Assembly, Annotation, and Proteogenomic Mapping of *Rotylenchulus reniformis*

William S. Sanders[1], Satish Ganji[1], Jennifer P. Arnold[1], Mark A. Arick II[1], Kurt C. Showmaker[1,2], Benjamin D. Bartlett[3], Amanda M. Cooksey[4], Justin T. Watts[5], Scott Knudstrup[6], Zenaida V. Magbanua[1,3], Daniel G. Peterson[1,3], Martin J. Wubben[2,7,]

[1]Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, MS, USA, [2]Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University, MS, USA, [3]Department of Plant and Soil Sciences, Mississippi State University, MS, USA, [4]School of Animal and Comparative Biomedical Sciences & BIO5 Institute, University of Arizona, AZ, USA, [5]School of Medicine, University of Mississippi Medical Center, MS, USA, [6]Department of Mathematics, University of Michigan, MI, USA, [7]United States Department of Agriculture, Agricultural Research Service (USDA-ARS), Mississippi State, MS, USA

*-wss2@igbb.msstate.edu

Abstract

Rotylenchulus reniformis (reniform nematode) is a plant-parasitic nematode that infects a number of agriculturally important crops including cotton where it accounts for losses estimated at $130,000,000 annually. Little molecular characterization of *R. reniformis* has been conducted to this point. We have constructed a set of genomic sequences for *R. reniformis* by isolating and sequencing the genomic DNA from a pooled population of *R. reniformis* nematodes and assembling and annotating these sequences. In addition, we isolated proteins from *R. reniformis* eggs, subjected them to high-throughput liquid chromatography mass spectrometry, and searched the resulting peptide spectra against four databases – specifically, (i) the *Caenorhabditis elegans* proteome, (ii) the predicted proteome of *Meloidogyne incognita*, (iii) available *R. reniformis* expressed sequence tags translated in all 6 reading frames, and (iv) our assembled *R. reniformis* genomic sequences translated in all 6 reading frames. The resulting spectra identified 569 proteins from the three publicly available databases, while providing confirmation of 179 of our assembled genomic sequences when searched against the database of *R. reniformis* genomic sequences. For poorly characterized non-model organisms, comparative proteomic analysis is a powerful tool that can be utilized to quickly gain insight into the molecular infrastructure of the organism. In *R. reniformis*, we have shown the expression of a number of parasitism genes or gene targets as early as the egg stage and have provided confirmation that these genes are translated into proteins. Such parasitism genes/proteins may serve as targets for future applied research aimed at limiting crop damage caused by *R. reniformis*

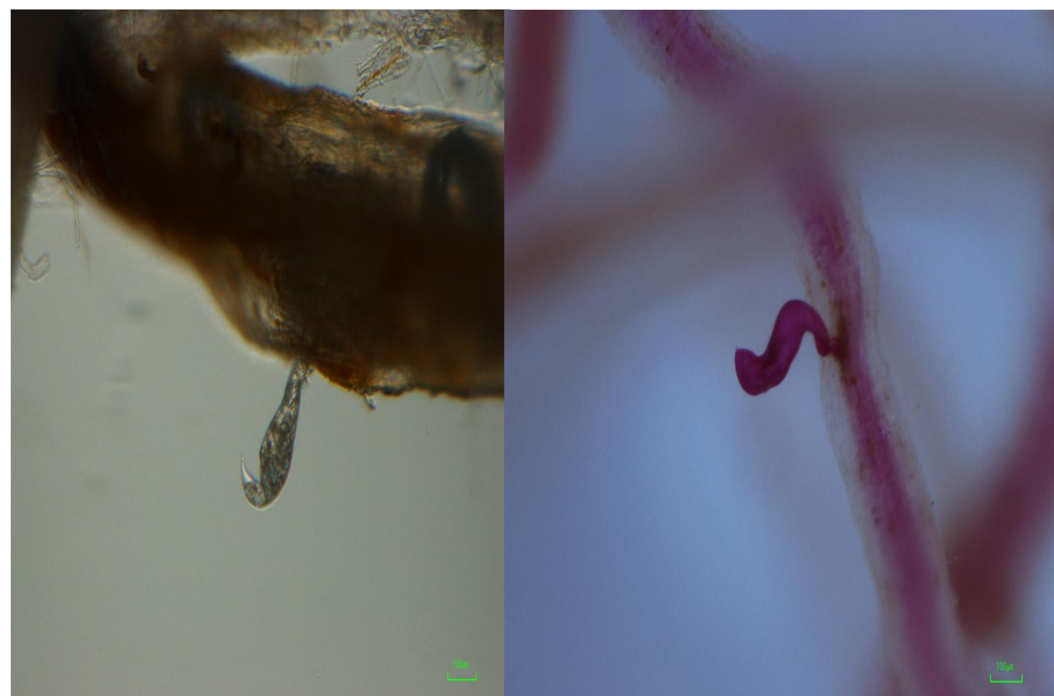


Figure 1. *R. reniformis* sedentary female with an established feed site on the root of a cotton plant.

Infection by *R. reniformis* can result in reduction of plant weight, a reduction in the number of pods per plant, a reduction in the chlorophyll content of leaves, and a decrease in the bulk density of stem parts.

Challenges When Working with *R. reniformis*

- Lack of a close reference genome (Fig. 2).
- Quality DNA isolation for sequencing is problematic – either from a pooled sample of nematodes or from a single egg subjected to whole genome amplification
- Parasitic relationship requires culturing the nemaode with its associated host plant and soil

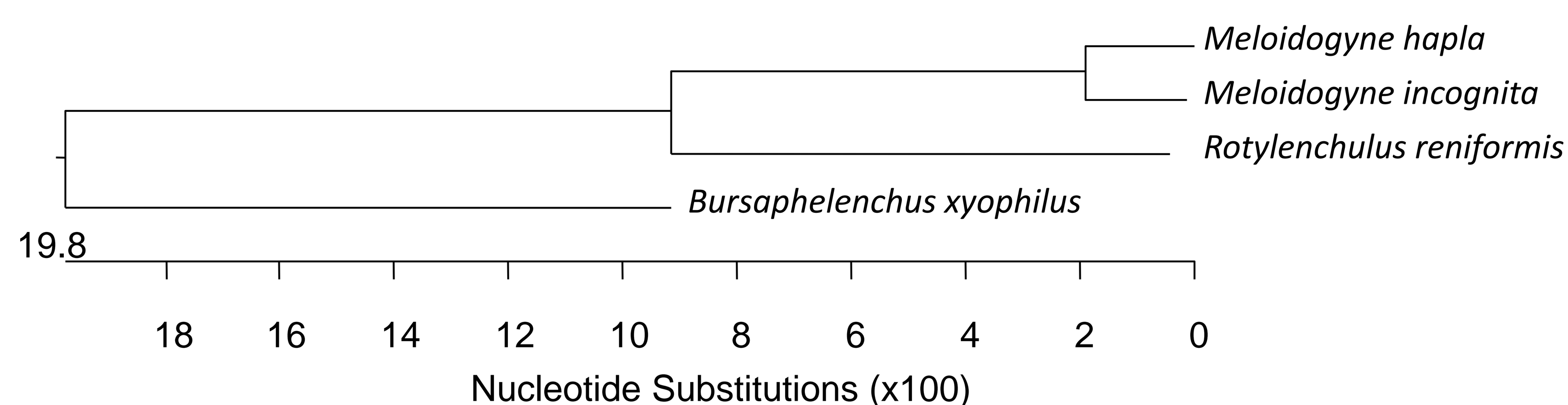
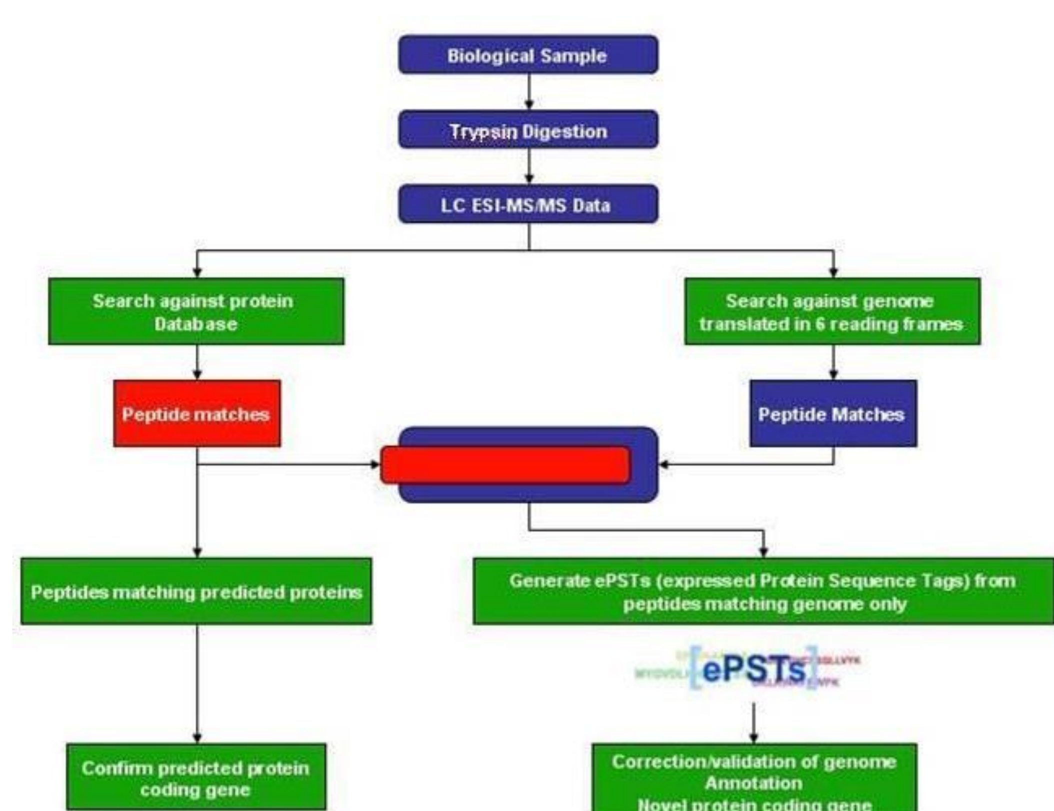


Figure 2. *R. reniformis* 18S rRNA (EU306342) aligned using CLUSTALW with other sequenced plant parasitic nematodes, *M. hapla* (AY593892), *M. incognita* (AY284621), and *B. xylophilus* (FJ235886).

Proteogenomic Mapping

Proteogenomic mapping is the process of utilizing peptides identified through high-throughput mass spectrometry to aid in the structural genome annotation of an organism with a poorly annotated and characterized genome sequence. Given the poor genome assembly currently for *R. reniformis*, we hope gain a better knowledge of the proteins in the species by mapping reniform peptides back to the *C. elegans* genome.



Sanders WS, Wang N, Bridges SM, Malone BM, Dandass YS, McCarthy FM, Nanduri B, Lawrence ML, Burgess SC. *The Proteogenomic Mapping Tool*. BMC Bioinformatics 2011, 12:115.

Genome Sequencing & Assembly

A total of 32,748,541,391 bp have been sequenced using combined Illumina and Roche 454 technologies.

Based on an estimated genome size of 190MB (flow cytometry, not shown), we have sequenced in the range of ~172X genome coverage.

Table 1. Sequencing Statistics for *R. reniformis*.

Pooled Population Sequencing:		# Sequences	Total Length (bp)
Illumina GAIIx:	2x100bp (450bp fragments)	49,839,506	4,983,950,600
	2x100bp (550bp fragments)	53,151,682	5,315,168,200
Roche 454 GS:	Single End	462,611	162,227,569
	Paired End (8kb inserts)	439,886	76,239,398
Single Individual, WGA:			
Illumina MiSeq:	2x250 (400bp fragments)	14,345,666	3,531,549,194
	2x250 (600bp fragments)	15,816,150	3,927,126,030
	2x250 (850bp fragments)	49,174,268	14,752,280,400
Total		**183,229,769**	**32,748,541,391**

Assembly Statistics

# contigs > 500bp	N80	N50	N20	max contig length	sum of all contigs
108,546	958	2,344	6,583	42,678	181,300,000

Sequences were trimmed, quality filtered, and then assembled using ABySS version 1.3.6 (available at: http://www.bcgsc.ca/platform/bioinfo/software/abyss) with a k-mer of 77. The sum of all contig lengths is 181.3 MB and the genome size predicted by flow cytometry is 190 MB, so we are capturing and assembling a majority of the genomic sequence.

The combined assembly contained 22,546 contigs > 2,000 bp in length (N50 = 4,932bp) covering a total of 100.2 Mb of genomic sequence.

Protein Isolation

A 50mg sample of R. reniformis eggs were pipetted into a TC 12X12 round bottom tube(Covaris). The samples were then processed using 500µL of a tissue homogenization buffer containing 100mM Tris-HCL(pH 7.5-8.0), 150mM sodium chloride, 1mM dithiothreitol and 3.5 mM SDS. Samples were then vortexed and loaded independently into the Covaris S20 instrument and lysed at 4ºC using the Tissue Homogenization program settings (Peak Power: 200.0; Duty Factor: 20.0: Cycles/Burst: 500; Time: 60 seconds). Samples were then transferred to a 0.1µm UltraFree vV filter unit and was spun for 30 minutes at 4 ºC and at 13,000G. The filtrate was collected and protein concentration was measured using the DC Protein Assay(Bio-Rad). Equal amounts of protein were then trypsin digested for subsequent LC-MS/MS analysis. Two experimental duplicates were executed.

Trypsin Digestion

Tryptic digestion of equivalent protein amounts was performed using standard protocol. 10mM dithiothreitol (DTT) in 50mM ammonium bicarbonate was added for reduction and incubated for 15 minutes at 80ºC, followed by an alkylation by 100mM iodoacetamide in 50mM ammonium bicarbonate at room temperature for 30 minutes with a 9:1 ratio of DTT to iodoacetamide. Samples were then digested at an enzyme to substrate ratio of 1:50, overnight at 37ºC, using trypsin. Peptides were lyophilized using a vacuum centrifuge and dissolved in 2% acetonitrile and 0.1% formic acid for downstream LC-MS/MS analysis.

Nanospray LC/MS

Data was collected with the use of an Orbitrap LTQ Velos mass spectrometer (Thermo Fisher Scientific) with Xcalibur version 2.1.0 coupled with an UltiMate 3000 nano flow HPLC system (Dionex). The peptides were separated with a reverse phased fused silica C18 column measuring 75µm by 150µm(Thermo Fisher Scientific). Peptides were eluted during a 120 minute multi-step gradient with a constant flow rate of 0.3µL per minute followed by a column wash with 95% solvent B(100% acetonitrile and 0.1% formic acid) for a duration of 30 minutes. An 25 minute equilibration of the column was performed using 2% solvent B. Peptide analysis was performed using the linear trap mass spectrometer run in a data dependent acquisition (DDA) mode. Identification of peptides was performed by using the top 18 collision scan events (CID) with a dynamic exclusion time of 30 seconds and a normalized collision energy at an activation time of 40 ms. The ion trap was used to analyze fragment masses at a normal mass range (300-2,000amu).

Peptide Identification & Filtering

Peptide spectra were identified using X!Tandem (version 13-02-01-1, available at http://www.thegpm.org/tandem/) and filtered using an e-value cut-off of 0.01.

Proteogenomic Mapping

Peptides were mapped onto the C. elegans genome translated in all 6 open reading frames using the Proteogenomic Mapping Tool (version 1.0, available at http://www.agbase.msstate.edu/tools/pgm/) with standard settings.

Current Progress

The 22,546 contigs > 2,000 bp in length were used to predict 107,729 coding regions with GeneMark, and these regions are being subjected to further annotation. As the genome improves more work will be performed with a variety of eukaryotic gene predictions tools to structurally annotate the reniform genome.

74.94% of RNA-Seq reads of our sequenced *R. reniformis* transcriptome mapped to the genome assembly with at least 1 reported alignment.

Additionally, using our proteomics data and a comparative proteogenomic mapping approach, we have identified a number of genes expressed in the early life stages of the reniform nematode –far-1, cathepsin Z1, calreticulin (crt-1), peroxidedoxin, thioredoxin, and transthyretin.

Future Work

- More sequencing incorporating mitochondrial DNA removal and non-specific whole genome amplification of *R. reniformis* DNA (larger insert sizes, mate-pair libraries, etc.)
- Further refinement of the assembly using alternative assembly algorithms and visualization tools
- Further structural and functional annotation – identification of ncRNAs, repeat elements, GO annotation
- Submission of curated sequences to repositories
- Incorporation of transcriptome sequences to help further refine our predicted gene models

Funding
USDA ARS 6402-21310-003-18S, USDA ARS 6402-21310-003-20S, and USDA NIFA 2010-34609-21206