Abstract

CROMER, JOHN PATRICK. Genome Annotation of *Meloidogyne hapla*. (Under the direction of David Bird and Charles Opperman).

Ascribing function to the genes in a species is one of the most instructive ways to learn about the underlying mechanisms an organism employs to overcome its specific environmental challenges. Following the genomic sequencing and assembly of the root-knot nematode *Meloidogyne hapla*, a gene set was constructed to provide a foundation for understanding the specific means by which plant parasitism is achieved. The utilization of manual computational gene finding for a subset of the genome, paired with automated gene-finding programs enhanced the accuracy of the resulting gene set. Broad trends in the genome-lifestyle relationship can be uncovered by analyzing groups of genes with similar structures. Through the comparison of the G-Protein Coupled Receptor, Nuclear Hormone Receptor, and collagen gene families found in the model free-living bacterivore *C. elegans* to *M. hapla*, a clearer relationship of gene structure and function to nematode lifestyle mode emerges. For all three families, a *C. elegans* expansion and/or an *M. hapla* reduction was observed, emphasizing the breadth in ability (through larger gene families) required for a free-living lifestyle and the efficient dependency on host resources (through more compact gene families) needed for a parasitic lifestyle. Prior discoveries of horizontal gene transfer and protein mimicry in nematodes inspired a broad screen for *M. hapla*-encoded plant peptide hormones. This analysis resulted in the identification of a putative *M. hapla* Rapid Alkalinization Factor (RALF) plant hormone mimic, which presumably plays a key role in host development events critical for the survival of the plant parasite.

Genome Annotation of *Meloidogyne hapla*

by
John Patrick Cromer

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

Bioinformatics

Raleigh, North Carolina

2010

APPROVED BY:

_____	_____
Dr. David Bird Committee Chair	Dr. Charles Opperman Committee Co-chair
_____	_____
Dr. Laura Mathies	Dr. Eric Stone

Biography

I completed a B.S. in Biological Sciences from NC State University in 2001. Concurrent with several years of laboratory experience in an agricultural biotechnology company, I pursued complementary formal education leading to fulfillment of the Master of Crop Science degree (with a Statistics minor) from NC State University in 2005. Through that experience and education, I developed both an appreciation for and an interest in computational methods applied to biological problems. The Bioinformatics program at NC State University provided the opportunity for rigorous training in this area. The exposure to a wide variety of challenges in the field of nematode genome annotation (under the direction of Dr. David Bird, Dr. Charles Opperman, Dr. Laura Mathies, and Dr. Eric Stone) further cultivated my interest, skills, and knowledge in this union of genomics, statistics, and computer science. Through this training, and through my contributions, I was awarded a Ph.D. in Bioinformatics in 2010.

Table of Contents

List of Tables

List of Figures

List of Boxes

Chapter 1: Manual Gene Finding in *Meloidogyne hapla*

Note: This chapter is an expanded version of my contribution to published work (Opperman et al., 2008).

Introduction

Given a blueprint underlying any design, one can learn about details that would never be obvious barring meticulous disassembly. Given multiple blueprints of designs, hypotheses can be formed about general construction principles based on drawing parallels between external features and underlying construction details. In the genomic sciences, such blueprints exist in genomes, where the entire genetic blueprints of organisms are stored in their DNA sequences. These sequences are composed using only four building blocks corresponding to the bases adenine (A), cytosine (C), guanine (G), and thymine (T). Species differ greatly in the lengths of their genomes, from a few hundred thousand basepairs (pairings of hydrogen-bonded nucleotides on complementary strands) seen in some bacteria to over 600 billion in the amoeboid *Polychaos dubium* (Parfrey et al., 2008). An ever-present goal of genomic scientists is to use these diverse blueprints to understand the mechanisms around which life revolves, and to exploit this understanding for the advancement of knowledge supporting ongoing progress in a wide range of fields such as agriculture, ecology, genetics, medicine, plant pathology, and zoology.

A prime target for genomic analysis and annotation is the phylum Nematoda, comprising both a highly diverse and critically important collection of organisms. Nematodes account

for approximately 80% of all individual animals on Earth, and can be grouped into an estimated 100,000 to 1 million extant species (Parkinson et al., 2004). Their occupancy in lifestyle mode and ecological niche is equally vast, including herbivore, predator of meiofauna, free-living microbivore, and plant and animal parasite. They can be organized into five major clades, based on phylogenetic inferences of small subunit ribosomal RNA: Dorylaimia (clade I), Enoplia (clade II), Spirurina (clade III), Tylenchina (clade IV), and Rhabditina (clade V). All major clades include parasites, as parasitism is believed to have independently appeared at several points during the course of nematode evolution. It is these nematode parasites of both plants and animals that make them so important to the human experience. Animal parasites cause a wide range of diseases, with an estimated 40-50% of the human population infected, and resulting human deaths approaching 100,000 annually (Chan, 1997). An equally substantial number of livestock animals are lost every year due to nematodes. Plant parasitic nematodes inflict comparable damage to crop plants, responsible for over $120 billion US dollars in crop losses annually worldwide (Barker et al., 1994; Koenning et al., 1999).

Once a genome is sequenced (the stage at which the informational content of a genome is extracted from its chemical composition and recorded), and assembled (pieced together based on overlapping segments into portions of maximum length), meaning must be ascribed through the general processes of annotation. Genome annotation attempts to assess three broad issues: where a particular feature of interest occurs (nucleotide-level), what is produced by the genome (protein (or ncRNA)-level), and how these products interact to

result in a biological mechanism (process-level) (Stein, 2001). The most essential products of annotation are arguably the accurate sets of predicted gene and protein sequences, which serve as the foundation for higher-order biological inferences. If nucleotides are letters on a page, genes might be thought of as sentences. They are surrounded by cryptic punctuation (intergenic space, promoters, terminators), and when expressed, have a resulting meaning in the context of a genome (the book). Eukaryotic genes are transcribed in the nucleus from DNA to pre-mRNA; the pre-mRNA is processed into mRNA (these processes include capping, splicing, and polyadenylation of the transcript); finally, the mRNA is sent to the cytoplasm for translation into a protein (Zhang, 2002). Within genes, exons might be thought of as words, while introns might be thought of as space between the words – that is, only the exons impart meaning into the sentence. Some genes consist of a single exon; other genes consist of multiple exons. Identifying and enumerating gene sequences within the genome is the focus of the stage known as "gene finding". In simple prokaryotic genomes (which do not contain introns), brute-force six-way translations with a minimum-length filter yields a fairly large proportion of the complete gene set (Stein, 2001). In contrast, less than one-quarter of the free-living nematode *Caenorhabditis elegans* genome is in coding regions, making the signal-to noise ratio of true gene to non-gene significantly lower. When alternative splicing is also considered, the challenge of identifying the exact beginning and ending positions for every exon in a gene looms even more formidably (Stein, 2001).

Given the large number of gene and protein predictions that must be made for the vast majority of eukaryotic species, automatic de novo prediction programs are often employed.

De novo prediction is classified as an "intrinsic methodology", as it relies only on information present in the individual sequence analyzed (Windsor and Mitchell-Olds, 2006). Gene prediction algorithms include FGENESH (Salamov and Solovyev, 2000), GeneMark.hmm (Besemer and Bordovsky, 1999), Genie (Reese et al., 2000b), GENSCAN (Burge and Karlin, 1997), GlimmerHMM (Salzberg et al., 1999; Majoros et al., 2004), Grail (Uberbacher and Mural, 1991), HEXON (Solovyev at el., 1994), HMMGene (Krogh, 1997), MZEF (Zhang, 1997), and VEIL (Henderson et al., 1997). The algorithms underlying these prediction programs commonly rely on the identification of specific statistical properties of the DNA or motifs through the use of sensors (Stein, 2001). An example of a potentially useful feature affirming the possibility of a eukaryotic coding start region detected by a sensor is a Kozak sequence – the pattern (GCC)RCCATGG (where R is a purine (A or G)) (Kozak 1986; Kozak 1987). While the minority of gene finders (e.g., HEXON, MZEF) only seek the presence of one feature type, the most common strategy relies on the incorporation of information from several sensor types to construct a whole gene model (Stein, 2001). Multiple features are assessed and incorporated into the model using underlying frameworks such as rule-based systems, neural networks (Grail), linear and quadratic discriminant analysis methods, perceptron methods, hexamer-coding measures, weight matrix and weight array methods, maximal-dependence decomposition donor matrices, decision trees, or hidden Markov models (FGENESH, Genie, GeneMark.hmm, GENSCAN, GlimmerHMM, HMMGene, VEIL) (Stein, 2001; Zhang, 2002). Of these frameworks, hidden Markov models have proven to be the most popular and versatile. Hidden Markov models are generative models representing systems as sets of discrete states ("hidden" refers to states that cannot be

directly observed) and the respective transitions (with individually settable probabilities for each transition) between those states. Transitions occur stochastically from state to state, with a single symbol emitted from each state according to the emission probabilities of that state (Majoros et al., 2004). Essentially, for applications where basic principles are known (for instance, there is at least one exon in every eukaryotic gene, while some eukaryotic genes also contain introns, promoters, and poly-A signals), the basic structure of a hidden Markov model can often be sketched before any training data is seen, cognizant of all possible combinations of emissions and transitions within any potential training data set. As training data is run through the model, each of the transition probabilities are adjusted to reflect what has been learned. The trained model can then be run on test data (for instance, a newly-sequenced genome which was not part of the training data set) in order to identify regions corresponding to what the model believes are genes (Stein, 2001). Of course, the choice of training data is crucial to a successful outcome of hidden Markov model applications. Generally, the more specific the training data is to the test data, and the greater the quantity of this training data, the more accurate the test data predictions are. While automatic gene-finding algorithms have been very helpful in the identification of genes, sensitivity (true positive detection ability) and specificity (false positive discrimination ability) of algorithms are significantly better for smaller gene regions (one nucleotide to one exon) than they are for entire gene structures (correctly calling the start and stop position of every exon within a gene) (Reese et al., 2000a; Stein, 2001). For example, in the human genome, 80% of genes are accurately predicted at the nucleotide level (in other words, 80% of protein-coding nucleotides are correctly predicted), compared to 45% at the exon level (where both splice

sites are correct), and 20% at the whole-gene level (where the entire gene structure is correct) (Zhang, 2002). Typically, sensitivity is higher than specificity across test regions of any size for ab initio gene finders (Windsor and Mitchell-Olds, 2006).

To give greater perspective into how ab initio gene finders work, it is instructive to discuss in additional detail the methodology behind the two popular hidden Markov model-based gene finders GlimmerHMM and FGENESH. GlimmerHMM employs a generalized hidden Markov model (GHMM) that emits complete gene features in each state (Majoros et al., 2004). GHMMs provide greater flexibility than a minimalist hidden Markov model in the exploration of alternative gene-finding approaches, as each state can be associated with a different type of feature, such as a splice donor, splice acceptor, exon, or intron. Multiple types of sub-models can be used at each state, and feature states can be retrained separately. The prediction of the best gene models in a GHMM requires the identification of the most probable paths through the model topologies given the sequences. That is, the objective for each gene model is to maximize the likelihood of the most probable path through the GHMM topology given a sequence. GlimmerHMM also incorporates splice site models modified from GeneSplicer (Pertea et al., 2001), a decision tree refined from GlimmerM (Salzberg et al., 1999), the Maximal Dependence Decomposition technique used to improve splice site identification specificity (Burge 1997) and dynamic programming for the efficient optimization of finding the best gene models. Variable-length features states (such as intergenic regions, introns, and exons) in GlimmerHMM are implemented utilizing N^{th} order interpolated Markov models (IMM) (Delcher et al., 1999) for N=8. Thus, the status predicted

for a given base depends on up to eight bases prior to that base in the sequence being analyzed. GlimmerHMM makes several assumptions in the prediction of genes from DNA sequences. Specifically, it is assumed that the coding region of every gene begins with the start codon ATG, a gene has no in-frame stop codons except for the very last codon, and each exon maintains a consistent reading frame with the previous exon. Such assumptions improve the efficiency of optimal gene model computation by restricting the required search space of the underlying GHMM. However, true frame shifts cannot be detected as a result (Salzberg et al., 1999; Majoros et al., 2004). The user is responsible for collecting as many high-quality training sequences specific to the intended target organism genome. Parameters related to estimates of average intergenic region size, upstream and downstream untranslated region size, flanking region size, and false positive and negative rates are optionally specifiable (Majoros et al., 2004).

FGENESH is a hidden Markov model sister program to FGENES, and allows organism-specific parameters for human, *Drosophila*, plants, yeast, and nematodes to be chosen (Salamov and Solovyev, 2000). Its underlying foundation is similar to those in Genie and GENSCAN, but FGENESH places a greater weight on "signal terms" (including start site scores and splice sites) relative to "content terms" (for instance, coding potentials). This preference is intended to reflect the biological significance of the signals, as the environment of the sites (as opposed to the conservation of nucleotides) could be considered a better indicator of true gene status. Bayes theorem is used to assess the a priori probabilities of exons simultaneous with the computation of exon coding scores. As a consequence, coding

scores of the potential exons are typically lower than those computed from GENSCAN (Salamov and Solovyev, 2000).

Information beyond that of the aforementioned de novo (or ab initio) gene finder programs is available for making gene predictions. In fact, the most powerful predictor of whether a genomic sequence is transcribed is similarity of a sequence that is already known to be transcribed with a genomic region (Stein, 2001). Examples of evidence based on sequence similarity include cDNA and EST data from the same species and BLAST matches to genes in other species. Similarity-based gene prediction algorithms (classified as "extrinsic methodologies" because they directly rely on data other than the test sequence itself (Windsor and Mitchell-Olds, 2006)) face their share of challenges, however. Pseudogenes, relatives of known genes which fail to code for a gene product, may exhibit great sequence similarity to a true gene, and present one such challenge. Evidence based on pseudogenes, or conversely, the attempt to find a gene within a pseudogene-rich environment, are annotation traps leading to the possible identification of false positives. Biases towards single-exon genes resulting from requirements of some similarity-based gene prediction software that splices maintain an in-phase open reading frame, alternative splicing of genomic DNA, repetitive elements found in cDNAs, genomic-DNA contaminated ESTs, chimerism, and automated sequencing lane-tracking errors are additional challenges in using sequence similarity as true gene evidence. Further, databases containing similarity data are very rarely complete, and will not contain representatives of transcripts that are expressed in small quantities or under uncommon conditions (Stein, 2001). Examples of extrinsic methodology

(database search-and-alignment or synteny-alignment) programs commonly used in gene finding include BLASTX (Gish and States, 1993; Altschul et al., 1997), Eannot (Ding et al., 2004), SIM4 (Florea et al., 1998), SLAM (Alexandersson et al., 2003), Twain (Majoros et al., 2005), and TwinScan (Korf et al., 2001).

As might be expected, gene prediction calls made using a combination of de novo and similarity-based techniques show greater accuracy than those calls made using only one technique. Using the example of annotating the draft of the human genome, Celera (Venter et al., 2001) and the Human Sequencing Consortium (Lander et al., 2001) represent two major efforts which arrived at similar results (about 30,000 human genes) using two contrasting methods (Stein, 2001). Celera (using the OTTO pipeline for automatic evidence integration) relied on well-characterized human genes from the RefSeq library (Pruitt et al., 2000) and the SWISS-PROT database (Bairoch and Apweiler, 2000) as targets for which to query the human genome against, followed by a splicing pattern refinement phase using GENSCAN. The Human Sequencing Consortium, in contrast, started with GENSCAN de novo gene predictions, refined these predictions using similarity databases, and merged these predicted models with Genie EST and RefSeq (Stein, 2001). There have also been a number of attempts to integrate the ab initio and similarity search phases of gene finding automatically (Gelfand et al., 1996; Xu and Uberbacher, 1996; Kulp et al., 1997; Krogh 2000; Birney and Durbin, 2000; Gotoh, 2000), with one direction centered on splice alignment programs. These programs were inspired by empirical data showing that exon boundary detection in a gene can be optimized if a similar protein homolog for that particular

gene exists (Zhang, 2002). While usually highly accurate, such programs can be computationally intensive, and are largely dependent on an extensive set of homologous sequences in order to identify full-length genes.

Meloidogyne hapla (Northern root-knot nematode) has been established as a tractable model plant-parasitic nematode, whose genome has been sequenced and assembled (Opperman et al., 2008). With a compact genome, it serves as a valuable subject for genomic interrogation of parasitism, especially in light of increasingly abundant nematode reference genomes. Through the use of forward and reverse genetic experimental approaches coupled with the *M. hapla* genetic map, bioinformatic analyses applied to the *M. hapla* sequence offer immense annotation potential. An essential foundation of the annotation effort is the set of full-length *M. hapla* genes, which must be derived from the genomic sequence.

The de novo gene finding programs GlimmerHMM and FGENESH, in conjunction with the alignment stitching software PASA (Program to Assemble Spliced Alignments) (Haas et al., 2003), tylenchid EST data, and the *M. hapla* genome, were used in the Plant Nematode Genomes Group to construct a predicted set of *M. hapla* protein-coding genes. PASA is a dynamic programming algorithm which identifies alternative splice variants through the assembly of cDNA alignments. It accomplishes this through the creation of unique maximal assemblies derived through the merging of compatible overlapping EST and cDNA alignment sets. The original application of PASA was to maximize and consolidate *Arabidopsis* full-length cDNA untranslated regions and to incorporate EST evidence in areas

where no full-length cDNA alignments existed. Doing so greatly improved the existing *Arabidopsis* annotation by providing maximal-length substrates. Several examples of specific alterations made by incorporating EST evidence into the previous *Arabidopsis* annotation include the extension of untranslated regions, the extension of protein sequences, the modification of internal gene structures, the establishment of alternate splicing isoforms, and the identification of novel gene annotations (Haas et al., 2003). ESTs used in PASA were required to align with genomic DNA at a minimum 90% identity and 90% coverage (Opperman et al., 2008).

Manually-identified genes can also serve as training or validation sets for de novo gene finding programs, and were used in order to increase confidence in finding bona fide full-length *M. hapla* genes. My role in the construction of the *M. hapla* gene set (the foundation upon which I made subsequent annotation contributions) was to manually identify a set of full-length genes for both training and validation of automatic gene finders, which were employed to make gene predictions for the entire genome. In order to generate a reference gene set for validating prediction calls made from GlimmerHMM and FGENESH, I followed a "random search" strategy for finding full-length genes from *M. hapla* genomic contigs 11-85. In addition, I validated 64 full-length genes (dispersed throughout a much wider range of *M. hapla* genomic contigs) previously identified by other annotators using a similar strategy. The manual gene-finding procedure entailed a number of steps emphasizing data collection for the dual purposes of deciding whether a full-length gene exists and subsequent training of gene-finding software.

Step One: Decide start and stop points for random manual gene-finding

This most basic of steps simply requires deciding where in the assembled genomic contig set to begin the random manual gene-finding process. This decision depends partly on the intended goals of the process. In most cases, any set of verified full-length genes is acceptable for training prediction software, so equal representation of gene families or gene processes is not important. Conversely, if an even sampling of general classes of genes (or some other representative criterion) is required, a more tactical approach would be in order. Under one hypothesis, it makes no difference where one starts because we have no *a priori* knowledge of the composition of contigs or the distribution of genes across contigs, so we are equally likely to pick a representative sampling of genes regardless of starting point. One natural starting point under this hypothesis is at the beginning of assembled genomic contig 0. Under an alternative hypothesis, it would be better to pick multiple, randomly-generated starting points in order to sample from a broader and more representative set of genes over the whole collection of assembled genomic contigs. This latter hypothesis might entail the use of a random number generator mapped to the acceptable range of assembled genomic contigs (for instance, given 2500 assembled genomic contigs, an annotator might end up sampling from contigs 56, 240, 249, 522, 666, 1101, 1729, 1786, 2034, and 2291). The selection of "end points" for each starting point will be largely dependent on time limitations of human annotators. Given unlimited time, an annotator would ideally manually find all genes in the entire genome of interest. However, in most cases, such a goal would be unreasonably expensive, both in terms of time and labor, to achieve. Therefore, a more practical approach would involve the assignment of either a preset number of assembled

genomic contigs (e.g., 30 contigs) or a preset window of time (e.g., one month) to which manual annotation shall be dedicated.

Based on discussions with collaborators, because other annotation work had been done starting with genomic contig 0, I decided that a good starting point would be contig 11. No specific end-point contig goal was given; rather, a rough time window of two months was dedicated to the collection of data for every EST (and associated dependent genomic feature and analysis) encountered starting with genomic contig 11.

Step Two: Collection of EST names and approximate locations

The second step of the manual gene finding process serves as the foundation for all other steps: the use of EST evidence as initial reference points for candidate full-length genes. In a graphical genome viewer such as Gbrowse, several 'tracks' of genomic information are aligned in parallel to allow for visual inspection of available evidence for any assembled genomic contig (Figure 1). The reference point for all evidence is the genomic contig number and position (usually expressed in kilobases) within that genomic contig. Starting at the chosen genomic contig numbers from Step One, I scanned each genomic contig for the existence of EST evidence. Because Gbrowse only displays EST and protein details for regions less than 50kb, it was necessary to both scroll and zoom using windows smaller than 50kb. The more crowded a genomic contig region is in terms of available data, the more zooming is necessary. The presence of EST data is a minimum requirement for the additional investigation of full-length genes, as it indicates a particular region of the genome is

transcribed. However, the mere existence of EST data at a given location is insufficient to call that region a full-length gene.

An example of EST evidence on Contig 30:50000-55000 is shown in Figure 2. I recorded the names and locations (to the nearest tenth of a kilobase) of all ESTs encountered in the process of successively scrolling and zooming through contigs using Gbrowse. Based on the genomic positioning ruler in Figure 2, I recorded 50.2kb and 55.0kb as the approximate start and stop locations for EST contig 1542.

Step Three: Collection and examination of support from potentially homologous proteins

Once the existence of EST data, and hence, a potentially transcribed region, is established, it is useful to examine potentially homologous protein sequences in other species in order to evaluate the likelihood that the corresponding genomic region is translated. Sequences similar to the genomic region with EST evidence can be found using the genomic reference sequence as the query in a BLASTX analysis (Gish and States, 1993; Altschul et al., 1997), following manual removal of introns from the genomic sequence using the EST sequence as an exon guide. Alternatively, TimeLogic GeneDetective (Active Motif, Inc., Carlsbad, CA) (Box 1) may be employed to automatically TeraBLAST genomic DNA against target protein sequences by applying additional logic (proprietary alignment algorithms) to detect splice sites and map the proteins to exonic regions. Three useful databases to query against include the “non-redundant” gene set from NCBI (NR) (http://www.ncbi.nlm.nih.gov/), The

Universal Protein Resource (UNIPROT) (http://www.uniprot.org/), and the more specific nematode sequence database (Wormpep) (http://www.wormbase.org/). Analyses of GeneDetective using protein databases as targets may be displayed in the Gbrowse viewer. The first major test that I applied at this stage was a preponderance of agreement of intron and exon structure between genomic DNA, EST sequence, and at least one (preferably many) proteins. In Figure 2, fair agreement of exon structure (50% of exons are in agreement, while a slight extension of the proteins from databases into the C-terminal direction would lead to at least 75% agreement) is seen between a portion of the EST sequence and several proteins (53.2kb to 54.2kb of genomic contig 30). By clicking on any individual entry in one of the GeneDetective tracks, more detailed information is displayed from NCBI. Additional information for 'gi|193209526|ref|NP_001123107.1' accessed by clicking on its link is shown in Figure 3. The next major test I assessed was the confidence (curation level) of the potentially homologous annotated protein. Similar protein sequences with a known function and 'confirmed' status (rather than 'hypothetical', 'putative', or 'predicted' status) lend greater weight to subsequent annotation steps of the query. Because hsp-43 has a known general function and is not explicitly annotated as 'hypothetical', 'putative', or 'predicted', greater weight is given to analyses dependent on this protein (Figure 3).

Following this strategy, I recorded potentially homologous protein sequence information (such as protein name, accession number, species, and function) from the GeneDetective vs. NR, UNIPROT, and Wormpep databases, for each EST contig with protein evidence. This involved following a link from a protein exhibiting inferred consensus or near-consensus

intron and exon structure with other proteins from each of the three databases used in the GeneDetective analyses to access each of their descriptive pages. Maximum e-value cutoffs for results to be shown in Gbrowse from WormPep, NR, and UNIPROT GeneDetective analyses were e^{-15}, e^{-15}, and e^{-5}, respectively. Recall that e-value cutoffs indicate the number of results exceeding the minimum inclusion criteria that will occur through chance alone. In terms of a BLAST analysis, an e-value of 1 means that for every true query match to a database, we expect another match to occur simply due to chance. An e-value of e^{-15} indicates that for every 10 quadrillion matches, it is expected that one equally good match (in terms of similarity) of query to target will occur by chance alone. Thus, e-values give an estimate of statistical significance in terms of counts, as opposed to probabilities. Lower e-values equate to better statistical significance.

Step Four: Obtain exact EST and genomic exon positioning information using SIM4

To fully elucidate the coding regions of the genomic DNA for the purposes of full-length gene identification and for constructing accurate training sets for automated gene-finding programs, I found it necessary to utilize more powerful tools. SIM4 allows for the accurate mapping of EST sequences to the exons of genomic DNA, providing exact EST length in basepairs, genomic coordinates of the EST alignment (exact exon/EST positioning information), percent identity of each genomic exon to the EST sequence, and the genomic strand (reference or complement) to which the EST aligns.

To use SIM4, individual genomic and EST sequences must be in separate files. Either through text editor finding, copying, and pasting functions or through the use of a Perl script, these sequences are ready to use in SIM4 once they are extracted from their large master files. SIM4 may be run simply by entering the string "sim4 seq1 seq2_db" at a UNIX prompt, where seq1 is the filename of genomic sequence in FASTA format, and seq2_db is a single sequence or a file of FASTA formatted EST sequences. Following the example of EST contig 1542 and genomic contig 30, I typed "sim4 Mh10g200708_Contig30 Contig1542 > sim.Contig1542" at the UNIX prompt to capture the output seen in Figure 4 in the file 'sim.Contig1542'. So as to repeat these steps for each and every EST encountered (and save all SIM4 output files), I created a UNIX directory for each genomic contig, and within each genomic contig directory, I established a directory for each EST contig. This produced 207 EST contig directories in total, nested within directories for each genomic contig encountered with any EST evidence. Additionally, I recorded the EST length, the genomic coordinates to which the EST aligns (providing exon locations), and the genomic strand to which the EST aligns (reference or complement) for every EST with protein support, to a table.

Step Five: Prediction of EST ORFs and their translations

Once exon positions of the genomic contig region of interest have been inferred, it is important to also predict the regions of those exons which are translated. In order to compare the protein product of the genomic region of interest to other reference proteins, it is necessary to find the open reading frames of the transcript and translate those regions. EditSeq (from the DNAstar, Inc. package) provides such functionality. Following the

example of EST contig 1542 aligning to genomic contig 30, I copied and pasted the FASTA-formatted DNA sequence of EST contig 1542 (saved from Step Four) into a new EditSeq DNA sequence window (File→New→New DNA). To find open reading frames, I selected the corresponding function from EditSeq (Search→Find ORF→Find Next). The result of this action is shown in Figure 5. I determined that the ORF started at the 58th basepair and stopped at the 1209th basepair, making its total length 1152 basepairs. With some ESTs, multiple ORFs exist, which must be found by repeatedly stepping through the "Find ORF" function. In cases where the EST sequence is predicted to align to the complement strand of the genomic contig sequence, the "reverse complement" of the DNA sequence must be checked (Goodies→Reverse Complement). I recorded all ORF locations for all ESTs with protein support. To obtain the translated sequence of the open reading frame of EST contig 1542, I chose the 'translate' function from the appropriate menu (Goodies→Translate DNA) to obtain the output seen in Figure 6. I saved this protein sequence to a Lasergene protein file (.pro) (File→Save As…), just as I saved all other translated EST ORFs, so that it may be used in the alignment procedure outlined in the next step.

Step Six: Comparison of translated EST ORF to known protein

At this point in the manual gene-finding process, I had deduced protein sequences from each EST mapped to known exonic locations of the genomic contig, as well as a potential ortholog which maps to at least a portion of this same region of the genomic contig. I now wished to assess the degree of coverage and the degree of similarity both proteins share in order to bolster the support for a translated region at the genomic location of interest, to determine if

the EST captures the whole length of the gene, and to possibly ascribe function to the putatively identified full-length gene. The most efficient route to accomplishing these objectives is to first construct an alignment between the two protein sequences.

Using MegAlign (from the DNAstar, Inc. package) protein alignments can be constructed. Continuing with the example of EST contig 1542 aligned to genomic contig 30, I obtained the protein sequence of the *C. elegans* hsp-43 protein from WormMart (found within www.wormbase.org) and aligned it to the *M. hapla* protein using the Clustal W alignment function. In order to create consistent and meaningful alignment reports for all protein alignments, several options must be chosen (and re-chosen for every session of MegAlign). Under Options→Alignment Report Contents, I chose "Show Sequences", "Show Sequence Names", "Show Sequence Positions", and the "Break Alignment" option to break the alignment every 80 residues. Under Options→New Decoration, I chose "Shade", "residues that match", the name of the protein to compare (rather than "the Consensus"), and the degree of shade and color. Clicking "View→Alignment Report" generated an alignment like the one shown in Figure 7. I used GIMP to take a screenshot of every alignment constructed in the manual gene-finding process.

The protein alignment shown in Figure 7 aids in the assertion of the existence of a full-length gene at *M. hapla* genomic contig 30:50231-54959 because it reveals similarity between the protein derived from the *M. hapla* EST and the *C. elegans* hsp-43 protein in excess of that which would be expected by chance alone. The most important criteria used in the

assessment of a protein alignment for purposes of predicting full-length genes is not only an adequate degree of similarity between the two proteins, but also the overlap or near-overlap of the subject protein with the N- and C-termini of the reference protein. Overlap of both reference protein termini establishes 'full-length' status.

Other examples of protein alignments supporting the existence of full-length genes in the manual annotation process are shown in Figures 8 through 12. These figures aid in the identification of full-length genes located at *M. hapla* genomic contigs 20:8152-7100, 30:9682-8603, 31:40163-41414, 53:66263-67429, and 78:35216-36948, respectively. Based on similarity to potentially homologous proteins, predicted functions of these genomic regions are ribosomal subunit formation, RNA polymerase III formation, signal peptidase activity, ubiquitin ligase activity, and oxygen binding, respectively.

Step Seven: Deciding whether a bona fide full-length gene exists at region of interest

With all appropriate evidence corresponding to one region of a genomic contig collected, the decision must be made regarding whether a full-length gene exists at the genomic region of interest. While I considered few metrics absolutely essential in the decision making process, I assessed a number of important criteria. Full-length genes must be supported by an EST unigene sequence, and genomic DNA must cover the entire EST unigene sequence. Similarity between EST sequence and genomic exons typically should exceed 95%. Similar inferred intron and exon structure must exist for a putative orthologous protein at the genomic region of interest (while the nucleotide sequences of introns typically evolve

rapidly, their locations within genes are often highly conserved between modestly-diverged species (Roy and Gilbert, 2005)). The putative orthologous protein must align to the protein derived from the EST unigene with fairly good similarity in at least a few key regions (more than could be expected by chance alone), and the protein derived from the EST sequence must cover (or almost cover) both termini of the orthologous protein. Even greater weight is given to putatively orthologous proteins with a known function, and a 'confirmed' (rather than 'hypothetical', 'putative', or 'predicted') status. Protein support from multiple species in multiple databases all exhibiting similar inferred intron/exon structures and with similar function also increases the confidence of the full-length gene call.

To finish the example of genomic contig 30:50200-55000, EST unigene evidence exists, protein support from a non-putative protein is revealed from three separate databases, the EST aligns with a high degree of similarity to corresponding genomic exons (forming a complete ORF), and the putative orthologous proteins align sufficiently well to label this genomic region a full-length gene.

Gene calls were made for every genomic region with EST evidence. I found a total of 25 full-length genes in *M. hapla* genomic contigs 11 through 85 (Table 1). More comprehensive data for full-length and non-full-length genes collected in the manual gene-finding process are found in Table 2 (data from the validation procedure of previously found full-length genes is not shown). Other examples of full-length genes identified as homologs to *C. elegans* include *nlp-12*, *smi-1*, and *vha-14*. These genes encode a neuropeptide-like protein, a survival of

motor neuron protein interactor, and a vacuolar H ATPase domain subunit, respectively. That these genes encode apparently unconnected functions is not surprising given the 'random' gene-finding approach of a relatively small subset of the genome, and more importantly given that genome position in eukaryotes is not correlated with specific gene function (excluding intra-contig position correlations).

The examples I have presented thus far were selected to lend perspective to the relatively straightforward cases where one EST maps to a particular genomic region, with unambigious open reading frames and convincing protein alignments to potential orthologs. More complex situations exist and occasionally arise. For instance, a single true protein coding gene might require the manual assembly of multiple EST unigenes, in which case intron/exon boundaries of the full-length gene would be determined by the composite intron/exon structure of the EST unigenes. In some cases, multiple (sometimes dubious) open reading frames are detected in one (relatively short) DNA sequence, requiring judgment to select the most likely true open reading frame. The prediction of strand to which the EST sequence aligns to the genomic contig could reveal a conflict (e.g., exons from the same EST sequence predicted to lie on different strands of the genomic sequence, due to short low-complexity EST-to-genomic matches), or the similarity of EST to predicted genomic exons could be questionable (e.g., in the 70-90% similarity range), making gene prediction difficult. Further, ambiguity occasionally exists in the assessment of protein alignments between ESTs and potential orthologs (e.g., convincing similarity may exist at the N-terminal and C-terminal thirds of the proteins, while the interior third reveals very poor similarity). Judgment is again

required to decide whether a poor alignment could be due to divergence between query and reference sequences or due to a particular annotation error. In practice there is a tradeoff between depth of analyses applied to putative full-length gene detection and the number of full-length genes that can be detected and annotated. As required certainty for full-length gene status increases, the number of full-length genes detected in a given amount of time decreases.

Conclusions

The power of the random manual gene-finding process lies in the degree of effort and attention to detail that human annotators can provide. Due to its time- and labor-intensive limitations, it is rarely practical as an exhaustive stand-alone method of gene finding or genome annotation; rather, it is better suited as one of several supporting pillars of evidence or in the collection of small-scale gene training sets. It is equally suitable as a tool to be used in a directed fashion toward the verification of a handful of putative genes identified using other automated methods. When used in a more directed manner, even greater effort can be focused on a larger number of annotation dimensions depending on the specific goals (e.g., signal sequence detection for predicting protein export patterns, full motif scans and GO-term hierarchical analysis for answering specific functional questions, secondary structure predictions for addressing protein folding issues). Thus, the specific steps outlined in this chapter should by no means be considered canonical or ideal for all annotation goals. They may simply be considered collectively as an adequate procedure for evaluating the fundamental qualifying aspects of full-length genes.

In the context of larger annotation objectives, while significant progress has been made in gene finding and annotation over the last decade, there remain gaping omissions of complete open reading frames for higher eukaryote genomes (Brent, 2005). If the goal is to identify every open reading frame in a genome, Brent argues that information sources and methods relied upon over the past decade will be insufficient; namely, the use of EST and mRNA sequences from randomly-selected cDNA clones, alignments of expressed sequences to loci other than those from which they were originally transcribed, synteny with additional genome sequences, and manual annotation using human curators. Instead, large-scale PCR amplification of select cDNAs followed by amplicon sequencing should be relied upon more heavily, as should cDNA-to-genome alignment programs based on pair hidden Markov model frameworks (Brent, 2005). However, the fundamental principles required for more complete annotations are clear: a higher degree of automation in sequencing, better sequencing technologies and a more complete pooling of transcripts from various expression conditions, combined with even more adaptable computational models better utilizing existing data from all sources. As these principles are realized through the natural progression of advances in the genomic sciences, we might one day asymptotically approach completeness in annotation accuracy, breadth, and dimension.

References

Alexandersson M, Cawley S, and Pachter L (2003). SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.* 13:496-502.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

Bairoch A and Apweiler R (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28:45-48.

Barker KR, Hussey RS, and Krusberg LR (1994). *Plant and Soil Nematodes: Societal Impact and Focus on the Future* (Committee on National Needs and Priorities in Nematology, Society of Nematologists, Marceline, Missouri, USA).

Besemer J and Borodovsky M (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Res*. 27:3911-3920.

Birney E and Durbin R (2000). Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* 10:547-548.

Brent M (2005). Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res.* 15:1777-1786.

Burge C (1997). Identification of genes in human genomic DNA. PhD Thesis, Stanford University, CA.

Burge C and Karlin S (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268:78-94.

Chan MS (1997). The global burden of intestinal nematode infections – fifty years on. *Parasitol. Today* 13:438-443.

Delcher AL, Harmon D, Kasif S, White O, and Salzberg SL (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27:4636-4641.

Ding L, Savo A, Berkowicz N, Meyer RR, Shotland Y, Johnson MR, Pepin KH, Wilson RK, and Spieth J (2004). Eannot: a genome annotation tool using experimental evidence. *Genome Res.* 14:2503-2509.

Florea L, Hartzell G, and Zhang Z (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8:967-974.

Gelfand MS, Mironov A, and Pevner P (1996). Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci. USA* 93:9061-9066.

Gish W and States DJ (1993). Identification of protein coding regions by database similarity search. *Nature Genet.* 3:266-272.

Gotoh O (2000). Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps. *Bioinformatics* 16:190-202.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, and White O (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31:5654–5666.

Henderson J, Salzberg S, and Fasman K (1997). Finding genes in human DNA with a hidden Markov model. *J. Computat. Biol.* 4:127-141.

Koenning SR, Overstreet C, Noling, JW, Donald PA, Becker JO, Fortnum BA (1999). Survey of crop losses in response to phytoparastic nematodes in the United States for 1994. *J. Nematol.* 31: 587-618.

Korf I, Flicek P, Duan D, and Brent MR (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics* 17:S140-S148.

Kozak M (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44:283-292.

Kozak M (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* 15:8125-8148.

Krogh A (1997). Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5:179-186.

Krogh A (2000). Using database matches with HMMgene for automated gene detection in *Drosophila*. *Genome Res.* 10:523-528.

Kulp D, Haussler D, Reese MG, and Eeckman FH (1997). Integrating database homology in a probabilistic gene structure model. *Pacif. Symp. Biocomput.* 232-244.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N,

Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang HM, Yu J, Wang J, Huang GY, Gu J, Hood L, Rowen L, Madan A, Qin SZ, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan HQ, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JGR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K,

Jang WH, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz JR, Slater G, Smit AFA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, and Morgan MJ (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

Majoros WH, Pertea M, and Salzberg SL (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20:2878-2879.

Majoros WH, Pertea M, and Salzberg SL (2005). Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. *Bioinformatics* 21:1782-1788.

Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, Cromer J, Diener S, Gajan J, Graham S, Houfek TD, Liu W, Mitros T, Schaff J, Schaffer R, Scholl E, Sosinski BR, Thomas VP, and Windham E (2008). Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. *PNAS* 105:14802-14807.

Parfrey LW, Lahr DJG, and Katz LA (2008). The dynamic nature of eukaryotic genomes. *Mol. Biol. Evol.* 25:787-794.

Parkinson J, Mitreva M, Whitton C, Marian T, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH, McCarter JP, and Blaxter ML (2004). A transcriptomic analysis of the phylum Nematoda. *Nat. Genet.* 36:1259-1267.

Pertea M, Lin X, and Salzberg SL (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* 29:1185-1190.

Pruitt K, Katz K, Sicotte H, and Maglott D (2000). Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* 16:44-47.

Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, and Lewis SE (2000a). Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10:483-501.

Reese MG, Kulp D, Tammana H, and Haussler D (2000b). Genie – gene finding in *Drosophila melanogaster*. *Genome Res.* 10:529-538.

Roy SW and Gilbert W (2005). Rates of intron loss and gain: Implications for early eukaryotic evolution. *PNAS* 102:5773-5778.

Salamov AA and Solovyev VV (2000). Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* 10:516-522.

Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H (1999). Interpolated Markov models for eukaryotic gene finding. *Genomics* 59:24–31.

Solovyev V, Salamov A, and Lawrence C (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* 22:5156-5163.

Stein L (2001). Genome annotation: from sequence to biology. *Nature Rev.* 2:493-503.

Uberbacher E and Mural R (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA* 88:11261-11265.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman

TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J,

Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, and Zhu X. (2001). The sequence of the human genome. *Science* 291:1304-1351.

Windsor AJ and Mitchell-Olds T (2006). Comparative genomics as a tool for gene discovery. *Curr. Opin. Biotech.* 17:161-167.

Xu Y and Uberbacher EC (1996). Gene prediction by pattern recognition and homology search. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 4:241-251.

Zhang M (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA* 94:565-568.

Zhang MQ (2002). Computational prediction of eukaryotic protein-coding genes. *Nature Rev. Genet.* 3:698-709.

Figures

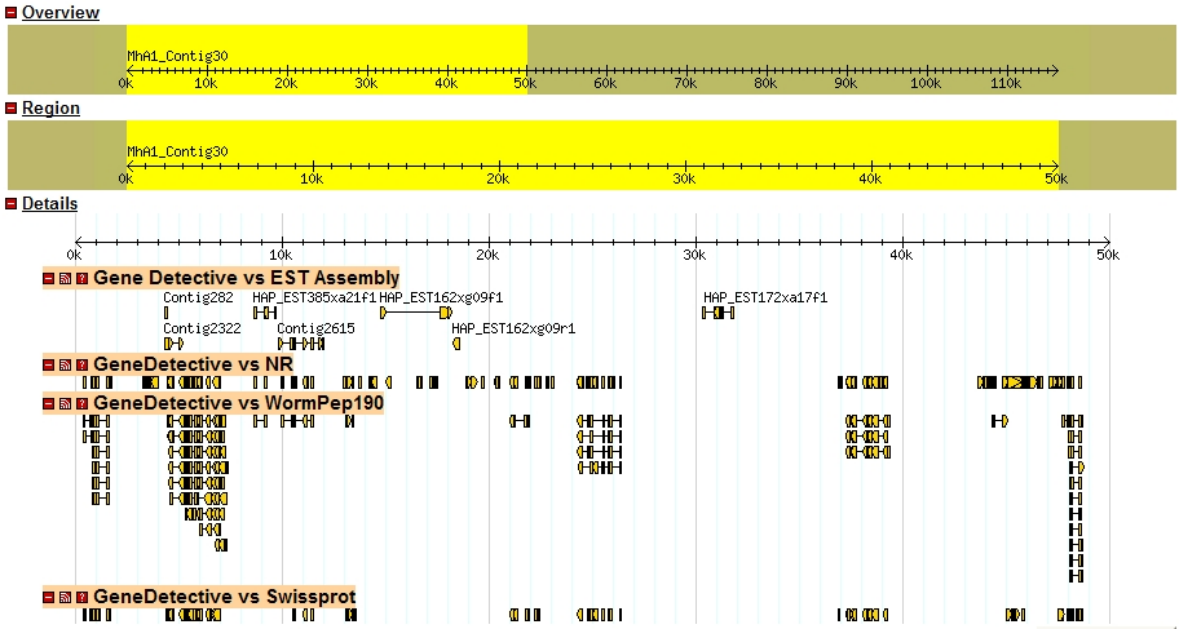

Figure 1. Gbrowse view of MhA1_Contig30:1-50000. 'Overview' and 'Region' places the 'Details' tracks into perspective (yellow shading) in relation to the larger contig context. GeneDetective analyses of EST data and targets from various protein databases are shown under the 'Details' section. Orange regions represent inferred exons of the genomic DNA, while thin black lines represent inferred introns. "GeneDetective vs NR" and "GeneDetective vs WormPep" tracks were generated based on e-value cutoffs of e-15. An older version of WormPep than the version shown was used at the time of manual annotation. "GeneDetective vs Swissprot" was labeled "GeneDetective vs UNIPROT" at the time of manual annotation, and was generated based on an e-value cutoff of e^{-5}.

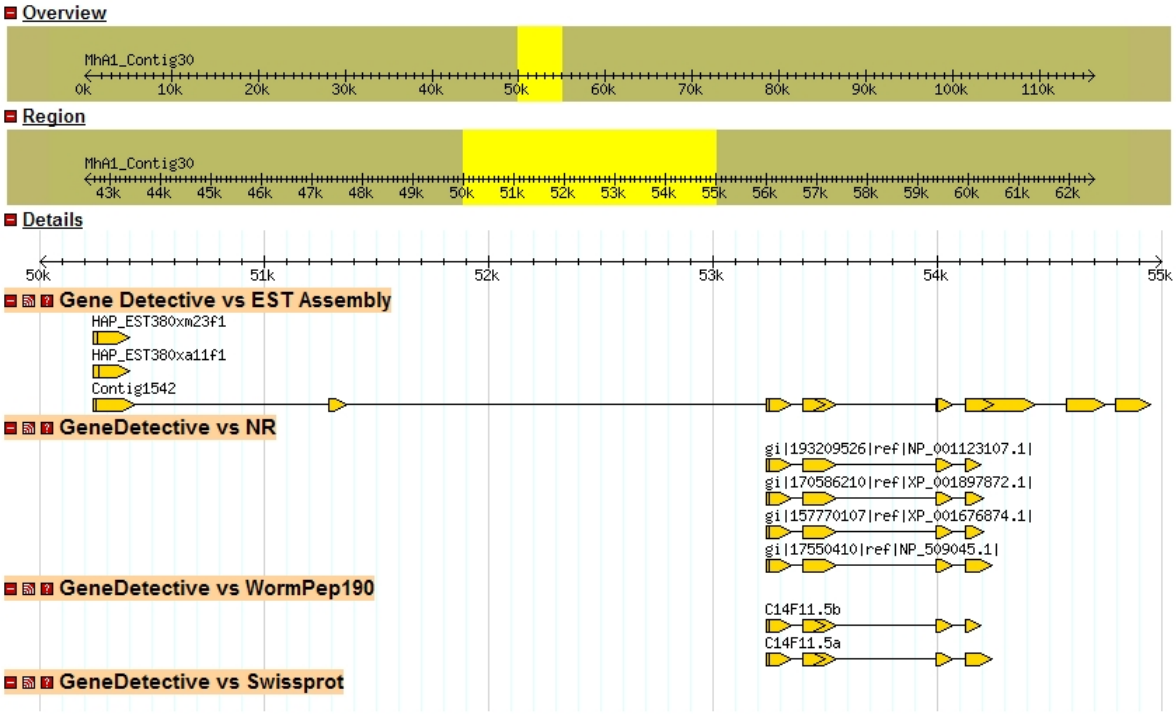

Figure 2. Gbrowse view of MhA1_Contig30:50000:55000. Some inferred exons from EST evidence align with inferred exons from protein databases, suggestive of possible transcribed and translated genomic regions.

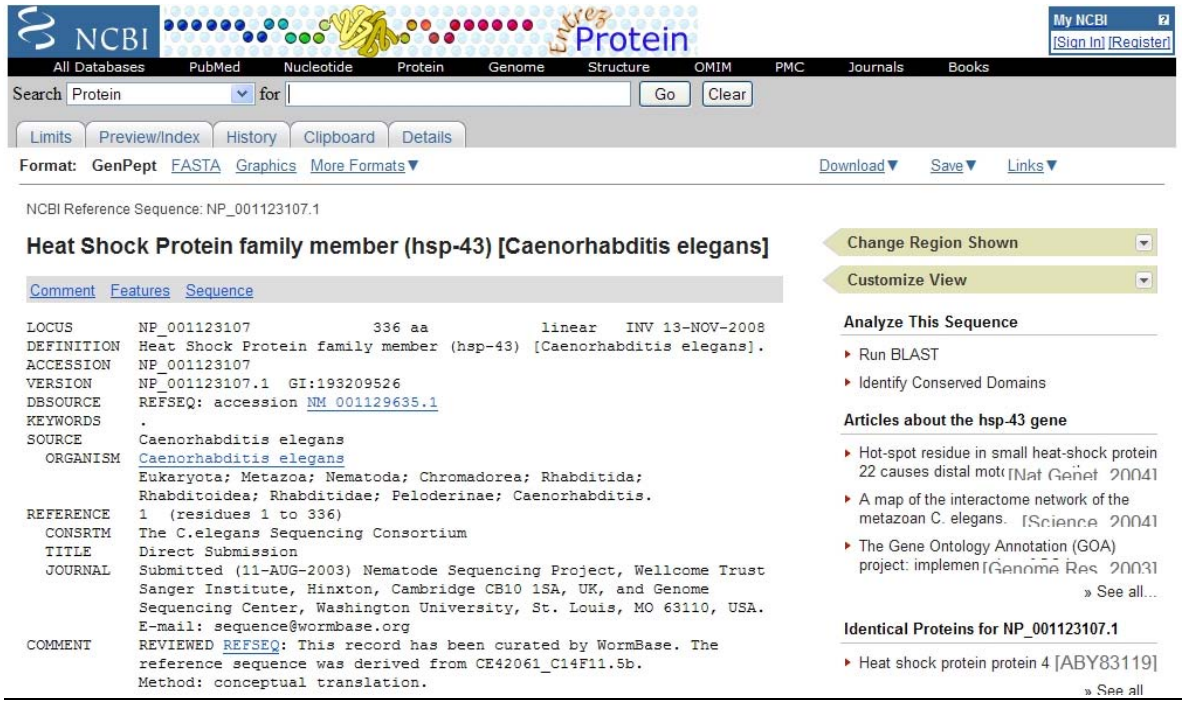

Figure 3. Additional information for the *C. elegans* hsp-43 protein is available at its NCBI page, accessible by clicking on its Gbrowse link. More weight is given to analyses dependent on this protein because it is not explicitly annotated as 'hypothetical', 'putative', or 'predicted'.

```
seq1 = Mhl0g200708_Contig30, 116350 bp
seq2 = Contig1542 (Contig1542), 1315 bp

50231-50423  (28-219)   98% ->
51291-51368  (220-297)   100% ->
53240-53350  (298-408)   100% ->
53400-53552  (409-561)   100% ->
53997-54068  (562-633)   100% ->
54123-54440  (634-951)   100% ->
54573-54750  (952-1129)   100% ->
54797-54959  (1130-1292)   96%
```

Figure 4. SIM4 output indicating the coordinates, in basepairs, of genomic contig 30 corresponding to EST contig 1542, as well as the percent identity between the two sequences and the matching strand direction. Each line corresponds to a different exon of the genomic subsequence. For instance, at genomic contig 30:53240-53350 (the third exon of this genomic contig subsequence), EST contig 1542:298-408 aligns with 100% identity to the reference strand of the genomic contig.

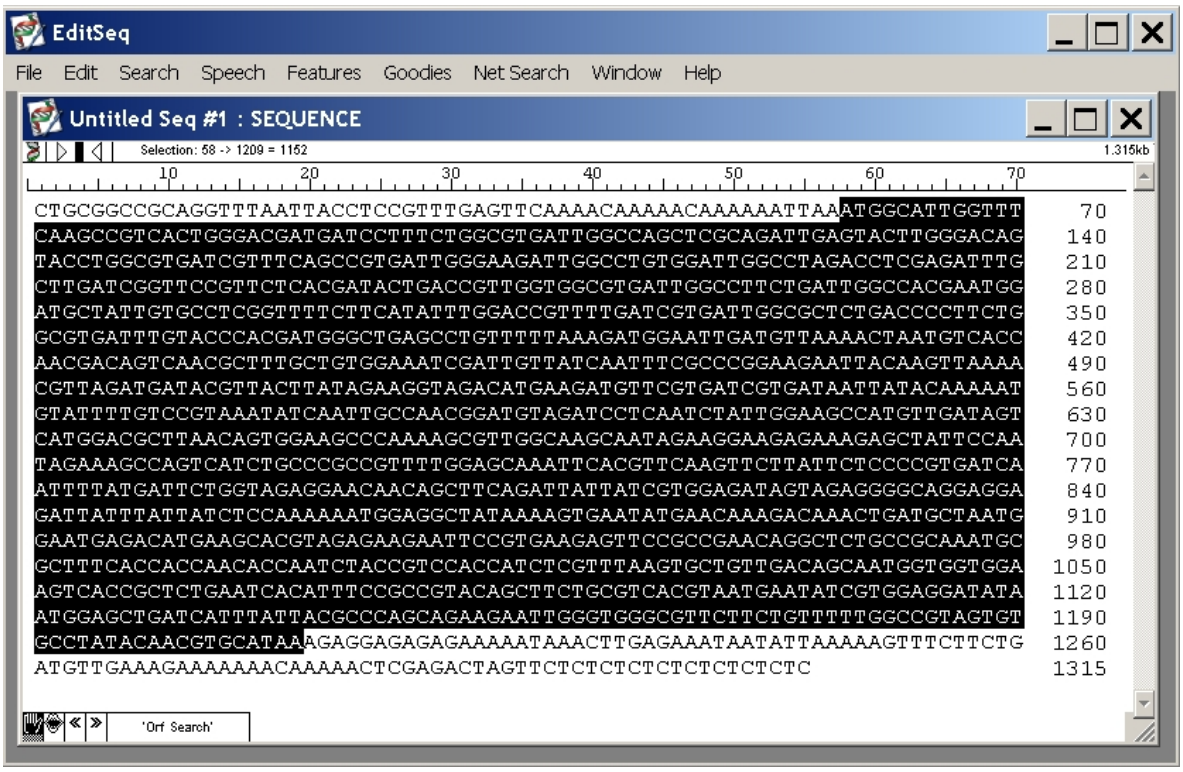

Figure 5. The identification of an open reading frame (ORF) in EST contig 1542 using EditSeq. As expected, the ORF begins with “ATG” and ends with “TAA”. The position of the ORF is indicated as “58-1209”, and its length is determined to be 1152 bp.

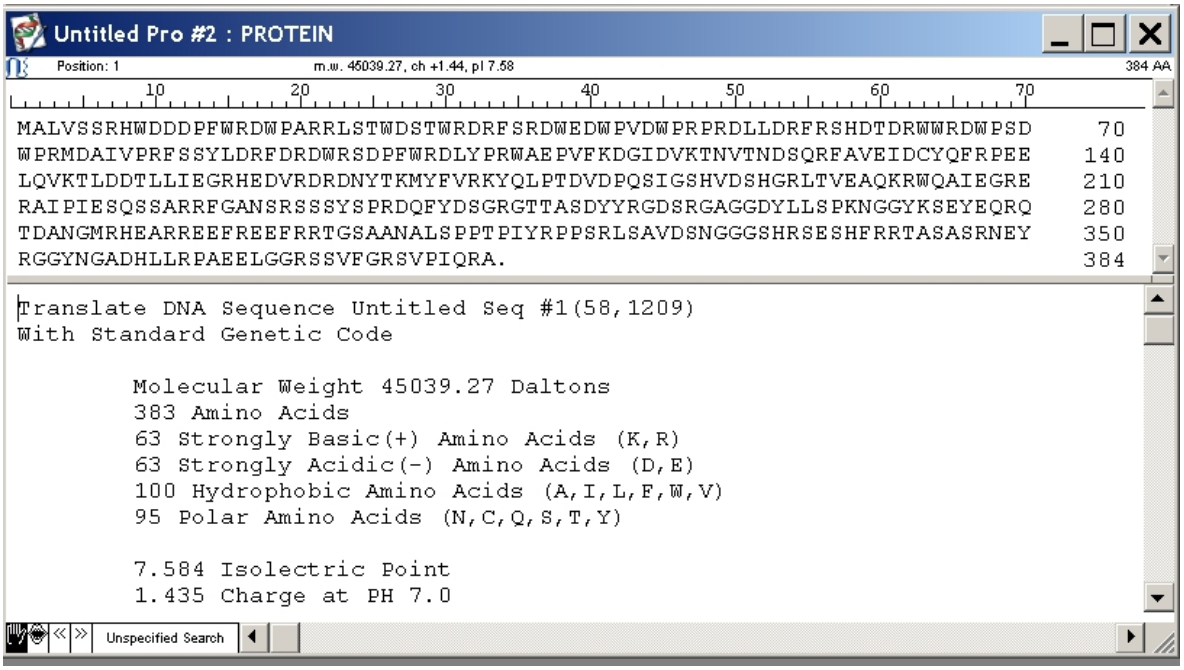

Figure 6. The translated product of the ORF shown in Figure 5. As indicated, this protein is 383 amino acids. Other useful statistics are provided (not all are shown), such as molecular weight, charge characteristics, origin sequence base counts and codon usage, and melting temperature.

Figure 7. Pairwise alignment of the protein derived from the translated ORF of *M. hapla* EST contig 1542 and the *C. elegans* hsp-43 protein. Matching amino acids are shaded black. This alignment serves as one supporting piece of evidence for the prediction of a full-length gene at *M. hapla* genomic contig 30:50231-54959.

Figure 8. Pairwise alignment of the protein derived from the translated ORF of *M. hapla* EST contig 1038 and the *C. elegans* rpl-32 protein. Matching amino acids are shaded black. This alignment serves as one supporting piece of evidence for the prediction of a full-length gene at *M. hapla* genomic contig 20:8152-7100.

Figure 9. Pairwise alignment of the protein derived from the translated ORF of *M. hapla* HAP_EST385xa21f1 and *C. elegans* RNA polymerase III subunit C11. Matching amino acids are shaded black. This alignment serves as one supporting piece of evidence for the prediction of a full-length gene at *M. hapla* genomic contig 30:9682-8603.

Figure 10. Pairwise alignment of the protein derived from the translated ORF of *M. hapla* EST contig 874 and a *C. elegans* signal peptidase-like protein. Matching amino acids are shaded black. This alignment serves as one supporting piece of evidence for the prediction of a full-length gene at *M. hapla* genomic contig 31:40163-41414.

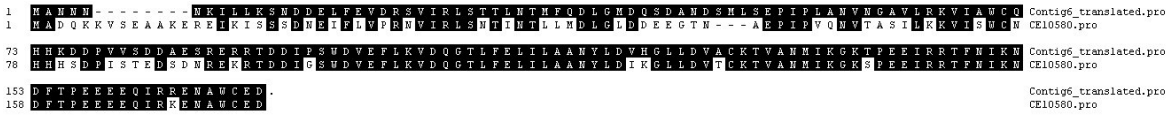

Figure 11. Pairwise alignment of the protein derived from the translated ORF of *M. hapla* EST contig 6 and a *C. elegans* ubiquitin-ligase complex component family member (skr-1). Matching amino acids are shaded black. This alignment serves as one supporting piece of evidence for the prediction of a full-length gene at *M. hapla* genomic contig 53:66263-67429.

Figure 12. Pairwise alignment of the protein derived from the translated ORF of *M. hapla* EST contig 2629 and a *C. elegans* oxygen-binding protein. Matching amino acids are shaded black. This alignment serves as one supporting piece of evidence for the prediction of a full-length gene at *M. hapla* genomic contig 78:35216-36948.

Boxes

1a. TeraBLAST is used to map ESTs, proteins, or Hidden Markov Models to genomic DNA, or
1b. Intron-spanning Hidden Markov Model Framesearch is used to map protein Hidden Markov Models to genomic regions.
2. Hits are used as guides to determine relevant genomic regions.
3. A variant of Smith Waterman (which is intron-tolerant or frameshift tolerant for nucleotide or protein sequences, respectively) is used to realign the sequences in order to generate optimal sequence alignments.
4. Canonical splice consensus sequences are used to identify splice sites.
5. A newly generated multiple sequence alignment guides the detection of splice variants.
6. The gene model is shown in a zoomable graphical display.

Box 1. The methodology underlying TimeLogic GeneDetective. From

http://www.timelogic.com/downloads/GeneDetective_2008.pdf.

Tables

Table 1. Summary of full-length genes found in *M. hapla* genomic contigs 11 through 85 using the manual gene-finding approach.

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	Annotation (based on similarity to putative orthologous gene)
Mh10g200708_Contig18	Contig501	11.5	12.5	zinc-finger protein
Mh10g200708_Contig20	Contig1038	7.1	8.2	large ribosomal subunit protein
Mh10g200708_Contig20	HAP_EST382xo03f1	84.7	85.8	subunit of N alpha-acetyltransferase
Mh10g200708_Contig25	Contig2653	45.8	48.2	methyltransferase
Mh10g200708_Contig25	Contig2567	54.1	55.3	membrane protein
Mh10g200708_Contig30	HAP_EST385xa21f1	8.6	9.7	RNA polymerase III subunit C11
Mh10g200708_Contig30	Contig1542	50.2	54.9	heat shock protein
Mh10g200708_Contig30	Contig2805	80.2	82.8	survival of motor neuron protein interactor family member (smi-1)
Mh10g200708_Contig31	Contig874	40.1	41.5	signal peptidase I
Mh10g200708_Contig42	Contig1544	27.2	28.8	unknown
Mh10g200708_Contig48	Contig351	41.7	44.2	vacuolar H ATPase family member (vha-14)
Mh10g200708_Contig50	HAP_EST320xi15f1	38.8	39.6	ubiquitin-like protein required for embryogenesis and terminal hypodermal differentiation (ned-8)
Mh10g200708_Contig50	Contig857	47.7	48.7	calponin homolog (cpn-3); involved in cytoskeleton
Mh10g200708_Contig53	Contig6	66.2	67.5	ubiquitin ligase complex component (skr-1)
Mh10g200708_Contig53	Contig1122	82.2	82.4	transcription factor activity
Mh10g200708_Contig55	Contig1122	61.4	63.7	transcription factor activity
Mh10g200708_Contig56	HAP_EST328xa21f1	21	21.7	ribosomal protein large subunit family member (rpl-33)
Mh10g200708_Contig56	Contig1122	38.9	39.2	transcription factor activity
Mh10g200708_Contig68	Contig1395	30.7	31.4	ribosomal protein large subunit family member (rpl-39)
Mh10g200708_Contig74	Contig75	25.6	27.6	thioredoxin-like protein; post-translational modifications
Mh10g200708_Contig74	Contig1479	42.9	45	unknown; possibly related to development and aging
Mh10g200708_Contig78	Contig864	6.6	7.5	neuropeptide-like protein family member (nlp-12)
Mh10g200708_Contig78	Contig2452	27.3	28.5	unknown
Mh10g200708_Contig78	Contig2874	28.5	29.8	unknown; integral to membrane
Mh10g200708_Contig78	Contig2629	35.2	37	globin-related (glb-1); oxygen binding

Table 2. More comprehensive data from the manual gene-finding approach applied to *M. hapla* genomic contigs 11 through 85.

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig11	Contig539	57.5	58.5	CE29008 - sqd-1 - (homologous to Drosophila SQD (squid) protein) (WBGene00022235)	heterogeneous nuclear ribonucleoprotein D isoform a [Mus musculus]. (NCBI NP_001070733)	Heterogeneous nuclear ribonucleoprotein D0 (hnRNP D0) (AU-rich element binding protein 1) (NCBI Q14103)	574	57475-57658, 57702-57871, 57917-57959, 57998-58138, 58616-58651	reference	34-574	No
Mh10g200708_Contig18	HAP_EST218xo07f1	8.4	9.6	CE06473 - (T21C9.4 gene) Enhancer of rudimentary [KOG1766]; [OMpre_WH010177] (WBGene00011892)	XP_001119900 PREDICTED: similar to enhancer of rudimentary CG1871-PA, isoform A [Apis mellifera]	Q94554 Protein enhancer of rudimentary	490	8448-8650, 9052-9139, 9184-9279, 9523-9572	complement	78-302	No
Mh10g200708_Contig18	Contig501	11.5	12.5	CE07665 - B0035.1 gene, Zn finger protein [KOG2893] (WBGene00007105)	EAT42753 conserved hypothetical protein [Aedes aegypti]. (yellow fever mosquito)	Q178Y7 Hypothetical protein	670	11563-11707, 11757-11981, 12095-12273, 12361-12481	reference	105-670	Yes
Mh10g200708_Contig18	Contig1006	44.7	45.1	(no info)	(no info)	(no info)					No
Mh10g200708_Contig18	HAP_EST260xm17f1	58.4	59.3	CE36163 - F52B11.2 is orthologous to the human gene PHOSPHOMANNOMUTASE 2 (PMS2; OMIM:601785), which when mutated leads to congenital disorder of glycosylation, type Ia. (WBGene00009925)	BAB29001 unnamed protein product [Mus musculus].	O35621 Phosphomannomutase 1 (PMM 1). Mus musculus (house mouse)	584	58384-58589, 58748-58948, 58996-59118, 59168-59221	reference	69-584	No
Mh10g200708_Contig18	HAP_EST174xc23f1	65.2	66	(no info)	(no info)	(no info)					No
Mh10g200708_Contig20	Contig1038	7.1	8.2	CE03709 - rpl-32 encodes a large ribosomal subunit L32 protein (WBGene00004446)	ABG81994 putative ribosomal protein L32e [Diaphorina citri].	Q1W2A3 Putative ribosomal protein 49 (Graphocephala atropunctata)	621	7100-7259, 7306-7398, 7841-8152	complement	80-484	Yes
Mh10g200708_Contig20	Contig1891	13.8	15.6	CE22580 - The nft-1 gene encodes an ortholog of human FHIT, which when mutated is associated with head and neck cancers (OMIM:601153). (WBGene00003594)	XP_001115539 PREDICTED: similar to nitrilase 1 [Macaca mulatta]. (Rhesus monkey)	O76463 Nitrilase and fragile histidine triad fusion protein NitFhit[Includes: Bis(5'-adenosyl)-triphosphatase (Diadenosine 5',5'''-P1,P3-triphosphate hydrolase) (Dinucleosidetriphosphatase) (AP3A hydrolase) (AP3Aase); Nitrilase homolog] (Caenorhabditis elegans)	822	13799-14132, 14635-14813, 14857-14964, 15200-15322, 15454-15531	complement	27-668	No
Mh10g200708_Contig20	Contig2331	22.8	23.7	CE18363 - crs-1 encodes a predicted cytoplasmic cysteinyl-tRNA synthetase (CysRS), a class I aminoacyl-tRNA synthetase that catalyzes the attachment of cysteine to its cognate tRNA and is thus required for protein biosynthesis; CRS-1 is essential for embryonic and larval development, and is required for molting and a normal rate of postembryonic development (WBGene00000800)	EAT44292 cysteinyl-tRNA synthetase [Aedes aegypti].	Q61X48 Hypothetical protein CBG04128 (Caenorhabditis briggsae)	1046	22880-22987, 23027-23192, 23456-23607	reference	(23-157), 303-737	No
Mh10g200708_Contig20	HAP_EST241xm15r1	48.1	48.8			O44970 Hypothetical protein (C. elegans)					No
Mh10g200708_Contig20	HAP_EST279xg05r1	56.9	57.7	CE01810 - rps-3 encodes a small ribosomal subunit S3 protein that contains a KH RNA-binding domain and by homology,is predicted to function in protein biosynthesis; in C. elegans, RPS-3 activity is required for embryonic and germline development, as well as the overall health of the animal. (WBGene00004472)	BAB27761 unnamed protein product [Mus musculus].	Q53G83 Ribosomal protein S3 variant	423	56945-57065, 57171-57355, 57522-57612	complement	3-191, (222-371, 256-423)	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig20	HAP_EST279xg05f1	57.7	58.6	CE01810 - rps-3 encodes a small ribosomal subunit S3 protein that contains a KH RNA-binding domain and by homology,is predicted to function in protein biosynthesis; in C. elegans, RPS-3 activity is required for embryonic and germline development, as well as the overall health of the animal. (WBGene00004472)	CAF94963 unnamed protein product [Tetraodon nigroviridis].	Q801H1 Ribosomal protein S3 - Scyliorhinus canicula (smaller spotted catshark)	471	57772-57831, 58070-58175, 58221-58394, 58441-58571	complement	49-459	No
Mh10g200708_Contig20	Contig2152	60	60.6	(no info)	(no info)	(no info)					No
Mh10g200708_Contig20	Contig2138	61.4	63.1	CE33830 - T11G6.8 gene Predicted RNA-binding protein (RRM superfamily) (WBGene00011722)	EAA09240 AGAP004509-PA [Anopheles gambiae str. PEST].	Q22412 Hypothetical protein. (C. elegans)	558	61455-61724, 62735-62897, 62945-63069	complement	18-558	No
Mh10g200708_Contig20	HAP_EST104xi21f1	81.5	82.5	CE18673 - F38E11.5 encodes a beta' (beta-prime) subunit of the coatomer (COPI) complex; in mass RNAi assays, F38E11.5 is required for fertility and general health. (WBGene00009542)	NP_068533 coatomer protein complex, subunit beta 2 (beta prime) [Rattus norvegicus].	P35605 Coatomer subunit beta' (Beta'-coat protein) (Beta'-COP) (p102).Bos taurus (cattle)	534	81545-81710, 82080-82327, 82369-82488	reference	70-534	No
Mh10g200708_Contig20	HAP_EST382xo03f1	84.7	85.8	CE18023 - K07H8.3 gene Subunit of the major N alpha-acetyltransferase [KOG3235]; (WBGene00019505)	XP_001120648 PREDICTED: similar to Ard1 CG11989-PA, isoform A [Apis mellifera]. Apis mellifera (honey bee)	O61219 Hypothetical protein. (C. elegans)	624	84751-84909, 84960-85156, 85341-85491, 85610-85713	reference	68-574	Yes
Mh10g200708_Contig20	HAP_EST121xg19f1	109.2	110.4	CE05163 - B0035.5 is orthologous to the human gene GLUCOSE-6-PHOSPHATE DEHYDROGENASE (G6PD; OMIM:305900), which when mutated leads to disease. (WBGene00007108)	NP_502129 B0035.5 [Caenorhabditis elegans].	Q27464 Glucose-6-phosphate 1-dehydrogenase (G6PD).	545	109268-109294, 109340-109440, 109817-109889, 109935-110141, 110188-110324	complement	(292-372), 399-545	No
Mh10g200708_Contig20	Contig594	113.1	114.5	CE06362 - T05E11.3 gene Endoplasmic reticulum glucose-regulated protein (GRP94/endoplasmin), HSP90 family [KOG0020]; (WBGene00011480)	EAT34979 endoplasmin [Aedes aegypti].	Q16KZ2 Endoplasmin. Aedes aegypti (Stegomyia aegypti)	1026	113147-113207, 113250-113474, 113517-113633, 113676-114161, 114332-114460	complement	(91-198[rc]), 384-1026[rc]	No
Mh10g200708_Contig20	HAP_EST235xg21f1	114.5	115.4	CE06362 - T05E11.3 gene Endoplasmic reticulum glucose-regulated protein (GRP94/endoplasmin), HSP90 family [KOG0020]; (WBGene00011480)	NP_999808 heat shock protein gp96 [Strongylocentrotus purpuratus].	Q868Z7 Heat shock protein gp96. (Strongylocentrotus purpuratus)	592	114489-114588, 114724-114789, 114835-115002, 115051-115268, 115312-115351	complement	92-271, (268-592)	No
Mh10g200708_Contig20	Contig2020	120	121	CE34140 - vps-29 - (related to yeast Vacuolar Protein Sorting factor) via paper evidence: Prasad BC and Clark SG(WBGene00014234)	CAB66549 hypothetical protein [Homo sapiens].	Q3T0M0 Similar to Vacuolar protein sorting 29 (Vesicle protein sorting 29). Bos taurus (cattle)	736	119980-120177, 120227-120284, 120333-120353, 120395-120455, 120498-120622, 120657-120718, 120761-120861, 120908-121018	complement	(109-246), 213-677	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig20	HAP_EST335xe03f1	153.4	154.2	CE07853 - tax-6 - (abnormal CHEmotaxis) via person evidence: David Dusenbery - tax-6 encodes an ortholog of calcineurin A that is required for inhibition and adaptation of several sensory neurons and for the normal regulation of egg-laying by serotonin; TAX-6 stimulates transcription of rcn-1, and binds RCN-1 protein if free Ca[2+] is present; in general, calcineurin positively regulates calcium-dependent signaling. (WBGene00006527)	NP_523373 Protein phosphatase 2B at 14D CG9842-PA [Drosophila melanogaster].	Q27889 Serine/threonine-protein phosphatase 2B catalytic subunit 2, (Calmodulin-dependent calcineurin A2 subunit). Drosophila melanogaster (fruit fly)	634	153437-153862, 153917-154060, 154098-154127	reference	285-634	No
Mh10g200708_Contig25	Contig709	9	10.1	CE31178 NP_508526 TWiK family of potassium channels family member (twk-16) [Caenorhabditis elegans].GI:25147273 WBGene00006670	NP_508526 TWiK family of potassium channels family member (twk-16) [Caenorhabditis elegans].GI:25147273 WBGene00006670	Q20673 Twik family of potassium channels protein 16 (Potassium channel subunit n2P16). GI:74964761 WBGene00006670	1143	8990-9419, 9463-9732, 9769-9810, 9855-10003	complement	(149-226[rc], 299-409[rc], 609-758[rc]), 795-1010[rc]	No
Mh10g200708_Contig25	Contig11	9.3	10.1	CE31178 NP_508526 TWiK family of potassium channels family member (twk-16) [Caenorhabditis elegans]. WBGene00006670	NP_508526 TWiK family of potassium channels family member (twk-16) [Caenorhabditis elegans]. WBGene00006670	Q20673 Twik family of potassium channels protein 16 (Potassium channel subunit n2P16). GI:74964761 WBGene00006670	583	9329-9419, 9463-9810, 9855-9998	complement	(46-195[rc]), 232-483[rc]	No
Mh10g200708_Contig25	Contig2077	20.9	21.6	CE16895 NP_492798 C34B2.7 [Caenorhabditis elegans]. WBGene00016392	AAH11301 Sdha protein [Mus musculus]. BC011301.1	Q09508 Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial precursor (FP) (Flavoprotein subunit of complex II). C. elegans. gi:22096345	525	20893-21063, 21107-21220, 21266-21460, 21504-21548	complement	135-362	No
Mh10g200708_Contig25	Contig1957	22.6	22.7	NP_508964 Regulator of G protein Signaling family member (rgs-6) [Caenorhabditis elegans] WBGene00004349	NP_508964 Regulator of G protein Signaling family member (rgs-6) [Caenorhabditis elegans] WBGene00004349	Q18563 Regulator of G-protein signaling rgs-6.	560	22605-22704	reference	15-98, (343-423, 454-560)	No
Mh10g200708_Contig25	HAP_EST137xg17f1	23.2	24.1	NP_509446 Succinate DeHydrogenase complex subunit A family member (sdha-1) [Caenorhabditis elegans]. WBGene00015391	AAB34901 succinate-ubiquinone oxidoreductase; fumarate reductase [Dirofilaria immitis]. (dog heartworm nematode)	Q36744 Succinate-ubiquinone oxidoreductase Dirofilaria immitis	577	23276-23363, 23408-23480, 23559-23608, 23672-23725, 23767-24077	complement	198-527	No
Mh10g200708_Contig25	Contig2657	24.6	24.9	(no info)	(no info)	(no info)					No
Mh10g200708_Contig25	Contig1985	31.4	33.8	(no info)	(no info)	(no info)					No
Mh10g200708_Contig25	Contig129	42.9	44	(no info)	(no info)	(no info)					No
Mh10g200708_Contig25	HAP_EST111xk01f1	43.2	44.1	(no info)	(no info)	(no info)					No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig25	HAP_EST117xi07f1	44.8	45.5	(no info)	(no info)	(no info)					No
Mh10g200708_Contig25	Contig2653	45.8	48.2	CE38615 NP_001022224 H20J04.9 [Caenorhabditis elegans]. WBGene00044310	XP_971597 PREDICTED: similar to CG31322-PA [Tribolium castaneum]. XM_966504.1	Q0V8Q7 Methionine-tRNA synthetase 2 Bos taurus (cattle)	1714	45864-46008, 46047-46118, 46233-46341, 46433-46608, 46651-46725, 46819-47010, 47085-47566, 47615-47916, 47964-48092, 48138-48169	reference	92-1030	Yes
Mh10g200708_Contig25	HAP_EST148xc13f1	52	52.8	(no info)	(no info)	(no info)					No
Mh10g200708_Contig25	HAP_EST148xc13r1	53.2	53.5	(no info)	(no info)	(no info)					No
Mh10g200708_Contig25	Contig2567	54.1	55.3	(no info)	(no info)	CE02743 - Q20516 UPF0057 membrane protein F47B7.1 Caenorhabditis elegans GI:3025316	1027	54171-54729, 54796-54835, 54880-54907, 54952-55325	reference	506-685	Yes
Mh10g200708_Contig25	HAP_EST173xc09f1	56.2	57.3	(no info)	(no info)	(no info)					No
Mh10g200708_Contig26	HAP_EST146xi01f1	7.6	8.4	(no info)	(no info)	Q22787 Hypothetical protein WBGene00020810					No
Mh10g200708_Contig26	Contig1998	8.4	8.6	CE36522 NP_498913 FLI-I (Drosophila flightless) homolog family member (fli-1) [Caenorhabditis elegans] WBGene00001443	XP_790260 PREDICTED: similar to flightless I homolog variant [Strongylocentrotus purpuratus].	P34268 Protein flightless-1 homolog. Caenorhabditis elegans	803	8468-8600	unknown	(40-189), 156-623	No
Mh10g200708_Contig26	Contig1781	13.1	14.2	CE37417 NP_510020 W03G11.3 [Caenorhabditis elegans]. WBGene00012225	NP_997905 fucosidase, alpha-L-1, tissue [Danio rerio].	Q7ZW80 Fucosidase, alpha-L-1, tissue.Danio rerio (zebrafish)	694	13120-13194, 13236-13382, 13424-13505, 13555-13616, 13660-13758, 13852-13965, 14012-14119	complement	31-672	No
Mh10g200708_Contig26	Contig1905	18.5	19.3	CE05684 NP_001021405 asparagiNyl tRNA Synthetase family member (nrs-1) [Caenorhabditis elegans]. WBGene00003815	CAF88671 unnamed protein product [Tetraodon nigroviridis].	Q4TEL1 Chromosome undetermined SCAF5294, whole genome shotgun sequence.	687	18568-18947, 19003-19309	complement	(4-111), 179-667	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig26	HAP_EST124xi19r1	18.9	19.4	CE05684 NP_001021405 asparagiNyl tRNA Synthetase family member (nrs-1) [Caenorhabditis elegans]. WBGene00003815	EAT41839 aspartyl-tRNA synthetase [Aedes aegypti].	Q175Q3 Aspartyl-tRNA synthetase. Aedes aegypti (Stegomyia aegypti)	450	18961-19167, 19208-19311, 19357-19495	complement	145-447	No
Mh10g200708_Contig26	Contig753	28.5	30.9	CE21417 NP_001021722 FK506-Binding protein family member (fkb-2) [Caenorhabditis elegans] WBGene00001427	NP_001040382 FK506-binding protein [Bombyx mori].	P48375 12 kDa FK506-binding protein (FKBP) (Peptidyl-prolyl cis-trans isomerase) (PPIase) (Rotamase) (Macrolide-binding protein).	467	28559-28787, 28836-28910, 28960-29063, 30885-30944	complement	2-292	No
Mh10g200708_Contig26	Contig706	54.1	62.9	(no info)	(no info)	CE18705 O02655 S-adenosylmethionine decarboxylase proenzyme (AdoMetDC) (SamDC) [Contains: S-adenosylmethionine decarboxylase alpha chain; S-adenosylmethionine decarboxylase beta chain].	418	54185-54239, 54955-55086, 55256-55392, 62786-62879	reference	(3-95, 92-187), 144-323	No
Mh10g200708_Contig26	Contig624	55.7	56.4	CE18705 NP_493448 SAM Decarboxylase family member (smd-1) [Caenorhabditis elegans]. WBGene00004875	P79888 S-adenosylmethionine decarboxylase proenzyme (AdoMetDC) (SamDC) [Contains: S-adenosylmethionine decarboxylase alpha chain; S-adenosylmethionine decarboxylase beta chain].	P79888 S-adenosylmethionine decarboxylase proenzyme (AdoMetDC) (SamDC) [Contains: S-adenosylmethionine decarboxylase alpha chain; S-adenosylmethionine decarboxylase beta chain]. Xenopus laevis (African clawed frog)	582	55705-55887, 55941-56336	complement	125-582	No
Mh10g200708_Contig29	Contig2255	8.2	8.6	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	Contig2247	10.8	11.7	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	HAP_EST281xo01f1	20.2	20.4	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	Contig790	30.1	30.9	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	HAP_EST107xo07f1	31.5	32.1	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	Contig2712	32.3	32.7	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	Contig489	32.7	33.4	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	Contig1601	34.7	35	(no info)	(no info)	(no info)					No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig29	Contig129	38.3	39.4	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	HAP_EST111xk01f1	38.6	39.5	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	HAP_EST352xe03f1	52.5	53.5	CE04392 NP_508943 UNCoordinated family member (unc-97) [Caenorhabditis elegans]. WBGene00006826	XP_001129783 PREDICTED: similar to LIM and senescent cell antigen-like domains 1 [Homo sapiens].	Q1L0R5 UNC-97-like protein. Heterodera glycines (soybean cyst nematode)	673	52523-52986, 53229-53341, 53411-53493	reference	124-673	No
Mh10g200708_Contig29	Contig1721	65.4	66.8	NP_492210 K04G2.2 [Caenorhabditis elegans]. WBGene00010561	EAT45797 conserved hypothetical protein [Aedes aegypti].	Q7PQX3 ENSANGP00000021371 Anopheles gambiae str. PEST	826	65433-65584, 65766-65969, 66015-66070, 66329-66733	complement	150-826	No
Mh10g200708_Contig29	HAP_EST154xk17f2	66.9	67.9	(no info)	(no info)	Q5ZJH2 Nicalin precursor (Nicastrin-like protein). GI:67460780 Gallus gallus (chicken)					No
Mh10g200708_Contig29	HAP_EST154xk17r1	70.4	70.7	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	Contig1118	73.2	75.1	(no info)	(no info)	(no info)					No
Mh10g200708_Contig29	HAP_EST171xa11f1	83.7	86.4	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	Contig282	4.2	4.3	CE02342 NP_496236 T09F3.2 [Caenorhabditis elegans]. WBGene00011662	NP_608615 CG18317-PA [Drosophila melanogaster].	Q9VQ37 CG18317-PA (LP02521p).	774	4296-4423, 86808-86827	complement	337-606[rc]	No
Mh10g200708_Contig30	Contig2322	4.2	5.2	CE08110 NP_492485 Heat Shock Protein family member (hsp-70) [Caenorhabditis elegans]. WBGene00002026	Q3S4T7 78 kDa glucose-regulated protein precursor (GRP 78) (Heat shock 70 kDa protein 5). Spermophilus tridecemlineatus (thirteen-lined ground squirrel)	Q90593 78 kDa glucose-regulated protein precursor (GRP 78) (Heat shock 70 kDa protein 5) (Immunoglobulin heavy chain-binding protein) (BiP).Gallus gallus (chicken)	554	4296-4705, 5016-5159	complement	(2-202[rc]), 40-420[rc]	No
Mh10g200708_Contig30	HAP_EST385xa21f1	8.6	9.7	CE25620 NP_500076 Y77E11A.6 [Caenorhabditis elegans]. WBGene00022309	EAT47484 DNA-directed RNA polymerases III 12.5 kDa polypeptide, putative [Aedes aegypti].	Q17L91 DNA-directed RNA polymerases III 12.5 kDa polypeptide, putative.	395	8603-8731, 9143-9327, 9623-9682	complement	68-388	Yes
Mh10g200708_Contig30	Contig2615	9.7	12	CE23280 NP_507876 DeCapping Scavenger enzyme homolog family member (dcs-1) [Caenorhabditis elegans]. WBGene00000940	AAP47149 histidine triad protein member 5 [Danio rerio].	Q504K8 MRNA decapping enzyme.	908	9796-10026, 10408-10494, 10537-10654, 11009-11178, 11352-11503, 11778-11910, 11956-11976	complement	126-791[rc]	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig30	HAP_EST162xg09f1	17.5	18.2	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	HAP_EST162xg09r1	18.2	18.6	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	HAP_EST172xa17f1	30.3	31.9	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	Contig1542	50.2	54.9	CE03986 NP_509045 Heat Shock Protein family member (hsp-43) [Caenorhabditis elegans]. WBGene00002024	NP_509045 Heat Shock Protein family member (hsp-43) [Caenorhabditis elegans]. WBGene00002024	Q17992 Heat shock protein protein 43.	1315	50231-50423, 51291-51368, 53240-53350, 53400-53552, 53997-54068, 54123-54440, 54573-54750, 54797-54959	reference	58-1209	Yes
Mh10g200708_Contig30	HAP_EST380xa11f1	50.2	50.5	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	HAP_EST380xm23f1	50.2	50.5	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	HAP_EST161xm21r1	56	58	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	HAP_EST161xm21f1	59	67.4	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	Contig1	60.3	60.6	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	HAP_EST242xg05f1	62.4	64	CE07510 NP_505309 Germinal Center Kinase family member (gck-1) [Caenorhabditis elegans]. WBGene00001526	XP_001095834 PREDICTED: serine/threonine kinase 3 (STE20 homolog, yeast) [Macaca mulatta].	Q9NB31 Serine/threonine-protein kinase cst-1 (STE20-like kinase 1) (STE20-like kinase MST) (cMST).	578	62444-62493, 62883-63042, 63090-63201, 63716-63839, 63884-63927	reference	145-578	No
Mh10g200708_Contig30	HAP_EST151xa13f2	68.5	69.2	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	Contig2750	74	74.6	(no info)	(no info)	(no info)					No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig30	Contig2631	74.4	75.5	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	HAP_EST164xo03r1	78.5	78.8	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	Contig2805	80.2	82.8	CE32964 NP_001022847 SMN (survival of motor neuron) protein Interactor family member (smi-1) [Caenorhabditis elegans]. WBGene00004886	AAI22059 LOC100124782 protein [Xenopus tropicalis].	Q5ZJH1 Hypothetical protein.	857	80204-80279, 80571-80764, 81070-81099, 81149-81301, 81637-81723, 81769-81939, 82601-82746	complement	6-821	Yes
Mh10g200708_Contig30	HAP_EST102xe01f1	94.1	94.7	(no info)	(no info)	(no info)					No
Mh10g200708_Contig30	Contig904	96.3	98	CE39282 NP_491666 T19B4.1 [Caenorhabditis elegans]. WBGene00020556	AAO92288 peptidylglycine alpha-hydroxylating monooxygenase [Heterodera glycines].	Q86D91 Peptidylglycine alpha-hydroxylating monooxygenase.	1053	96308-96511, 96562-96649, 96908-97026, 97082-97267, 97400-97566, 97614-97667, 97715-97949	reference	123-680	No
Mh10g200708_Contig30	HAP_EST102xe01r1	96.4	98.1	(no info)	(no info)	Q95XM2 Probable peptidylglycine alpha-hydroxylating monooxygenase Y71G12B.4 precursor (PHM).					No
Mh10g200708_Contig30	Contig427	106.9	107.9	NP_509997 F55G7.2 [Caenorhabditis elegans]. WBGene00010120	XP_394544 PREDICTED: similar to DOT1-like, histone H3 methyltransferase [Apis mellifera].	Q8TEK3 Histone-lysine N-methyltransferase, H3 lysine-79 specific (Histone H3-K79 methyltransferase) (H3-K79-HMTase) (DOT1-like protein).	551	106939-106994, 107037-107293, 107661-107767, 107818-107850	reference	159-551[rc]	No
Mh10g200708_Contig31	HAP_EST116xi11f1	17.5	18.7	CE18958 NP_492638 Eukaryotic Initiation Factor family member (eif-3.C) [Caenorhabditis elegans]. WBGene00001226	XP_787192 PREDICTED: similar to Eukaryotic translation initiation factor 3, subunit 8 [Strongylocentrotus purpuratus].	O02328 Probable eukaryotic translation initiation factor 3 subunit 8 (eIF3 p110) (eIF3c). Caenorhabditis elegans	537	17558-17699, 18041-18250, 18358-18522, 18622-18641	complement	176-537 [rc]	No
Mh10g200708_Contig31	HAP_EST358xc11f1	20.9	22	(no info)	(no info)	(no info)					No
Mh10g200708_Contig31	HAP_EST247xa19f1	26.2	27.7	CE29602 NP_492433 HomoGentisate Oxidase family member (hgo-1) [Caenorhabditis elegans]. WBGene00001843	YP_631004 homogentisate 1,2-dioxygenase [Myxococcus xanthus DK 1622]. Myxococcus xanthus DK 1622	Q48GS7 Homogentisate 1,2-dioxygenase (Homogentisicase) (Homogentisate oxygenase) (Homogentisic acid oxidase). Pseudomonas syringae pv. phaseolicola 1448A	595	26298-26319, 26387-26462, 26508-26560, 26718-26826, 26873-26977, 27165-27291, 27576-27677	reference	3-595	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig31	HAP_EST246xo05r1	27.5	28.4	CE29602 NP_492433 HomoGentisate Oxidase family member (hgo-1) [Caenorhabditis elegans]. WBGene00001843	XP_001122563 PREDICTED: similar to homogentisate 1,2-dioxygenase CG4779-PA, partial [Apis mellifera]. Apis mellifera (honey bee)	Q9Y041 Homogentisate 1,2-dioxygenase (Homogentisicase) (Homogentisate oxygenase) (Homogentisic acid oxidase). Caenorhabditis elegans	349	27578-27757, 28238-28406	reference	93-338 [rc]	No
Mh10g200708_Contig31	Contig874	40.1	41.5	NP_491092 Y54E10BR.5 [Caenorhabditis elegans]. WBGene00021844	AAH57885 Sec11a protein [Mus musculus].	Q1LVK8 Novel protein similar to vertebrate signal peptidase complex (18kD) (SPC18).	695	40163-40307, 40562-40806, 41059-41154, 41204-41414	reference	26-571	Yes
Mh10g200708_Contig31	HAP_EST124xc11r1	41.5	41.6	(no info)	(no info)	(no info)					No
Mh10g200708_Contig32	HAP_EST169xo07f1	11	12	CE22460 NP_491085 Y54E10BL.5 [Caenorhabditis elegans] WBGene00021839	NP_491085 Y54E10BL.5 [Caenorhabditis elegans].	Q61WZ8 Hypothetical protein CBG04188.	507	11003-11125, 11173-11306, 11719-11968	reference	104-454[rc]	No
Mh10g200708_Contig32	HAP_EST159xe07f1	25.1	25.8	CE25674 NP_741758 C.Elegans Homeobox family member (ceh-18) [Caenorhabditis elegans]. WBGene00000441	CAC34943 Oct-1L [Mus musculus].	P14859 POU domain, class 2, transcription factor 1 (Octamer-binding transcription factor 1) (Oct-1) (OTF-1) (NF-A1).	538	25198-25437, 25489-25786	complement	230-364	No
Mh10g200708_Contig32	HAP_EST205xc23f1	29.9	30.6	(no info)	(no info)	(no info)					No
Mh10g200708_Contig32	HAP_EST385xo21f1	30.8	31	(no info)	(no info)	(no info)					No
Mh10g200708_Contig33	HAP_EST230xi05f1	13.7	15.2	(no info)	(no info)	(no info)					No
Mh10g200708_Contig33	HAP_EST230xi05r1	15.1	15.6	(no info)	(no info)	(no info)					No
Mh10g200708_Contig33	Contig76	20.5	21.2	(no info)	(no info)	(no info)					No
Mh10g200708_Contig33	Contig2435	31.9	32.9	(no info)	(no info)	(no info)					No
Mh10g200708_Contig34	HAP_EST249xi01r1	17.1	17.7	(no info)	(no info)	(no info)					No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig34	HAP_EST261xk01f2	17.6	27.8	(no info)	(no info)	(no info)					No
Mh10g200708_Contig34	Contig1717	17.9	18.2	(no info)	(no info)	(no info)					No
Mh10g200708_Contig34	HAP_EST332xo03f1	97.1	97.8	(no info)	(no info)	(no info)					No
Mh10g200708_Contig34	HAP_EST267xk23r1	97.8	98.4	CE23635 NP_509703 ERD (yeast endoplasmic reticulum retention defective) related family member (erd-2) [Caenorhabditis elegans]. WBGene00001331	NP_509703 ERD (yeast endoplasmic reticulum retention defective) related family member (erd-2) [Caenorhabditis elegans]. WBGene00001331	P48583 ER lumen protein retaining receptor.	384	97838-97963, 98009-98108, 98157-98207, 98249-98355	complement	35-283[rc]	No
Mh10g200708_Contig34	HAP_EST338xi09f1	99.9	101.1	CE26820 NP_510459 EPsiN (endocytic protein) homolog family member (epn-1) [Caenorhabditis elegans]. WBGene00001329	XP_397121 PREDICTED: similar to liquid facets CG8532-PA, isoform A [Apis mellifera].	Q2M123 GA21142-PA.	660	99951-100058, 100100-100232, 100276-100377, 100753-101071	reference	21-660	No
Mh10g200708_Contig36	Contig1593	9.3	9.6	(no info)	(no info)	(no info)					No
Mh10g200708_Contig36	Contig1818	62.7	63	CE34454 NP_498926 ZK370.4a [Caenorhabditis elegans]. WBGene00022718	NP_498926 ZK370.4a [Caenorhabditis elegans]. WBGene00022718	Q627I7 Hypothetical protein CBG00653.	915	53429-53690	complement	262-915	No
Mh10g200708_Contig36	Contig715	91.1	94	(no info)	(no info)	(no info)					No
Mh10g200708_Contig41	Contig1092	27	28.5	(no info)	(no info)	O61973 Cystatin-like protease inhibitor protein 1, isoform a (Cystatin C). GI:74960064					No
Mh10g200708_Contig41	HAP_EST139xi19f1	32.9	33.1	CE31374 NP_001021762 Y47G6A.19a [Caenorhabditis elegans]. WBGene00021645	EAT32315 zinc carboxypeptidase [Aedes aegypti].	Q19121 Hypothetical protein.	545	32967-33100	complement	43-363[rc]	No
Mh10g200708_Contig41	Contig574	39.1	40.1	CE40163 NP_506558 Y49A3A.1 [Caenorhabditis elegans]. WBGene00013024	NP_506558 Y49A3A.1 [Caenorhabditis elegans]. WBGene00013024	Q16NW6 Ethanolaminephosphotransferase.	512	39162-39482, 39548-39595, 39634-39714, 39975-40036	complement	36-359[rc]	No
Mh10g200708_Contig41	HAP_EST261xo17r1	39.2	40.1	CE40163 NP_506558 Y49A3A.1 [Caenorhabditis elegans]. WBGene00013024	NP_506558 Y49A3A.1 [Caenorhabditis elegans]. WBGene00013024	Q28ZT2 GA19298-PA. Drosophila pseudoobscura	476	39220-39482, 39548-39576, 39634-39714, 39975-40077	complement	5-205[rc]	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig41	Contig33	40.4	42.1	CE40163 NP_506558 Y49A3A.1 [Caenorhabditis elegans]. WBGene00013024	NP_506558 Y49A3A.1 [Caenorhabditis elegans]. WBGene00013024	Q1NZ15 Hypothetical protein.	915	40456-40523, 40565-40649, 40700-40802, 41019-41057, 41116-41263, 41309-41368, 41520-41589, 41637-41685, 41732-41825, 41868-42065	complement	29-915	No
Mh10g200708_Contig41	HAP_EST114xk17f1	43	43.2	(no info)	(no info)	(no info)					No
Mh10g200708_Contig41	HAP_EST272xe21r1	43.3	43.9	CE01431 NP_498109 anNEXin family member (nex-1) [Caenorhabditis elegans]. WBGene00003588	NP_001036881 annexin B13 [Bombyx mori].	Q8WPG9 Annexin B13b.	362	43335-43454, 43574-43693, 43759-43880	complement	90-281[rc]	No
Mh10g200708_Contig41	HAP_EST255xm19f2	43.8	45	CE00661 NP_497903 C28A5.3 [Caenorhabditis elegans]. WBGene00003590	NP_569100 annexin A7 [Rattus norvegicus].	Q6IRJ7 Annexin A7.	657	43816-43876, 43915-43979, 44024-44125, 44172-44358, 44427-44458, 44506-44652, 44871-44909	complement	(86-229), 248-657	No
Mh10g200708_Contig42	HAP_EST108xm19f1	11	11.9	(no info)	(no info)	Q5DX48 Sensory axon guidance protein 2, isoform a. Caenorhabditis elegans					No
Mh10g200708_Contig42	Contig1544	27.2	28.8	CE00838 NP_498608 B0361.9 [Caenorhabditis elegans]. WBGene00015163	NP_498608 B0361.9 [Caenorhabditis elegans]. Caenorhabditis elegans WBGene00015163	Q616B1 Hypothetical protein CBG15350.	844	27210-27355, 27693-28032, 28411-28574, 28618-28789	complement	53-817	Yes
Mh10g200708_Contig42	Contig365	47.9	49.7	CE18512 NP_499530 C18D11.1 [Caenorhabditis elegans]. WBGene00007679	NP_499530 C18D11.1 [Caenorhabditis elegans]. WBGene00007679	Q61BM9 Hypothetical protein CBG13296. Caenorhabditis briggsae	1065	47924-48359, 48507-48732, 49238-49476, 49518-49681	reference	347-1018	No
Mh10g200708_Contig42	Contig1001	70.8	74.6	CE31867 NP_498245 related to Islet cell Diabetes Autoantigen family member (ida-1) [Caenorhabditis elegans]. WBGene00002048	NP_498245 related to Islet cell Diabetes Autoantigen family member (ida-1) [Caenorhabditis elegans]. WBGene00002048	Q09962 Related to islet cell diabetes autoantigen protein 1 (IDA-1 protein) (IA2).	690	70894-71072, 73683-73993, 74366-74565	reference	41-690	No
Mh10g200708_Contig42	Contig366	76.1	78.1	CE31867 NP_498245 related to Islet cell Diabetes Autoantigen family member (ida-1) [Caenorhabditis elegans]. WBGene00002048	EAA12380 AGAP008127-PA [Anopheles gambiae str. PEST].	Q7Q404 ENSANGP00000010449	1292	76195-76578, 76626-76732, 76955-77116, 77282-77448, 77493-77551, 77601-77660, 77752-78076	reference	199-1083	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig42	HAP_EST115xg17f1	89.6	90.6	CE00205 NP_499032 INtegrin Alpha family member (ina-1) [Caenorhabditis elegans]. WBGene00002081	NP_499032 INtegrin Alpha family member (ina-1) [Caenorhabditis elegans]. WBGene00002081	Q03600 Integrin alpha ina-1 precursor.	571	89598-89745, 90021-90261, 90363-90543	reference	(95-211), 344-565	No
Mh10g200708_Contig42	Contig119	91.5	93.6	(no info)	(no info)	Q61QZ5 Hypothetical protein CBG06805.					No
Mh10g200708_Contig42	HAP_EST248xe01f1	114.6	115.8	CE29940 NP_001022985 ZK1128.6a [Caenorhabditis elegans]. WBGene00014232	XP_001094742 PREDICTED: tubulin tyrosine ligase-like family, member 4 isoform 3 [Macaca mulatta].	Q8IGW4 RE12363p. Drosophila melanogaster (fruit fly)	585	114658-114785, 115141-115249, 115289-115364, 115457-115726	complement	230-585	No
Mh10g200708_Contig42	Contig1911	118.5	120.2	(no info)	(no info)	Q61IN7 Hypothetical protein CBG10136. Caenorhabditis briggsae					No
Mh10g200708_Contig42	HAP_EST331xo13f1	124.7	126	CE31583 NP_497665 R148.5a [Caenorhabditis elegans]. WBGene00020104	EAT48681 conserved hypothetical protein [Aedes aegypti].	O17265 Hypothetical protein R148.5.	689	124683-124806, 124852-124922, 125155-125398, 125704-125790, 125837-125999	reference	314-689	No
Mh10g200708_Contig42	Contig2139	157.2	160.9	NP_001022682 M01A8.2a [Caenorhabditis elegans]. WBGene00010796	NP_002947 restin isoform a [Homo sapiens].	Q17RS4 RSN protein.	834	157298-157316, 158311-158549, 158616-158799, 160355-160504, 160549-160691, 160792-160890	complement	15-834[rc]	No
Mh10g200708_Contig43	Contig2727	23.7	25	CE19602 NP_001022796 SET (trithorax/polycomb) domain containing family member (set-1) [Caenorhabditis elegans]. WBGene00004781	NP_650354 CG3307-PA, isoform A [Drosophila melanogaster].	Q297V5 GA17259-PA.	680	23705-23800, 23846-23990, 24282-24469, 24715-24937	reference	104-565	No
Mh10g200708_Contig44	Contig910	8.4	9	CE19602 NP_001022796 SET (trithorax/polycomb) domain containing family member (set-1) [Caenorhabditis elegans]. WBGene00004781	NP_650354 CG3307-PA, isoform A [Drosophila melanogaster].	Q297V5 GA17259-PA.	427	8466-8624, 8672-8742, 8795-8926	reference		No
Mh10g200708_Contig45	HAP_EST151xk15r1	33.6	34.5	(no info)	XP_001116304 PREDICTED: similar to CG17593-PA isoform 1 [Macaca mulatta].	Q9D024 Coiled-coil domain-containing protein 47 precursor (Adipocyte-specific protein 4).					No
Mh10g200708_Contig45	HAP_EST338xc23f1	46.4	48	CE09593 NP_506574 F23B12.1 [Caenorhabditis elegans]. WBGene00009079	XP_392943 PREDICTED: similar to protein phosphatase 1, catalytic subunit, gamma isoform isoform 1 [Apis mellifera].	Q3T0E7 Serine/threonine-protein phosphatase PP1-alpha catalytic subunit (PP-1A).	687	46490-46552, 46601-46832, 47161-47291, 47341-47457, 47745-47842, 47896-47941	reference	9-687	No
Mh10g200708_Contig45	Contig1998	54.4	55.3	CE36522 NP_498913 FLI-I (Drosophila flightless) homolog family member (fli-1) [Caenorhabditis elegans]. WBGene00001443	XP_790260 PREDICTED: similar to flightless I homolog variant [Strongylocentrotus purpuratus].	P34268 Protein flightless-1 homolog. Caenorhabditis elegans	803	54397-54459, 54500-54665, 55124-55205, 55254-55290	reference	156-623	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig45	Contig371	56.9	58.3	CE36522 NP_498913 FLI-I (Drosophila flightless) homolog family member (fli-1) [Caenorhabditis elegans]. WBGene00001443	NP_498913 FLI-I (Drosophila flightless) homolog family member (fli-1) [Caenorhabditis elegans]. WBGene00001443	P34268 Protein flightless-1 homolog. Caenorhabditis elegans	722	56989-57066, 57326-57538, 57586-57655, 57912-58006, 58049-58313	complement	242-553[rc]	No
Mh10g200708_Contig46	HAP_EST164xk09f1	8.5	9.6	(no info)	(no info)	Q8I100 Hypothetical protein rabx-5. Caenorhabditis elegans					No
Mh10g200708_Contig46	Contig2720	10.1	10.8	CE02041 NP_498094 T10F2.2 [Caenorhabditis elegans]. WBGene00020422	EAT34099 mitochondrial ornithine transporter [Aedes aegypti].	Q10050 Hypothetical protein. Caenorhabditis elegans	381	10153-10361, 10406-10483, 10600-10633, 10680-10720, 10764-10784	complement	108-281[rc]	No
Mh10g200708_Contig46	Contig1584	10.8	11.8	CE02041 NP_498094 T10F2.2 [Caenorhabditis elegans]. WBGene00020422	XP_788407 PREDICTED: similar to MGC108450 protein, partial [Strongylocentrotus purpuratus].	Q60RJ1 Hypothetical protein CBG21322.	593	10867-10973, 11028-11082, 11127-11174, 11281-11409, 11456-11543, 11589-11641, 11685-11797	complement	32-593	No
Mh10g200708_Contig46	Contig1311	13.2	14.2	CE38104 NP_498096 T10F2.4 [Caenorhabditis elegans]. WBGene00020423	XP_001084826 PREDICTED: PRP19/PSO4 pre-mRNA processing factor 19 homolog [Macaca mulatta].	Q5ZMA2 Pre-mRNA-processing factor 19 (PRP19/PSO4 homolog). Gallus gallus (chicken)	725	13270-13406, 13460-13645, 13688-13823, 13866-13909, 13953-14144	complement	67-725	No
Mh10g200708_Contig46	Contig2072	18	19.5	CE24022 NP_499457 T28D6.6 [Caenorhabditis elegans]. WBGene00012126	CAE56931 Hypothetical protein CBG24776 [Caenorhabditis briggsae].	Q60IX6 Hypothetical protein CBG24776.	885	18016-18063, 18114-18180, 18231-18313, 18360-18565, 18612-18661, 18707-18898, 19164-19204, 19249-19444	complement? (see conclusions)	168-842[rc]	No
Mh10g200708_Contig47	HAP_EST233xk19f1	16.9	18.4	CE27935 NP_498621 F08F8.8 [Caenorhabditis elegans]. WBGene00017273	NP_498621 F08F8.8 [Caenorhabditis elegans]. WBGene00017273	Q4SYZ0 Chromosome 10 SCAF11883, whole genome shotgun sequence. Tetraodon nigroviridis	603	16936-17048, 17092-17178, 17624-17766, 17814-17917, 18263-18408	complement	236-442[rc]	No
Mh10g200708_Contig48	HAP_EST233xk19f1	3	3.2	CE27935 NP_498621 F08F8.8 [Caenorhabditis elegans]. WBGene00017273	NP_498621 F08F8.8 [Caenorhabditis elegans]. WBGene00017273	Q4SYZ0 Chromosome 10 SCAF11883, whole genome shotgun sequence. Tetraodon nigroviridis	603	59-189	complement		No
Mh10g200708_Contig48	Contig335	10.6	12.1	CE37548 NP_001022726 R144.12 [Caenorhabditis elegans]. WBGene00043060	CAE70087 Hypothetical protein CBG16528 [Caenorhabditis briggsae].	Q612X2 Hypothetical protein CBG16528.	364	10601-10630, 10675-10776, 11447-11583, 11639-11662, 12017-12071	complement	39-364[rc]	No
Mh10g200708_Contig48	Contig2244	31.2	32.2	(no info)	(no info)	Q3V5J3 Hypothetical protein.					No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig48	Contig351	41.7	44.2	CE00209 NP_499094 Vacuolar H ATPase family member (vha-14) [Caenorhabditis elegans]. WBGene00010130	NP_499094 Vacuolar H ATPase family member (vha-14) [Caenorhabditis elegans].	P34462 Probable vacuolar ATP synthase subunit D (V-ATPase subunit D) (Vacuolar proton pump subunit D).	832	41787-42022, 42491-42644, 43122-43231, 43279-43306, 43353-43475, 43883-44000, 44046-44108	complement	17-790	Yes
Mh10g200708_Contig50	HAP_EST259xo23f1	10.1	10.6	CE19464 NP_001021496 UNCoordinated family member (unc-73) [Caenorhabditis elegans]. WBGene00006805	CAE66729 Hypothetical protein CBG12078 [Caenorhabditis briggsae].	Q61EG8 Hypothetical protein CBG12078.	367	10092-10183, 10228-10282, 10328-10400, 10445-10590	complement	171-320[rc]	No
Mh10g200708_Contig50	HAP_EST320xi15f1	38.8	39.6	CE10552 NP_492717 F45H11.2 [Caenorhabditis elegans]. WBGene00003587	AAF73908 polyprotein [bovine viral diarrhea virus-1 strain CP821].	P61282 NEDD8 precursor (Ubiquitin-like protein Nedd8) (Neddylin).	437	38830-38925, 38978-39070, 39380-39579, 98109-98132	reference	10-258	Yes
Mh10g200708_Contig50	Contig857	47.7	48.7	CE09767 NP_491282 F28H1.2 [Caenorhabditis elegans]. WBGene00000779	NP_491282 F28H1.2 [Caenorhabditis elegans].	O01542 Calponin protein 3.	652	47772-47961, 48179-48254, 48294-48356, 48403-48471, 48517-48706, 98109-98132	reference	3-437	Yes
Mh10g200708_Contig50	Contig983	47.7	48.3	CE09767 NP_491282 F28H1.2 [Caenorhabditis elegans]. WBGene00000779	NP_491282 F28H1.2 [Caenorhabditis elegans].	O01542 Calponin protein 3.	264	47774-47961, 48179-48225	reference	1-243	No
Mh10g200708_Contig50	HAP_EST366xm23f1	81.9	83.2	CE07741 NP_493695 B0432.3 [Caenorhabditis elegans]. WBGene00015185	NP_493695 B0432.3 [Caenorhabditis elegans]. WBGene00015185	Q61QH4 Hypothetical protein CBG07017.	696	81943-82193, 82509-82665, 82893-83163	reference	85-627	No
Mh10g200708_Contig53	Contig6	66.2	67.5	CE10580 NP_492513 SKp1 Related (ubiquitin ligase complex component) family member (skr-1) [Caenorhabditis elegans]. WBGene00004807	AAL34093 SKR-1 [Caenorhabditis elegans].	Q8WSZ9 SKR-1	670	66263-66434, 66556-66675, 66726-66838, 67171-67429	reference	64-579	Yes
Mh10g200708_Contig53	HAP_EST263xe21f2	66.2	67.3	CE10580 NP_492513 SKp1 Related (ubiquitin ligase complex component) family member (skr-1) [Caenorhabditis elegans]. WBGene00004807	AAL34093 SKR-1 [Caenorhabditis elegans].	Q5SUR3 OTTMUSP00000005802 (13 days embryo liver cDNA, RIKEN full-length enriched library, clone:2510015J15 product:S-phase kinase-associated protein 1A, full insert sequence) (Bone marrow macrophage cDNA, RIKEN full-length enriched library, clone:I830081B01 product:S-phase kinase-associated protein 1A, full insert sequence).	611	66263-66675, 66726-66838, 67171-67255	reference	64-192, (302-611)	No
Mh10g200708_Contig53	Contig1122	82.2	82.4	CE40717 NP_491262 T20F5.3 [Caenorhabditis elegans]. WBGene00020625	XP_966879 PREDICTED: similar to CG4447-PA [Tribolium castaneum].	Q16WG4 Mitochondrial ribosome recycling factor.	1043	83223-83400	complement	219-1001	Yes
Mh10g200708_Contig53	Contig126	92.5	92.8	CE20547 NP_503425 UNCoordinated family member (unc-60) [Caenorhabditis elegans]. WBGene00006794	NP_503425 UNCoordinated family member (unc-60) [Caenorhabditis elegans].	Q07749 Actin-depolymerizing factor 2, isoform c (Uncoordinated protein 60).	827	92565-92596, 92660-92730, 93041-93091	reference	80-346	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig53	Contig1340	125.3	125.8	CE03762 NP_496441 SR Protein (splicing factor) family member (rsp-2) [Caenorhabditis elegans]. WBGene00004699	NP_496441 SR Protein (splicing factor) family member (rsp-2) [Caenorhabditis elegans]. WBGene00004699	Q4SYY0 Chromosome 21 SCAF11909, whole genome shotgun sequence.	540	125736-125887, 125934-126027, 126226-126307, 126349-126463, 127045-127090, 127281-127316	reference	42-540	No
Mh10g200708_Contig53	Contig499	145.5	146.7	CE13740 NP_491663 T19B4.3 [Caenorhabditis elegans]. WBGene00020557	P54363 Adenine phosphoribosyltransferase (APRT).	P54363 Adenine phosphoribosyltransferase (APRT).	894	145387-145824, 145993-146181, 146223-146332, 146545-146700	complement	74-601	No
Mh10g200708_Contig53	HAP_EST168xm05f1	150.9.	153.2	CE34364 NP_491976 T10B11.3 [Caenorhabditis elegans]. WBGene00020399	(no info)	Q4ST31 Chromosome undetermined SCAF14310, whole genome shotgun sequence.	586	150961-151020, 151236-151375, 151489-151573, 152462-152655, 152689-152736, 153055-153109	complement	130-586	No
Mh10g200708_Contig55	Contig260	44.4	45.6	CE04946 NP_741809 Beta Carbonic Anhydrase family member (bca-1) [Caenorhabditis elegans]. WBGene00000245	XP_001121676 PREDICTED: similar to CG11967-PA [Apis mellifera].	Q22460 Beta carbonic anhydrase protein 1, isoform a.	556	44403-44506, 44596-44817, 45267-45327, 45373-45541	complement	4-556	No
Mh10g200708_Contig55	HAP_EST329xa23f1	46.1	46.7	CE27453 NP_503032 Y116A8C.30 [Caenorhabditis elegans]. WBGene00013807	NP_001041016 Y116A8C.30a [Caenorhabditis elegans]. WBGene00013807	Q9U2S7 Hypothetical protein.	365	46167-46253, 46300-46420, 46532-46661	complement	9-194, (191-365)	No
Mh10g200708_Contig55	Contig1122	61.4	63.7	CE40717 NP_491262 T20F5.3 [Caenorhabditis elegans]. WBGene00020625	XP_966879 PREDICTED: similar to CG4447-PA [Tribolium castaneum]. Tribolium castaneum (red flour beetle)	Q16WG4 Mitochondrial ribosome recycling factor.	1043	61444-61647, 61774-61844, 61893-62079, 62119-62269, 62484-62531, 62580-62747, 63409-63622	complement	219-1001	Yes
Mh10g200708_Contig56	Contig2055	9.2	10.4	CE20317 NP_493554 Y54E5A.6 [Caenorhabditis elegans]. WBGene00013201	NP_493554 Y54E5A.6 [Caenorhabditis elegans]. WBGene00013201	Q60V02 Hypothetical protein CBG19749	801	9168-9280, 9352-9453, 9501-9600, 9649-9857, 10044-10112, 10159-10366	complement	116-757[rc]	No
Mh10g200708_Contig56	HAP_EST163xo19f1	14.7	15.2	CE06577 NP_492708 Ribosomal Protein, Small subunit family member (rps-7) [Caenorhabditis elegans]. WBGene00004476	NP_492708 Ribosomal Protein, Small subunit family member (rps-7) [Caenorhabditis elegans]. WBGene00004476	Q0IEM1 40S ribosomal protein S7.	552	14714-14789, 14842-14970, 15018-15113	complement	54-552	No
Mh10g200708_Contig56	Contig2075	15.3	16.1	(no info)	XP_795944 PREDICTED: hypothetical protein [Strongylocentrotus purpuratus].	Q16VW9 Hypothetical protein.					No
Mh10g200708_Contig56	HAP_EST174xi03f1	16.3	16.7	NP_497263 Ribosomal Protein, Small subunit family member (rps-29) [Caenorhabditis elegans]. WBGene00004498	ABI52655 ribosomal protein S29 [Argas monolakensis].	P90983 Ribosomal protein, small subunit protein 29.	229	16380-16422, 16468-16653	reference	none!	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig56	HAP_EST328xa21f1	21	21.7	CE04362 NP_495468 Ribosomal Protein, Large subunit family member (rpl-33) [Caenorhabditis elegans]. WBGene00004447	NP_067087 ribosomal protein L35a [Rattus norvegicus].	Q506L9 Ribosomal protein L35a. Cirrhinus molitorella (mud carp)	570	21021-21343, 21388-21471, 21520-21640	complement	23-298	Yes
Mh10g200708_Contig56	HAP_EST336xo05f1	21	21.6	CE04362 NP_495468 Ribosomal Protein, Large subunit family member (rpl-33) [Caenorhabditis elegans]. WBGene00004447	NP_067087 ribosomal protein L35a [Rattus norvegicus].	Q506L9 Ribosomal protein L35a. Cirrhinus molitorella (mud carp)	489	21017-21343, 21388-21471, 21520-21599	complement	147-365	No
Mh10g200708_Contig56	HAP_EST204xi01f2	28.6	29.3	CE25963 NP_510555 R03A10.6 [Caenorhabditis elegans]. WBGene00010986	NP_510555 R03A10.6 [Caenorhabditis elegans]. WBGene00010986	Q21659 Hypothetical protein	472	28686-28792, 28838-29086, 29141-29233	complement	79-472[rc]	No
Mh10g200708_Contig56	Contig1122	38.9	39.2	NP_491262 T20F5.3 [Caenorhabditis elegans]. WBGene00020625	XP_966879 PREDICTED: similar to CG4447-PA [Tribolium castaneum].	Q16WG4 Mitochondrial ribosome recycling factor.	1043	38933-39127	complement	219-1001	Yes
Mh10g200708_Contig61	HAP_EST120xi09f1	13.8	14.5	(no info)	XP_689436 PREDICTED: similar to myosin heavy chain [Danio rerio].	Q1RL61 Zinc finger protein. Ciona intestinalis (class:preliminary)					No
Mh10g200708_Contig64	HAP_EST114xc23f1	57.5	59.3	(no info)	CE01640 NP_495688 T05H10.3 [Caenorhabditis elegans]. WBGene00011508	Q61BB3 Hypothetical protein CBG13422. Caenorhabditis briggsae	566	57561-57687, 58331-58476, 58518-58691, 58736-58758, 59132-59227	reference	40-566	No
Mh10g200708_Contig65	Contig24	25	27.6	CE02733 NP_494846 F41C3.5 [Caenorhabditis elegans]. WBGene00018271	NP_001042933 Os01g0332800 [Oryza sativa (japonica cultivar-group)].	Q5W727 Putative serine carboxypeptidase II (Os05g0158500 protein).	1040	24955-25104, 25156-25175, 25599-25683, 25729-25856, 26234-26317, 26362-26397, 26449-26608, 26652-26704, 27127-27262, 27373-27567	complement	129-965[rc]	No
Mh10g200708_Contig65	Contig908	48	49.4	CE06116 NP_505661 K07C5.6 [Caenorhabditis elegans]. WBGene00010629	XP_001084674 PREDICTED: similar to step II splicing factor SLU7 isoform 1 [Macaca mulatta].	Q3ZBE5 Pre-mRNA-splicing factor SLU7.	701	48056-48303, 48665-48727, 48954-49344	reference	148-678	No
Mh10g200708_Contig65	HAP_EST232xi23f1	53.7	65.6	CE02733 NP_494846 F41C3.5 [Caenorhabditis elegans]. WBGene00018271	NP_001042933 Os01g0332800 [Oryza sativa (japonica cultivar-group)].	Q5W727 Putative serine carboxypeptidase II (Os05g0158500 protein).	570	53734-53859, 64743-64881, 65095-65169, 65374-65582, 65633-65653	complement	12-570	No
Mh10g200708_Contig65	HAP_EST242xg05f1	56.2	56.5	CE07510 NP_505309 Germinal Center Kinase family member (gck-1) [Caenorhabditis elegans]. WBGene00001526	XP_001095834 PREDICTED: serine/threonine kinase 3 (STE20 homolog, yeast) [Macaca mulatta].	Q9NB31 Serine/threonine-protein kinase cst-1 (STE20-like kinase 1) (STE20-like kinase MST) (cMST).	578	56294-56405	complement	145-578	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig68	HAP_EST102xm11f1	22.1	22.5	CE02255 NP_495811 Ribosomal Protein, Large subunit family member (rpl-5) [Caenorhabditis elegans]. WBGene00004416	XP_971947 PREDICTED: similar to CG17489-PA.3 [Tribolium castaneum].	Q59GX9 Ribosomal protein L5 variant.	547	22111-22411	unknown	212-457[rc]	No
Mh10g200708_Contig68	Contig1395	30.7	31.4	NP_505006 Ribosomal Protein, Large subunit family member (rpl-39) [Caenorhabditis elegans]. WBGene00004453	ABI83747 60s ribosomal protein L39 [Anopheles funestus]. Anopheles funestus (African malaria mosquito)	Q1HRB5 60S ribosomal protein L39. Aedes aegypti (Stegomyia aegypti)	405	30792-30977, 31260-31397	complement	92-247, (324-405)	Yes
Mh10g200708_Contig68	HAP_EST377xi07r1	30.8	31.4	(no info)	(no info)	Q4GX84 Ribosomal protein L39e.					No
Mh10g200708_Contig70	Contig228	55.1	60.9	(no info)	AAR37368 putative esophageal gland cell secretory protein 37 [Meloidogyne incognita].	Q5QJ72 Putative esophageal gland cell secretory protein 37.					No
Mh10g200708_Contig70	Contig91	59.3	60.9	(no info)	AAR37368 putative esophageal gland cell secretory protein 37 [Meloidogyne incognita].	Q5QJ72 Putative esophageal gland cell secretory protein 37.					No
Mh10g200708_Contig70	HAP_EST116xa17f1	121.3	122.1	CE20735 NP_505733 yeast Glc Seven-like Phosphatases family member (gsp-1) [Caenorhabditis elegans]. WBGene00001747	NP_001042978 Os01g0349400 [Oryza sativa (japonica cultivar-group)].	O82734 Serine/threonine-protein phosphatase PP1 isozyme 8.	537	121343-121551, 121619-121695, 121737-121767, 121811-121939, 121987-122051, 122093-122106	reference	(144-269), 281-537	No
Mh10g200708_Contig70	HAP_EST168xm13f1	142.4	143.8	CE25482 NP_500034 COP-9 SigNalosome subunit family member (csn-4) [Caenorhabditis elegans]. WBGene00000816	AAH56527 Cops4 protein [Danio rerio]. ZDB-GENE-030131-4317	Q9BT78 COP9 signalosome complex subunit 4 (Signalosome subunit 4) (SGN4) (JAB1-containing signalosome subunit 4).	666	142448-142475, 142748-142907, 143033-143163, 143204-143340, 143570-143673, 143718-143800	complement	234-666	No
Mh10g200708_Contig74	Contig460	8.3	9	NP_001022356 T23G7.2a [Caenorhabditis elegans]. WBGene00011965	CAE57888 Hypothetical protein CBG00934 [Caenorhabditis briggsae].	Q17WR2 Integral membrane protein with phosphohydrolyase domain.	468	8363-8476, 8574-8927	complement	none!	No
Mh10g200708_Contig74	HAP_EST204xc17f2	9	9.8	CE03704 NP_495955 T23G7.3 [Caenorhabditis elegans]. WBGene00011966	AAG18009 hepatocellular carcinoma-related putative tumor suppressor [Homo sapiens].	Q28H41 PIN2-interacting protein 1.	613	9068-9242, 9350-9578, 9630-9790	complement	82-504[rc]	No
Mh10g200708_Contig74	Contig75	25.6	27.6	NP_491127 Y54E10A.3 [Caenorhabditis elegans]. WBGene00021826	EAT39951 conserved hypothetical protein [Aedes aegypti].	Q2LZA9 GA18927-PA. Drosophila pseudoobscura	1027	25653-25935, 26209-26487, 26700-26859, 26907-26964, 27266-27363, 27408-27555	complement	55-903	Yes
Mh10g200708_Contig74	Contig85	25.6	25.9	CE24129 NP_499652 Y111B2A.20 [Caenorhabditis elegans]. WBGene00013740	NP_499652 Y111B2A.20 [Caenorhabditis elegans]. WBGene00013740	Q627X2 Hypothetical protein CBG00501.	637	25624-25810	complement	156-482	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig74	Contig1479	42.9	45	CE31044 NP_492029 M05B5.2 [Caenorhabditis elegans]. WBGene00010870	NP_492029 M05B5.2 [Caenorhabditis elegans]. WBGene00010870	Q21516 Hypothetical protein.	894	42983-43088, 43649-44002, 44240-44319, 44617-44949, 53088-53104	complement	89-574	Yes
Mh10g200708_Contig74	Contig2610	46.7	49.6	CE17477 NP_501545 RADiation sensitivity abnormal/yeast RAD-related family member (rad-26) [Caenorhabditis elegans]. WBGene00007761	NP_957438 RAD54-like [Danio rerio].	Q7ZV09 RAD54-like (S. cerevisiae).	743	46685-46877, 48791-48977, 49030-49353, 49539-49577	complement	274-654	No
Mh10g200708_Contig74	HAP_EST168xm13f1	51.3	52.8	CE25482 NP_500034 COP-9 SigNalosome subunit family member (csn-4) [Caenorhabditis elegans]. WBGene00000816	AAH56527 Cops4 protein [Danio rerio]. ZDB-GENE-030131-4317	Q9BT78 COP9 signalosome complex subunit 4 (Signalosome subunit 4) (SGN4) (JAB1-containing signalosome subunit 4).	666	51321-51428, 51476-51579, 51869-52005, 52044-52174, 52281-52440, 52696-52723	reference	234-666	No
Mh10g200708_Contig75	HAP_EST124xk15f1	10.1	10.8	CE20860 NP_490798 F54A5.2 [Caenorhabditis elegans]. WBGene00018787	NP_490798 F54A5.2 [Caenorhabditis elegans]. WBGene00018787	Q61UW0 Hypothetical protein CBG05129.	562	10163-10383, 10435-10776	complement	222-562	No
Mh10g200708_Contig75	HAP_EST162xc05f1	25.8	27.2	CE13096 NP_491248 T03F1.1 [Caenorhabditis elegans]. WBGene00020184	NP_001046919 Os02g0506500 [Oryza sativa (japonica cultivar-group)].	P91430 Hypothetical protein.	561	25840-25901, 25941-26034, 26071-26157, 26477-26577, 26624-26743, 27048-27144	complement	2-561	No
Mh10g200708_Contig76	Contig1899	6.1	6.3	CE24418 NP_497107 Y53F4B.39 [Caenorhabditis elegans]. WBGene00013176	NP_497107 Y53F4B.39 [Caenorhabditis elegans].	Q7S4V1 Hypothetical protein NCU02376.1.	600	6198-6246	complement	(225-377), 390-548	No
Mh10g200708_Contig76	Contig2440	16.1	17.2	CE41207 NP_491707 P-GlycoProtein related family member (pgp-2) [Caenorhabditis elegans]. WBGene00003996	AAC38987 P-glycoprotein [Haemonchus contortus].	O61301 P-glycoprotein. Haemonchus contortus	586	16127-16151, 16532-16719, 16785-16899, 16941-17197	complement	144-586	No
Mh10g200708_Contig76	HAP_EST124xe23f1	31.3	32.1	(no info)	XP_001105785 PREDICTED: apoptotic chromatin condensation inducer 1 isoform 1 [Macaca mulatta].	Q1JQ01 Acin1a protein. Danio rerio (zebrafish)					No
Mh10g200708_Contig78	Contig864	6.6	7.5	CE12268 NP_490908 Neuropeptide-Like Protein family member (nlp-12) [Caenorhabditis elegans]. WBGene00003750	NP_490908 Neuropeptide-Like Protein family member (nlp-12) [Caenorhabditis elegans].	O01970 Neuropeptide-like protein protein 12.	684	6627-6974, 7127-7463	complement	37-492	Yes
Mh10g200708_Contig78	Contig1553	15.4	15.7	CE12424 NP_506318 CathePsin Z family member (cpz-2) [Caenorhabditis elegans]. WBGene00000789	AAH72275 MGC82409 protein [Xenopus laevis].	P91771 Cysteine protease.	968	15424-15503, 15569-15670	reference	36-950	No
Mh10g200708_Contig78	HAP_EST116xe17f1	16.5	17.6	CE07615 NP_505074 COLlagen family member (col-43) [Caenorhabditis elegans]. WBGene00000620	NP_505074 COLlagen family member (col-43) [Caenorhabditis elegans]. WBGene00000620	Q23364 Collagen protein 43.	541	16546-16584, 16637-16732, 16999-17043, 17089-17284, 17422-17586	complement	97-541	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig78	Contig2452	27.3	28.5	CE05458 NP_510226 C49F8.3 [Caenorhabditis elegans]. WBGene00008215	NP_510226 C49F8.3 [Caenorhabditis elegans]. WBGene00008215	Q18709 Hypothetical protein. Caenorhabditis elegans	782	27385-27867, 27927-28004, 28078-28201, 28329-28426	complement	153-770[rc]	Yes
Mh10g200708_Contig78	Contig2874	28.5	29.8	CE33329 NP_504033 T27C4.1 [Caenorhabditis elegans]. WBGene00020854	NP_504033 T27C4.1 [Caenorhabditis elegans]. WBGene00020854	O61906 Hypothetical protein T27C4.1.	1043	23259-23278, 28540-28801, 28852-29089, 29164-29284, 29375-29516, 29563-29789	complement	22-888	Yes
Mh10g200708_Contig78	Contig2629	35.2	37	NP_498974 ZK637.13 [Caenorhabditis elegans]. WBGene00014030	AAL56426 cuticle globin [Syngamus trachea].	Q8WS57 Cuticle globin. Syngamus trachea	907	35216-35397, 35446-35537, 35593-35651, 36274-36348, 36397-36792, 36845-36948	reference	69-554, (160-309, 333-554, 405-554, 480-554, 641-757, 775-858)	Yes
Mh10g200708_Contig78	Contig191	42.7	43.9	CE03567 NP_510191 ASpartyl Protease family member (asp-4) [Caenorhabditis elegans]. WBGene00000217	ABC88426 cathepsin D-like aspartic proteinase preproprotein [Meloidogyne incognita].	Q1A2Z2 Cathepsin D-like aspartic proteinase preproprotein.	826	42739-42995, 43046-43250, 43291-43414, 43484-43590, 43635-43693, 43759-43832	reference	151-826	No
Mh10g200708_Contig78	HAP_EST236xe03r1	44.3	44.9	CE03567 NP_510191 ASpartyl Protease family member (asp-4) [Caenorhabditis elegans]. WBGene00000217	ABC88426 cathepsin D-like aspartic proteinase preproprotein [Meloidogyne incognita].	Q1A2Z2 Cathepsin D-like aspartic proteinase preproprotein. Meloidogyne incognita (southern root-knot nematode)	467	44322-44518, 44568-44637, 44680-44879	complement	2-388	No
Mh10g200708_Contig79	Contig827	17.1	17.9	CE27416 NP_503408 Y75B7AR.1 [Caenorhabditis elegans]. WBGene00022287	NP_503408 Y75B7AR.1 [Caenorhabditis elegans]. WBGene00022287	Q61RW0 Hypothetical protein CBG06429. Caenorhabditis briggsae	561	17182-17256, 17327-17407, 17446-17849	complement	23-208, (466-561)	No
Mh10g200708_Contig79	HAP_EST159xe09f1	30.6	31.4	CE00491 NP_499119 C48B4.1 [Caenorhabditis elegans]. WBGene00008167	NP_499119 C48B4.1 [Caenorhabditis elegans]. WBGene00008167	O62140 Hypothetical protein. Caenorhabditis elegans	534	30690-30938, 30983-31056, 31100-31260, 31304-31354	reference	34-534	No
Mh10g200708_Contig79	Contig896	53.5	53.7	NP_509944 STOmatin family member (sto-4) [Caenorhabditis elegans]. WBGene00006066	CAE68847 Hypothetical protein CBG14809 [Caenorhabditis briggsae].	Q617V8 Hypothetical protein CBG14809.	667	53592-53743	unknown	177-608	No
Mh10g200708_Contig81	Contig266	35.6	36.1	CE23779 NP_490755 F56A6.2 [Caenorhabditis elegans]. WBGene00002040	NP_490755 F56A6.2 [Caenorhabditis elegans].	Q61XP6 Hypothetical protein CBG03901.	537	35548-35793, 35843-36016	unknown	194-307, (379-468)	No
Mh10g200708_Contig81	HAP_EST124xc23r1	37.5	38.3	CE34519 NP_498712 C02C2.5 [Caenorhabditis elegans]. WBGene00015334	EAT38359 iodotyrosine dehalogenase [Aedes aegypti].	Q6PHW0 Iodotyrosine dehalogenase 1 precursor (IYD-1).	531	37596-37817, 37856-37951, 38014-38100, 38175-38300	complement	25-501[rc]	No
Mh10g200708_Contig81	Contig2134	41.9	43.2	CE20445 NP_500372 OSMotic avoidance abnormal family member (osm-9) [Caenorhabditis elegans]. WBGene00003889	BAC23088 HrTRPV [Halocynthia roretzi].	Q8IU33 HrTRPV. Halocynthia roretzi	555	41990-42034, 42086-42206, 42251-42361, 42413-42535, 42583-42698, 43102-43140	complement	96-555	No

Table 2 (continued)

Genomic Contig	EST Contig	Start location (kb)	Stop location (kb)	BLASTX vs Wormpep	BLASTX vs NR	BLASTX vs UNIPROT	EST length (bp, from SIM4)	Genomic Coordinates of EST (from SIM4)	Ref/Com Strand (from SIM4)	ORFs of predicted gene (from EditSeq)	Confident Full-Length Gene?
Mh10g200708_Contig83	Contig998	37.7	40.9	CE19071 NP_502694 Lateral Signaling Target family member (lst-4) [Caenorhabditis elegans]. WBGene00003086	NP_502694 Lateral Signaling Target family member (lst-4) [Caenorhabditis elegans]. WBGene00003086	O95061 WASP interactor protein.	694	37714-37847, 38602-38723, 38908-39148, 40584-40676, 40754-40859	complement	(185-361[rc]), 183-694	No
Mh10g200708_Contig83	HAP_EST135xa21f1	46.6	47.9	CE16735 NP_503060 ZK550.4 [Caenorhabditis elegans]. WBGene00013998	NP_503060 ZK550.4 [Caenorhabditis elegans]. WBGene00013998	Q628I6 Hypothetical protein CBG00367.	570	46651-46976, 47246-47465, 47877-47900	reference	68-570	No
Mh10g200708_Contig84	HAP_EST132xe09f1	11.2	13	CE30244 NP_001021942 Protein Tyrosine Phosphatase family member (ptp-3) [Caenorhabditis elegans]. WBGene00004215	A48758 protein-tyrosine-phosphatase (EC 3.1.3.48), receptor-linked form P1 precursor - rat.	Q4JFL7 Receptor-linked protein tyrosine phosphatase. Rattus norvegicus (Norway rat)	557	11250-11278, 11612-11803, 11856-11994, 12798-12994	reference	18-557	No
Mh10g200708_Contig85	HAP_EST392xi19f1	3.2	4.2	CE26429 NP_498226 B0336.13 [Caenorhabditis elegans]. WBGene00015150	EAT46700 transcription initiation factor IIA (TFIIA), gamma chain [Aedes aegypti].	Q17J10 Transcription initiation factor IIA (TFIIA), gamma chain.	491	3239-3416, 3514-3637, 3931-4058, 4112-4141	complement	89-325	No

ORF numbers in parentheses indicate ORF predictions less likely to be true than those not in parentheses.
[rc], reverse-complement ORF.

Chapter 2: Analysis of Three Major *Meloidogyne hapla* Gene Families

Note: This chapter is an expanded version of my contribution to published work (Opperman et al., 2008).

Introduction

Once a genome has been sequenced and assembled, and a gene set has been constructed, hypotheses can be proposed for the function of individual genes thus identified. As similarity in structure often equates to similarity in function, large groups of genes with highly similar structures can often be predicted to enable similar functions in an organism. Therefore, classifying genes into groups (or families) by computational means is a valuable step in conserving experimental resources, since the function of many genes can be inferred from the experimental evidence of a significantly fewer number of reference genes. It is at this stage where some of the most revealing discoveries are made in the study of a species. Gene families which are overrepresented in a given species compared to an appropriate reference species may suggest an elevated importance in the genes composing these families, and hence, a more highly-utilized or diverse biological process, cellular component, or molecular function underlying these genes than in the reference species. Likewise, gene families which are underrepresented in a given species relative to the same reference species would imply a tempered need for diversity, complexity, or utilization in the underlying processes, components, or functions of the genes which are members of this family, compared to the reference species.

Gene family classification is a rich field, diverse in ideology of objectives and methodology. It is necessary to classify genes and proteins into families in order to infer possible clues of gene and protein action and interaction in the cell, and to integrate existing knowledge of structure, function, and evolution at the level of the genome (Ouzounis et al., 2003). At the most basic level of distinction, classification schemes can be coarsely divided into automated versus curated schemes, with a focus on function versus structure, in a spectrum of combinations (i.e., semi-automatic with some manual effort required, or preliminary structural classification followed by functional validation). Automated methods benefit from high-throughput, scalable and reproducible results based on metrics that may not be obvious to the human eye, with the drawback of potentially spurious clustering. Curated methods benefit from human expertise and high-quality clustering, yet their end result might not be reproducible or scalable to large amounts of data. The number of classes in curated methods is 'supervised', as it is determined by experts. In contrast, automatic classification methods yield 'unsupervised' classes, as their number is determined by the underlying algorithm, and might not be fixed. Structural classification methods attempt to extract measures of molecular similarity exclusively on the basis of protein structure at various levels. Functional classification methods ignore structure, but incorporate other lines of evidence, such as pathway information, cell cycle phases and networks, genome evolutionary constraints, and keywords from literature searches (Ouzounis et al., 2003). To preface further discussion of gene and protein classification schemes, it is important to note that most automated structure-based methods principally analyze similarity – the degree to which two traits correspond to one another in some way (Schwarz 2005). Similarity can be directly observed through

protein comparisons. In contrast, homology is the property of two traits in different organisms derived from a common trait in a common ancestor, and can only be indirectly discerned through similarity. Similarity can arise from divergent homology (a gradual loss of similarity due to cumulative changes from a common starting point) or convergent evolution (similarity due to independent instances of transformation towards a common ending point). Homology can arise through orthology (speciation), paralogy (intragenomic duplication within a single species), or xenology (horizontal gene transfer). In short, similarity is a useful proxy for other evolutionary relationships, but is insufficient as the sole criterion to base definitive phylogenetic conclusions. With these points in mind, we can consider phylogenetic-based schemes as the third major protein classification methodology (joining structural and functional methodologies). Methods based directly on evolution will undeniably and unsurprisingly overlap the structural and functional-based methods to a large extent, as similarity in structure and function can often be explained by common events in evolutionary histories (Schwarz 2005).

Structural classification schemes can be further divided by primary, secondary, and tertiary-level protein similarity (Ouzounis et al., 2003). Several examples of protein sequence (or primary structure) classification schemes, which rely on accurate definitions of protein families provided by homologous sequences and motif databases, exist. These include the Bio-Dictionary resource (Rigoutsos et al., 2002), the BLOCKs database (Henikoff and Henikoff, 1991; Henikoff et al., 1999), ClustR (Kriventseva et al., 2001, 2003), COGS (Tatusov et al., 1997; Wheeler et al., 2003), the eMOTIF collection (Nevill-Maning et al.,

1998; Huang and Brutlag, 2001), MetaFam (Silverstein et al., 2001; Shoop et al., 2001), Pfam (Sonnhammer et al., 1997; Bateman et al., 2002), Protein Fingerprints (PRINTS) (Attwood et al., 1994, 2003), the ProDom database (Corpet et al., 1998, 2000), the PROSITE protein families and domains database (Bairoch, 1991; Falquet et al., 2002), ProtoMap (Yona et al., 2003, 2005), the Simple Modular Architecture Research Tool (SMART) (Schultz et al., 1998; Letunic et al., 2002), the SYSTERS database (Krause et al., 2002), the TIGRFAMS protein-family database (Haft et al., 2001, 2003), and TRIBES (Enright et al., 2003). The Bio-Dictionary resource finds sequence patterns using an unsupervised combinatorial pattern-discovery algorithm, and has the ability to automatically annotate proteins containing those patterns. BLOCKS extracts gap-free protein family alignments from PROSITE. ClustR bases protein family classification on the Smith-Waterman dynamic programming local sequence alignment algorithm output. COGS archives the membership of orthologous genes across whole genomes. The eMOTIF database takes elements from the BLOCKS and PRINTS databases, converting amino acid similarity and pattern definitions into a unified language. MetaFam can be thought of a higher-order database that makes the integration and querying of other family databases possible. Pfam relies on seed alignments which are extended using Hidden Markov Model (HMM) queries against the full protein database. PRINTS employs a thoroughly curated motif database which is used to determine functional properties of proteins at varying degrees of sequence similarity. ProDom contains sequence database-derived protein domain families. PROSITE, one of the oldest curated motif databases, depends on community contributions to extend its definitions of sequence profiles and patterns. ProtoMap creates protein families based on similarity graphs constructed using

a variety of sequence-similarity analysis methods. SMART employs multiple alignments of sensitive database searches to describe mobile domains. SYSTERS relies on automated clustering to classify families and superfamilies into non-overlapping sets. TIGRFAMS consists of manually-curated protein families expressed in HMM form. TRIBEs contains protein families detected by the TRIBE-MCL algorithm in a database. Additionally, six of the motif and domain collections (Pfam, PRINTS, ProDom, PROSITE, SMART, and TIGRFAMS) are integrated into the InterPro database (Mulder et al., 2003). Other methods focused on primary sequence examine hydrophobic profiles, compositional biases, and size ranges to group proteins into families (Schwarz, 2005). At the secondary structure level, protein structure databases include the Dictionary of Secondary Structures of Proteins (DSSP) database (Kabsch and Sander, 1983) and the Homology-derived Secondary Structure of Proteins (HSSP) database (Sander and Schneider, 1991; Dodge et al., 1998). DSSP gathers three-dimensional input coordinates to enumerate secondary-structure elements based on hydrogen-bonding patterns and other features. HSSP extends DSSP-derived secondary structure definitions by employing a position-weighted dynamic programming sequence profile alignment method to incorporate sequence similarity information. Tertiary structure classification examples include the Class/Architecture/Topology/Homology (CATH) database (Orengo et al., 1997; Pearl et al., 2003), the Fold Classification based on Structure-Structure Alignment of Proteins (FSSP) database (Holm et al., 1992; Holm and Sander, 1998), and the Structural Classification of Proteins (SCOP) (Murzin et al., 1995; Lo Conte et al., 2002). CATH classifies proteins based on basic secondary structure elements, orientations of these secondary-structure elements, topology, and homology classes. FSSP is

a database constructed from exhaustively-permutated protein structure comparisons. SCOP is similar to CATH, relying on visual protein fold inspection, manual curation into subgroups, and hierarchical multi-level (fold, family, and superfamily) classification (Ouzounis et al., 2003).

Functional classification methods examine cellular localization, functional roles, enzyme reaction mechanisms, and biochemical pathway participation similarity (Ouzounis et al., 2003). The Enzyme Commission (EC) hierarchical classification (Bairoch 1993, 2000) is one of the oldest functional classification schemes. It groups enzyme reactions into clusters (and assigns EC numbers) with similar properties, and subclassifies them based on reaction specificities, reactants, products, and mechanisms. The Yeast Proteome Database (YPD) (Garrels 1996; Hodges et al., 1999) is a highly-curated database which has been expanded to contain entries on other species, and catalogs subcellular localization information (it has been acquired by Incyte Genomics). The What Is There? (WIT) resource (Overbeek et al., 2000), commercialized by Integrated Genomics Inc. and renamed ERGO, contains metabolic reconstruction and conserved gene cluster information. The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Mellor et al., 2002) and Predictome (Yanai et al., 2001) both provide for genome-wide association of function detected using comparative genomics techniques. A more complex classification scheme, the Gene Ontology (GO) classification (Ashburner et al., 2000), analyzes proteins for their membership in biological process, cellular component, and molecular function classes. It provides a flexible and elegant method to categorize proteins into classes at varying specificities of knowledge.

Metabolic pathway participation and association provides an alternative way to functionally classify genes and proteins. Examples of metabolic pathway-based functional classification schemes include the *E. coli* gene and metabolism encyclopedia known as the EcoCyc database (Karp et al., 1996) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kenehisa et al., 2002; Ogata et al., 1999). Proteins can also be dynamically classified (that is, classes are not necessarily defined, and are arbitrarily formed) based on cellular localization and protein interaction information. Examples of functional classification schemes using this method are the BIND database (Bader et al., 2003), Database of Interacting Proteins (DIP) (Xenarios et al., 2002), and the Molecular INTeraction database (MINT) (Zanzoni et al., 2002). All three are public-domain resources which incorporate protein interactions from multiple species.

Phylogenetic-based classification methods attempt to group proteins based on the intangible property of evolutionary history – a much more difficult facet to observe, infer, and quantify. These methods are inextricably linked to structural and functional-based methods due to the fact that the foundation for phylogenetic inferences is structural and functional evidence. Earlier methods required significant effort focused on one individual gene (or protein), requiring manual protein isolation and gene cloning based on limited reference data (Schwarz, 2005). The sequencing of whole genomes has alleviated some of the burden in obtaining data, but manually drawing phylogenetic inferences becomes an overwhelming task at the whole-genome level. One automatic approach for making such inferences is to generate homolog clusters (achievable via the PFAM and InterPro methods, discussed

earlier) while leaving orthology and paralogy of the members undefined (Schwarz 2005). Multiple-genome BLAST searches incorporated into Bayesian matrices provide another homology-based method (Enright et al., 2002). Homology-only methods can easily dissect complex proteins into multiple domains, but ignore more intricate evolutionary histories, often lumping proteins into misleadingly large groups. Orthology-based methods include the construction of KOGs (euKaryotic Orthologous Groups), TWOGs (candidate TWo-species Orthologous Groups) and LSEs (Lineage-Specific Expansions) by using triplets of mutual best BLAST hits from a limited number of species (Tatusov et al., 1997, 2003). Using species pairs rather than larger species multiples, InParanoid generates pairwise orthologs and paralogs (Remm et al., 2001). Orthologous groups have also been generated from PFAM using HOPS (Storm and Sonnhammer, 2003) and RIO (Zmasek and Eddy, 2002). A recurrent challenge of automatic evolutionary-based classification methods is defining and discerning protein motifs from protein families. There is a high level of disagreement in definition across broad classification schemes regarding these terms. InterPro defines motifs as independent structural units which can be found either in isolation or with other domains or repeats; families are defined as groups of proteins with shared domain or repeat architecture (Mulder et al., 2005).

Given the immense number of structural, functional, and evolutionary classification schemes, an ongoing goal in the field of gene and protein classifications is to develop one unified, natural classification scheme with a complete set of strictly-defined, formal definitions for all elements in the scheme. However, the underlying fundamental conceptual framework for

such a scheme has still not been fully defined, and much more consistency in definition (e.g., protein domain, gene family) needs to be applied before this goal can be accomplished (Ouzounis et al., 2003). Until this unified scheme is developed and agreed-upon as the standard, the term "gene family" will have a meaning whose portability is unpredictable, largely dependent on seemingly arbitrary precedents.

The most highly-studied nematode species hails from the Rhabditina clade, the non-parasitic, free-living, bacterivorous model organism *Caenorhabditis elegans*. While much has been gained across diverse medical fields including aging, neurobiology, cancer, and other diseases through the study of *C. elegans*, the understanding of other nematode species is only recently emerging to reach a comparable level (Parkinson et al., 2004). Over 22,000 proteins (with an excess of 75% of these experimentally verified) result from the complete genome of *C. elegans*. A phylum-wide EST study of nematodes revealed that out of 30 species (distributed across the major clades) analyzed, between 35% and 70% of each species' genes had significant similarity to proteins from *C. elegans* (Parkinson et al., 2004). Over half of the putative genes were unique to Nematoda, with 23% unique to the species from which they were derived. Aside from framing the great diversity in nematodes, these gene sets provide candidate nematode-specific drug targets with a reduced risk of toxicity to host or nontarget species (Parkinson et al., 2004). Based on localization predictions from the primary sequence of *C. elegans* proteins, approximately 60% are cytosolic, 30% are transmembrane-embedded, and 10% are secreted (Schwarz, 2005). Roughly 90% of *C. elegans* proteins range from 100 to 1000 amino acids in length, with an overrepresented spike at about 340

amino acids, corresponding to the large group of seven-transmembrane receptors. A heavily-lopsided distribution of members to gene families further exists in *C. elegans*. Out of roughly 15,200 genes with an NCBI affiliation, half fall into only 372 (out of 5,290) NCBI families and one-quarter of the genes fall into only 51 NCBI families. These results are in contrast to the 3,341 NCBI families that are encoded by a mere one gene each in *C. elegans*. The most gene-populated *C. elegans* NCBI protein families include heterotrimeric GTP-binding protein coupled receptors (GPCRs), C-type lectins, nuclear hormone receptors (NHRs), collagens, casein kinases, zinc-fingers, as well as the uncharacterized set of proteins (Schwarz, 2005).

While *C. elegans* is one of the most highly-studied nematode species, other nematode species have a much larger impact on humans due to their negative effects on agricultural crops. Most agricultural crop damage is a consequence of the sedentary root-knot (*Meloidogyne* spp.) and cyst (*Globodera* and *Heterodera* spp.) nematodes (Bird and Kaloshian, 2003). Root-knot nematode species not only cause billions of US dollars worth of annual crop losses, but their host range encompasses almost all cultivated plants (Agrios, 1997; Sasser and Freckman, 1987). The life cycle of root knot nematodes (RKN) begins at the surface of the plant root, where females deposit hundreds of eggs within a proteinaceous matrix (Bird and Kaloshian, 2003). Motile second-stage larvae (L2) follow the first egg molt, where they hatch in the soil to reinfect the plant. At the L2 stage, *Meloidogyne* spp. are in a long-lived, non-feeding state where further development is temporarily arrested (in a similar fashion as *C. elegans* dauer larva, whose duration in this stage is dependent on nematode population

density and food signals). Entry of the *Meloidogyne* L2 into the plant follows destructive penetration of the root (concurrent with protein secretion by the larvae). Once in the plant, the nematode migrates intercellularly through the vascular cylinder, where a feeding site (known as the giant cell) is established through a partially-understood process involving interactions between nematode pharyngeal gland proteins and plant host genes. It is at this point when development in the L2 resumes. After the nematode feeds (through use of the stylet) for approximately twelve days, three molts transition it into an adult (Bird and Kaloshian, 2003).

The GPCR, NHR, and collagen gene families are the focus of the gene family membership analyses in this chapter. GPCRs and NHRs comprise not only the largest of *C. elegans* gene families; they also represent gene families with very noteworthy lineage-specific expansion and significant divergence of primary sequence (Schwarz, 2005). Due to these factors, reliably determining gene membership to each of these families has required an enormous effort from experts, and revisions to these families are still being made.

GPCRs, synonymous with G protein-linked receptors (GPLRs), seven-transmembrane domain receptors, 7TM receptors, serpentine receptors, and heptahelical receptors, form a large family of membrane-bound receptors that sense various molecules outside of the cell, activate signal transduction pathways, and elicit a cellular response. A wide range of ligands are capable of binding these receptors, including neurotransmitters, hormones, odors, pheromones, and light-sensitive compounds. GPCRs have roles in many common diseases

(e.g., diabetes, hypo- and hyper-thyroidism, retinitis pigmentosa, fertility disorders), and as such, are targets for between one-third to one-half of all modern therapeutic drugs (Klabunde and Hessler, 2002). As its synonyms suggest, GPCRs possess seven transmembrane domains, and are associated with a cytoplasmic heterotrimeric G protein consisting of three subunits (α, β, and γ) (Bargmann, 1998). The α subunit is capable of binding a guanosine triphosphate (GTP) or guanosine diphosphate (GDP), whereby GTP or GDP exchange and dissociation of the Gα subunit from the Gβγ subunit subsequently occurs. Compared to ligand-gated ion channels, GPCRs tend to generate slower and more persistent changes in neuronal excitability. As the nervous system contains roughly one-third of all somatic cells in *C. elegans*, the long *C. elegans* GPCR gene roster (over 1,000; roughly 5% of its genes) is not surprising (Bargmann, 1998). GPCRs can be organized into two groups: those with clear and defined relationships with receptors previously identified in other animals (approximately 100 genes; thought to encode neurotransmitter receptors), and those without similarity to other animals: worm-specific "orphans" (approximately 1000 genes; thought to encode chemoreceptors). Given the free-living lifestyle of *C. elegans*, a wide range of microenvironments could be encountered, requiring a keen ability to detect volatile and water-soluble cues associated with food, threats, predators, and potential mates (Bargmann, 2006a). Thus, such a strong representation in chemosensation-related genes is unremarkable (Bargmann, 1998). The *C. elegans* senses of taste and smell are made possible by GPCRs: most of these are expressed in chemosensory neurons (contained within the amphid chemosensory organs), but approximately 20% are expressed in other tissues (Bargmann, 2006b). A paucity of knowledge exists regarding recognition specificity of *C. elegans*

chemoreceptors. One GPCR was identified as the receptor for diacetyl, a volatile odorant, using forward genetic screens for olfactory mutants: ODR-10. Mutants of *odr-10* exhibit a 100-fold diacetyl sensitivity reduction, yet are sensitive to all other odorants, implying that some *C. elegans* receptors only respond to a narrow range of specific chemicals (Bargmann, 2006b). In spite of their critical role in chemosensation, it should be noted that other gene families (aside from GPCRs) are also used for chemosensation in *C. elegans* (Bargmann, 2006b). At the eukaryote level, the GPCR superfamily can be divided into six classes according to the GPCR database (GPCRDB) classification scheme (Horn et al., 2003): Class A rhodopsin-like (over 80% of GPCRs fall into this class), Class B secretin-like, Class C metabotropic glutamates, Class D pheromones, Class E cAMP receptors, and Class F frizzled/smoothened family (Davies et al., 2007). *C. elegans* GPCRs can further be broken into smaller subfamilies: *str* (which includes the *odr-10* receptor), *sra*, *srab*, *srb*, *srbc*, *srd*, *sre*, *srg*, *srh*, *sri*, *srj*, *srm*, *srr*, *srsx*, *srt*, *sru*, *srv*, *srw*, *srx*, *srxa*, and *srz* (Robertson and Thomas, 2006). Annotation of *C. elegans* GPCRs is a much more complex task than mammalian GPCR annotation, due to the additional gene structure complexity entailing anywhere from one to eight embedded introns, and highly disparate intron locations between gene subfamilies. The application of EST data offers limited further guidance, due to uniformly low expression levels. As a consequence of all of these factors, confidence in the functionality of specific genes is typically low. *C. elegans* GPCR genes display an unusual degree of within- and between-family chromosome clustering, with the most pronounced clustering occurring on chromosome V. A striking level of chemoreceptor subfamily membership divergence is seen in comparative studies between *C. elegans* and *C. briggsae*,

emphasizing a common theme of *C. elegans* subfamily amplification through frequent duplications and gene loss. Because of the highly dynamic state of the GPCR family within nematodes, it is difficult to draw conclusions on orthology between species; few persuasive one-to-one ortholog examples exist (Robertson and Thomas, 2006).

NHRs, also commonly referred to as simply nuclear receptors (NRs), are one of the most abundant classes of metazoan-specific proteins found within the cell interior that sense the presence of steroid and thyroid hormones, regulate gene expression, and control diverse physiological functions such as embryonic development, organ physiology, cell differentiation, and homeostasis (Novac and Heinzel, 2004). Belonging to a much larger superfamily of transcription factors, this gene family is involved in several broad genetic processes, including transcriptional activation, chromatin remodeling, histone modification, transcriptional machinery cross-talk, and transcriptional repression. With such broad biological activities reliant on NHRs, it naturally follows that a broad assortment of pathological processes are also involved with NHRs, including arthritis, diabetes, cancer, asthma, and hormone resistance syndromes (Novac and Heinzel, 2004). NHRs function as cell type- and promoter-specific regulators of gene transcription, through the binding of specific DNA regulatory elements, and can selectively activate or repress the activity of specific genes depending on cellular content, physiological context, and gene context. Ligands for known NHRs are typically small lipophilic molecules that can penetrate biological membranes with ease, including steroids, thyronines, retinoic acids, eicosanoids, and fatty acids; however, a number of orphan NHRs exist for which the cognate ligand is not

known. NHRs bind to regulatory elements containing two copies of a Purine-GGTCA core sequence arranged as a direct or inverted repeat. In addition to DNA regulatory elements, promoters can be regulated by NHRs by means of DNA-independent protein-protein interactions (Novac and Heinzel, 2004). NHRs target specific promoter or enhancer DNA regulatory elements through the use of the N-terminal DNA-binding domain (DBD), which consists of two highly conserved cysteine-rich zinc finger motifs. Within the first zinc finger, the P-box and DR-box protein motifs have essential roles in determining NHR target DNA sequence specificity (Novac and Heinzel, 2004). Similarly, the P-loop in the first zinc finger determines specific sequence binding, depending on the specific NHR class. The D-loop, located between the first and second cysteine of the second zinc finger motif, forms part of the dimerization interface in steroid receptors. Non-steroid receptors also depend on the dimerization interface in the ligand-binding domain (LBD) region (Novac and Heinzel, 2004). The area between DBD and LBD, referred to as the flexible hinge region, often contains a nuclear localization signal. As its name suggests, the C-terminal ligand-binding domain is responsible for binding of NHR to a specific ligand, and displays a lower degree of conservation in NHRs compared to the DBD. NHRs represent an ancient family of genes, fulfilling the more complex needs of multicellular metazoan organisms through the successful union of nuclear receptor DBD and LBD (Novac and Heinzel, 2004). They can be grouped into four loose evolutionary-neutral classes: receptors which heterodimerize with RXR and recognize direct repeats as a response element, ligand-induced homodimers (which include steroid hormone receptors) that bind to inverted repeat DNA recognition sites, receptors which bind primarily to direct repeats as homodimers, and receptors which bind to

extended core sites as monomers. A newer classification scheme based on phylogeny has been approved, splitting the NHRs into seven subfamilies numbered zero through six (Novac and Heinzel, 2004). A striking difference exists between human and *C. elegans* NHR sets (Maglich et al., 2001; Zhang et al., 2004). In 2001, there were 28 known human NHRs vs. 270 for *C. elegans*. Of the 270 *C. elegans* NHRs, 255 are diverged from those found in human, but it is unclear whether these divergent *C. elegans* NHRs represent new subfamilies or are significantly diverged members of one of the NHR subfamilies (Maglich et al., 2001). Many of the 270 *C. elegans* NHRs exist due to numerous apparent gene duplications, and some might be nonfunctional pseudogenes (Sluder and Maina, 2001). In spite of this possibility, EST and GFP-reporter evidence suggests that most are functional (Miyabayashi et al., 1999; Sluder and Maina, 2001). Results from basic membership analyses using *C. elegans*, *C. briggsae*, and *Brugia malayi* ESTs encoding NHR DBDs imply that rapid evolutionary change occurring in a subset of nematode NHRs has occurred differentially among the branches of nematode phylogeny. Moreover, such results are in agreement with the hypothesis that NHR diversification has continued through nematode evolution, allowing for specific adaptations for particular lifestyles through the contributions of distinct NHRs. Examples of NHR biological roles in *C. elegans* deduced from phenotypic screens include chemosensory neuron pair function (*odr-7*), post-embryonic remodeling of motor neuron synaptic specificity (*unc-55*), and neurotransmitter expression and axon pathfinding (*fax-1*). Reverse genetic methods have also been applied to uncover probable roles in resistance to the colchicine and choloroquine toxins (*nhr-8*), molting (*nhr-23*), and epidermal morphogenesis during embryogenesis (*nhr-25*) (Sluder and Maina, 2001; Gissendanner et al., 2004).

Additional NHRs have been identified in *C. elegans*, bringing the total to 284 as of 2004 (Gissendanner et al., 2004). Increased biochemical activity complexity arises through the interplay of alternatively spliced NHR mRNAs, in which some isoforms contain extended N-terminal regions or truncated DBDs. Those NHRs only containing LBDs are capable of dimerizing with other NHRs, downregulating their transcriptional activity (Gissendanner et al., 2004). The identification of putative ligands for the large number of nematode NHRs is an ongoing area of research (Sluder and Maina, 2001).

Collagens are a group of naturally occurring structural proteins, found only in animals, which serve as the principal protein of connective tissue (Adams, 1978). This gene family has been the subject of vast multidisciplinary literature, due to its roles in animal tissue development and differentiation, human diseases (disorders of skin, tendon, vessels, ligaments, muscles, organs, bones, cartilage, and hair) and treatments (dental repair, cosmetic surgery, skin substitutes for burns, orthopedic applications), and industrial uses (gelatins and adhesives) (Adams 1978; Buehler 2006; Fratzl 2008). It is characterized at the primary protein sequence level by containing repeats of $(Gly\text{-}X\text{-}Y)_n$, where Gly represents glycine, X often represents proline, and Y sometimes represents hydroxyproline; however, marked differences in X and Y exist both between species and within different tissue types of the same species (Adams 1978). At the secondary and tertiary structural levels, these characteristic collagen repeats form a triple-helix of three different collagen polypeptide strands through the linkage of hydrogen bonds. Invertebrates, and more specifically, nematodes, have been a prime focus of collagen research, due to their ease of culture (Adams, 1978). The *C. elegans* cuticle is

composed primarily of small covalently cross-linked collagens (comprising about 80% of the soluble protein released following extraction in reducing agents (Page and Johnstone, 2007)), in turn creating a multi-layered elastic external covering, analogous to human skin (Cox et al., 1981; Politz and Edgar, 1984; von Mende et al., 1988). The cuticle serves as a flexible and resilient exoskeleton, allowing for growth through the process of molting, while simultaneously enabling locomotion through its attachment to muscle (Page and Johnstone, 2007). Through the study of morphological mutants with affected gross morphology (e.g., long, small, blistered, roller, dumpy (dpy), and squat), a clearer picture of cuticular structural protein roles has emerged (Kusch and Edgar, 1986). For instance, *dpy-13* was classified as a collagen and determined to play a unique role in cuticle growth, both through sequence analysis and through the observation that body length is shortened in tandem with the loss-of-function mutation in this gene (von Mende et al., 1988). The associated phenotypes of over 20 other *C. elegans* cuticle collagen genes have been uncovered (Page and Johnstone, 2007). A significant degree of overall similarity exists between *C. elegans* collagen family members, suggestive of significant allowable collagen cross-substitution in the cuticle assembly process, a property which would make the detection of individual collagen genes difficult using genetic methods due to the masking of loss-of-function mutants by genes with related function (von Mende et al., 1988). Indeed, extensive genetic screens have not been successful in detecting mutants of most cuticle collagen genes, purportedly due to the extremely subtle phenotypic differences proposed even in the absence of cross-substitution (Page and Johnstone, 2007). The *C. elegans* collagen gene family contains over 160 members with a characteristic structure of short interrupted blocks of Gly-X-Y sequence surrounded by

conserved cysteine residues. Within the scope of whole proteins, three domains are typically present: two non-collagenous amino and carboxy domains, and a central Gly-X-Y collagen domain. The non-collagenous amino-terminal domain is often removed through proteolysis prior to cuticle incorporation. Further study of these collagen genes and their elaborate interactions leading to cuticle formation will undoubtedly reveal critical clues toward successful organ culture systems and the treatment of debilitating diseases (Page and Johnstone, 2007).

In order to reveal high-level clues leading to a refined understanding of gene classes most important to the plant parasitic nematode lifestyle, I analyzed *Meloidogyne hapla* genes for membership in the three major aforementioned *C. elegans* gene families (GPCRs, NHRs, and collagens) in this chapter. This entailed the tedious manual collection of a large number of *C. elegans* query sequences (due to a dearth of complete, authoritative, and user-friendly *C. elegans* gene family databases), the determination of appropriate analytical methods and parameters (due to unique challenges encountered in each of the gene families), the execution of these methods followed by subsequent filters to remove unexpected or spurious results (due to a wide-ranging degree of query and target similarity and the difficulty in assigning one-to-one relationships), the organization and consolidation of large output result sets (due to large query sets, high sensitivity, and inconsistent and/or synonymous gene naming conventions found in query sequences), and the interpretation of these results. Gross disparities in gene family representation between the free-living *C. elegans* and the plant-parasitic *M. hapla* might suggest radical differences in gene diversity requirements,

nonfunctional evolutionary vestiges, or a compromise between these possibilities, between these two lifestyle modes.

Methods

I collected *C. elegans* G-Protein Coupled Receptor (GPCR) protein sequences from http://www.gpcr.org/7tm/. I manually obtained query sequence data by following a trail of links for each entry until FASTA resources could be found. Entries with the following SWISS-PROT identifiers were skipped, because their FASTA sequences could not be found at the time of sequence collection: O45338_CAEEL, O45339_CAEEL, O45967_CAEEL, O45984_CAEEL, O62076_CAEEL, O62506_CAEEL, P91439_CAEEL, Q18687_CAEEL, Q21690_CAEEL, Q23265_CAEEL, Q3Y401_CAEEL, Q7M3K7_CAEEL, Q7M3K9_CAEEL, Q8MTW9_CAEEL, Q963E7_CAEEL, Q965H7_CAEEL, and YWO4_CAEEL. The *C. elegans* GPCR query protein sequence dataset contained 1,011 sequences.

I collected *C. elegans* Nuclear Hormone Receptor (NHR) protein sequences from http://www.ncbi.nlm.nih.gov/. I eliminated duplicate entries and entries representing proteins from non-unique genomic locations. The *C. elegans* NHR protein sequence data set contained 284 sequences.

I used the *C. elegans* GPCR and NHR query protein sequences in a Tera-BLASTP analysis against *M. hapla* gene freeze 1 translated in all six frames at E values of $10e^{-5}$, $10e^{-10}$, $10e^{-15}$, and $10e^{-20}$ with and without a low-complexity filter on. I imported the resulting BLAST results sets into a spreadsheet where the number of *C. elegans* query and *M. hapla* target hits and non-hits were tabulated. Because the same *M. hapla* target region could be repeatedly hit by multiple *C. elegans* queries, it was necessary to remove duplicate or overlapping target hits, sort the resulting list of unique regions, and enumerate target hits using the sort and word count functions provided in UNIX in order to yield the appropriate number of *M. hapla* genes which are members of the GPCR and NHR gene families. To further classify the unique target regions, I used the corresponding query names to place the targets into categories based on common keywords. In the event that a target region was hit by multiple queries, I used the name of the most significant query. I placed any protein name lacking descriptive text or containing the phrase ‘‘Temporarily Assigned Gene name family member’’ into the ‘‘UNCHARACTERIZED/OTHER’’ category.

I collected *C. elegans* collagen protein sequences from http://www.wormbase.org/. After I eliminated duplicate entries and entries from non-unique genomic locations, the *C. elegans* collagen protein sequence data set contained 182 sequences. I further reduced this data set by removing all sequences which did not contain a minimum of one instance of a $(Gly\text{-}X\text{-}Y)_3$ motif (that is, a triplet of Gly-X-Y, where Gly represents Glycine, and X and Y represent any amino acid). My manual removal of these entries resulted in a set of 165 *C. elegans* collagen protein sequences. Using *M. hapla* HapPep freeze 1 as a target, I employed a number of

methods to determine the true number of putative collagen proteins in *M. hapla*. First, I wrote a Perl script which counts the number of sequences with a minimum of one instance of an uninterrupted $(Gly\text{-}X\text{-}Y)_n$ motif, where Gly represents glycine, X represents any amino acid (including glycine), and Y represents any amino acid except glycine. The restriction on Y was imposed in order to avoid finding sequences with only a very low complexity glycine repeat (i.e. GGGGGGGGG). I tabulated the number of sequences meeting the "minimum of one instance of an uninterrupted $(Gly\text{-}X\text{-}Y)_n$ motif" criterion for three protein sets: a *C. elegans* protein set from http://www.sanger.ac.uk/Projects/C_elegans/Science98/, October1998 (19099 proteins total); the *C. elegans* collagen protein set from Wormbase (http://www.wormbase.org/) which was reduced to remove redundant entries as well as those not meeting the "minimum of one instance of a $(Gly\text{-}X\text{-}Y)_3$" criterion (165 proteins total); and *M. hapla* HapPepf1v1 (13336 proteins total). Finding the maximum value for n by which all (or almost all) proteins from the *C. elegans* reference set of 165 qualified as collagens and applying this same value to the *M. hapla* protein set filter to enumerate the number of qualifying *M. hapla* proteins was one method I used in the estimation of the true number of *M. hapla* collagen proteins. Second, I performed a manual (visual) screen using n={4,5,6,7}, partitioning the proteins into "true collagens" and "questionable/false collagens". I used the *C. elegans* collagen protein set of 165 as a reference for the visual comparison. Third, I conducted a Tera-BLASTP analysis using the *C. elegans* reference set of 165 collagen proteins as the query, and *M. hapla* HapPep1 as the target, using e-values of e^{-5}, e^{-10}, e^{-15}, e^{-20}, and e^{-25}, with and without a low complexity filter on. I compared the results against the manually-screened set of *M. hapla* proteins that were initially filtered

down by using the "minimum of one instance of an uninterrupted $(Gly\text{-}X\text{-}Y)_n$ motif, where Gly represents Glycine, X represents any amino acid (including Glycine), and Y represents any amino acid except Glycine" Perl script with n=4. I inspected those proteins identified by Tera-BLASTP which were not identified by the manual screen to determine whether they were a predicted true collagen. I used the combined methods to arrive at a confident number and list of true *M. hapla* collagen proteins from HapPep freeze 1.

Results

C. elegans and *M. hapla* GPCR, NHR, and collagen gene family membership is compared in Figure 1. *M. hapla* exhibits drastic gene reductions in all three families, with the most drastic reduction in GPCRs (1,011 *C. elegans* genes versus 147 *M. hapla* genes). The differences within NHRs (284 *C. elegans* genes versus 76 *M. hapla* genes) and collagens (165 *C. elegans* genes versus 81 *M. hapla* genes) are also significant. In order for *M. hapla* to maintain the same proportion of members to total protein-coding genes as *C. elegans*, it would need to contain 688, 193, and 112 GPCR, NHR, and collagen genes, respectively. Conversely, in order for *C. elegans* to maintain the same proportion of members to total protein-coding genes as *M. hapla*, it would need to contain 216, 112, and 119 GPCR, NHR, and collagen genes, respectively.

Within the *M. hapla* GPCR family, GPCR subclasses are also greatly reduced compared to *C. elegans* (see Opperman et al., 2008 Figure S2). Many of the GPCRs have an unknown function, while others encode nervous system components with roles in signal transduction.

Of those GPCRs characterized based on query keywords, the largest groups function as receptors for dopamine, serotonin, octopamine, and acetylcholine.

The number of proteins in *C. elegans* and *M. hapla* meeting various collagen motif filter criteria are shown in Table 1. It is noteworthy that using a minimum collagen repeat number of 7 yields almost the same sensitivity as a minimum collagen repeat number of 4 (80 true collagens versus 81 true collagens, respectively), but with a greatly improved specificity (0 false collagens versus 64 false collagens, respectively) (Table 1). Examples of collagen true and false positives are given in Figures 2 and 3, respectively.

Discussion

G Protein Coupled Receptors comprise the largest family of genes in *C. elegans*, but this family is greatly reduced in *M. hapla*. One possible explanation for the *M. hapla* reduction is the evolution towards a more specialized niche inside the plant, requiring more specialized olfactory receptors for a smaller range of stimuli emitted from within the protected microenvironment of the host. It is also possible, but less likely, that *M. hapla* is able to achieve environmental awareness with a fewer number of more generalized receptors each capable of detecting a broader class of compounds. Conversely, the free-living *C. elegans* must traverse a wide range of microenvironments during its life, thus requiring the ability to detect and respond to a wide range of potentially toxic chemicals (such as those with high osmolarity or acidic pH, heavy metals, bitter alkaloids, or detergents (Bargmann, 2006a)), odors signaling proximity to food sources, and other stimuli via the chemosensory GPCRs,

particularly given the absence of visual and auditory systems (Robertson and Thomas, 2006). This theme of diversity as a means by which to cope with highly variable environments has been previously noted at the genome-level, where diversity in sensory recognition, neuronal excitability, information transfer, and cell-cell recognition emerge as integral properties of *C. elegans*, enabled through olfactory receptors, potassium channels, neurotransmitter receptors, and gap junctions, respectively (Bargmann, 1998). Furthermore, after finding the host root and migrating through the vascular cylinder to ultimately cultivate its giant cell food source, *M. hapla* is nonmotile, so requires little feedback between immediate external environment and its cuticle, in contrast to *C. elegans*, which must respond to both external and internal stimuli even for seemingly simple feats of movement. Indeed, in addition to the detection of external cues, it is likely that some of the *C. elegans* GPCRs have been recruited to roles such as monitoring of internal chemistry (Robertson and Thomas, 2006). *C. elegans* prefers an oxygen environment between 4% and 12%, monitored by GPCRs (Bargmann, 2006a), while the oxygen concentration preferred by *M. hapla* is not known. However, it is possible that because *M. hapla* no longer has control over its oxygen environment after it feeds from the giant cell, any GPCRs responsible for oxygen monitoring would be useless, leading to the loss of *M. hapla* oxygen-detecting GPCRs, and possibly accounting for a small portion of the GPCR count difference between species. Chemosensation also has a major role in the entry into and the exit from the *C. elegans* alternative dauer larva stage: temperature, density, and food cues are monitored both to decide when to form dauer larvae and when to resume normal development (Bargmann, 2006a). Differences in dauer-related monitoring needs between *C. elegans* and *M. hapla* may account for minor differences in GPCR family sizes,

but it is not completely clear why *M. hapla* would require significantly fewer receptors for this process. Determining whether a gene family expansion through multiple duplications was involved to generate the expansive *C. elegans* GPCR gene family, or whether a reduction in *M. hapla* GPCRs due to weak selective pressure of receptors capable of detecting rarely-encountered compounds from the external environment (leading to nonfunctional pseudogenes and progressive gene loss) occurred, will require a more thorough and precise understanding of the phylogenetic history of these two species. Chemosensory GPCRs are often found in *C. elegans* clusters containing ten or fifteen closely related genes (with heaviest concentration on the arms of chromosome V) – an extreme density compared to most other *C. elegans* genes, where clusters of more than three are unusual (Bargmann, 1998). This observation serves as intriguing support for the hypothesis that because clustering provides an opportunity for gene addition or loss through unequal sister chromatid exchange, a rapid expansion of *C. elegans* GPCRs may have occurred by repeated unequal exchanges of chromatids containing GPCR clusters. GPCR gene positions are consistent with a history of rare large chromosomal rearrangements (with most suspected to be inversions) and frequent local gene duplications (Robertson and Thomas, 2006). Regardless of the specific means by which the *C. elegans* GPCR family evolved to greatly outnumber the *M. hapla* GPCR family, the results of this gene family comparison underscores the likely pronounced increased requirement for chemical compound-detection provisions in the free-living nematodes over the root-knot nematodes, whose shelter is provided by the plant host for much of its life.

The Nuclear Hormone Receptors have similarly undergone either a great reduction in *M. hapla*, or a great expansion in *C. elegans*. Much of this difference might potentially be attributed to the key roles of NHRs in integrating the complexities of homeostasis and development (Maglich et al., 2001). Given that *C. elegans* is exposed to a much wider range of solutes, ions, and solutions in its traversals through various environments than *M. hapla*, which is largely restricted to a plant host (which is itself in homeostasis), *C. elegans* likely has a greater need for the regulation of gene transcription events to mobilize reactive responses to fluctuating conditions than *M. hapla*. As noted previously, *nhr-8* has a role in resistance to the colchicines and chloroquine toxins, and likely represents one of several NHR genes in *C. elegans* necessary for ameliorating chemical threats. These threats are likely fewer, rarer, and more predictable within the environment of the host plant. Some aspects of nematode development are broadly regulated by diffusible signals; for instance, specific external stimuli may trigger the production of ecdysis-regulating chemicals through the action of NHRs (Sluder and Maina, 2001). Examples of external ligands eliciting a response in the nematode would tend to further expand the candidate list of *C. elegans* cognate receptors over those of *M. hapla*, due to the insular host microenvironment of *M. hapla*. On the other hand, plant parasitic nematodes could potentially recruit NHRs for the purposes of responding to the hormonal state of their hosts (Sluder and Maina, 2001). Over the course of its life, *C. elegans* is a much more motile organism than *M. hapla*, with a larger, more complex frame supporting greater neuron development. Greater neuron complexity relative to *M. hapla* would explain the larger NHR family in *C. elegans*, as a major role of NHRs is neural development and function (Sluder and Maina, 2001). In addition, *C. elegans*

is more likely to encounter microenvironments and periods where continuous growth is not favored. As a solution, one NHR, DAF-12, is employed for the development of the dauer larva (an alternative diapaused larval stage of variable duration), which enables *C. elegans* to ration its energy reserves until conditions improve (Sluder and Maina, 2001). However, *daf-12* orthologs have been observed in parasitic nematodes, where it is thought they function similarly to trigger development through infective stages. As seen with GPCRs, numerous gene duplications have contributed to the NHR abundance in *C. elegans* (Maglich et al., 2001), so redundancy or non-functionality in some genes is an additional possibility in gene family membership disparity.

The collagen gene family has additionally experienced either a contraction in *M. hapla* or an expansion in *C. elegans*, of a less dramatic, but still noteworthy, nature. A normalization of the two species' collagen counts based on total number of protein-coding genes yields a less impressive difference. The most parsimonious explanation for the remainder of the biologically significant difference in collagen family membership lies in the increased requirement of *C. elegans* cuticle function and complexity compared to *M. hapla*. The *C. elegans* cuticle must accommodate movement in a wide variety of terrain (e.g., abrasive soil particles, rough vegetable matter, bacterial biofilm deposits) over much of its life, requiring both sufficient strength to offer protection from hostile environments, and flexibility to allow for rapid escape from toxins and predators. In contrast, *M. hapla* is only exposed to similar conditions in stages prior to host infection (roughly one-quarter of its life), while enjoying the protection and homeostatic environment of its plant host, in a highly sedentary state

(requiring little cuticle flexibility), for the remainder of its life. Therefore, *M. hapla* would be expected to require fewer collagens due to its lower lifestyle mean cuticle strength, flexibility, and mobility. It is likewise possible that due to more integral importance of the *C. elegans* cuticle (and hence, collagens) relative to *M. hapla*, a greater degree of genetic redundancy exists in order for the species to withstand mutants leading to defects in individual collagen genes. Evidence for redundancy (or at least, very subtle differences in phenotype or function) has been gleaned from the highly-researched *C. elegans* body morphology defects (von Mende et al., 1998; Page and Johnstone, 2007). As collagen mRNA lengths have previously been found inadequate to individually account for the mass of individual cuticle collagen proteins (von Mende et al., 1988), an alternative explanation could be that a disparity exists in the degree by which individual collagen chains are cross-linked between the two species due to factors indirectly correlated with lifestyle. For instance, the smaller body size of *M. hapla* might allow for a more efficiently constructed cuticle, with fewer crosslink types required for structural integrity, than *C. elegans*, due to universal engineering principles. Moreover, gene expression levels likely differ between the various collagen genes at different points in cuticle construction for both species, so abundant collagen proteins may be explained by an increased gene dose. Regardless of the underlying reasons for collagen disparity between the species, collagens undoubtedly play a critical role in their specific lifestyle modes. The sequencing of additional nematode species and the in-depth analysis of their corresponding collagen gene families will enhance the understanding of how specific biological factors account for collagen gene family membership.

Through the generation of multiple complete, draft, and partial nematode genome sequences, nematodes have been shown to possess greater molecular diversity than was previously recognized (Mitreva et al., 2005). A recurrent question in gene family analysis between any two species is which species experienced a reduction, or gene loss, and which species experienced an expansion, or gene gain. While clues to the answer can often be found through the incorporation of additional related species, some facets of the question may never be fully answerable due to complex evolutionary histories which preclude disentanglement of molecular events. In such cases, we can only offer our best unbiased guesses of the larger evolutionary undercurrents based on knowledge of the underlying biology of the species. While a tendency exists in evolutionary analyses to believe that gene gain is the sole mechanism of increased fitness, gene loss through mutation is another avenue by which fitness can increase. In fact, Olson (1999) proposes that loss of gene function may represent a common evolutionary response of populations experiencing an environmental shift (and a change in selective pressure patterns). This idea is often referred to as the "less is more" hypothesis, and holds that adaptive loss of gene function may occur frequently and spread rapidly through small populations. Extrapolating this idea to whole genomes, studies on yeast gene ablation reveal that 85% of yeast genes can be deleted without an effect on haploid viability (Goebl and Petes, 1986), although the complete battery of environmental tests has not been (and probably can never be) fully mimicked in the laboratory to fully evaluate the complete effect of such a large number of deletions. Some yeast strains with disrupted genes have also exhibited better growth than the "wild type" strain on a rich medium (Goebl and Petes, 1986). Therefore, specialized ecological niches may exist in which selection favors

fewer than the complete set of functional genes (Olsen, 1999). Such sentiment might very well be appropriately directed towards a possible reduction in the three *M. hapla* gene families analyzed. Once within the confines of a hospitable plant host, an increase in fitness through the loss of genes previously required for independent foraging in more hostile environments would be fairly unremarkable. On the other hand, previous research would also support the converse view (of gene gains in *M. hapla,* if *C. elegans* and *M. hapla* results were transposed), due to the continuous coevolution between plant host immune response and parasite infection tactics in *M. hapla* that drive accelerated molecular evolutionary change (Mitreva et al., 2005). An even more likely and balanced synopsis of evolutionary forces explaining gene family memberships between the two species is a combination of gene gains and gene losses, with a net *C. elegans* gain in the three analyzed families. This explanation is reflective of the required emphasis on diverse *C. elegans* functions (along with redundancy via duplication) over the rapidly evolving *M. hapla* functions in each of the gene families. As a greater number of nematode genomes are sequenced, annotated, and analyzed in the coming years, a more direct relationship will undeniably emerge between lifestyle and composition of the large gene families.

References

Adams E (1978). Invertebrate collagens: marked differences from vertebrate collagens appear in only a few invertebrate groups. *Science* 202:591-597.

Agrios GN (1997). Plant diseases caused by nematodes. In *Plant Pathology*. Ed. Agrios GN. New York: Academic Press: 565-597.

Ashburner MA, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G (2000). Gene ontology: tool for the unification of biology. *Nature Genet.* 25:25-29.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, and Zygouri C (2003). PRINTS and its automatic supplement, prePRINTS. *Nucl. Acids Res.* 31:400-402.

Bader GD, Betel D, and Hogue CW (2003). BIND: the Biomolecular Interaction Network Database. *Nucl. Acids Res.* 31:248-250.

Bairoch A (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* 19:2241-2245.

Bairoch A (1993). The ENZYME data bank. *Nucl. Acids Res.* 22:3626-3627.

Bairoch A (2000). The ENZYME database in 2000. *Nucl. Acids Res.* 28:304-305.

Bargmann CI (1998). Neurobiology of the *Caenorhabditis elegans* genome. *Science* 282:2028-2033.

Bargmann CI (2006a). Chemosensation in *C. elegans* (October 25, 2006), *WormBook*, ed. The *C. elegans* Research Community, WormBook, http://www.wormbook.org.

Bargmann CI (2006b). Comparative chemosensation from receptors to ecology. *Nature* 444:295-301.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, and Sonnhammer ELL (2002). The Pfam protein families database. *Nucl. Acids Res.* 30: 276-280.

Bird DMcK and Kaloshian I (2003). Are roots special? Nematodes have their say. *Physiological and Mol. Plant Pathol.* 62:115-123.

Buehler MJ (2006). Nature designs tough collagen: Explaining the nanostructure of collagen fibrils. *PNAS* 103:12285-12290.

Corpet F, Gouzy J, and Kahn D (1998). The ProDom database of protein domain families. *Nucl. Acids Res.* 26:323-326.

Corpet F, Servant F, Gouzy J, and Kahn D (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucl. Acids Res.* 28:267-269.

Cox GN, Kusch M, and Edgar RS (1981). Cuticle of *Caenorhabditis elegans*: its isolation and partial characterization. *J. Cell Biol.* 90:7-17.

Davies MN, Gloriam DE, Secker A, Freitas AA, Mendao M, Timmis J, and Flower DR (2007). Proteomic applications of automated GPCR classification. *Proteomics* 7:2800-2814.

Dodge C, Schneider R, and Sander C (1998). The HSSP database of protein structure-sequence alignments and family profiles. *Nucl. Acids Res.* 26:313-315.

Enright AJ, Kunin V, and Ouzounis CA (2003). Protein families and TRIBEs in genome sequence space. *Nucl. Acids Res.* 31:4632-4638.

Enright AJ, Van Dongen S, and Ouzounis CA (2002). An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* 30:1575-1584.

Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJA, Hofmann K, and Bairoch A (2002). The PROSITE database, its status in 2002. *Nucl. Acids Res.* 30:235-238.

Fratzl P (2008). Collagen: Structure and mechanics. New York: Springer.

Garrels JI (1996). YPD – a database for the proteins of *Saccharomyces cerevisiae*. *Nucl. Acids Res.* 24:46-49.

Gissendanner CR, Crossgrove K, Kraus KA, Maina CV, and Sluder AE (2004). Expression and function of conserved nuclear receptor genes in *Caenorhabditis elegans*. *Dev. Biol.* 266:399-416.

Goebl MG and Petes TD (1986). Most of the yeast genome sequences are not essential for cell growth and division. *Cell* 26:983-992.

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, and White O (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucl. Acids Res.* 29:41-43.

Haft DH, Selengut JD, and White O (2003). The TIGRFAMs database of protein families. *Nucl. Acids Res.* 31:371-373.

Henikoff S and Henikoff JG (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* 19:6565-6567.

Henikoff S, Henikoff JG, and Pietrokovski S (1999). Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15:471-479.

Hodges PE, McKee AH, Davis BP, Payne WE, and Garrels JI (1999). The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucl. Acids Res.* 27:69-73.

Holm L, Ouzounis C, Sander C, Tuparev G, and Vriend G (1992). A database of protein structure families with common folding motifs. *Protein Sci.* 1:1691-1698.

Holm L and Sander C (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* 26:316-319.

Horn F, Lau AL, and Cohen FE (2003). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Nucl. Acids Res.* 31:294-297.

Huang JY and Brutlag DL (2001). The EMOTIF database. *Nucl. Acids Res.* 29:202-204.

Kabsch W and Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.

Karp PD, Riley M, Paley SM, and Pellegrini-Toole A (1996). EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* 24:32-39.

Kanehisa M, Goto S, Kawashima S, and Nakaya A (2002). The KEGG databases at GenomeNet. *Nucl. Acids Res.* 30:42-46.

Klabunde T and Hessler G (2002). Drug design strategies for targeting G-protein coupled receptors. *Chem. Bio. Chem.* 3:928-944.

Krause A, Haas SA, Coward E, and Vingron M (2002). SYSTERS, GeneNet, SpliceNest: exploring sequence space from genome to protein. *Nucl. Acids Res.* 30:299-300.

Kriventseva EV, Fleischmann W, Zdobnov EM, and Apweiler R (2001). CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucl. Acids Res.* 29:33-36.

Kriventseva EV, Servant F, and Apweiler R (2003). Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucl. Acids Res.* 31:388-389.

Kusch M, and Edgar RS (1986). Genetic studies of unusual loci that affect body shape of the nematode *Caenorhabditis elegans* and may code for cuticle structural proteins. *Genetics* 113:621-639.

Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, and Bork P (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucl. Acids Res.* 30:242-244.

Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, and Murzin AG (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res.* 30: 264-267.

Maglich JM, Sluder A, Guan X, Shi Y, McKee DD, Carrick K, Kamdar K, Willson TM, and Moore JT (2001). Comparison of complete nuclear receptor sets from the human, *Caenorhabditis elegans* and *Drosophila* genomes. *Genome Biol.* 2:research0029.1-0029.7.

Mellor JC, Yanai I, Clodfelter KH, Mintseris J, and DeLisi C (2002). Predictome: a database of putative functional links between proteins. *Nucl. Acids Res.* 30:306-309.

Mitreva M, Blaxter ML, Bird DM, and McCarter JP (2005). Comparative genomics of nematodes. *TRENDS Genet.* 21:573-581.

Miyabayashi T, Palfreyman MT, Sluder AE, Slack F, and Sengupta P (1999). Expression and function of members of a divergent nuclear receptor family in *Caenorhabditis elegans*. *Dev. Biol.* 215:314-331.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleishmann W, Griffith-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJA, Vaughan R, and Zdobnov EM (2003). The InterPro database, 2003 brings increased coverage and new features. *Nucl. Acids Res.* 31:315-318.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lousdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowell J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJA, Silventoinen V, Studholme DJ, Vaughan R, and Wu CH (2005). InterPro, progress and status in 2005. *Nucl. Acids Res.* 33:D201-D205.

Murzin AG, Brenner SE, Hubbart T, and Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536-540.

Nevill-Maning CG, Wu TD, and Brutlag DL (1998). Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA* 95:5865-5871.

Novac N and Heinzel T (2004). Nuclear receptors: overview and classification. *Curr. Drug Targets – Inflammation & Allergy* 3:335-346.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, and Kanehisa M (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.* 27:29-34.

Olson MV (1999). When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* 64:18-23.

Ouzounis CA, Coulson RMR, Enright AJ, Kunin V, and Pereira-Leal JB (2003). Classification schemes for protein structure and function. *Nature Rev. Genet.* 28:508-519.

Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, Cromer J, Diener S, Gajan J, Graham S, Houfek TD, Liu W, Mitros T, Schaff J, Schaffer R, Scholl E, Sosinski

BR, Thomas VP, and Windham E (2008). Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. *PNAS* 105:14802-14807.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, and Thornton JM (1997). CATH – a hierarchic classification of protein domain structures. *Structure* 5:1093-1108.

Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr, Kyrpides N, Fonstein M, Maltsev N, and Selkov E (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucl. Acids Res.* 28:123-125.

Page AP, and Johnstone IL (2007). The cuticle (March 19, 2007), *WormBook*, ed. The *C. elegans* Research Community, WormBook, http://www.wormbook.org.

Parkinson J, Mitreva M, Whitton C, Marian T, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH, McCarter JP, and Blaxter ML (2004). A transcriptomic analysis of the phylum Nematoda. *Nat. Genet.* 36:1259-1267.

Pearl FMG, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, and Orengo CA (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucl. Acids Res.* 31:452-455.

Politz JC, and Edgar RS (1984). Overlapping stage-specific sets of numerous small collagenous polypeptides are translated in vitro from *Caenorhabditis elegans* RNA. *Cell* 37:853-860.

Remm M, Storm CE, and Sonnhammer EL (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314:1041-1052.

Rigoutsos I, Huynh T, Floratos A, Parida L, and Platt D (2002). Dictionary-driven protein annotation. *Nucl. Acids Res.* 30:3901-3916.

Robertson HM and Thomas JH (2006). The putative chemoreceptor families of *C. elegans* (January 06, 2006). *WormBook*, ed. The *C. elegans* Research Community, WormBook, http://www.wormbook.org.

Sander C and Schneider R (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9:56-68.

Sasser JN and Freckman DW (1987). A world perspective on nematology: the role of the society. In *Vistas on Nematology*. Ed. Veech JA and Dickson DW. Hyattsville: Society in Nematology: 7-14.

Schultz J, Milpets F, Bork P, and Ponting CP (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA* 95:5857-5864.

Schwarz EM (2005). *Genomic classification of protein-coding gene families* (September 23, 2005). *WormBook*, ed. The *C. elegans* Research Community, WormBook, http://www.wormbook.org.

Shoop E, Silverstein KA, Johnson JE, and Retzel EF (2001). MetaFam: a unified classification of protein families. II. Schema and query capabilities. *Bioinformatics* 17:262-271.

Silverstein KA, Shoop E, Johnson JE, and Retzel EF (2001). MetaFam: a unified classification of protein families. I. Overview and statistics. *Bioinformatics* 17:249-261.

Sluder AE and Maina CV (2001). Nuclear receptors in nematodes: themes and variations. *TRENDS Genet.* 17:206-213.

Sonnhammer EL, Eddy SR, and Durbin R (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28:405-420.

Storm CE and Sonnhammer EL (2003). Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res.* 13:2353-2362.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, and Natale DA (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41.

Tatusov RL, Koonin EV, and Lipman DJ (1997). A genomic perspective on protein families. *Science* 278:631-637.

von Mende N, Bird DM, Albert PS, and Riddle DL (1988). dpy-13: a nematode collagen gene that affects body shape. *Cell* 55:567-576.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, and Yaschenko E (2006). Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* 00 (Database issue):D1-D8

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S, and Eisenberg D (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* 30:303-305.

Yanai I, Derti A, and DeLici C (2001). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl Acad. Sci. USA* 98:7940-7945.

Yona G, Linial N, and Linial M (1999). ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins* 37:360-378.

Yona G, Linial N, and Linial M (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucl. Acids Res.* 28:49-55.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, and Cesareni G (2002). MINT: a Molecular INTeraction database. *FEBS Lett.* 513:135-140.

Zhang Z, Burch PE, Cooney AJ, Lanz RB, Pereira FA, Wu J, Gibbs RA, Weinstock G, and Wheeler D (2004). Genomic analysis of the nuclear hormone receptor family: new insights into structure, regulation, and evolution from the rat genome. *Genome Res.* 14:580-590.

Zmasek CM and Eddy SR (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3:14.

Tables

Table 1. Number of Wormbase-annotated *C. elegans* collagen proteins and *M. hapla* HapPep1 proteins meeting various collagen motif filters.

Number of tandem repeats of (GXY), where G represents Glycine, X represents any amino acid, and Y represents any amino acid except for Glycine	*C. elegans* non-redundant collagen-annotated proteins from Wormbase containing this motif	*M. hapla* HapPep1 proteins containing this motif	*M. hapla* HapPep1 true positives (based on manual screen)	*M. hapla* HapPep1 false positives (based on manual screen)
3	165	830	N/A	N/A
4	165	145	81	64
5	164	89	80	9
6	164	81	80	1
7	164	80	80	0
8	164	80	N/A	N/A
9	164	80	N/A	N/A
10	164	78	N/A	N/A
11	162	77	N/A	N/A
12	156	75	N/A	N/A
13	144	67	N/A	N/A
14	138	64	N/A	N/A
15	133	63	N/A	N/A
16	131	63	N/A	N/A
17	128	60	N/A	N/A
18	126	60	N/A	N/A
19	124	59	N/A	N/A
20	119	54	N/A	N/A

Figures

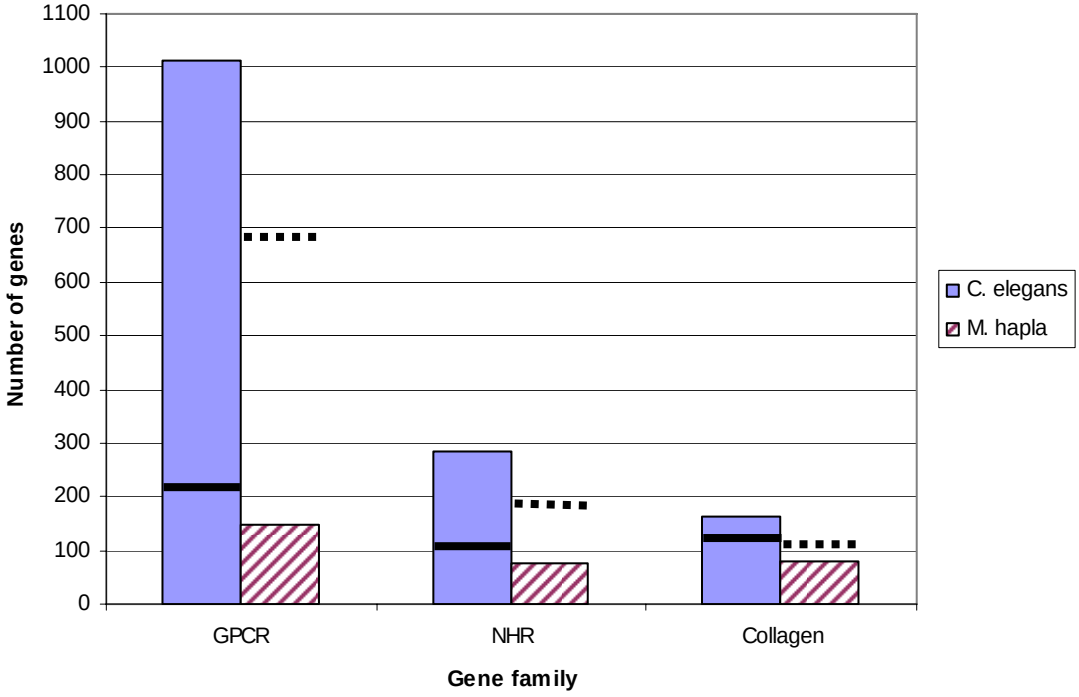

Figure 1. *C. elegans* and *M. hapla* GPCR, NHR, and collagen gene family membership comparison. The thick solid line indicates the number of *C. elegans* members for each family that would be needed to maintain the same proportion of members to total genes as *M. hapla*. The thick dotted line indicates the number of *M. hapla* members for each gene family that would be needed to maintain the same proportion of members to total genes as *C. elegans*.

>Mh10g200708_Contig78:15801..17727
MKVLASATTSSEGTIGNQQRTIPRRNLKHEIKNGKNNADGEWEEDENEDEEAGRDTAYKLVSFTAVGFSLLAVL
SVVICMPIVYNFVEHVQRQTRRELDFCKGSARDIMIEVDDKQRRLAAAAQIAKNQTDFASLITHIQKRAKREAS
GGSCDSCCIPGHAGRPGQPGPSGTHGTPGAPGRPGNPGRPPIICEEVEIPPCNPCPPGPPGRVGPNGQQGSPGR
PGASGRPGNNGLPGAPGLSGPPGMPGNAGADGEKGEPGRPAASSPSSPGEPGPQGEQGPQGLPGADGSPGREGS
PGQSGPPGLPGGSGKEKDRDCPQGPPGQPGPPGSPGVAGQQGERGICPKYCALDGGVFFEDGTRR

>Mh10g200708_Contig1040:24309..25842
MLETRLVIWLACGSSLLAVLATIVGIHQLYSDFSQLTHRIRGGVQAFRVDTDSAWIDLMDLQMAVTPLTKPREN
PFQSIFRPKRKVGSHCGCKARNSHCPPGPQGPPGTPGFPGQPGQAGQNGQPGMPGQTGPSCPVPSYDCQRCPAG
APGKQGAPGLPGPPGRGGAPGMPGKSSGIGPAGPPGPQGPPGPKGNCGGSGLPGQPGKDANINPVLPGPKGVSG
SPGQIGPPGPPGKPGAHGLQGPIGPPGKPGPQGNPGAPGDDGEDGLPGQPGQDAQYCPCPERSAMAKAAKGA

>Mh10g200708_Contig2152:2162..3535
MSIICSEGRVTGFIAISTTIIIGNLWTDLLVNTKKRVSVLSLSSDYDGRIKKRQASNPNCDCINTAQCPPGPPG
PAGQPGIDGEHGKPGEPGPAGEQALAANTQAPVDESCRVCPLGPQGLRGYPGPAGPPGAPGGPGLPGLAGKPGH
PGQIGSPGEMGEQGMAGKEGMPGPPGLEGVRGTSGRQGEPGLAGMPGPKGYSGAPGRPGMPGLQGMKGEAGPEG
PPGQPGWDGMRGQFGENGQPGPDASYCKCPQRLSLTGEAPTITSSKNNVSIQPNPSPPSYGPKETNASSTTEEM
SSKEKEGYVS

Figure 2. Examples of *M. hapla* collagen true positives based on manual screening. Glycines and prolines are highlighted in green and yellow, respectively.

>Mh10g200708_Contig88:217673..220137
MGGSSSHMQPPAMLEHHQLPPSLIHDTTPFQTQQQQYQFPDMNQIFKGIPKMLQVADDMHRMSNYRLDDQSEQS
SLSVSRPSNNYRHYHSAKPIDNWTHKIEMDRMLDSTSTRHTNLTPQHGAHGTVGSHGGSHGAIMMNNKSALASS
SSPEHQHHYNSLPLHEQHQTSSVVATPSHTPTQIITGHSHQTNGQLKRLDPVKLIVDEMPYADNEHEQ

>Mh10g200708_Contig224:8573..13241
MTVLMNSELVQRLTTQLLDPKCVLNLDGLLANEQLVKSLIECRISDINFRVIKVIGRGAFGEVQLVRNAQNNQV
YAMKMLDKDKMIRRSDAAFFWEERNIMAYSNSEWIVKLHYAFQDLRNLYMVMEFMPGGDMVNLMANHDIPEDWA
QFYTAELVLALDAIHSLGYIHRDVKPDNMLISADGHIKLADFGTCIRMNKDGLVRCSMAVGTPDYISPEVLRSQ
GGEGVYGREVDWWAVGVFLYEMLFGETPFYAESLVQTYSKIMDHNKQLKFPGDVKVSSSAKDIIRSFLSEPNIR
LGRDGISGIRGHAFFKNSVWTFDTIQKSRPPYVPDLVGDDDTSHFDDVEPADPVDAESFQIPKAFTGNQLPFIG
FTYSNEFGPLDLIKKQLEEAVLRPIVNGNNNLLDTPMVNGTCVDGNNCHLEEKIDELTNENEVLQQQLKNIKEQ
LETETNLAKTEIDLLAKEREILEMKCAEMKQNQEMIESEVDSLKQEKESLTIRVKELEIEVGKLAPTKQEAEQL
KETNKMLEIHLQKCKEQLEQQQQINSHHLNNNNNTAATPAVPPPRLSHEVCFLHLLREALLVGELKYQRRVLEN
KIQDLQEQLEGQKRLFSDCQQRLEAEQKLSGLFKIEIEARTKDNEETEKKFQNMGQQIEEFVLTLERERVSHQH
LQNSFVDLQKDKAFLETELRAAMQRHEQEYKSLTASLNLASKKENELNSQNKQLQEEINQIRKHSHGPMGLAPV
YGGTNTLSSAVSTSSLVSATNSFASVQLHNNNNSILSSSGDLEMLSRDQLIHRCQREIKLKEQVIEKLAYLGRQ
KGINDDSAGRANSKKKKKHVDEKRVRDLVEFQMEQERKKYLQKIAQLEEIIQDLHQNLVEEQRKKADLIDEVNL
YKKGGAEHIDLNSAQTFRDRRVPKRRQGGGGQYQNLENDNDPV

>Mh10g200708_Contig276:23145..24803
MKKILISSWRNVKKKRWPPKFSNSHFNISNDSKEKYRNKFIIILLLFLKILIPVRGQAFGPVMNQCPMGQVADG
CRPNQVFQQCPPGTFQDGFCPITGICFDRRETCFNGLCCQTIGQGQIPSPNQIYPPPNQIYPPNQIYPPNQIYP
PPNQIYPPPNQIYPPPNQIYPPNQIIPPNQIFPPNQIYPPRQIYPQNQIFPPNQIYPPNQIFPPNQIFPQNQIY
PNNIIPYPRQQLPGIGGYQPYPVGGLGGMGGFGIRPGGGGMSYGGLGGLGTGGLGGIGGLGAGGLGGMGGLGFG
GIGGMGGIGGIGFCIDYYPNCGFYTTYCFFSTIAEACMDSCGLCDLYKRK

Figure 3. Examples of *M. hapla* collagen false positives based on manual screening.

Glycines and prolines are highlighted in green and yellow, respectively.

Chapter 3: Identification of a RALF plant peptide hormone mimic in *Meloidogyne hapla*

Abstract

The plant-parasitic root-knot nematode *Meloidogyne hapla* is one of the primary instigators of worldwide crop damage. Understanding its mechanisms of host plant targeting, entry, and manipulation are integral objectives in a pest management program. I have identified one possible mechanism in the discovery of the first nematode-encoded Rapid Alkalinization Factor (RALF) plant peptide hormone. While the function of this nematode-encoded RALF is not completely understood, previous studies have demonstrated the inhibitory effect of exogenously-applied RALF on root growth and development. I speculate that an *M. hapla*-secreted RALF serves as a key driver in the series of events *M. hapla* triggers during host infection.

Introduction

Plant parasitic nematodes cause more than $120 billion in annual worldwide crop losses, as the greatest biotic stressors of plants (Koenning et al., 1999; Bird, 2004). Impacting all crops to some degree, they collectively utilize all parts of the plant in an elaborate manipulative scheme to derive sustenance. The preponderance of damage is caused by the approximately 80 species of the sedentary endoparasitic tylenchid genus *Meloidogyne*. Hatching in the soil, they follow cues to the host root, penetrating the zone of elongation. The apoplast serves as a migration route to the vascular cylinder, where permanent feeding sites known as giant cells

are formed. Root-knot nematodes feed exclusively from giant cells, but do so extracellularly (Bird, 2004).

Small lipophilic plant hormones have been long known to play roles in internal plant interactions. Across many developmental stages, hormones such as abscisic acid, auxins, brassinosteroids, cytokinins, ethylene, gibberellins, and jasmonates mediate comprehensive cell-to-cell communication (Matsubayashi and Sakagami, 2006). In the past decade, peptide signals (both secretory and nonsecretory) have also been uncovered as key players in plant growth regulation, including the specific areas of callus growth, defense responses, leaf-shape regulation, meristem organization, nodule development, organ abscission, and self-incompatibility (Farrokhi et al., 2008). These plant peptide hormones provide one avenue through which root-knot nematodes may manipulate their hosts. A key discovery in plant nematology within the last nine years has been that root-knot and cyst nematodes secrete peptides that mimic the clavata-like element (CLE) class of plant signaling ligands (Wang et al., 2001; Gao et al., 2001, 2003; Huang et al., 2003). Functional analysis performed by several groups has demonstrated these nematode-encoded hormone mimics to be biologically active in plants (Wang et al., 2005; Huang et al., 2006b; Mitchum et al., 2008), making them strong candidates for roles in establishing feeding sites involved in parasitic interactions (Huang et al., 2006a).

A 49 aa plant peptide hormone known as Rapid Alkalinization Factor (RALF) was originally identified in the purification process of tobacco systemins (Pearce et al., 2001). Its etymology

is based on its ability to rapidly induce an alkalinization of plant suspension culture medium. While its specific role in growth has not been fully elucidated, a synthesized tomato homolog of RALF was shown to cause an arrest of root growth and development in tomato and *Arabidopsis* seedlings. A 115 aa-encoding tobacco cDNA was isolated with identity to RALF at its C-terminus, affirming that RALF occurs as a processed peptide with an N-terminal signal prior to cleavage (Pearce et al., 2001; Olsen et al., 2002). RALF precursor genes are widely found throughout the plant kingdom with highly conserved homologs, and not restricted to any particular plant cell type, suggestive of an essential physiological role (Pearce et al., 2001; Wu et al., 2007). A putative RALF receptor, consisting of two protein species, has been identified in plant plasma membranes (Scheer et al., 2005). Molecular cloning and loss-of-function mutational analysis using this RALF receptor should reveal additional clues leading to the determination of plant RALF function (Matsubayashi and Sakagami, 2006).

Canonical RALF features include an N-terminal signal peptide and a conserved C-terminal motif containing two cysteines. Less commonly conserved are a mono- or dibasic motif approximately four residues upstream from, and a YIXY motif positioned a few residues downstream of, the N-terminus of the mature region (Germain et al., 2005). An acidic region is often found between the signal peptide and mature RALF peptide, which is possibly involved in signaling (Pearce et al., 2001; Farrokhi et al., 2008). Some RALF sequences are characterized by C-terminal extensions, similar to the CLE family in *Arabidopsis,* and could determine binding specificity to different receptors (Olsen et al., 2002). Two pairs of

conserved cysteines are found in Tobacco RALF (which have been demonstrated to form disulfide bonds (Pearce et al., 2001)), as well as *Arabidopsis* and rice RALF-like sequences, indicative of a similar disulfide bonding pattern and overall structure (Olsen et al., 2002). Secondary structure predictions for RALF-like sequences paint a landscape of coil-dominated structure peppered with beta-strands in the most highly conserved regions (Olsen et al., 2002).

Through screening for a diverse group of plant peptide hormones using Smith-Waterman local sequence alignment, I report the first identification of RALF in any nematode species. By use of thorough bioinformatics predictions, the diversity of tactics parasitic nematodes employ to manipulate their hosts appears even more expansive than what was recently believed.

Results

Preliminary identification of RALF in M. hapla

The first indication of a putative RALF mimic in *M. hapla* was gleaned through a Smith-Waterman analysis of assorted RALF queries against translated *M. hapla* genomic contigs. *Arabidopsis thaliana* RALF-LIKE 28 was a statistically significant hit (e-value $8.6e^{-7}$) to a region in *M. hapla* genomic contig 2062 (MhC2062).

Reverse-BLASTing of MhC2062 back to NR

The second logical step in the investigation was to BLASTP an extended region (delineated by the nearest surrounding stop codons) of the initial significant Smith-Waterman match to NR in order to affirm the initial alignment. *A. thaliana* RALF-LIKE 2 was the most significant hit (but its e-value was poor (2.8)). BLASTP of this same *M. hapla* region against UNIPROT returned three uncharacterized *A. thaliana* proteins of a length similar to previously annotated plant RALFs (85-97 aa). All three *A. thaliana* sequences contained both the YIXY motif as well as the highly conserved C-terminal domain.

MhC2062 exhibits similarity to Tobacco RALF protein

In a second Smith-Waterman analysis (with altered parameters), the canonical mature Tobacco RALF sequence was used as a query against translated *M. hapla* genomic contigs to further assess sequence similarity. MhC2062 was the only significant target (e-value 0.006138) of Tobacco RALF. Considering the general sequence dissimilarity between tobacco and nematodes (regardless of possible convergent evolution (i.e. protein mimicry) of this specific *M. hapla* sequence), an aligned sequence with an e-value less than 0.01 is potentially promising.

To further bolster the support for the existence of a RALF mimic in *M. hapla*, the same canonical mature Tobacco RALF sequence was used as a Smith-Waterman query against translated *M. incognita* genomic contigs (with altered parameters). Translated sequences derived from genomic contigs 165 (e-value 0.001663) and 186 (e-value 0.004313) were the

only significant targets to the query. In contrast, no significant targets were detected when this same analysis was repeated on translated *C. elegans* genomic contigs.

An annotated alignment of the translated *M. hapla* contig 2062 sequence, translated *M. incognita* contig 165 sequence, Tobacco RALF protein sequence, Arabidopsis RALF-LIKE 2 protein sequence, and Arabidopsis RALF-LIKE 28 protein sequence is shown in Figure 1. Predicted signal peptide cleavage sites occur at the same distance from the N-terminus for both the MhC2062 protein and for Tobacco RALF. Presumed mature regions of all proteins begin with an alanine, although this presents a conflict for the two *A. thaliana* RALF-LIKE proteins, where the predicted signal peptide cleavage site occurs immediately after this alanine. All proteins contain an XIXY motif (except the *M. incognita* contig 165 sequence, which contains XVXY (note that isoleucine (I) and valine (V) are both neutrally-charged hydrophobic residues)) 2-4 aa after the first alanine of the mature region (*M. incognita* lacks this alanine). Further, all proteins contain a highly-conserved C-terminal domain beginning with a tyrosine and possessing two cysteine residues presumably critical for the formation of a key disulfide bridge (the formation of a bridge at this region has been demonstrated in Tobacco RALF (Pearce et al., 2001; Olsen et al., 2002)). Four out of five proteins possess a "tail" C-terminal to this highly-conserved domain.

Secondary structure predictions

Secondary structure predictions in the N-terminal half for both MhC2062 and Tobacco RALF precursors were predominantly of alpha-helix structure, with a small (4-7 aa) beta-

strand component. However, in the C-terminal half (and particularly in the conserved C-terminal domain region), overall confidence in secondary structure predictions was lower, with a coiled structure (or other nonclassic structure) serving as the primary prediction.

Additional validating procedures

To further confirm the predictions of a RALF-LIKE sequence in *M. hapla* contig 2062, a number of additional methods were employed. Using queries of both the *M. hapla* sequence and Tobacco RALF, InterProScan detected the same RALF domain (IPR008801/PF05498) at the C-terminus of both sequences (e-values $6.1e^{-5}$ and $5.6e^{-39}$, respectively). InterProScan taxonomic coverage reveals that no RALF nematode sequence had been previously reported.

The automated gene prediction program GlimmerHMM predicts a "terminal exon" in the *M. hapla* genome corresponding to the predicted RALF mature frame and region, but does not make any prediction for the RALF precursor region N-terminal to the mature region. Conflicting information surrounds the region preceding the mature region: GeneDetective predicts both an intron and a frameshift within this area, while a manual inspection suggests the possibility of a mutation of a leucine to a stop codon. Nucleotides from the stop codon and subsequent alanine comprise the acceptor splice site sequence of "ag", which also complicates the prediction of exons and introns. Whichever scenario is true, the most critical regions of N-terminal signal peptide and mature C-terminal regions are at least moderately conserved amongst the *M. hapla* sequence and RALF and RALF-LIKE sequences.

Discussion

With the discovery that plant CLE mimics are encoded in nematode genomes, a backdrop has been set for other sinister and devious methods root-knot nematodes recruit in their plot to subserviate their hosts. The intriguing discovery in this work of a *M. hapla*-encoded RALF plant peptide mimic represents another tantalizing piece of the nematode-plant interaction puzzle. It is noteworthy that, like the plant CLEs, which are characterized by the conservation of a 14-amino acid sequence close to the C-terminus in plant species (Matsubayashi and Sakagami, 2006), the plant and nematode RALF (and RALF-LIKE) sequences similarly appear to be characterized by a 13-14 amino acid sequence near the C-terminus (Figure 1), suggesting that core sequence elements of plant peptide hormones may be sufficient to confer a functional role. While the specific function of this RALF mimic in *M. hapla* is not well-understood, it likely plays a key role in modifying plant development for the benefit of *M. hapla* survival. Exogenously-applied RALF has previously been shown to have an inhibitory effect on root growth and development, which implies a conservation of energy reserves within the host. It is possible that a secreted *M. hapla* RALF mimic modulates an integral component of the plant development cascade, shuttling energy and nutrients from the plant root source to the giant cell sink. A successful parasitism strategy from a genetic standpoint is one in which the host plant is kept alive at least long enough for one new generation of nematodes to arise. Therefore, the negative effects on plant root development (and the putative positive effects on giant cell induction) due to the secretion of a *M. hapla* RALF mimic is an evolutionarily justifiable side-effect of the overall strategy that *M. hapla* utilizes for parasitism.

Methods

Initial detection of sequence similarity

Starting with a query of assorted RALF and RALF-LIKE proteins collected from NCBI, the TimeLogic hardware-accelerated Smith-Waterman local sequence alignment algorithm was used with a database consisting of six-way translated *M. hapla* genomic contigs. The following parameters were used: similarity matrix = BLOSUM62, gap opening penalty = -12, gap extension penalty = -2, low complexity filter off, e-value cutoff = $10e^{-5}$. Resulting alignments were visually screened and subjected to additional analysis.

Reverse-BLASTing of M. hapla sequence back to NR and UNIPROT

An extended region (delineated by surrounding stop codons) of the *M. hapla* contig 2062 protein returned from the initial Smith-Waterman analysis was manually retrieved and used as a BLASTP query against NR and UNIPROT, with default parameters. CLUSTALX was used to construct a multiple sequence alignment for manual inspection of key domains.

Similarity of M. hapla sequence to canonical mature Tobacco RALF

Using the canonical mature Tobacco RALF as a query, the TimeLogic hardware-accelerated Smith-Waterman local sequence alignment algorithm was again used with a database consisting of six-way translated *M. hapla* genomic contigs. The following parameters were used: similarity matrix = BLOSUM65, gap opening penalty = -6, gap extension penalty = -2, low complexity filter off, e-value cutoff = 0.01.

Validation using M. incognita and C. elegans

Using the canonical mature Tobacco RALF as a query, the TimeLogic hardware-accelerated Smith-Waterman local sequence alignment algorithm was again used with databases consisting of six-way translated *M. incognita* and *C. elegans* genomic contigs. The following parameters were used: similarity matrix = BLOSUM62, gap opening penalty = -12, gap extension penalty = -2, low complexity filter off, e-value cutoff = 0.01.

Multiple sequence alignment of M. hapla sequence to M. incognita sequence, Tobacco RALF protein, and A. thaliana RALF-LIKE proteins

GeneDoc was employed to construct a multiple sequence alignment of the *M. hapla* Contig 2062 sequence of interest, *M. incognita* Contig 165 sequence of interest, Tobacco RALF, and *A. thaliana* RALF-LIKE 28 and RALF-LIKE 2 proteins. Multiple sequence alignment parameters were set as follows: constant cost length = 20, gap opening penalty = -11, gap extension penalty = -1. Manual modifications were subsequently made to the alignment.

Signal peptide predictions

SignalP V3.0 (http://www.cbs.dtu.dk/services/SignalP/) was used to check for the presence of an N-terminal signal peptide and its most likely cleavage site. TargetP V1.1 (http://www.cbs.dtu.dk/services/TargetP/) was subsequently used to predict most likely localization sites of proteins containing N-terminal signal peptides.

Secondary structure prediction

Secondary structure predictions were made using both Jpred3 (http://www.compbio.dundee.ac.uk/www-jpred/) and PredictProtein (http://www.predictprotein.org/).

Independent detection of RALF domain and associated features

As an independent verification step for the presence of a RALF domain, InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/) was used with default parameters. The gene finder GlimmerHMM was available via Gbrowse track in an in-house system, and was utilized to predict exon structure of *M. hapla* Contig 2062. In addition, GeneDetective (local TimeLogic installation) predicted introns, exons, and frameshifts.

References

Bird DMcK (2004). Signaling between nematodes and plants. *Curr. Opin. Plant Biol.* 7:372-376.

Farrokhi N, Whitelegge JP, and Brusslan, JA (2008). Plant peptides and peptidomics. *Plant Biotech. J.* 6:105-134.

Gao B, Allen R, Maier T, Davis EL, Baum TJ, and Hussey RS (2001a). Molecular characterisation and expression of two venom allergen-like protein genes in *Heterodera glycines*. *Int J Parasitol.* 31:1617-1625.

Gao B, Allen R, Maier T, Davis EL, Baum TJ, and Hussey RS (2001b). Identification of putative parasitism genes expressed in the esophageal gland cells of the soybean cyst nematode *Heterodera glycines*. *Molec Plant Microbe Interacts.* 14:1247-1254.

Gao B, Allen R, Maier T, Davis EL, Baum TJ, and Hussey RS (2003). The parasitome of the phytonematode *Heterodera glycines*. *Molec Plant Microbe Interact.* 16:720-726.

Germain H, Chevalier E, Caron S, and Matton DP (2005). Characterization of five RALF-like genes from *Solanum chacoense* provides support for a developmental role in plants. *Planta* 220:447-454.

Huang G, Gao B, Maier T, Allen R, Davis EL, Baum TJ, and Hussey RS (2003). A profile of putative parasitism genes expressed in the esophageal gland cells of the root-knot nematode *Meloidogyne incognita. Mol Plant Microbe Interact.* 16:376-381.

Huang G, Allen R, Davis EL, Baum TJ, and Hussey RS (2006a). Engineering broad root-knot resistance in transgenic plants by RNAi silencing of a conserved and essential root-knot nematode parasitism gene. *Proc Natl Acad Sci USA* 103:14302–14306.

Huang G, Dong R, Allen R, Davis EL, Baum TJ, and Hussey RS (2006b). A root-knot nematode secretory peptide functions as a ligand for a plant transcription factor. *Molec Plant Microbe Interact* 19:463–470.

Koenning SR, Overstreet C, Noling JW, Donald PA, Becker JO, Fortnum BA (1999). Survey of crop losses in response to phytoparastic nematodes in the United States for 1994. *J. Nematol.* 31:587-618.

Matsubayashi Y and Sakagami Y (2006). Peptide hormones in plants. *Annu. Rev. Plant Biol.* 57:649-674.

Mitchum MG, Wang X, and Davis EL (2008). Diverse and conserved roles of CLE peptides. *Cur Opin Plant Biol*. 11:75-81.

Olsen AN, Mundy J, and Skriver K (2002). Peptomics, identification of novel cationic *Arabdidopsis* peptides with conserved sequence motifs. *In Silico Biol.* 2:441-451.

Pearce G, Moura DS, Stratmann J, and Ryan CA Jr (2001). RALF, a 5-kDa ubiquitous polypeptide in plants, arrests root growth and development. *Proc. Nat. Acad. Sci.* 98:12843-12847.

Scheer JM, Pearce G, and Ryan CA (2005). LeRALF, a plant peptide that regulates root growth and development, specifically binds to 25 and 120 kDa cell surface membrane proteins of *Lycopersicon peruvianum*. *Planta* 221:667-674.

Wang X, Allen R, Ding X, Goellner M, Maier T, de Boer JM, Baum TJ, Hussey RS, and Davis EL (2001). Signal peptide-selection of cDNA cloned directly from the esophageal gland cells of the soybean cyst nematode *Heterodera glycines*. *Molec Plant Microbe Interact.* 14:536-544.

Wang X, Mitchum MG, Gao BL, Li C, Diab H, Baum TJ, Hussey RS, and Davis EL (2005). A parasitism gene from a plant-parasitic nematode with function similar to *CLAVATA3/ESR (CLE)* of *Arabidopsis thaliana*. *Mol Plant Pathol.* 6:187-191.

Wu J, Kurten EL, Monshausen G, Hummel GM, Gilroy S, and Baldwin IT (2007). NaRALF, a peptide signal essential for the regulation of root hair tip apoplastic pH in *Nicotiana*

attenuata, is required for root hair development and plant growth in native soils. *Plant J.* 52: 877-890.

Figures

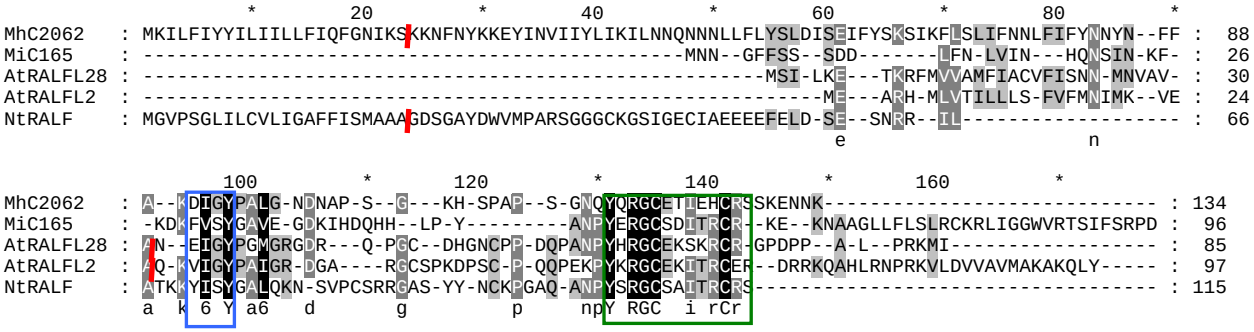

Figure 1. Alignment of *M. hapla* Contig 2062 translated region of interest, *M. incognita* Contig 165 translated region of interest, *A. thaliana* RALF-LIKE 28 and RALF-LIKE 2, and *N. tabacum* RALF. Presumed mature region of sequences based on conserved alanine assumption is found in the second alignment row. SignalP-predicted signal peptide cleavage sites are denoted by thick red vertical lines. XIXY motif (XVXY motif for *M. incognita*) is shown in blue box. Conserved C-terminal motif is shown in green box.

Chapter 4: Afterword

For centuries, scientists have sought clues leading to a fuller understanding of life on Earth. While a significant body of knowledge has been amassed based on visual clues alone, a wealth of valuable information lies encrypted within the genome of an organism, interpretable only through the processes of sequencing, assembly, and annotation. The ascription of meaning to genomic regions involves not only finding where genes are encoded, but also what is encoded by these genes (proteins), and how these gene products result in biological effects (interactions). The extent to which hypotheses can be tested and questions answered depends largely on the integration of all available genomic, proteomic, and protein-protein interaction data, and a thorough analysis of all resulting dimensions of this data, from single-nucleotide mutations to systems level relationships.

I have outlined three distinct applications of genome annotation in this dissertation. In Chapter 1, a training and validation gene set was sought for automatic gene finding programs, enabling the required foundation (a comprehensive gene set) to be constructed for use in subsequent analyses. Clues to recurrent lifestyle challenges and their respective solutions were uncovered in Chapter 2, where the membership of three large gene families was compared between the free-living *C. elegans* and the plant parasite *M. hapla*. Chapter 3 represented a much more directed annotation effort in the multiple analyses employed to support the production of a single protein product (the RALF plant peptide hormone mimic) by the *M. hapla* genome. All of these projects demonstrate the value of incorporating both

manual and automated annotation methods, the utility of employing multiple computational tools and techniques, and the role that reference genomes play in drawing inferences.

While the future of genome annotation will undoubtedly become increasingly automated, the disparate nature of research goals and the specific peculiarities of study organisms will simultaneously require the continuous development and refinement of annotation strategies. Due to this requirement, genome annotation beckons both the power of collaborative efforts and the creativity of individual minds. At the frontier of genomic discovery, it remains an invigorating field with enormous potential for unlocking the mysteries of life.