# A chromosomal assembly of the soybean cyst nematode genome

Rick Masonbrink[1], Tom Maier[1], Matthew Hudson[2], Andrew Severin[1], and Thomas Baum[1]

[1]Iowa State University
[2]University of Illinois System

March 10, 2021

**Abstract**

The soybean cyst nematode (Heterodera glycines) is a sedentary plant parasite that exceeds a billion dollars in yield losses annually. It has spread across the soybean-producing world, emerging as the primary pathogen of soybeans. This problem is exacerbated by H. glycines populations overcoming the limited sources of natural resistance in soybean and by the lack of effective and safe alternative treatments. Although there are genetic determinants that render soybean plants resistant to certain nematode genotypes, resistant soybean cultivars are increasingly ineffective because their multi-year usage has selected for virulent H. glycines populations. Successful H. glycines infection relies on the comprehensive re-engineering of soybean root cells into a syncytium, as well as the long-term suppression of host defenses to ensure syncytial viability. At the forefront of these complex molecular interactions are effectors, the proteins secreted by H. glycines into host root tissues. The mechanisms that control genomic effector acquisition, diversification, and selection are important insights needed for the development of essential novel control strategies. As a foundation to obtain this understanding, we developed a nine scaffold, 158Mb pseudomolecule assembly of the H. glycines genome using PacBio, Chicago, and Hi-C sequencing. An annotation of 22,465 genes was predicted using a Mikado pipeline informed by published short- and long-read expression data. Here we present results from our assembly and annotation of the H. glycines genome.

## Introduction

Almost all major crops surrender yield losses to parasitic nematodes with annual damages exceeding US one billion worldwide [1]. The most well-known and complex group of these plant parasites comprises root-knot (*Meloidogyne* spp) and cyst nematodes (*Globodera*spp. and *Heterodera* spp), which manipulate the host to develop a long-term feeding site. The soybean cyst nematode (*Heterodera glycines* ) is of particularly great economic importance due to prominent role in reducing soybean yields worldwide [1-4]. Overcoming this crop pest requires scrutinizing the *H. glycines* lifestyle and the molecular exchange at the core of this problem.

*H. glycine* s' lifecycle begins as an egg that is queued to hatch. The emerged juvenile nematode migrates to the host root zone, where it penetrates the outer layers of roots using a combination of mechanical and enzymatic processes, and eventually induces a single root cell near the vascular cylinder to form a feeding site, known as a syncytium [5]. The syncytium becomes metabolically active and expands to incorporate hundreds of adjacent cells through cell wall breakdown and protoplast fusion. The syncytium matures to an efficient nutrient sink with enlarged host nuclei and pronounced cytoplasmic streaming [6, 7].

Successful feeding site development depends upon the parasite's ability to manipulate a complex interaction with its host via the transfer of nematode gland cell-produced effector proteins into or around host root cells [8-10]. During juvenile nematode migration within the root, plant cell walls are digested by an abundance of secreted enzymes including cellulases, pectate lyases and other hydrolases [11-13]. In later parasitic stages,

the nematode manipulates plant metabolism [14], development [15-18], and elicits a dramatic and long-term suppression of host defenses (reviewed by [9, 10, 19, 20]). While the functional mechanisms of many effector proteins remain elusive, a variety of functions have been attributed to previously characterized effector proteins secreted from the esophageal gland cells of root-knot and cyst nematodes [11, 19-33]. For example, a chorismate mutase protein, typically absent in animals, is secreted by root-knot and cyst nematodes to manipulate the plant host's shikimate pathway, a pathway involved in producing aromatic amino acids, plant hormones, cell wall components, and plant defense metabolites [14, 34-36]. Signaling peptides, like CLAVATA3 plant peptide mimics, can affect plant developmental pathways [16, 17, 33, 37-39]. While these effectors have led to a better understanding of plant-nematode interactions, only a small portion have been functionally characterized.

Understanding the totality of effector proteins in the nematode genome and how they manipulate the host will shed light on this molecular interplay, inspiring the development of novel mechanisms to defend plants from these important pests. To accomplish this goal for the soybean cyst nematode, two annotated genome assemblies were published from two different nematode strains: a partially virulent TN10 line [40] and a highly virulent X12 line [41]. Here we improve upon the current genomes by reassembling the TN10 PacBio reads and scaffolding with Chicago and Hi-C reads to obtain the highest quality plant-parasitic nematode genome assembly to date with nine complete pseudomolecule chromosomes and zero unscaffolded contigs. We went to great lengths to ensure the integrity of the assembly, as shown by 97% of input reads mapping back to the assembly and by a high degree of synteny to related species. Though 30-39% of the genome is repetitive, 28 and 58% of the newly assembled genome is syntenic to the X12 assembly and TN10 draft, respectively. While large rearrangements exist between the TN10 and X12 pseudomolecule assemblies, technological improvements in Hi-C scaffolding software (Lachesis vs Juicer) revealed that these differences can be attributed to many small and a few large chromosomal misjoins in the X12 assembly. Though the X12 and this latest TN10 assembly have similar assembly metrics and size, 141 vs 158Mb, choices in gene prediction created a large disparity in gene frequency between annotations. Here we have attempted to bridge this gap with an extensive gene annotation that uses multiple prediction pipelines and lines of evidence to generate an annotation that is complete and comparable to other parasitic nematode species. We also limited our gene homology input to include only genes of related Tylenchida species to prevent the homology-driven over-simplification of gene structure when using more distant and nonparasitic relatives. Fortuitously, this resulted in gene counts (22,465) that were neatly positioned between the two previous assemblies' gene counts [29,769 (TN10) and 11,882 (X12)]. Even with catering to bias in gene structure from parasitism, the evaluation of universal single copy orthologs with BUSCO is still higher in our latest TN10 assembly than in previous assemblies at 83% (Table 1). Using this vastly improved genomic resource, we explore the nature of previously published effectors and other secreted proteins to address the heart of *H. glycines* genomics, to understand the adaptive evolution involved in the constant battle between host resistance and parasite virulence.

# Methods

## Dovetail Chicago library preparation and sequencing:

A Chicago library was prepared as described previously [42]. Briefly, ˜500ng of high molecular weight (HMW) gDNA (mean fragment length = 75 kbp) was reconstituted into chromatin in vitro and fixed with formaldehyde. Fixed chromatin was digested with*DpnII* , the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ˜350 bp mean fragment size, and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X to produce 500 million 151 base paired-end reads.

## Dovetail Hi-C library preparation and sequencing:

A Dovetail Hi-C library was prepared in a similar manner as described previously [43]. Chromatin was fixed in the nucleus with formaldehyde, extracted, and *DpnII* digested. 5' overhangs were filled with biotinylated nucleotides, and then free blunt ends were ligated. Crosslinks were reversed for DNA purification from proteins. Purified DNA was treated to remove biotin outside of ligated fragments. The DNA was then sheared to ˜350 bp, and libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotinylated fragments were isolated with streptavidin beads before PCR enrichment of libraries and sequenced on an Illumina HiSeq X to produce 531 million 2x151 bp paired end reads.

## Genome Assembly

The *H. glycines* genome was assembled with Falcon using previously deposited Pacbio sequencing (SRX2692203 - SRX2692222). Falcon unzip 0.4.0 [44] was then used to reduce the heterozygosity in the assembly. Dovetail Genomics scaffolded this assembly with Chicago and Hi-C reads using a modified SNAP read mapper (http://snap.cs.berkeley.edu) and an iterative assembly with HiRise. This Dovetail assembly was further scaffolded with the previously mentioned Pacbio subreads using Sspace-longread v1.1 [45]. Gmcloser 1.6.2 [46] was used to fill gaps using PacBio circular consensus (CCS) reads. Pilon 1.22 [47] was then used to polish the assembly using ˜142 million 260bp PE Illumina reads, which were processed with Trim Galore 0.4.5 [48], Hisat2 2.1.0 [49], and Samtools 1.9 [50]. The genome was then scaffolded using the previously mentioned Hi-C reads using Juicer 1.7.6 [43], 3D-DNA 180419 [51], and manually corrected using Juicebox 1.9.8 [52]. Once pseudomolecules were assembled, the genome was polished with Pilon 1.23 [47] using CCS reads, then with the 142 million Illumina PE reads, and then with 38 iterations of polishing using Pacbio CCS reads with Pilon 1.23. A final round of polishing was performed with the 142 million Illumina 250bp reads (SRR8381095) using Pilon 1.23 [47].

## Gene Prediction

Genes were predicted using a Mikado 1.24 pipeline [53] that picked consensus transcripts from seven transcriptome assemblies and gene predictions. First, the genome was masked using RepeatModeler 1.0.11 [54] and RepeatMasker 4.0.9 [55]. Previously published Illumina RNA-seq reads (SRX3339090- SRX3339098) were processed with Trim Galore 0.4.5 [48], Hisat2 2.1.0 [49], and Samtools 1.9 [50] on both a masked and an unmasked genome. Previously published NCBI expressed sequence tags (downloaded 06-17-19) and IsoSeq (SRX3702373) were aligned using Gmap (version 2018-03-25) [56]. These data were utilized with Braker 2.1.0 [57] using all three data sources, annotating both an unmasked assembly and a masked assembly to compensate for parasitism-related CNV genes. Transcriptomes were assembled using the guidance of a masked genome with Trinity 2.3-.2 [58, 59], Class2 2.1.7 [60], Stringtie 1.3.4a [61, 62], and Spades 3.13.1 [63]. This first Mikado prediction was utilized in a second round of Mikado, supplemented with masked braker prediction and a Maker 2.31.10 [64] gene annotation from a 368-scaffold version of the assembly. All resulting predictions from the second round of Mikado were collapsed into gene loci via using shared intron/exon borders with Cufflinks gffread (Cufflinks 2.2.1) [65].

The Maker annotation mentioned previously was run over four rounds, with Maker's internal algorithm first, then Augustus 3.2.1, then Snaphmm 2006-07-28, followed by GeneMark-ES 4.32. Repeatmodeler 1.0.11 and RepeatMasker 4.0.9 were used to perform the softmasking used in the annotation. Maker utilized all transcripts and proteins from related species genomes [66, 67] and UniProt [68], including:*Bursaphelenchus xylophilus [69]* , *Ditylenchus destructor* , *Globodera pallida [70]* , *Globodera rostochiensis [71]* , *Globodera ellingtonae [72]* ,*Meloidogyne floridensis [73]* , *Meloidogyne hapla [74], Meloidogyne incognita [75]* ,*Parastrongyloides trichosuri* , *Rhabditophanes_KR3021* ,*Strongyloides papillosus* , *Strongyloides ratti* ,*Strongyloides stercoralis* ; all *H. glycines* ESTs from NCBI [67], and a Braker 1.9 [76] annotation on this unmasked assembly using published RNA-seq (SRX3339090- SRX3339098).

### Functional Gene Annotations

Gene annotations were compiled from Interproscan 5.27-66.0 [77] and BLAST [78] searches to NCBI NT and nr databases downloaded on 10-23-19 [67], as well as swissprot/uniprot databases downloaded on 12-09-2019 [68]. Genes encoding transposable element-associated proteins were identified using Bedtools 2.27.1 [79] with exon overlaps to Repeatmodeler-predicted transposable elements.

### Differential Gene Expression

The strandedness of the RNA-seq was evaluated with RseQC V4.0 [80, 81], followed by alignment to the genome with HiSat (2.2.0) [49], and converted to bam with Samtools (1.1.0) [50]. Read counts were calculated with FeatureCounts from Subread package (1.6.0) [82], followed by Deseq2 (1.20.0) [83] with P-value cutoffs at 0.05 to determine differential expression between the samples.

### BUSCO analysis

Universal single copy orthologous genes were assessed using BUSCO 3.0.2 [84-86] on both the predicted proteins and the genome against the nematoda ODB9 dataset. Missing genes were verified with BLAST [78] to the predicted protein sequences using a 0.01 e-value and 1.6x -0.4x length cutoff (S table 1).

### Effector gene mapping

Effector proteins were mapped to the predicted proteome using Diamond 0.9.23 [87]. Effector genes were mapped to the genome using Gmap (2018-03-25) [56]. Secreted proteins were identified with SignalP 5.0 [88] on the predicted proteome.

### Repeat Prediction

Multiple repeat predictions were pursued to finely detail genome structure. To comprehensively predict the structure of transposable elements in the genome with Extensive de-novo TE Annotator, EDTA 1.7 [89]. Tandem repeat finder 4.0.9 [90] was run on the genome to identify tandem repeats. A repeat prediction sensitive to copy number variation was also pursued with RepeatModeler 1.0.11 [54] and RepeatMasker 4.0.9 [55].

### Synteny

Genome alignments were performed using Mummer3 [91] and merged for display in Circos 0.69-6 [92] using Bedtools merge [79]. By inferring gene orthology from primary mapping sites of the predicted transcripts from our genome with Gmap 2018-03-25 [56], we inferred gene-based synteny with iAdHoRe 3.0.01 [93].

## Results and Discussion

### Genome Quality Metrics

The *H. glycines* genome assembly comprises 2,109 contigs, all of which were incorporated into the expected 9 pseudomolecule scaffolds using the Juicer pipeline, in agreement with cytological observations [41, 94]. The genome size of the new assembly is 157,982,452 bp, within the expected range for this clade of species (Table 1). However, this increased genome size may be due to the incorporation of repetitive haplotigs, an assumption supported by the inflation of repeats (61.4 Mb) compared to the previous TN10 *H. glycines* draft (42.1 Mb). Still, total repeat content (38.9%) is within the published range (34-47.7%) [40, 41], and yet maintains a lower repeat content than the X12 assembly (67.3 Mb) (Supplementary Table 1).

To assess quality and completeness, the input sequences were aligned to the assembly. High rates of alignment: 97.3%, 97.2%, and 73.5%, were observed for Pacbio subreads, Pacbio CCS reads, and 260bp PE Illumina reads, respectively. To evaluate the genic complement of the annotation, we ran BUSCO3 (Benchmarking Universal Single Copy Orthologs) [86] on both the genome and its predicted proteins. Of the 982 possible Nematoda BUSCO genes, 634 (64.6%) and 674 (68.6%) were complete, 46 (4.7%) and 122 (12.4%) were duplicated, and 86 (8.8%) and 88 (9%) were fragmented, respectively. A stringent BLAST of missing BUSCO proteins on the predicted proteins found 141 of the missing BUSCO proteins (median e-value of 5.8e-18), achieving a possible complete rate of 83% [95]. Overall, the high proportion of input read mapping, high BUSCO scores, and complete incorporation of all contigs, suggests this latest *H. glycines* assembly is of high quality.

## Improvements over existing soybean cyst nematode assemblies

This pseudomolecule assembly is a massive step forward in the genomics of plant-parasitic nematodes, increasing the ability of interspecies comparisons. To assess the contiguity and accuracy of our new assembly we used gene-based synteny with BLAST and i-ADHoRe [93] as well as direct chromosome alignments using Mummer3 (Figure 1). With the gene-based approach, 67Mb and 31.7Mb of synteny was found to the TN10 draft and the X12 genome assemblies, respectively. Using Mummer, these assessments rose to 92.4Mb and 43.4Mb, respectively. Considering that more than 61Mb of repeats are in the genome, synteny to 42-58% and 20-27% of the genome in the TN10 draft and X12 assemblies, respectively, is high.

Assignment of contigs to chromosomes was improved in this assembly compared to existing X12 assembly. These differences resulted in the identification of a number of large chromosomal misjoins in the X12 assemby: including multiple interchromosomal translocations and the misassignment of chromosome 9 (Figure 1; Supplemental Table 2). Surprisingly, after adjusting for these large chromosomal misjoins in X12, there were very few chromosomal rearrangements between the two lines of a highly adaptable species (Figure 1).

## Gene Annotation

The gene annotation resulted in 22,465 gene models, encoding 23,933 transcripts with an average gene length of 4,569bp, values that are comparable to related species (Table 2). While the frequency of genes is substantially larger than the previously published X12 annotation (11,882), the propensity for parasites to duplicate genes involved in host-parasite interactions requires a novel approach to gene prediction. To prevent the obliteration of parasitism genes thought to be maintained at high copies in the nematode, we developed an annotation approach was taken to predict all transcribed elements in the genome, including repetitive elements. A genome without repeat masking was used to allow highly similar, high-copy number genes to be identified. However, because repetitive elements frequently reside in noncoding regions of genes, multiple genome-guided transcriptomes and gene predictions enabled the dissection of high-confidence gene models (Supplemental Table 3). This improvement in gene prediction is indicated by our total gene count (22,265). Our analyses included known parasitism genes and repeats missing from X12 (11,882) and produced a more highly contiguous genome than the previous TN10 assembly (29,769). Our average and median gene and transcript lengths are the largest among the compared species, while exon count per transcript has also increased relative to earlier annotations of *H. glycines* and other related species. Another line of evidence to support these gene predictions lies with the high proportion of genes that have functional annotations with 85.2% of predicted proteins or transcripts having homology to sequences in Interpro, Swissprot, NCBI NR, or NCBI NT databases (Supplemental Table 4; Supplemental Table 5).

## Effector gene prediction

With a complete genome, we can now better understand the molecules that are exchanged between the parasite and host. The first step to identifying these molecules lies in characterizing transcripts that produce proteins with signal peptides and without transmembrane domains, which partitioned 1,514 transcripts from

the 23,933 total transcripts, and which were attributed to 1,421/22,465 genes (Table 3). A second elementary step in identifying these molecules lies with attributing the previously published effectors to genes in the genome [11, 19]. Using DNA sequence similarity, 125 potential effector genes were identified with a minimum query alignment and sequence identity of 50%. Using the same parameters with protein query length and identity with Diamond, 362 effector-like protein-coding genes were identified. However, only 44/125 and 117/362 of these putative effectors encode secreted proteins, indicating that genes may be variable among SCN lines in their propensity to be secreted, a variability that may contribute to SCN virulence.

### Expression of Effector genes

In the hopes of further resolving the genes important to the host-parasite exchange, we leveraged existing *H. glycines* RNA-seq (SRP122521). All possible comparisons were made between pre-parasitic (i.e., before root penetration) second-stage juvenile nematodes (PP), second-stage parasitic (i.e., after root penetration) nematodes on a susceptible host (C for compatible), and second-stage parasitic nematodes on a resistant host (IC for incompatible) (Supplemental Data 1). These data were integrated into a tabular database of genes, functional annotations, sequences, and differential expression. Using this database, we filtered genes in the *H. glycines* genome for key traits of genes involved in parasitism, including differential expression at a p-value of 0.05, >1 log2 fold change, signal peptide presence, and the absence of a transmembrane domain. Of the 1,421 genes we assessed, 61 genes were differentially expressed between the PP and C, 392 genes between PP and IC, and 609 in novel comparisons between C and IC samples (Table 3). Among these comparisons, we assessed which genes may be involved in host nucleus reprogramming by annotating NLS signals in these differentially expressed genes. We only found four and fourteen genes encoding proteins with NLS signals in the C and IC vs PP comparisons, respectively. However, comparisons between C and IC revealed 113 differentially expressed genes that were also secreted and nuclear targeted.

While effector genes upregulated in the preparasitic stages are likely to be associated with the migratory phase of parasitism, getting a list of candidate effectors for parasitic stages was less complete. Only four published effectors were upregulated in C vs PP, and three were upregulated in IC vs PP (Table 3). However, by comparing effector genes that were downregulated in parasitic IC samples but also upregulated in C samples vs PP, we discovered a number of effector genes that were downregulated when encountering resistance: 6E07[11], 4G06 (ubiquitin extension)[11], 4D06[11], three versions 45D07 type effectors (chorismate mutase) [14], 30C02 (defense suppressor) [29], four 2D01 type effectors (interacts with plant LRR)[96], 20E03[11], 12H04[11], 5A08 (RAN-binding, interacts with soybean LRRs [97]), and Gland14 (endopeptidase) [19]. While interesting, these decreases in expression could also be interpreted as the early stages of nematode death on a resistant host (IC), warranting further investigation into the mechanisms for these expression changes.

## Conclusion

In summary, we present the most complete *H. glycines* assembly, with a consensus gene prediction pipeline sensitive to the prediction of high-copy parasitism-related genes. We confirm this with a high percentage of synteny to previous assemblies, high read mapping rates, and the complete integration of all contigs into nine pseudomolecules. Using currently available data, we compiled a comprehensive resource that extensively annotates *H. glycines* genes, a critical resource for the development of advanced technology to combat this pest. This resource will be integrated into SCNBase.org, which further extends the transparency and availability of *H. glycines* genomic data. To demonstrate the utility of this new resource, we assessed the conservation of previously published effectors and leveraged published RNAseq and gene features to further explore effector expression during the host-parasite exchange.

## Data availability

All scripts and notes used to prepare this genome are available at Dovetail2SCNGenome@github.com. The genome, annotation, and Hi-C reads were uploaded to Genbank and SRA under the Bioproject PRJNA603076 and SRR8381095. All genome track data and annotations will also be hosted on SCNBase.org.

## Author Contributions

Conceptualization – REM, AJS, TM, TB; Data curation – REM, MH; formal analysis – REM, AS, TM; funding acquisition – AJS, TM, TB; investigations – REM, AJS, TM, TB; methodology – REM, TM, AJS, TB; resources – AJS, TM, MH, TB; Software – REM, AJS; Validation – REM; Visualization – REM; Writing – REM; Review and Editing – REM, TM, MH, AJS, TB

# References

1. Nicol, J., et al., *Current nematode threats to world agriculture* , in *Genomics and molecular genetics of plant-nematode interactions* . 2011, Springer. p. 21-43.

2. Davis, E.L. and G.L. Tylka, *Soybean cyst nematode disease.* The plant health instructor, 2000.

3. Jones, J.T., et al., *Top 10 plant-parasitic nematodes in molecular plant pathology.* Molecular plant pathology, 2013.**14** (9): p. 946-961.

4. Koenning, S.R. and J.A. Wrather, *Suppression of soybean yield potential in the continental United States by plant diseases from 2006 to 2009.* Plant Health Progress, 2010. **10** .

5. Pogorelko, G., et al., *A cyst nematode effector binds to diverse plant proteins, increases nematode susceptibility and affects root morphology.* Molecular plant pathology, 2016. **17** (6): p. 832-844.

6. Abad, P. and V.M. Williamson, *Plant nematode interaction: a sophisticated dialogue* , in *Advances in botanical research* . 2010, Elsevier. p. 147-192.

7. Lilley, C.J., H.J. Atkinson, and P.E. Urwin, *Molecular aspects of cyst nematodes.* Molecular Plant Pathology, 2005. **6** (6): p. 577-588.

8. Eves-van den Akker, S., et al., *The Feeding Tube of Cyst Nematodes: Characterisation of Protein Exclusion.* PLOS ONE, 2014.**9** (1): p. 1-9.

9. Hewezi, T. and T.J. Baum, *Manipulation of plant cells by cyst and root-knot nematode effectors.* Molecular Plant-Microbe Interactions, 2013. **26** (1): p. 9-16.

10. Mitchum, M.G., et al., *Nematode effector proteins: an emerging paradigm of parasitism.* New Phytologist, 2013. **199** (4): p. 879-894.

11. Gao, B., et al., *The parasitome of the phytonematode Heterodera glycines.* Molecular Plant-Microbe Interactions, 2003.**16** (8): p. 720-726.

12. Rai, K.M., et al., *Genome wide comprehensive analysis and web resource development on cell wall degrading enzymes from phyto-parasitic nematodes.* BMC plant biology, 2015. **15** (1): p. 187.

13. De Boer, J.M., et al., *Cloning of a putative pectate lyase gene expressed in the subventral esophageal glands of Heterodera glycines.* Journal of nematology, 2002. **34** (1): p. 9.

14. Bekal, S., T.L. Niblack, and K.N. Lambert, *A chorismate mutase from the soybean cyst nematode Heterodera glycines shows polymorphisms that correlate with virulence.* Molecular Plant-Microbe Interactions, 2003. **16** (5): p. 439-446.

15. Mitchum, M.G., X. Wang, and E.L. Davis, *Diverse and conserved roles of CLE peptides.* Current opinion in plant biology, 2008.**11** (1): p. 75-81.

16. Wang, J., et al., *Dual roles for the variable domain in protein trafficking and host-specific recognition of Heterodera glycines CLE effector proteins.* New Phytologist, 2010. **187** (4): p. 1003-1017.

17. Wang, J., et al., *Identification of potential host plant mimics of CLAVATA3/ESR (CLE)-like peptides from the plant-parasitic nematode Heterodera schachtii.* Molecular plant pathology, 2011.**12** (2): p. 177-186.

18. Matthews, B.F., et al., *Arabidopsis genes, AtNPR1, AtTGA2 and AtPR-5, confer partial resistance to soybean cyst nematode (Heterodera glycines) when overexpressed in transgenic soybean roots.* BMC plant biology, 2014. **14** (1): p. 96.

19. Noon, J.B., et al., *Eighteen new candidate effectors of the phytonematode Heterodera glycines produced specifically in the secretory esophageal gland cells during parasitism.* Phytopathology, 2015.**105** (10): p. 1362-1372.

20. Hewezi, T., *Cellular signaling pathways and posttranslational modifications mediated by nematode effector proteins.* Plant physiology, 2015. **169** (2): p. 1018-1026.

21. Elling, A.A. and J.T. Jones, *Functional characterization of nematode effectors in plants* , *Plant-Pathogen Interactions* . 2014, Springer. p. 113-124.

22. Eves-van den Akker, S., et al., *Identification and characterisation of a hyper-variable apoplastic effector gene family of the potato cyst nematodes.* PLoS pathogens, 2014. **10** (9).

23. Gao, B., et al., *Identification of putative parasitism genes expressed in the esophageal gland cells of the soybean cyst nematode Heterodera glycines.* Molecular Plant-Microbe Interactions, 2001.**14** (10): p. 1247-1254.

24. Maier, T.R., et al., *Isolation of whole esophageal gland cells from plant-parasitic nematodes for transcriptome analyses and effector identification.* Molecular Plant-Microbe Interactions, 2013.**26** (1): p. 31-35.

25. Vanholme, B., et al., *Detection of putative secreted proteins in the plant-parasitic nematode Heterodera schachtii.* Parasitology research, 2006. **98** (5): p. 414-424.

26. Wang, X., et al., *Signal peptide-selection of cDNA cloned directly from the esophageal gland cells of the soybean cyst nematode Heterodera glycines.* Molecular Plant-Microbe Interactions, 2001.**14** (4): p. 536-544.

27. Chronis, D., et al., *A ubiquitin carboxyl extension protein secreted from a plant-parasitic nematode Globodera rostochiensis is cleaved in planta to promote plant parasitism.* The Plant Journal, 2013.**74** (2): p. 185-196.

28. Haegeman, A., et al., *Functional roles of effectors of plant-parasitic nematodes.* Gene, 2012. **492** (1): p. 19-31.

29. Hamamouch, N., et al., *Τηε ιντεραστιον οφ τηε νοελ 30ϒ02 ϛψστ νεματοδε εφφεστορ προτειν ωιτη α πλαντ β-1, 3-ενδογλυϛανασε μαψ συππρεσς ηοστ δεφενϛε το προμοτε παρασιτισμ.* Journal of experimental botany, 2012. **63** (10): p. 3683-3695.

30. Hewezi, T., et al., *Arabidopsis spermidine synthase is targeted by an effector protein of the cyst nematode Heterodera schachtii.* Plant physiology, 2010. **152** (2): p. 968-984.

31. Lee, C., et al., *The novel cyst nematode effector protein 19C07 interacts with the Arabidopsis auxin influx transporter LAX3 to control feeding site development.* Plant Physiology, 2011.**155** (2): p. 866-880.

32. Patel, N., et al., *A nematode effector protein similar to annexins in host plants.* Journal of Experimental Botany, 2010.**61** (1): p. 235-248.

33. Wang, X., et al., *A parasitism gene from a plant-parasitic nematode with function similar to CLAVATA3/ESR (CLE) of Arabidopsis thaliana.* Molecular Plant Pathology, 2005. **6** (2): p. 187-191.

34. Lu, S.-W., et al., *Alternative splicing: a novel mechanism of regulation identified in the chorismate mutase gene of the potato cyst nematode Globodera rostochiensis.* Molecular and biochemical parasitology, 2008. **162** (1): p. 1-15.

35. Noon, J.B. and T.J. Baum, *Horizontal gene transfer of acetyltransferases, invertases and chorismate mutases from different bacteria to diverse recipients.* BMC evolutionary biology, 2016.**16** (1): p. 74.

36. Vanholme, B., et al., *Structural and functional investigation of a secreted chorismate mutase from the plant-parasitic nematode Heterodera schachtii in the context of related enzymes from diverse origins.* Molecular plant pathology, 2009. **10** (2): p. 189-200.

37. Lu, S.-W., et al., *Structural and functional diversity of CLAVATA3/ESR (CLE)-like genes from the potato cyst nematode Globodera rostochiensis.* Molecular Plant-Microbe Interactions, 2009.**22** (9): p. 1128-1142.

38. Olsen, A.N. and K. Skriver, *Ligand mimicry? Plant-parasitic nematode polypeptide with similarity to CLAVATA3.* Trends in plant science, 2003. **8** (2): p. 55-57.

39. Replogle, A., et al., *Nematode CLE signaling in Arabidopsis requires CLAVATA2 and CORYNE.* The Plant Journal, 2011. **65** (3): p. 430-440.

40. Masonbrink, R.E., et al., *The genome of the soybean cyst nematode (Heterodera glycines) reveals complex patterns of duplications involved in the evolution of parasitism genes.* bioRxiv, 2018: p. 391276.

41. Lian, Y., et al., *Chromosome-level reference genome of X12, a highly virulent race of the soybean cyst nematode Heterodera glycines.*Molecular ecology resources, 2019. **19** (6): p. 1637-1646.

42. Putnam, N.H., et al., *Chromosome-scale shotgun assembly using an in vitro method for long-range linkage.* Genome research, 2016.**26** (3): p. 342-350.

43. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.*science, 2009. **326** (5950): p. 289-293.

44. Chin, C.-S., et al., *Phased diploid genome assembly with single-molecule real-time sequencing.* Nature methods, 2016.**13** (12): p. 1050-1054.

45. Boetzer, M. and W. Pirovano, *SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information.* BMC bioinformatics, 2014. **15** (1): p. 211.

46. Kosugi, S., H. Hirakawa, and S. Tabata, *GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments.* Bioinformatics, 2015. **31** (23): p. 3733-3741.

47. Walker, B.J., et al., *Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.* PloS one, 2014. **9** (11): p. e112963.

48. Krueger, F., *Trim galore.* A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 2015.

49. Kim, D., B. Langmead, and S. Salzberg, *HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes* . 2017.

50. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25** (16): p. 2078-2079.

51. Dudchenko, O., et al., *De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.* Science, 2017.**356** (6333): p. 92-95.

52. Durand, N.C., et al., *Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom.* Cell systems, 2016.**3** (1): p. 99-101.

53. Venturini, L., et al., *Leveraging multiple transcriptome assembly methods for improved gene structure annotation.* GigaScience, 2018. **7** (8): p. giy093.

54. Smit, A., R. Hubley, and P. Green, *RepeatModeler Open-1.0. 2008-2010.* Access date Dec, 2014.

55. Smit, A., R. Hubley, and P. Green, *RepeatMasker Open-4.0. 2013–2015.* Institute for Systems Biology. http://repeatmasker.org, 2015.

56. Wu, T.D., et al., *GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality.*Statistical Genomics: Methods and Protocols, 2016: p. 283-334.

57. Hoff, K.J., et al., *BRAKER2: incorporating protein homology information into gene prediction with GeneMark-EP and AUGUSTUS.* Plant and Animal Genomes XXVI, 2018.

58. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Nature protocols, 2013. **8** (8): p. 1494-1512.

59. Henschel, R., et al. *Trinity RNA-Seq assembler performance optimization* . in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond* . 2012.

60. Song, L., S. Sabunciyan, and L. Florea, *CLASS2: accurate and efficient splice variant annotation from RNA-seq reads.* Nucleic acids research, 2016. **44** (10): p. e98-e98.

61. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nature protocols, 2016. **11** (9): p. 1650.

62. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature biotechnology, 2015.**33** (3): p. 290.

63. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* Journal of computational biology, 2012. **19** (5): p. 455-477.

64. Cantarel, B.L., et al., *MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.* Genome research, 2008. **18** (1): p. 188-196.

65. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.*Nature protocols, 2012. **7** (3): p. 562.

66. Howe, K.L., et al., *WormBase ParaSite- a comprehensive resource for helminth genomics.* Molecular and biochemical parasitology, 2017. **215** : p. 2-10.

67. Coordinators, N.R., *Database resources of the national center for biotechnology information.* Nucleic acids research, 2016.**44** (Database issue): p. D7.

68. Consortium, U., *UniProt: a worldwide hub of protein knowledge.* Nucleic acids research, 2019. **47** (D1): p. D506-D515.

69. Kikuchi, T., et al., *Genomic insights into the origin of parasitism in the emerging plant pathogen Bursaphelenchus xylophilus.*PLoS pathogens, 2011. **7** (9): p. e1002219.

70. Cotton, J.A., et al., *The genome and life-stage specific transcriptomes of Globodera pallida elucidate key aspects of plant parasitism by a cyst nematode.* Genome biology, 2014. **15** (3): p. R43.

71. Eves-van den Akker, S., et al., *The genome of the yellow potato cyst nematode, Globodera rostochiensis, reveals insights into the basis of parasitism and virulence.* Genome biology, 2016.**17** (1): p. 124.

72. Phillips, W.S., et al., *The Draft Genome of Globodera ellingtonae.* Journal of nematology, 2017. **49** (2): p. 127.

73. Lunt, D.H., et al., *The complex hybrid origins of the root knot nematodes revealed through comparative genomics.* PeerJ, 2014.**2** : p. e356.

74. Opperman, C.H., et al., *Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism.*Proceedings of the National Academy of Sciences, 2008. **105** (39): p. 14802-14807.

75. Abad, P., et al., *Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita.* Nature biotechnology, 2008. **26** (8): p. 909.

76. Hoff, K.J., et al., *BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS.* Bioinformatics, 2015: p. btv661.

77. Finn, R.D., et al., *InterPro in 2017—beyond protein family and domain annotations.* Nucleic acids research, 2016. **45** (D1): p. D190-D199.

78. Madden, T., *The BLAST sequence analysis tool* , in *The NCBI Handbook [Internet]. 2nd edition* . 2013, National Center for Biotechnology Information (US).

79. Quinlan, A.R., *BEDTools: the Swiss-army tool for genome feature analysis.* Current protocols in bioinformatics, 2014: p. 11.12. 1-11.12. 34.

80. Wang, L., S. Wang, and W. Li, *RSeQC: quality control of RNA-seq experiments.* Bioinformatics, 2012. **28** (16): p. 2184-2185.

81. Wang, L., et al., *Measure transcript integrity using RNA-seq data.* BMC Bioinformatics, 2016. **17** (1): p. 58.

82. Liao, Y., G.K. Smyth, and W. Shi, *The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote.* Nucleic acids research, 2013. **41** (10): p. e108-e108.

83. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15** (12): p. 550.

84. Seppey, M., M. Manni, and E.M. Zdobnov, *BUSCO: assessing genome assembly and annotation completeness* , in *Gene Prediction* . 2019, Springer. p. 227-245.

85. Simão, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015: p. btv351.

86. Waterhouse, R.M., et al., *BUSCO applications from quality assessments to gene prediction and phylogenomics.* Molecular biology and evolution, 2017. **35** (3): p. 543-548.

87. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nature methods, 2015. **12** (1): p. 59-60.

88. Armenteros, J.J.A., et al., *SignalP 5.0 improves signal peptide predictions using deep neural networks.* Nature biotechnology, 2019. **37** (4): p. 420-423.

89. Ou, S., et al., *Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline.* Genome biology, 2019. **20** (1): p. 1-18.

90. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic acids research, 1999. **27** (2): p. 573-580.

91. Kurtz, S., et al., *Versatile and open software for comparing large genomes.* Genome biology, 2004. **5** (2): p. R12.

92. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome research, 2009. **19** (9): p. 1639-1645.

93. Proost, S., et al., *i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets.* Nucleic acids research, 2011. **40** (2): p. e11-e11.

94. Cotten, J., *Cytological investigations in the genus heterodera.* Nematologica, 1965. **11** (3): p. 337-342.

95. Weinstein, D.J., et al., *The genome of a subterrestrial nematode reveals adaptations to heat.* Nature communications, 2019.**10** (1): p. 1-14.

96. Lin, M., *Characterization of 16B09 and 2D01 effector proteins in cyst nematodes.* 2016.

97. Quentin, M., P. Abad, and B. Favery, *Plant parasitic nematode effectors target host defense and nuclear functions to establish feeding cells.* Frontiers in plant science, 2013. **4** : p. 53.

# Captions

Figure 1. Hi-C and synteny plot comparing *H. glycines* TN10 genome and X12 genome. A. Hi-C plot of the nine pseudomolecules in the TN10 genome. B. Gene-based synteny between TN10 and X12 *H. glycines*genomes.

Table 1. Relevant genome stats of related nematode genome assemblies.

Table 2. Gene, transcript, and exon stats for the TN10 pseudomolecule assembly and related species statistics

Table 3. Differentially expressed transcripts and genes with consecutive filtration by significant differential expression, greater than one log2fold change, signal peptide presence, and the lack of transmembrane domain. These filtered genes are further defined by the presence of a nuclear localization signal and if a gene was associated with a previously identified effector.

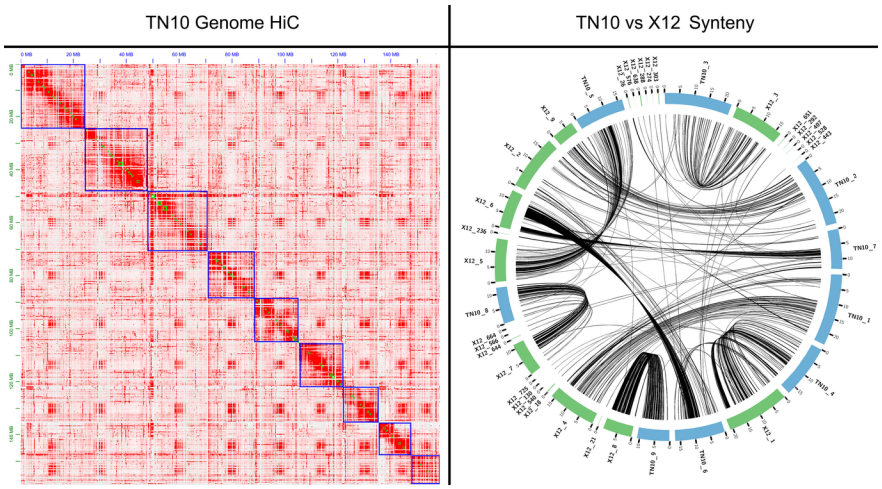
Supplemental Table 1. Genomic repeat content comparisons between the current assembly, the TN10 draft, and the X12 genome.

Supplementary Table 2. Chromosome size differences between the TN10 and X12 pseudomolecule assemblies.

Supplemental Table 3. Gene, transcript, and exon statistics for the final consensus gene annotation, and the input assemblies.

Supplemental Table 4. Genes and mRNAs in the TN10 genome that were annotated by a database.

Supplemental Table 5. Comparison of TN10 and X12 genome prediction through the annotation of various databases.




### Hosted file

`Tables_And_STables.xlsx` available at https://authorea.com/users/238645/articles/512958-a-chromosomal-assembly-of-the-soybean-cyst-nematode-genome