence that plant-like genes in *Chlamydia* species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the choloroplast. *Genome Res* 12:1159-1167.

Capela D., Barloy-Hubler F., Gouzy J., Bothe G., Ampe F., Batut J., Boistard P., Becker A., Boutry M., Cadieu E., *et al*. 2001. Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc Natl Acad Sci* 98:9877-9882.

Dayhoff M. O. 1978. *Atlas of Protein Sequence and Structure. Volume 5 Supplement 3.* Washington, DC: National Biomedical Research Foundation.

Dehal P., Satou Y., Campbell R. K., Chapman J., Degnan B., De Tomaso A., Davidson B., Di Gregorio A., Gelpke M., Goodstein D. M., *et al*. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 298:2157-2167.

Ditt R. F., Nester E. W., Comai L. 2001. Plant gene expression response to *Agrobacterium tumefaciens*. *Proc Natl Acad Sci* 98:10954-10959.

Downie J. A., Young J. P. W. 2001. Genome sequencing: the ABC of symbiosis. *Nature* 412:597-598.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783-791.

Felsenstein J. 1993. PHYLIP (Phylogeny Inference Package) version 3.6a2. Seattle, WA: Department of Genetics, University of Washington.

Göttfert M. 1993. Regulation and function of rhizobial nodulation genes. *FEMS Microbiol Rev* 10:39-64.

Hedges S. B. 2002 The origin and evolution of model organisms. *Nat Rev Genet* 3:838-849.

Huelsenbeck J. P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754-755.

International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. 2001. *Nature* 409:860-921.

Jain R., Rivera M. C., Lake J. A. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci* 96:3801-3806.

Jaubert S., Laffaire J.-B., Abad P., Rosso M.-N. 2002. A polygalacturonase of animal origin isolated from the root-knot nematode *Meloidogyne incognita*. *FEBS Lett* 522:109-112.

Jones D. T., Taylor W. R., Thornton J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275-282.

Koenning S. R., Overstreet C., Noling J. W., Donald P. A., Becker J. O., Fortnum B.A. 1999.Survey of crop losses in response to phytoparasitic nematodes in the United States for 1994. *J Nematol* 31:587-618.

Koltai H., Dhandaydham M., Opperman C., Thomas J., Bird D. McK. 2001. Overlapping plant signal transduction pathways induced by a parasitic-nematode and a rhizobial endosymbiont. *Mol Plant Microbe Interact* 14:1168-1177.

Koski L. B., Morton R. A., Golding G. B. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* 18:404-412.

Lambert K. N., Allen K. D., Sussex I. M. 1999. Cloning and characterization of an esophageal-gland-specific chorismate mutase from the phytoparasitic nematode *Meloidogyne javanica*. *Mol Plant Microbe Interact* 12:328-336.

Lange B. M., Rujan T., Martin W., Croteau R. 2000. Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proc Natl Acad Sci* 97:13172-13177.

Lawrence J. G., Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 44:383-397.

Lawrence J. G. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol* 2:519-523.

Ludwig R. A. 1980. Physiological roles of glutamine synthetase I and II in ammonium assimilation in *Rhizobium* sp. 32H1. *J Bacteriol* 141:1209-1216.

Mallatt J., Winchell C. J. 2002. Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. *Mol Biol Evol* 19:289-301.

McCarter J. P., Abad P., Jones J., Bird D. McK. 2000. Rapid gene discovery in plant parasitic nematodes via expressed sequence tags. *Nematology* 2:719-731.

McCarter J. P., Clifton S. W., Bird D. McK., Waterston R. 2002. Nematode gene sequences; Update for June 2002. *J Nematol* 34:71-74.

McCarter J. P., Mitreva M. D., Martin J., Dante M., Wylie T., Rao U., Pape D., Bowers Y., Theising B., Murphy C., *et al*. 2003. Analysis and functional classification of transcripts from the root-knot nematode *Meloidogyne incognita*. *Genome Biol* 4:R26.

Nematode.net (http://www.nematode.net)

Ochman H., Lawrence J. G., Groisman E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299-304.

Ochman H., Moran N. A. 2001. Genes lost and genes found: evolution of bacterial pathogensis and symbiosis. *Science* 292:1096-1098.

Popeijus H., Overmars H., Jones J., Blok V., Goverse A., Helder J., Schots A., Bakker J., Smant G. 2000. Degradation of plant cell walls by a nematode. *Nature* 406:36-37

Ray C., Abbott A. G., Hussey R. S. 1994. Trans-splicing of a *Meloidogyne incognita* mRNA encoding a putative esophageal gland protein. *Mol Biochem Parasitol* 68:93-101.

Rice P., Longden I., Bleasby A. 2000. EMBOSS: The European molecular biology open software suite. *Trends Genet* 16:276-277.

Royo J., Gimez E., Hueros G. 2000. CMP-KDO synthetase: A plant gene borrowed from Gram-negative eubacteria. *Trends Genet* 16:432-433.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406-425.

Sharp P. M., Li W.-H. 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281-1295.

Smant G., Stokkermans J. P. W. G., Yan Y., De Boer J. M., Baum T. J., Wang X., Hussey R. S., Gommers F. J., Henrissat B., *et al*. 1998 Endogenous cellulases in animals: Isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proc Natl Acad Sci* 95:4906-4911.

Stanhope M. J., Lupas A., Italia M. J., Koretke K. K., Volker C., Brown J. R. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 411:940-944.

Stephens R. S., Kalman S., Lammel C., Fan J., Marathe R., Aravind L., Mitchell W., Olinger L., Tatusov R. L., Zhao Q. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis. Science* 282:754-759.

Sullivan J. T., Ronson C. W. 1998. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci* 95:5145-5149.

Tandingan-De Ley I., De Ley P., Vierstraete A., Karssen G., Moens M., Vanfleteren J. 2002. Phylogenetic analyses of *Meloidogyne* small subunit rDNA. *J Nematol* 34:319-327.

Thompson J. D., Gibson T. J., Plewniak F., Jeanmougin F., Higgins D. G. 1997. The Clustal_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25:4876-4882.

Turner S. L., Young J. P. W. 2000. The glutamine synthetases of rhizobia: phylogenetics and evolutionary implications. *Mol Biol Evol* 17:309-319.

Van Rhijn P., Vanderleyden J. 1995. The *Rhizobium*-plant symbiosis. *Microbiol Rev* 59:124-142.

Veronico P., Jones J., Di Vito M., De Giorgi C. 2001. Horizontal transfer of a bacterial gene involved in polyglutamate biosynthesis to the plant-parasitic nematode *Meloidogyne artiellia*. *FEBS Lett* 508:470-474.

Wheeler D. L., Church D. M., Lash A. E., Leipe D. D., Madden T. L., Pontius J. U., Schuler G. D., Schriml L. M., Tatusova T. A., Wagner L., Rapp B. A. 2002. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* 30:13-16.

Wolf Y. I., Aravind L., Koonin E. V. 1999. Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. *Trends Genet* 15:173-175.

Wood D. W., Setubal J. C., Kaul R., Monks D. E., Kitajima J. P., Okura V. K., Zhou Y., Chen L., Wood G. E., Almedia Jr N. F., *et al*. 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* 294:2317-2323.

WU-BLAST (http://blast.wustl.edu)

Yan Y., Smant G., Stokkermans J., Qin L., Helder J., Baum T., Schots A., Davis E. 1998. Genomic organization of four β-1,4-endoglucanase genes in plant-parasitic cyst nematodes and its evolutionary implications. *Gene* 220:61-70.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39:306-314.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555-556.

Chapter 4

Conclusion

Introduction

Some have asserted that *Caenorhabditis elegans* has "limited usefulness in studying nematode adaptations for plant parasitism" (Hussey et al., 2002). The research contained within this dissertation suggests otherwise. Using expressed sequence tag (EST) datasets for eight different Heteroderidae species and the complete *C. elegans* genome we were able to initiate a program of comparative genomics to investigate evolution and parasitism in root-knot and cyst nematodes. Phylogenetic analysis of *Meloidogyne* was accomplished by compiling sets of genes in Heteroderidae with orthologues in *C. elegans* (Chapter 2). Putative xenologues were identified in *Meloidogyne* through a filtering process aimed, in part, at finding genes with no apparent *C. elegans* homologue (Chapter 3). Pinpointing these similarities and differences between *Meloidogyne* and their free-living cousins would not have been possible without the availability of the complete *C. elegans* genome.

One necessity in proper use of this raw data is the identification of orthologues across different species. A database of Clusters of Orthologous Groups of proteins (COGs) is currently available at the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov). However, this database only contains orthologue information for completely sequenced genomes. I contend that a complete genome is not needed to do genomics. Rather, clustering of orthologues can begin before a genome is complete. I have initiated this process for root-knot and cyst nematodes and have identified 47 core metabolism genes found in *Meloidogyne*, *Globodera* and *Heterodera*, predicated on the presence of a homologous gene in *C. elegans*. Ontological classification was assigned to individual COGs according to annotation of the complete *C. elegans*

genome. Indeed, this fully sequenced and well-annotated genome affords a base-line for the initial construction of such sets of genes spanning not only the Heteroderidae family, but also different nematode orders. This set of COGs has provided the fundamental data needed to determine the evolutionary relationships between the *Meloidogyne* as well as within the larger classification of Heteroderidae. With this data set, a robust phylogeny resolving the relationship between those plant-parasitic nematodes with the largest worldwide economic impact was reconstructed.

The fully sequenced and annotated *C. elegans* genome is useful not only for identifying orthologues and assigning them putative ontology, but also for exploring the differences between free-living nematodes and their parasitic cousins. Exploiting the *C. elegans* genome as well as the fully sequenced genome of the arthropod *Drosophila melanogaster* permitted creation a filtering system designed to identify putatively horizontally acquired genes in plant-parasitic nematodes. This filter was able to confirm identification of previously postulated horizontal gene transfer (HGT) candidates as well as suggest new candidates. The deduced protein products of the majority of these candidates are believed to be involved in the parasitic lifestyle in *Meloidogyne*. Indeed, homology searches and phylogenetic reconstruction suggests a rhizobial donor for at least four of the 12 HGT candidates, indicating a putative genetic link between these bacterial plant-symbionts and the *Meloidogyne* plant-parasites. It can be hypothesized that root nodulation in plants infected by *Meloidogyne* and by rhizobia is not coincidental, but rather the ability to nodulate roots is a trait acquired by an ancestral nematode species from an ancestral bacterial species.

These analyses were completed solely with publicly available data deposited in a variety of databases. The research contained herein represents only a small sample of the research that can be accomplished with such data.

Future Directions

Clusters of orthologous groups of proteins

Construction of clusters of orthologous groups of proteins (COGs) provides an insight into the relationship between species. We have constructed 55 nematode-specific COGs, 47 of which were used in the multi-gene phylogenetic analysis of *Meloidogyne*. The remaining eight COGs, with putatively paralogous members, were not included in this analysis but remain a legitimate set of orthologues.

The number of (ESTs) for the species included in our research has increased since completion of our analyses. ESTs for other *Meloidogyne* species are also now available. These new sequences should be used to augment the COG construction process, either by incorporating them as members of existing COGs or by establishing new groups of orthologues. As more sequences are made available, the COGs will ideally become complete, i.e. genes missing from a given species in a given COG will be added.

The COGs constructed for the phylogenetic analysis of *Meloidogyne* required the presence of a *C. elegans* homologue (Chapter 2). This was to allow post-assembly ontological classification of the individual COGs. However, assembly of COGs using genes that are specific to the Heteroderidae, with no member in the free-living *C. elegans*, are needed to identify genes likely to be more directly involved in the parasitic lifestyle. Summary statistics such as average percent identity and protein distance could be compiled and compared between both sets of COGS, those with members from *C. elegans* and those without. Additionally, phyletic information (information about gene presence and absence) can be deduced as more species are added to this set. As genomes

for these species reach completion, these COGs, as well as genes that do not fit into any COG, will give greater insight into the genetic differences between the root-knot and cyst nematodes. These differences will further aid in recognizing elements potentially responsible for distinct parasitism strategies across the Heteroderidae. Ultimately, the COGs will also provide targets for the identifying genetic elements responsible for the differences in host specificity and host response between each species within each genus.

The individual COGs can also be used to test hypotheses of differential selective pressure between various categorizations of species. For instance, the recovered species tree from Chapter 2 can be divided into nested, biologically relevant, sub-trees and estimates of selective pressure compared between the sub-trees. These subdivisions might include parasitic versus free-living nematodes, cyst nematodes versus root-knot nematodes versus free-living nematodes, or the individual species represented in a given COG. Correlations between differences in selective pressure and phylogenetic subdivision, ontology or chromosomal location can be determined. Work on these and similar questions is in progress.

Statistical support for horizontal gene transfer

Whilst hypotheses of gene acquisition via horizontal gene transfer (HGT) exist, and HGT has been observed in laboratory settings, indications of xenologous origins of genes in nature are mostly anecdotal. Claims of HGT have been supported through evidence such as unusual base composition (G+C content and codon bias) or incongruencies in phylogeny, where a gene tree does not reflect the expected species tree. However, these claims generally have no measure of statistical support.

Progress has been made on the development of algorithms for quantifying hypotheses of gene gain and gene loss. These algorithms mainly have applications in determining whether HGT is a most parsimonious explanation for a specific phyletic pattern. For example, Page (1994) developed a system for resolving conflicts between gene trees and species trees by determining the minimum number of gene duplication and gene loss events needed to explain the incongruence. He also suggested that HGT could be a viable alternative to duplication as a method of gene gain in this model. A slightly more sophisticated algorithm, based on the same concept used by Page but with differential weighting for gene gain and gene loss (Mirkin et al., 2003), includes horizontal gene transfer events by grouping them with gene duplication under the overall category of "gain". Though these methods can provide some support for the acceptance of a HGT event as a parsimonious explanation for phylogenetic incongruence, they do not provide a robust measure of statistical support.

A more rigorous approach to assigning statistical support to HGT events was suggested by Huelsenbeck et al. (2000) based on a modification of their stochastic model for host-parasite cospeciation. This model takes into account host switching, in which parasites change host specificity in conjunction with host speciation events. Host switching events and rates are estimated along a phylogeny using Bayesian methods. The authors have suggested a similar model may be applied to other phylogenetic associations such as scenarios where horizontal gene transfer can be considered analogous to host switching. A similar approach might involve modeling gene loss and gene acquisition as a birth-death process. Given a species tree and gene presence/absence information one could calculate the likelihood of seeing such a phyletic pattern under two alternative

hypotheses, one based on a combination of gene loss and gene gain and the other based on only gene loss. Such hypothesis tests could establish confidence statistics for claims of horizontal acquisition, though such models may be too reliant on complete genomes (to distinguish gene absence versus lack of sequence) to be plausible at this time.

Mapping HGT candidates in Meloidogyne

At least four of the HGT candidates found in *Meloidogyne incognita* appear to have a rhizobial ancestor. Determining the physical location of these four genes within the *M. incognita* genome could provide insight into whether these genes were acquired as a single cassette. Should all four map onto the same chromosome and colinearity is determined between these genes in *M. incognita* and *Mesorhizobium loti*, the hypothesis of a single transfer event from a rhizobial ancestor would gain support.

Furthermore, if these genes are located on one chromosome in *M. incognita*, sequencing the intervening span of this chromosome between these genes might provide more information regarding possible acquisition of a pathogenicity island from rhizobia. If genes identified in this area suggest preserved synteny with the *M. loti* chromosome, the already defined pathogenicity island in *M. loti* may need to be expanded (Sullivan and Ronson, 1998).

Horizontal gene transfer in other parasitic nematodes

The application of a phylogenetic filter for identifying HGT candidates, as described in Chapter 3, need not be limited to the plant-parasitic *Meloidogyne*. Sequencing projects for at least 15 parasitic nematodes from families other than Heteroderidae

(http://www.nematode.net/) are currently under way. Sequences for each of these parasites could be similarly run through our phylogenetic filtering system to identify putatively horizontally acquired genes. Those HGT candidates, which may be involved in the acquisition of traits necessary to a given species' mode of parasitism, could be mapped onto the Nematoda phylogeny to determine if horizontal gene transfer may have played a role in the multiple, independent acquisitions of parasitism across the phylum.

Suggested genome sequencing project

Without the complete sequence of the free-living nematode *C. elegans*, much of the research contained within this dissertation would not have been possible. Only by having a reference genome against which to make comparisons could the screening for HGT candidates take place. Since the inception of this research a second *Caenorhabditis* species, *C. briggsae*, has been completely sequenced and annotated. The animal-parasite *Brugia malayi* has a genome sequencing project under way and many other parasitic nematodes have increasing numbers of EST sequences available. It appears as though sequencing is mainly focused on those nematodes with a major impact on human health or economically important crops. However, without genomes of free-living species, the use of high-throughput comparative genomics to search for genes intimately involved in parasitism is more difficult, if not impossible.

The sequencing of other free-living nematodes from a wider variety of families should facilitate more informative genomic comparisons could be made. For example, the free-living nematode *Zeldia punctata* and the plant-parasitic Heteroderidae are all members of the same clade, Clade IVb (Blaxter et al., 1998). Presumably *Z. punctata*,

being more closely related to *Meloidogyne* than *C. elegans*, will share a larger percentage of homologous with *Meloidogyne* than *C. elegans*. However, as of December 2003, only one protein and 395 nucleotide sequences have been deposited with NCBI's GenBank. Similarly, *Panagrellus* spp., free-living nematodes more closely related to the animal-parasitic *Strongyloides* than *Caenorhabditis*, has a mere 13 nucleotide and 20 protein sequences deposited with GenBank. If a complete, annotated genome of another free-living nematode, preferably not from Clade V, were available, more conservative comparisons between the free-living and parasitic nematodes could be accomplished. These comparisons need not only be related to acquisition of horizontally acquired genes, but can also be adapted to map phyletic information onto a phylogeny to identify a large picture of gene presence and absence.

Conclusions

The link between horizontal gene transfer and parasitism has been well established in prokaryotes. With our high-throughput approach directed at identifying horizontal gene transfer candidates, a similar link has now been established between HGT and parasitism in eukaryotes. This link combined with a robust *Meloidogyne* phylogeny and continued sequencing of various *Meloidogyne* sequences can only aid in our understanding of the acquisition of traits that confer parasitism within the root-knot nematodes and answer questions about the evolution of host specificity within the genus. Analyses such as these need to be expanded upon to further our understanding of the multiple independent acquisition of parasitism and the different parasitic strategies employed in the phylum Nematoda as a whole.

References

Blaxter M. L., DeLey P., Garey J., Liu L. X., Scheldeman P., Vierstraete A., Vanfletern J., Mackey L. Y., Dorris M., Frisse L. M., *et al*. 1998. A molecular evolutionary framework for the phylum Nematoda. *Nature* 392:71-75.

Huelsenbeck J. P., Rannala B., Larget B. 2000. A Bayesian framework for the analysis of cospeciation. *Evolution* 54:352-364.

Hussey R. S., Davis E. L., Baum T. J. 2002. Secrets in secretions: genes that control nematode parasitism of plants. *Braz. J. Plant Physiol.* 14:183-194.

Mirkin B. G., Fenner T. I, Galpern M. Y., Koonin E. V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology* 3:2.

Page R. D. M. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* 43:58-77

Sullivan J. T., Ronson C. W. 1998. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci* 95:5145-5149.