Identification of protein coding genes in the plant parasitic nematode *Rotylenchulus reniformis* through comparative proteogenomic mapping with *Caenorhabditis elegans*

William S. Sanders[1,*], Satish Ganji[1], Jennifer P. Arnold[1], Mark A. Arick II[1], Kurtis C. Showmaker[1], Martin Wubben[2], Daniel Peterson[1,3]

[1]Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, USA; [2]USDA-ARS, Mississippi State, USA;; [3]Department of Plant and Soil Sciences, Mississippi State University, USA;
*-wss2@igbb.msstate.edu

Abstract

Peptides identified through high-throughput mass spectrometry can be utilized to complement traditional structural genome annotation methods through proteogenomic mapping and provide experimental evidence that a given gene is being transcribed and translated into a protein. In model organisms, proteogenomic mapping can aid researchers by identifying novel exons and aid in the identification of intron-exon boundary correction and discovery. In non-model organisms, the peptides identified through mass spectrometry can be mapped back to the genome sequence of a model organism, providing insight into genes conserved in the non-model species. *Rotylenchulus reniformis* is a plant parasitic nematode affecting US cotton production, and causes an estimated annual $130,000,000 USD crop loss. Proteins were isolated from *R. reniformis* eggs and subjected to proteolytic digestion and analysis through mass spectrometry. Genomic *R. reniformis* DNA was isolated and sequences were generated using two platforms, combining both Illumina Sequencing-by-Synthesis and Roche 454 Pyrosequencing technologies. Through the proteogenomic mapping of these peptides to the proteome and genome of the model nematode species, *Caenorhabditis elegans*, and the use of the *R. reniformis* genome sequence data, we have identified a set of gene sequences to serve as a validation set for our ongoing efforts to sequence and assembly the genome of *Rotylenchulus 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 bolls 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).
- DNA isolation typically done from a pooled sample of nematodes
- Parasitic relationship requires culturing the nemaode with its associated host plant and soil

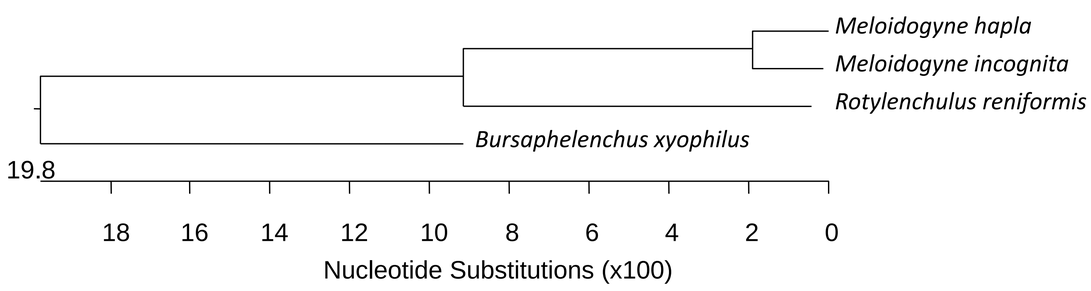

Figure 2. *R. reniformis* 18S rRNA (EU306342) aligned using CLUSTALW with other sequenced plant parasitic nematodes, *M. hapla* (AY593892), *M. incognita* (AY284621), and *B. xylophilus* (FJ235886).

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.

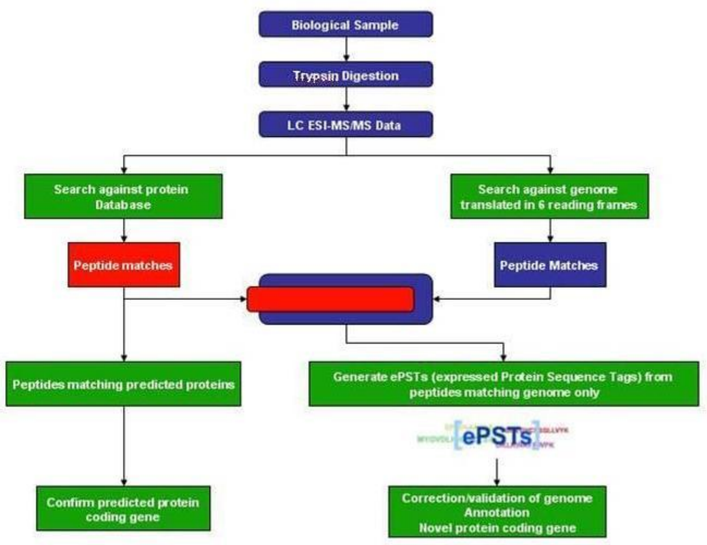

Figure 3. The Proteogenomic Mapping Pipeline.

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.

Materials & Methods

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 4°C 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°C 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.

Results & Discussion

By searching *R. reniformis* peptide spectra against a database of theoretical spectra generated from the *C. elegans* proteome, we were able to identify 355 proteins in *R. reniformis*.

Searching the *R. reniformis* spectra against a database of theoretical spectra made by translating the *C. elegans* genome into all 6 open reading frames, we identified an additional 1,144 proteins, or protein modifications present in *R. reniformis* that are not present in the set of annotated proteins for *C. elegans*.

A majority of the proteins and protein modifications identified by our searches agains the C. elegans proteome and genome were identified as proteins serving basic house-keeping roles within the organism – histones, telomerases, DNA polymerases, etc.

Previous work to identify genomic sequences in R. renifomis identifed 631 UniGene EST sequences (Wubben MJ, Callahan FE, Scheffler BS. *Transcript analysis of parasitic females of the sedentary semi-endoparasitic nematode Rotylenchulus renifomis.* Molecular Biochemistry & Parasitology. 172(1):31-40.), and combined with our 1,499 new sequences, there are a total of 2,130 sequences that can be used for the verification and validation of future genome and transcrptome assmeblies in the species.

Current Status of the *R. reniformis* Genome Project

A total of 11,155,121,426 bp have been sequenced using combined Illumina and Roche 454 technologies.

Based on an estimated minimum genome size of 58 Mb and an estimated maximum genome size of 144 Mb, we have sequenced in the range of 78X – 191X genome coverage.

In order to obtain a more accurate coverage estimate, we are currently utilizing flow cytometry analysis in order to get an estimate of genome size of *R. reniformis*.

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

	# Sequences	Total Sequence Length (bp)
Illumina 1x75	56,182,791	4,262,542,542
Illumina 1x100	15,613,588	1,561,358,800
Illumina 2x100 (250 bp insert)	28,920,210	2,892,021,000
Illumina 2x100 (350 bp insert)	21,594,142	2,159,414,200
	122,310,731	10,875,336,542
Roche 454 SE	462,594	161,287,004
Roche 454 PE (8kb insert)*	679,515	118,497,880
	1,142,109	279,784,884

Future Work

- More sequencing incorporating mitochondrial DNA removal and non-specific whole genome amplification of *R. reniformis* DNA to help with the genome assembly
- 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
- Proteomics (Proteogenomic Mapping) against other sequenced plant parasitic nematode species to help refine gene models during structural genome annotation
- Proteogenomic Mapping of different lifestages of R. reniformis against a higher quality R. reniformis genome sequence

Funding
USDA ARS 6402-21310-003-18S - Cotton, USDA ARS 6402-21310-003-20S - Biocomputing