BIOAG FINAL REPORT

TITLE: Sequencing the genomes of two critically-important biological control agents

PRINCIPAL INVESTIGATOR(S) AND COOPERATOR(S):

William E. Snyder, Professor, Department of Entomology; wesnyder@wsu.edu
Axel Elling, Assistant Professor, Department of Plant Pathology; elling@wsu.edu
Zhen "Daisy" Fu, postdoctoral scholar in the Department of Entomology (Snyder lab); zhen.fu@wsu.edu

Our project is in coordination with Andy Jensen, Research Director of the ID, WA and OR Potato Commissions; nematodes were collected on-farm in cooperation with potato growers across the Columbia Basin.

KEY WORDS: BIODIVERSITY; BIOLOGICAL CONTROL; GENOMICS; ECOLOGY; NATURAL PEST CONTROL

ABSTRACT

Potatoes are a valuable Washington crop that is threatened by devastating insect pests. We have found (1) that insect-killing “entomopathogenic” nematodes (EPNs) are key natural enemies of these insects, and (2) that organic farming greatly increases EPN genetic diversity. Indeed, genetically-diverse mixes of EPN strains are more lethal to insects than any single worm strain. We would like to identify the specific genes that allow different EPN strains to “complement” one another in killing pests. Unfortunately, our efforts have been limited by the lack of well-constructed reference genomes for the two most common species of EPNs in Washington, *Steinernema feltiae* and *Heterorhabditis bacteriophora*. We proposed to use the third-generation advanced sequencers now available at WSU (e.g., the PacBio RS) to construct high-coverage genomes of these two species. Insect-killing nematodes can be used as bio-pesticides. Understanding the traits that make worms lethal to pests will allow us to design effective bio-pesticide blends, and to conserve and enhance beneficial nematode biodiversity on farms. Thus, our project will develop *novel approaches to pest management* that *increase the sustainability of farming systems*.

PROJECT DESCRIPTION

Recent advances in genomics hold the potential to revolutionize plant breeding. Sequencing technology is getting both cheaper and more powerful, such that complete genomes exist for many important crops (e.g., Valesco et al. 2010). When genes associated with agronomically-important traits can be identified, targeted breeding programs can be designed to promote these genes (Perez de Castro et al. 2012). However, modern genomics approaches have not been used for the targeted development of biological control agents that possess desirable traits. If genes associated with biocontrol effectiveness could be identified, strategies could then be developed to enhance the frequency of these traits during natural enemy conservation or when biological control agents are deployed as bio-pesticides. Linking a natural enemy's pest-control effectiveness to underlying genes also would enhance our fundamental understanding of biological control.

The Colorado potato beetle, *Leptinotarsa decemlineata*, is a major pest of potato crops worldwide. The beetles quickly develop resistance to new insecticides, making biological control an attractive (and sometimes the only) control option. Potato beetles are attacked by insect-killing, “entomopathogenic” nematodes (EPNs). Working in Washington potato fields, we have found that two EPN species, *Steinernema feltiae* and *Heterhabditis bacteriophora*, are most common. EPNs are exceptionally abundant on organic potato farms, exerting natural pest control that could explain how organic potato farmers produce high yields without much insecticide use (Ramirez and Snyder 2009, Crowder et al. 2010). The nematodes can be sprayed as a "bio-pesticide"to improve beetle control (Grewal et al. 2005). Most recently, we have found that organic farming promotes greater genetic diversity within each EPN species, and that greater intraspecific genetic diversity improves pest control. For example, pairs of two genetically-different isolates of *S. feltiae* kill far more insects than any single strain. Thus, EPN strains "complement" one another. BIOAg funding in 2012 allowed us to conduct RAD-TAG sequencing to delineate genetic differences among complementary nematode strains. This sequencing approach involves digesting the genome with a restriction endonuclease, which after shearing is then sequenced. The short fragments (or "tags") of DNA that flank each digestion site are screened for genetic variation (i.e., single nucleotide polymorphisms or "SNPs"). Interpreting these sequences requires alignment with high-quality reference genomes. Unfortunately, we found that the *S. feltiae* and *H. bacteriophora* genomes initially provided to us by other researchers were of low quality, with poor genome coverage and thus weak alignment with our RAD-TAG sequences. This made it difficult to annotate our SNPs and correlate them with specific genes, such that we could not generate the convincing preliminary data that we need to pursue federal grants.

We proposed to make use of the third-generation sequencing technology then just-recently available at the genomics core facility at WSU, to construct high-quality reference genomes for *S. feltiae* and *H. bacteriophora*. This would allow us to use our RAD-TAG data to search for genes that underlie nematode complementarity. Additionally, we proposed to use new sequencing technology at U Idaho to compare gene expression profiles among pairs of complementary nematode strains, to provide an additional, and complementary, tool to search for complementarity-related EPN genes. Our ultimate goal is to engineer EPN bio-pesticides that combine strains with complementary modes of activity, and to gain a greater fundamental understanding of how EPN strains complement one another to kill more pests.

In summary, our project had 3 inter-related objectives:

(1) Sequence the genomes of the two most common insect-attacking nematode species in WA potato fields, *S. feltiae* and *H. bacteriophora*.

(2) Compare gene-activity patterns among pairs of different strains of *S. feltiae* that complement one another to kill more beetle pests than could any single strain alone.

(3) Use these data to identify genes that lead nematodes to complement one another.

OUTPUTS

- Work Completed:

We initially faced challenges in obtaining high-quality and high-molecular DNA. We first tried a phenol:chloroform extraction protocol, which gave the highest yield. However, inevitable phenol carryover caused partial failure of library preparation and consequently a low output of sequences from the PacBio sequencer. Because of these difficulties, we next took the approach of compromising yield but circumventing phenol carryover; this can be accomplished by following the protocol of the Qiagen Blood and Tissue kit. This second method resulted in carbohydrate carryover in the DNA sample, which would likely thwart any sequencing efforts. As an additional step to address this newest problem, we incorporated the Qiagen QIAshredder spin column into the Blood and Tissue kit to filter lysed cells, which efficiently removed carbohydrates and improved sample handling. This, finally, resulted in clean DNA.

A library of 15-20kb DNA fragments was constructed from nematode DNA and loaded into the sequencer. The PacBio sequencer generated 4 GB of data. All of these data were imported into the Hierarchical Genome Assembly Process (HGAP) 2.0, which is embedded in the single molecule real-time (SMRT) Portal software. HGAP2.0 is an integrated program that allows the executing of assembly steps in a web-based graphical user interface. Low quality reads and reads shorter than 500 bp were removed from the assembly. After assembly and consensus polishing, we obtained an assembly with 66 Mb, N50=5,170, consisting of over 15,000 contigs with the longest contig being 115,000 bp (Table 1).

We aligned our assembly (this will be referred as "assembly" hereafter) with a draft genome (will be referred as "draft genome" hereafter) of *S. feltiae*, which had been provided by collaborators, using the program Mauve (Darling et al. 2010) with default parameters. The alignment showed that the assembly shares a large number of homologous regions with the draft (Figure 1, 2). Intensive colored lines connecting the upper and lower panels depict Locally Collinear Blocks (LCBs), free of internal genome rearrangements. White areas in the lower panel demonstrate genome regions containing sequence elements specific to the draft genome. Also notice that our assembly is shorter than the draft genome, which implies our assembly is not complete (Figure 2). We conclude that the assembly is highly homologous to the draft genome, which verifies that the sequences we obtained and incorporated into our genome assembly reflect DNA from nematodes.

In the months since our last progress report, submitted in support of our request for a no-cost extension to the original award, we have made dramatic progress in each of the following areas:

1. Realignment of *H. bacteriophora* RAD-tag sequences to the reference genome

The release of the *Heterorhabditis bacteriophora* genome and predicted annotation has facilitated our progress with this species. We are able to relate each of the single nucleotide polymorphisms (SNP) identified in our RAD-tag sequences to specific genomic regions, whether in coding or non-coding regions. When the SNP was in the coding region, we employed a gene annotation program to identify the functional role of that particular genomic region. This is a major step towards our ultimate goal of linking specific genes to complementary effects of different nematode strains on their insect hosts.

Approximately 4 % of the *H. bacteriophora* genome was represented in our RAD-tag sequencing. We realigned the reads of our *H. bacteriophora* populations (these included 16 field-collected strains and one laboratory population) to the reference genome (Wormbase accession number: PRJNA13977) using Bowtie2 (http://bowtie-bio.sourceforge.net) deploying the "sensitive" preset. The resulting 17 BAM files were used as input in SAMtools (Li et al. 2009). The Mpileup program in SAMtools was used to generate a pileup file from the BAM alignment files. In the sequence alignment process, 76% - 87% of the reads were aligned to the reference genome. To avoid the short reads aligning to multiple genomic regions, we employed a high stringency approach. The mean coverage of the SNP among all the populations was 3283 x.

2. Single nucleotide polymorphism (SNP) identification and SNP effect

The generated pileup file was used as the input in VarScan (Koboldt et al. 2012) to identify each SNP. SNPs were further filtered by coverage, p-value, and minimum allele counts using in-house Perl scripts. In the end, we included 10,214 SNPs where all of the populations have reads covered. Meanwhile, we identified 227 SNPs that are population specific, meaning that not all of the populations have reads at those loci. It is highly likely that polymorphism at those loci cause non-recognition of the restriction enzyme; consequently, no RAD-tags were generated.

Filtered SNPs were annotated using SnpEff (Cingolani et al. 2012). Because of the unavailability of a database for *H. bacteriophora*, genome sequence fasta files and annotation files in gff3 format were used to create a new database locally. Annotated SNPs were further filtered and categorized by functional class and effect in SnpSift (Cingolani et al. 2012). With the detailed annotation of the *H. bacteriophora* genome, we were able to designate each SNP to one of the following genomic regions: (1) an intergenic region (regions > 5 kb from genes), (2) an upstream/downstream region (< 5 kb from genes), (3) a regulatory region (splice donor, with close proximity to the exons and introns), or (4) exons or introns. The number of SNPs identified in the intergenic region, down/upstream and intron regions, synonymous substitution in exons and regulatory regions, and nonsynonymous of exons were 2,957, 3,736, 935, and 2,586 respectively (Table 2). The 2,585 nonsynonymous SNP are located in 1,058 unique genes.

3. Population differentiation and population stratification

Principal component analysis (PCA) was employed to infer *H. bacteriophora* population structure through EIGENSTRAT (Price et al. 2006), and the top two components were used to plot population structure (Figure 1). Populations from the same field were found to be more genetically similar than those from different fields (Figure 3). The first component (x axis) separates populations of field NW6 and 11 (NW6 and 11 are from the same farm) from populations of Hintz 32 and the laboratory population. The second component (y axis) differentiates populations of Hintz, the laboratory population and that from field 32. Noticeably, two populations collected from NW6 are clustered with populations collected from field 11 on both first and second component projections, which might be due to similar ancestry as fields 11 and NW6 originated from the same farm, 1 km apart. The laboratory population appeared to share similar ancestry as populations from Hintz.

Fixation index or F-statistics (Fst), which measure population differentiation based on the variation of allele frequencies among populations, were calculated for each *H. bacteriophora* population pair. Fst values of nematode populations from different fields were found to be significantly higher than Fsts of nematode populations derived from the same field; this pattern was found at all genomic regions (Figure 4). The higher Fst indicates populations from distant fields are more divergent (Figure 4).

We conducted nonparametric tests on Fst means of nematode population pairs from the same field vs. different fields, at each nonsynonymous site. The P-value was significant at 499 out of 2568 nonsynonymous sites. Interestingly, for all of those sites, the Fst of pairs between the fields is higher than the pairs derived from the same field (Figure 5). Those 499 SNPs affect 100 unique genes, including a large number of genes that play important roles in biological process of the nematodes (Figure 6), e.g. responses to stimulus, locomotion and signaling. This provides us with a number of candidate genes, and candidate genes family, to examine in our search for genes related to strain complementarity.

- Publications, Handouts, Other Text & Web Products:

Jones, EI, Z Fu, DW Crowder, C Bates, R Jabbour, PA Hohenlohe, WE Snyder, and AA Elling. Manuscript. Emergent effects of intra- and inter-specific genetic diversity among natural enemies. To be submitted to *Molecular Ecology*.

- Outreach & Education Activities:

IMPACTS

- Short-Term: We are producing detailed genomes for beneficial insect-killing nematodes, and identifying candidate genes that could explain why different strains of the same nematode species complement one another in killing hosts.
- Intermediate-Term: We will successfully compete for funding to further investigate these genetic differences; identify genes that correspond to important traits tied to a worm-strains ability to kill hosts and develop the means to search for these traits in nematodes in potato fields; and, design and test bio-pesticides that combine beneficial and complementary nematode traits.
- Long-Term: Provide new commercial bio-pesticides that effectively control potato beetles, while also providing a model approach for understanding why natural enemies complement one another that can be applied to other pests and/or cropping systems.

ADDITIONAL FUNDING APPLIED FOR / SECURED

Our ability to compete for federal funding was initially thwarted by our lack of a high-quality nematode genome, due to the difficulties described above that we have now overcome. An additional challenge in this past year was the decision by our key collaborator, Axel Elling, to move to an industry position and leave academics (and WSU). Nonetheless, we now have the necessary preliminary data in support of grant applications to one or more of the following funding sources: (1) USDA Foundational Program: Entomology and Nematology; (2) the USDA OREI and/or ORG programs; and (3) the NSF Ecology program.

GRADUATE STUDENTS FUNDED

This project is supporting the research of 2 PhD students in PI Snyder's laboratory, Carmen Castillo-Carillo and Karol Krey. Carmen and Karol continue to make excellent progress toward their degrees, with graduation of both students expected in 2015.

RECOMMENDATIONS FOR FUTURE RESEARCH

Our RAD-TAG data are now allowing us to search for genes that underlie complementarity in the nematode *H. bacteriophora*. Several key steps remain. We would like to use new sequencing technology to compare gene expression profiles among pairs of complementary nematode strains, to provide an additional, and complementary, tool to search for complementarity-related EPN genes. Our ultimate goal is to engineer EPN bio-pesticides that combine strains with complementary modes of activity, and to gain a greater fundamental understanding of how EPN strains complement one another to kill more pests. This can be pursued by examining genes related to ecological differences among nematode strains, such as those related to host infection or nematode foraging, identified in our RAD-tag sequencing.

Table 1. Parameters of the genome assembly

Total number of bases	66,301,155
Total number of contigs	15,683
Max length (bp)	115,484
Min length (bp)	1,000
Mean length (bp)	4,228 +/- 3,492
N50 (bp)	5,170

Table 2. Summary of single nucleotide polymorphism characterized from 17 *Heterorhabditis bacteriophora* populations

Genomic regions		# of SNPs identified
Coding region	Synonymous	2586
	Nonsynonymous	935
Non-coding region	Intergenic region	2957
	Upstream/downstream of genes and intron	3736

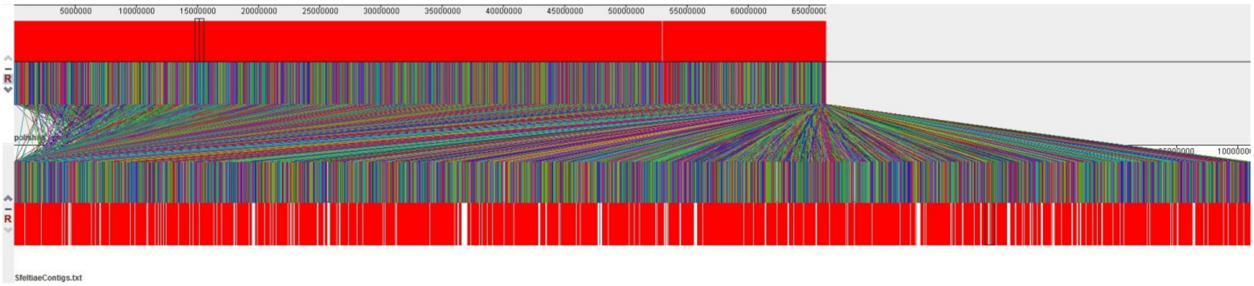

Figure 1. Overview of alignment of current *S. feltiae* assembly (upper panel) and *S. feltiae* draft genome (lower panel) provided by P. Sternberg at the California Institute of Technology. Lines connecting panels depict homologous regions, known as Locally Collinear Blocks (LCBs). Because recombination can cause genome rearrangements, homologous regions of one genome may be reordered or inverted relative to another genome (the “X” shaped lines at the very left of the figure). Areas that are completely white were not aligned and probably contain sequence elements specific to a particular genome. The graph was generated by the program Mauve (Darling et al. 2010).

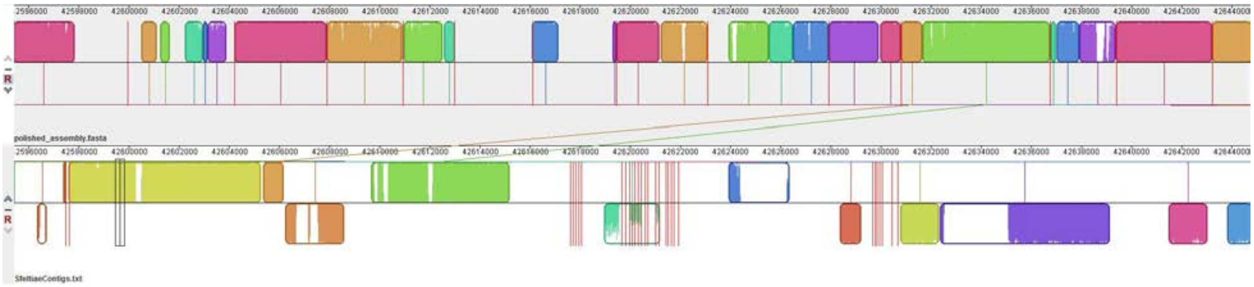

Figure 2. Detailed view of the comparison between our *S. feltiae* assembly (upper panel) and the draft genome (lower panel) provided by our collaborators. Each genome's panel contains the name of the genome, a scale showing the sequence coordinates, and a single black horizontal center line. Each colored block depicts genome sequence that aligned to part of the other genome, known as Local Collinear Blocks (LCBs). When a block lies above the center line, the aligned region is in the forward orientation relative to the first genome sequence. Blocks below the center line indicate regions that align in the reverse complement (inverse) orientation. Inside each block the program draws a similarity profile of the genome sequence. The height of the similarity profile corresponds to the average level of conservation in that region of the genome sequences. Areas that are completely white were not aligned and probably contain sequence elements specific to a particular genome. Red vertical lines are boundaries of contigs. Notice that the green lines and brown lines connecting the upper and lower panel demonstrate homologous regions.

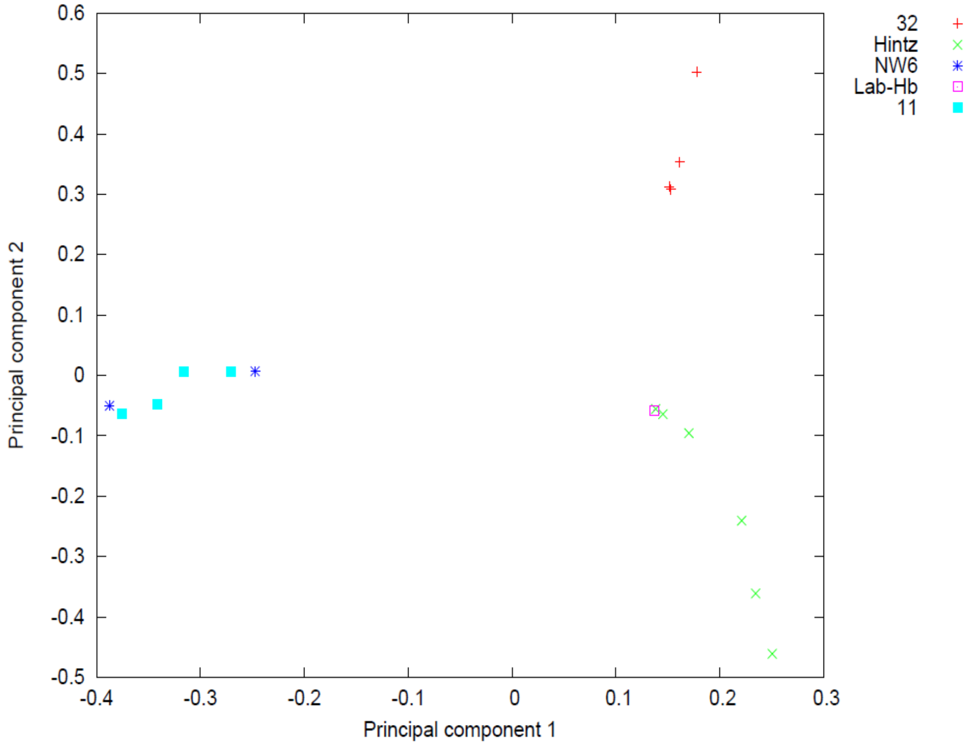

Figure 3. Projection of top two components from principal component analysis inferred from genome-wide genotyping of *Heterorhabditis bacteriophora* populations.

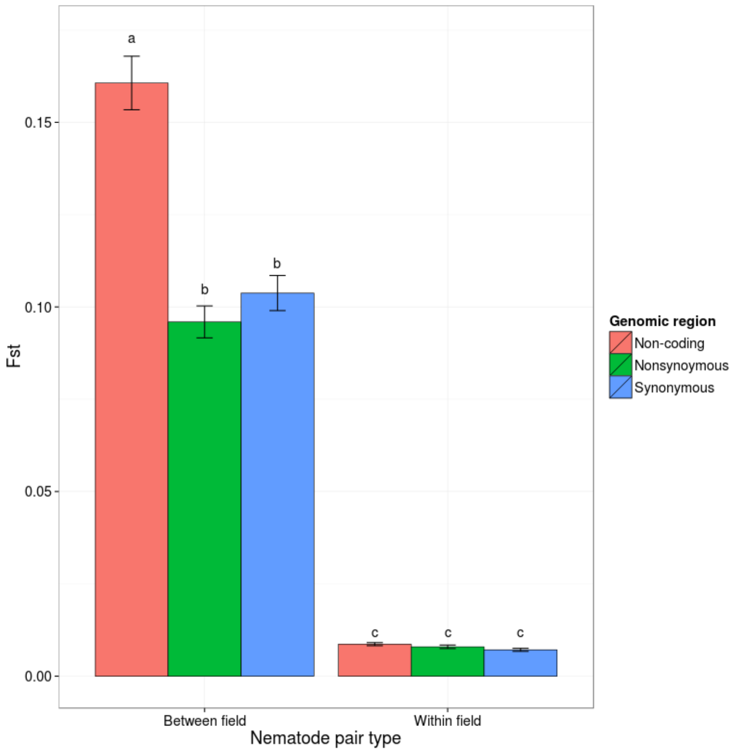

Figure 4. Pairwise population differentiation (Fst) of *Heterorhabditis bacteriophora* at the following genomic regions: (1) non-coding region, (2) synonymous substitution, and (3) nonsynonymous substitution. Different letters indicate statistical differences determined from a Tukey HSD mean comparison at *P* = 0.01.

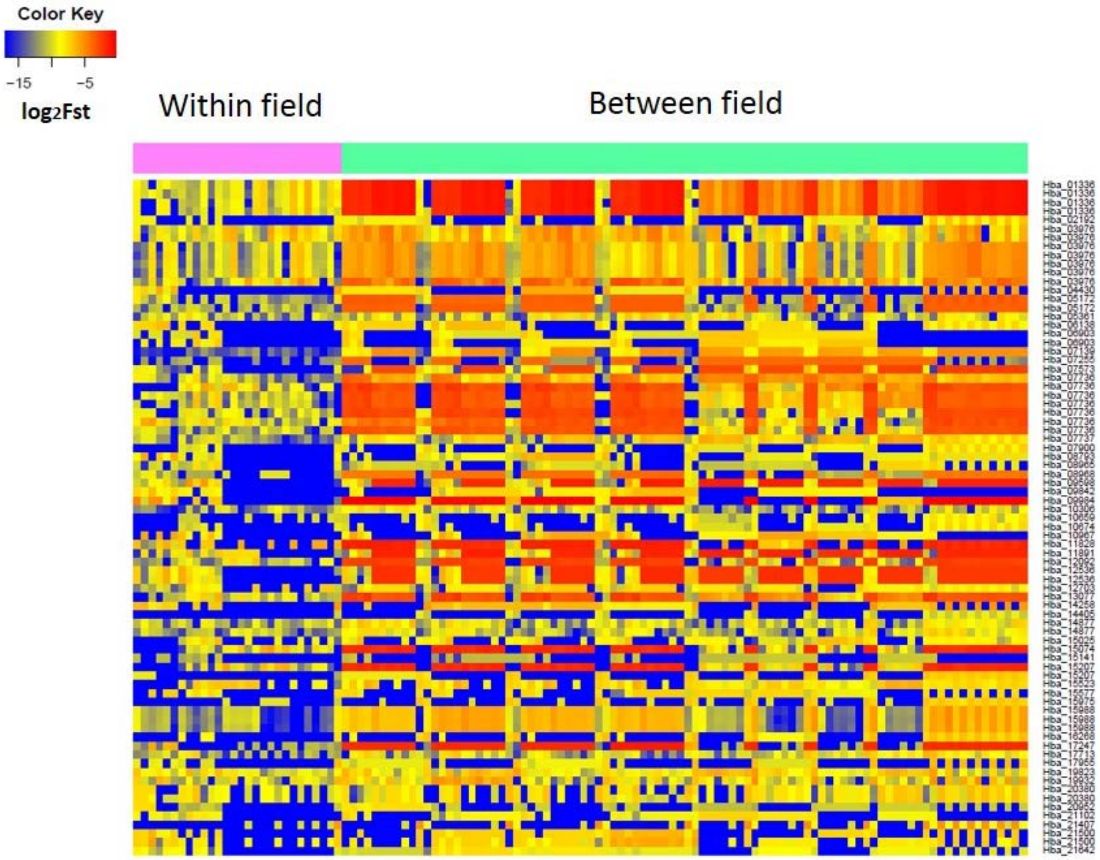

Figure 5. Fst of pairwise *Heterorhabditis bacteriophora* populations collected from the same field (within field) or from different fields (between fields). Fst was log 2 transformed. Each column of the above heat-map represents a pair of nematode populations. Each row represents a locus where the mean Fst of nematode pairs within fields vs. between fields was significantly different.

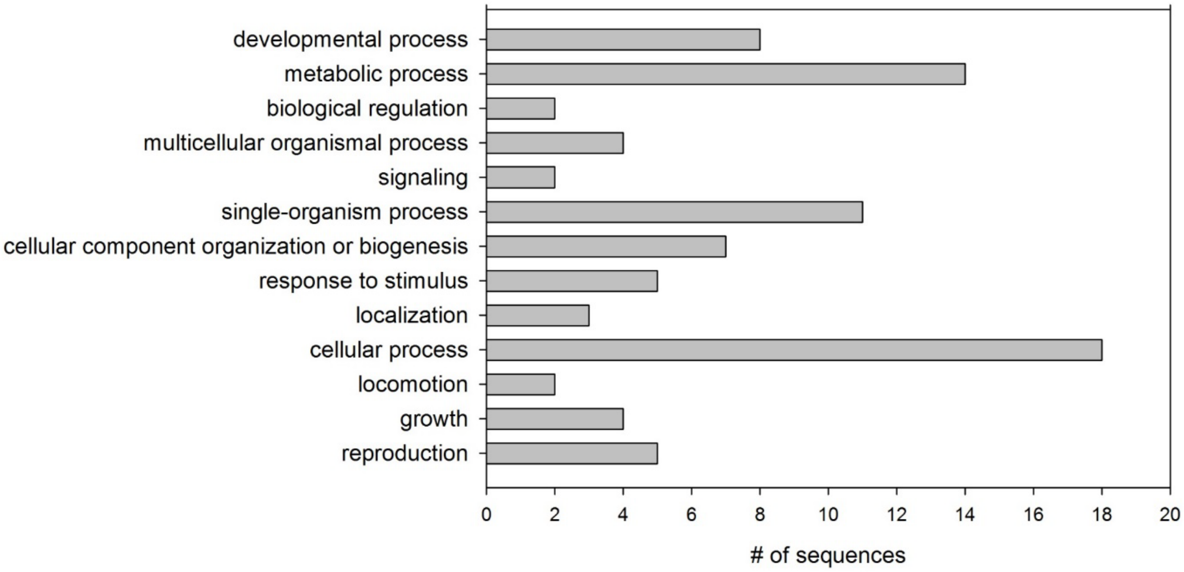

Figure 6. Putative gene functions of loci where mean Fsts of nematode population pairs within fields vs. between fields significantly differed.