academicJournals

Vol. 13(33), pp. 3332-3343, 13 August, 2014
DOI: 10.5897/AJB2014.13911
Article Number: 222BAC446614
ISSN 1684-5315
Copyright © 2014
Author(s) retain the copyright of this article
http://www.academicjournals.org/AJB

African Journal of Biotechnology

Full Length Research Paper

Expressed sequence tags (EST) analysis of a normalized full-length cDNA library from the pinewood nematode (*Bursaphelenchus xylophilus*)

Ha Young Chung[1†], Man-Jung Kang[1†], Hye Rim Han[2], Joon-Soo Sim[1], Byung-Ju Oh[1], Inchan Choi[1], Chang-Muk Lee[1], Sang-Hong Yoon[1] and Bum-Soo Hahn[1]*

[1]National Academy of Agricultural Science, Suwon 441-857, Korea.
[2]Korea Forest Research Institute, Dongademun-gu, Seoul 130-712, Korea.
[†]Both authors contributed equally to this work.

Received 15 May, 2014; Accepted 21 July, 2014

The pinewood nematode (*Bursaphelenchus xylophilus*) infects pine trees and causes pine wilt disease. To clarify the functions and subcellular localization of *B. xylophilus* genes/proteins transcribed and predicted from mixed stages (egg, J1, J2, J3 J4 and adult), we prepared a normalized full-length *B. xylophilus* cDNA library and analyzed expressed sequence tags (ESTs) using the Pendant-Pro sequence analysis suite. Most cDNAs inserted into the library ranged from 0.9 to 1.8 kb (average 1.5 kb). The 1,902 ESTs from *B. xylophilus* consisted of 286 clusters and 1,273 singletons. EST sizes ranged from 9 to 743 bp with a mean of 336 bp. The predicted protein length from *B. xylophilus* ESTs revealed that most proteins ranged from 50 to 149 amino acids. Enzyme nomenclature (EC) numbers were classified into 133 (8.5%) of 1,559 contigs using UniProt database hits by the EC numbers method. Transmembrane regions of 1559 clusters were predicted using the TMpred algorithm. The 1,559 contigs with transmembrane regions were annotated; 481 (30.8%) contigs were assigned 'above one' domain and 1,078 (69.1%) were assigned 'none.' Additionally, taxonomy was classified for 672 (43.1%) of 1,559 contigs. Of the 1,559 contigs, 685 (43.9%) were assigned gene ontology terms using the gomerger method of contigs, including singletons. Thirty-one (31) contigs of predicted proteins grouped by BLASTP identity values had significant homology to genes expressed in subcellular structures (for example, mitochondrion, plasma membrane, endoplasmic reticulum, nucleus and golgi). *B. xylophilus* ESTs provide the foundation for research information on related plant parasite nematodes and contribute to finding an important novel parasite control strategy.

Key words: Pinewood nematode, *Bursaphelenchus xylophilus,* pine wilt disease, expressed sequence tag, Pendant-Pro Sequence Analysis Suite

INTRODUCTION

The pine wood nematode (PWN), *Bursaphelenchus xylophilus* (Steiner and Buhrer, 1934) Nickle, a plant parasitic nematode, was first described in 1934 in Louisiana, and thus originated in North America. From there

*Corresponding author. E-mail: bshahn@korea.kr. Tel: 82-31-299-1624. Fax: 82-31-299-1622.

it was introduced into Japan, then spread to neighboring East Asian countries, such as China and Korea in 1982 and 1988, respectively, and it was found in Portugal in 1999 (Mota et al., 1999; Mota and Vieira, 2008) and in Spain in 2008 (Abelleira et al., 2011). PWN is the causal agent of pine wilt disease (PWD) in pine trees (Linit, 1988; Yan et al., 2012). PWD typically occurs in mature pine trees 20 or more years old. Nematodes feed on the cells surrounding resin ducts, which causes resin to leak into tracheids, resulting in tracheid cavitation or air pockets in the water transport system. Next, transpiration of the tree cannot be sustained, which eventually leads to pine death (Myers, 1988; Donald et al., 2003). The beetle *Monochamus alternatus* (Hope) was immediately designed as a principal vector of the causal agent and spread nematodes from tree to tree (Shibata, 1987; Sakai and Yamasaki, 1990; Fan et al., 2007). Species of *Monochamus* from conifers are the principal vectors of *B. xylophilus*, and of these *M. alternatus* is the major vector in Japan, whereas *M. carolinensis* and *M. scutellatus* are the major vectors in North America and in Europe it is *Monochamus galloprovincialis*. The dispersal fourth-stage dauber larvae of the pinewood nematode cause primary transmission to occur during maturation feedings by the pine sawyer (Fielding and Evans, 1996). Also, nematode activity in pine trees is similar to natural infection by dispersal fourth-stage dauber larvae transmitted from pine sawyers when a pine tree is infected with nematodes isolated from fungal cultures (Linit 1988; Sriwati et al., 2007).

In recent years, the GenBank entries for *B. xylophilus* and *B. mucronatus* revealed 13,327 and 3,193 ESTs, respectively (Kikuchi et al., 2007). Additionally, the complete *B. xylophilus* genome (genome size of 74.5 Mb with a GC content of 40.4%) sequence was annotated by Kikuchi and colleagues (Kikuchi et al., 2011), as well as the complete mitochondrial genome of *B. xylophilus* (14,778 bp) (Sultana et al., 2013). Also, plant parasitic nematode ESTs have recently been generated for analyzing the function of genes from several plant-parasitic nematode species (Scholl et al., 2003; Dubreuil et al., 2007; Kikuchi et al., 2007; Nagaraj et al., 2008; Rosso et al., 2008; Sultana *et al*., 2013). However, since the majority of ESTs were generated from nematode total RNA without normalization of a cDNA library, these EST libraries may contain multiplicative ESTs of the nematode.

In this study, we report an analysis of 1,902 ESTs from a normalized full-length cDNA library of *B. xylophilus* mixed stages (egg, J1, J2, J3 J4, and adult). Firstly, the ESTs from the *B. xylophilus* library were grouped into clusters that were analyzed by the most conserved nematode genes between *B. xylophilus* and other nematodes, classification of EC number, transmembrane regions, taxonomy, gene ontology (GO), and the identification of gene-related subcellular localization were done. These results provide the foundation for information studies focusing on understanding the gene function of *B. xylophilus* nematode, as well as the development of a novel parasite control strategy.

MATERIALS AND METHODS

Nematode propagation

B. xylophilus (Bx90) used in this study were isolated from the Gangneung area, Korea and was propagated on a lawn of *Botrytis cinerea* cultured on a potato dextrose agar plate at 28°C. Mixed stages (egg, J1, J2, J3, J4 and adult) were collected on 25 µm sieves. Eggs and nematodes were sterilized with 1% NaOCl and suspended in sterile deionized water. The samples were frozen in liquid nitrogen and ground with a mortar and pestle.

RNA preparation and cDNA library construction

Total RNAs from mixed stages were isolated using TRIzol reagent according to the manufacturer's instructions (Invitrogen, The Netherlands). A full-length cDNA library of *B. xylophilus* was prepared from total RNAs of mixed stages as described previously (Oh et al., 2003).

In brief, the pretreated total RNA with bacterial alkaline phosphatase (Roche Diagnostics, Switzerland) and tobacco acid pyrophosphatase (Wako, Japan) was ligated with a 5'-oligoribonucleotide using RNA ligase (TaKaRa, Japan).

First-strand cDNA synthesis and amplification from purified mRNA were performed as described by Maruyama and Sugano (1994). Amplified PCR products were then digested as described by Kang et al. (2010). The ligated cDNA was then transformed into *Escherichia coli* Top10F' (Invitrogen, USA) by electroporation (Gene Pulser II; Bio-Rad, USA). The *B. xylophilus* cDNA library consisted of 4.8 × 10^6 colonies. To construct a normalized *B. xylophilus* cDNA library, a single-stranded DNA library was prepared as described previously (Vieira and Messing, 1987). Finally, we obtained a normalized library of 2 × 10^6 colonies.

Sequencing of plasmids

Each colony from the *B. xylophilus* cDNA libraries was picked into 96-well plates containing 0.5 mL of Luria Bertani (LB) medium containing 75 µg/mL ampicillin. Plates were incubated overnight at 37°C. Plasmids were purified using FB glass-fiber plates (Millipore, USA) to remove protein and cellular debris. Plasmid inserts were sequenced from the 5' end using the primer in the pCNS vector (5'-GGT CTA TAT AAG CAG AGC TC-3') and the BigDye terminator ver. 3.1 kit (Applied Biosystems, USA) on an ABI 3730*xl* DNA sequencer (Applied Biosystems).

EST sequence clustering

Vector trimming and trimmed sequence cleaning for automated trimming, and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low complexity sequences were performed using the cross_match, SeqClean, and Lucy programs (Xie et al., 2010; Tae et al., 2012; Yang et al., 2012), respectively. The cleanup process included quality assessment, confidence reassurance, vector trimming, and vector removal. ESTs of at least 200 bp after both vector and low-quality trimming were regarded as "high-quality" ESTs. Clustering was performed using TGI clustering tools (TGICL), a software pipeline designed to automate clustering and assembly of a large EST/mRNA data set (Rensing et al., 2003; Menon et al., 2012). Sequence clustering was

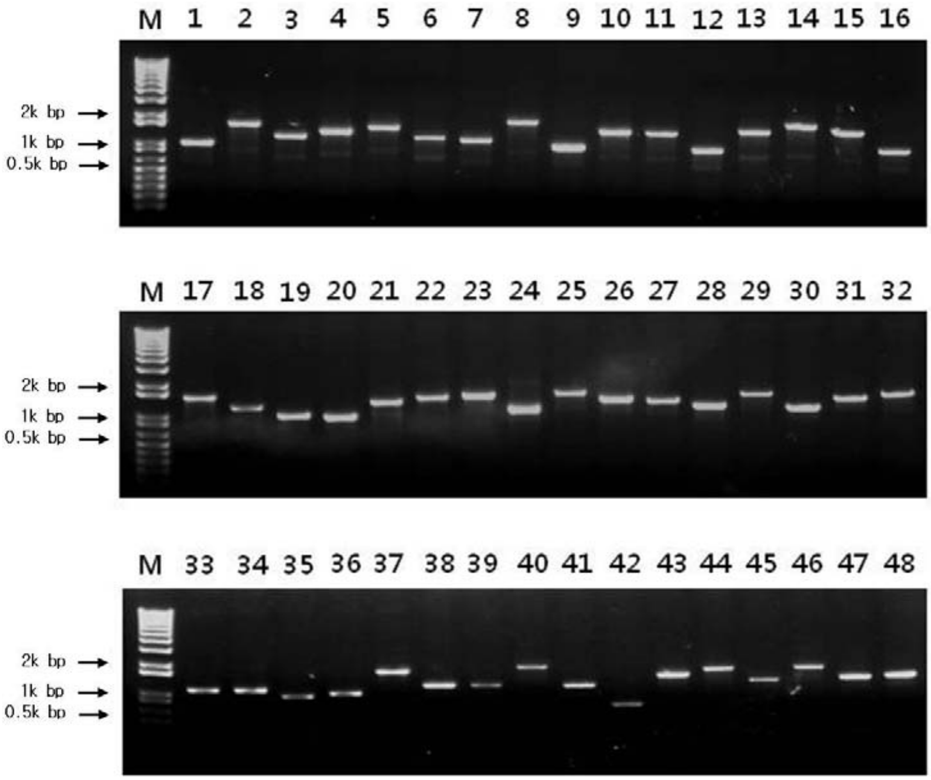

Figure 1. Evaluation of the cDNA library insert size on agarose gel. Colonies were picked at random and colony PCR was performed using T7 and SP6 promoter primers (lanes 1-48). M, DNA size marker (Invitrogen, USA)

was performed with a slightly modified version of NCBI's megablast, and the resulting clusters were then assembled using the CAP3 assembly program. TGICL began with a large multi-FASTA file (an optional peer quality values file) and the output assembly files, as produced by CAP3. Both clustering and assembly phases could be parallelized by distributing the searches and assembly jobs across multiple CPUs since TGICL can take advantage of either SMP machines or parallel virtual machine (PVM) clusters (Lee et al., 2005).

Contig analysis

Each contig sequence was evaluated for similarity and annotation analysis of nucleotide and protein sequences using the Pendant-Pro Sequence Analysis Suite. The Pendant-Pro genome database provides an exhaustive pre-computed analysis using a large variety of established bioinformatics tools. To investigate protein function, BLAST similarity was searched against the complete nonredundant protein sequence database (Altschul et al., 1997). Motifs were searched against Pfam (Bateman et al., 2002), BLOCKS (Henikoff et al., 1999), PROSITE (Falquet et al., 2002), and InterPro (Apweiler et al., 2001) databases. Predictions of cellular roles and functions based on high-stringency BLAST were searched against protein sequences with manually assigned functional categories according to the Functional Catalogue (FunCat), developed by MIPS and Biomax Informatics AG. Enzyme nomenclature (EC) numbers were predicted based on similarity. Keywords and superfamilies were extracted by similarity-based assignments from the PIR-International sequence database (Barker et al., 2000). Sequences were assigned to known clusters of orthologous groups (COGS) (Tatusov et al., 2001). To investigate protein structure, transmembrane regions were predicted using the TMpred algorithm against TMbase (Hofmann and Stoffel, 1993). Local low similarity regions and entire regions were identified using non-globular domains based on the SEG algorithm (Wootton and Federhen 1993). Coiled-coil motifs were also predicted (Lupas et al., 1991). In addition, sequence similarities were searched on the NCBI GenBank and RefSeq databases using BLASTN. BLAST similarity hits were classified based on their taxonomic origin.

RESULTS AND DISCUSSION

Evaluation of *B. xylophilus* cDNA library quality

To confirm and evaluate the insertion and quality of the *B. xylophilus* cDNA library, the range and average plasmid insert lengths were determined by PCR amplification. Forty-eight (48) cDNA clones were randomly picked from the library and colony PCR was performed with T7 and Sp6 promoter primers. Most cDNA inserts in the *B. xylophilus* libraries ranged from 0.9 to 1.8 kb with an average cDNA insert size of 1.5 kb (Figure 1) (1-48). These results are similar to those from our previous study in *Meloidogyne incognita*, as well as results from the library constructed from mixed-stage *B. xylophilus* vigorously growing on fungi (Kikuchi et al., 2007; Kang et al., 2010).

Table 1. Results of *Bursaphelenchus xylophilus* EST clustering.

Parameter	CAP3
Total sequences analyzed	1,902
Number of ESTs in cluster	629
Number of clusters	286
Number of singletons	1,273

The CAP3 program was performed with default parameters. Contigs and singletons were counted as clusters.

EST clustering and protein distribution analysis

ESTs from the *B. xylophilus* cDNA library were assembled into clusters to identify a set of genes. The 1,902 *B. xylophilus* ESTs could be classified into 286 clusters and 1,273 singletons (Table 1). Cluster sizes of *B. xylophilus* ESTs varied from a single EST (1,273 cases) to four ESTs (1 case) (Figure 2). Most clusters consisted of two to four ESTs, demonstrating the high quality of the normalized cDNA library. Also, these results show that abundant transcripts in total RNA during cDNA library preparation were removed and that the full-length cDNA library was successfully normalized. The quality of the cDNA library also showed more efficiency than the results from Kikuchi and colleagues (2011) due to gain probability of the contigs and singletons from analyzed ESTs of each library (our experiment and Kikuchi colleagues study were 82 and 50%, respectively; (Kikuchi et al., 2007; Kang et al., 2010). The sequence distribution of ESTs ranged from 9 to 743 bp with an average size of 336 bp (Figure 2A). The high frequency contigs ranged from 351 to 400 bp and represented 335 (21.5%) of 1,559 contigs. These results matched the consistent quality of the constructed *B. xylophilus* cDNA library in Figure 1.

In addition, protein length frequency values showed high-quality ESTs produced from *B. xylophilus*. The length of predicted proteins from *B. xylophilus* contigs revealed that most proteins ranged from 50 to 149 amino acids (88.1%). The longest proteins ranged from 150 to 299 amino acids and represented 113 (7.2%) of 1,559 total proteins (Figure 2B). In addition, the protein isoelectric point distribution indicated the high quality of ESTs produced from *B. xylophilus*. Putative proteins predicted from contig sequences were classified into nine groups with distinct isoelectric points: very strong basic proteins (>10.5 pI) represented 185 (11.9%) of 1,559 total proteins (Figure 2C); neutral proteins (6.5 < pI < 7.5) represented 153 (9.8%) of the total proteins; very strong acidic proteins (pI < 3.5) represented three (0.2%) of the total proteins (Figure 2C). However, these results only reflected isoelectric points predicted from part of the protein sequences. Furthermore, long-range sequencing of ESTs will need to be performed to determine isoelectric points of *B. xylophilus* full-length proteins.

Identification of highly conserved genes

Highly conserved genes between *B. xylophilus* and other nematodes were annotated by the BLASTP program using the UniProt database. Predicted proteins were classified by BLASTP identity values. The highly-conserved proteins (>90% identity) represented 1.2% of the total predicted proteins. Proteins without homology (0% identity) represented 56.3% of the total predicted proteins.

On the basis of the results of best protein match (BLASTP UniProt) with more than 50% identity, highly conserved transcripts in the *B. xylophilus* cDNA library represented genes encoding body morphogenesis (cuticlin protein), carbohydrate metabolism (pyruvate carboxylase 1, trehalose 6-phosphate synthase, and pectate lyase precursor), cytoskeletons (beta tubulin isotype 1, tubulin alpha chain, and actin-related protein 2), DNA-binding protein (ATP-dependent helicase DDX48), intracellular signaling (CBR-ARF-6 protein, serine/-threonine protein phosphatase, and CBR-TAG-210 protein), ion transporter (NADH-ubiquinone oxidoreductase 75-kDa subunit), molecular chaperones (heat shock protein 70, heat shock protein 90, and T-complex protein 1 subunit alpha), protein biosynthesis (elongation factor 1-alpha, 60S ribosomal protein L3, lysyl-tRNA synthetase, 40S ribosomal protein S3, 60S acidic ribosomal protein P0, 60S ribosomal protein L11 and CBR-RPS-18 protein), protein targeting (glycoprotein 25L2), protein transport (transport protein Sec61 alpha subunit), proteasomal and proteolytic degradation (proteasome subunit alpha type, 26S proteasome regulatory complex subunit p97, and 26S proteasome regulatory subunit rpn11), and rRNA processing (protein R74.7) (Table 2). These results confirmed that the expression of genes in our library is similar to previously reported cDNA clone sequences involved in the response to the *B. xylophilus* (Kikuchi et al., 2007). Notably, the EST data showed novel EST sequences including trehalose 6-phosphate synthase, transport protein Sec61 alpha subunit, uncharacterized protein gsk-3, SAM-dependant methyltransferase, lysyl-tRNA synthetase, 26S proteasome regulatory complex subunit p97, protein R74.7, T-complex protein 1 subunit alpha and acyl-CoA dehydrogenase.

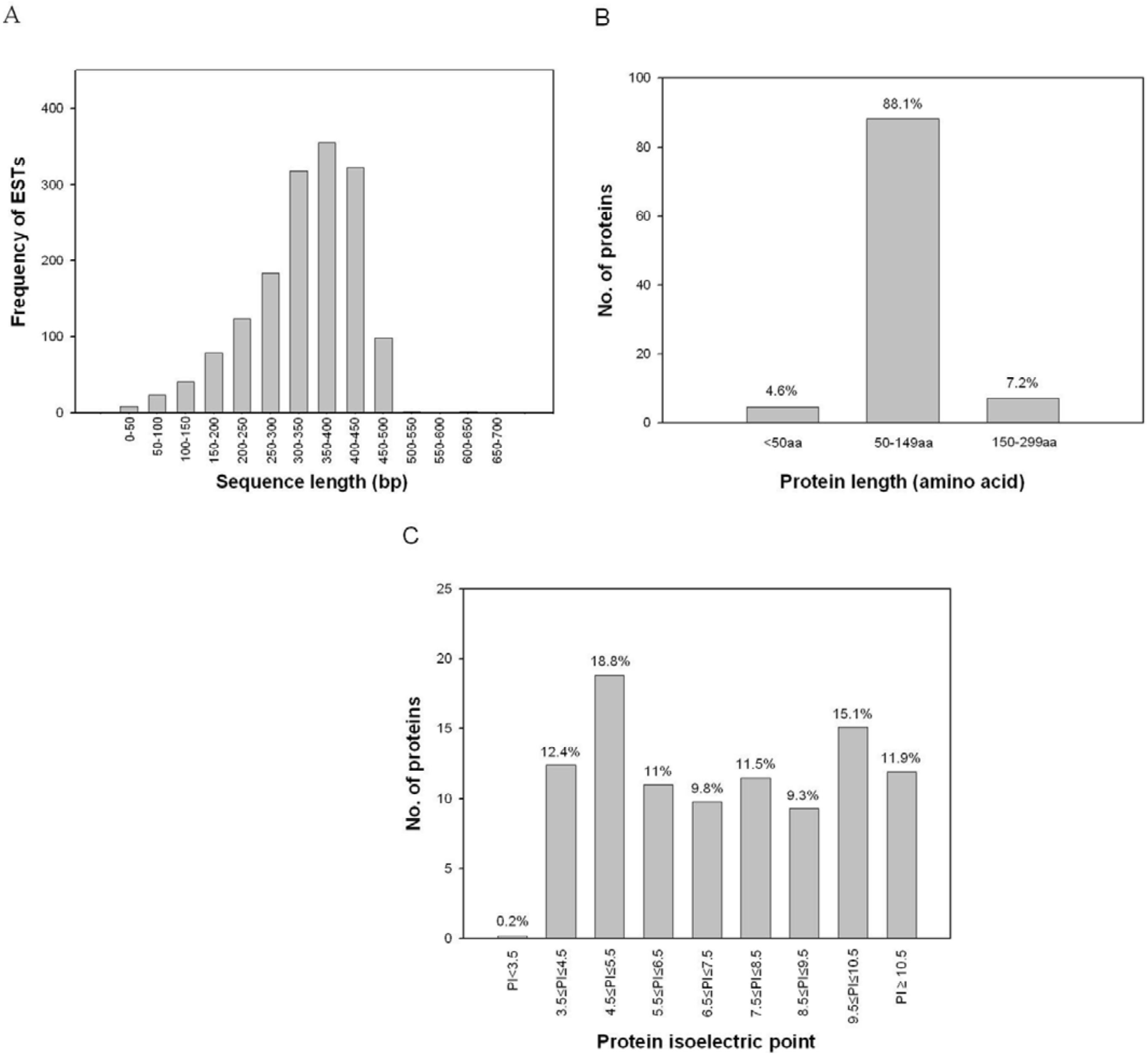

Figure 2. Length and isoelectric point distributions of *Bursaphelenchus xylophilus* cDNA inserts and proteins. Frequency versus sequence length of *B. xylophilus* cDNA inserts (A). The distributions of protein length (B) and isoelectric point (C) were analyzed from a total of 1,559 EST sequences of *B. xylophilus*.

Functional classification of proteins based on GO assignments

Gene ontology (GO) has been widely used to address the need for consistent descriptions of gene products in different databases. It has presented three-structured and controlled ontologies that describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner. Three individual memo methods, gobysimilarity, gofromeckw and gofrominterpro, were calculated to extract GO categories for a protein. The gomerger method integrates the results of the three algorithms (Martin et al., 2004). The gomerger method takes (as input) the GO term assignments for a given genetic element that are the outputs of the three predecessor methods: gofromeckw, gobysimilarity and gofrominterpro. Each GO assignment consists of the GO term number and quality score (BLAST E value or Interpro score). The latter is converted to a negative decimal logarithm. Each GO term number corresponds to a node on the GO tree. The gomerger method then

Table 2. Highly conserved nematode genes between *Bursaphelenchus xylophilus* and other nematodes (*Caenorhabditis* spp. and parasite nematodes).

Parameter	Contig	EST	Best Hit Code	Description	E-value
Body morphogenesis	CL155	1	Q6KFY9_DIRIM	P. Cuticlin protein - *Dirofilaria immitis*	6.80E-62
Carbohydrate metabolism	BXE_1535	1	PYC1_CAEEL	S. Pyruvate carboxylase 1 (EC 6.4.1.1) (Pyruvic carboxylase 1) (PCB 1) - *Caenorhabditis elegans*	2.60E-69
	BXE_421	1	Q5K2C1_APHAV	P. Putative trehalose 6-phosphate synthase (EC 2.4.1.15) - *Aphelenchus avenae*	3.20E-69
	BXE_86	1	Q33CQ4_BURXY	P. Pectate lyase precursor (EC 4.2.2.2) - *Bursaphelenchus xylophilus*	2.70E-59
Cytoskeleton	BXE_449	1	A2TF56_HAECO	P. Beta tubulin isotype 1 - *Haemonchus contortus*	5.90E-63
	BXE_1242	1	A8PYK7_BRUMA	P. Tubulin alpha chain-mouse, putative - *Brugia malayi*	4.20E-58
	BXE_880	1	ARP2_CAEEL	S. Actin-related protein 2 (Actin-like protein 2) (Actin-like protein C) - *Caenorhabditis elegans*	1.30E-52
DNA-binding protein	BXE_1728	1	A8P212_BRUMA	P. ATP-dependent helicase DDX48, putative (Fragment) - *Brugia malayi*	1.00E-57
Intracellular signaling	CL207	1	A8WMU9_CAEBR	P. CBR-ARF-6 protein - *Caenorhabditis briggsae*	4.60E-70
	BXE_1555	1	A8NZM1_BRUMA	P. Serine/threonine protein phosphatase (EC 3.1.3.16) - *Brugia malayi*	6.20E-69
	CL191	1	A8XXM7_CAEBR	P. CBR-TAG-210 protein - *Caenorhabditis briggsae*	1.10E-61
Ion transporter	BXE_1049	1	Q86S77_CAEEL	P. NADH-ubiquinone oxidoreductase 75 kDa subunit (EC 1.6.5.3) - *Caenorhabditis elegans*	1.90E-55
Molecular chaperone	BXE_1864	1	A4UU63_BURXY	P. Heat shock protein 90 (Fragment) - *Bursaphelenchus xylophilus*	3.80E-60
	BXE_747	1	Q8MUA7_HETGL	P. Heat shock protein 70-C - *Heterodera glycines*	6.40E-60
	BXE_1714	1	TCPA_CAEEL	S. T-complex protein 1 subunit alpha (TCP-1-alpha) (CCT-alpha) - *Caenorhabditis elegans*	1.90E-54
	BXE_1301	1	A8XHC9_CAEBR	P. Putative uncharacterized protein - *Caenorhabditis briggsae*	2.90E-53
Protein biosynthesis	CL223	1	A8PJ17_BRUMA	P. Elongation factor 1-alpha (EF-1-alpha), putative - *Brugia malayi*	1.90E-78
	BXE_1507	1	Q95ZQ3_CAEEL	P. Lysyl-tRNA synthetase (EC 6.1.1.6) - *Caenorhabditis elegans*	3.70E-60
	CL150	1	A8NXR7_BRUMA	P. 40S ribosomal protein S3, putative - *Brugia malayi*	5.30E-60
	BXE_271	1	RLA0_CAEEL	S. 60S acidic ribosomal protein P0 - *Caenorhabditis elegans*	1.40E-58

Table 2. Contd.

	BXE_960	1	A8NKQ0_BRUMA	P. 60S ribosomal protein L11, putative - *Brugia malayi*	5.90E-55
	BXE_1101	1	A8Y099_CAEBR	P. CBR-RPS-18 protein - *Caenorhabditis briggsae*	3.50E-54
Electron transport	CL284	2	GCDH_CAEEL	S. Probable glutaryl-CoA dehydrogenase, mitochondrial precurso (EC 1.3.99.7) (GCD) - *Caenorhabditis elegans*	5.50E-53
Protein targeting	CL255	2	A8PU74_BRUMA	P. Glycoprotein 25L2, putative - *Brugia malayi*	1.10E-58
Protein transport	CL110	1	A8Q009_BRUMA	P. Probable transport protein Sec61 alpha subunit, putative - *Brugia malayi*	4.70E-68
	BXE_632	1	A8QCC8_BRUMA	P. Proteasome subunit alpha type (EC 3.4.25.1) - *Brugia malayi*	4.40E-67
	BXE_1164	1	A8PPJ3_BRUMA	P. 26S proteasome regulatory complex subunit p97, putative - *Brugia malayi*	1.00E-59
	CL113	1	A8PAL6_BRUMA	P. 26S proteasome regulatory subunit rpn11, putative - *Brugia malayi*	1.30E-54
rRNA processing	BXE_1063	1	A8Q6F8_BRUMA	P. Protein R74.7, putative - *Brugia malayi*	1.10E-55

The sequence similarities of contigs were searched using the UniProt database.

searches the GO tree (which is kept in memory) upward and finds all ancestors of all GO nodes assigned. Scores from all preceding nodes are then added together, assigning GO categories to a protein sequence (Ashburner and Lewis, 2002; Martin et al., 2004). Using the gomerger method of contigs, we functionally assigned GO terms to 685 (43.9%) of 1,559 contigs (of 1,559 contigs, 1,105 align to InterPro domains and 58 to the keyword method). The 685 contigs with GO annotations from the *B. xylophilus* data set were further annotated, with 621 hits assigned a biological process (BP), 620 hits assigned molecular functions (MF), and 503 hits assigned a cellular component (CC) in GO terms. A summary of GO annotations by biological process, cellular component and molecular function is provided in Figure 3. Among the most common GO categories representing biological processes were cellular process (GO: 0009987) and metabolic process (GO: 0008152). The largest number of GO terms in cellular component was 433 contigs in the cell (GO: 0005623), 433 contigs in a cell part (GO: 0044464), and 399 intracellular contigs (GO: 0005622). The largest number of GO terms in molecular function was 467 contigs in binding (GO: 0005488), 348 contigs in catalytic activity (GO: 0003824), and 328 contigs in protein binding (GO: 0005515). The largest number of GO terms in biological process was 465 contigs in cellular process (GO: 0009987), 415 contigs in metabolic processes (GO: 0008152), and 374 contigs in cellular metabolic processes (GO: 00044237). These results were highly similar to a previous study analyzing ESTs in *Meloidogyne incognita* (Kang et al., 2010). However, in terms of reproduction, reproductive process, and response stimulate in biological processes, a higher frequency was found than in a previous study (Kang et al., 2009). The total transcriptome analysis of *B. xylophilus* will be needed for the complete GO analysis.

Localization of proteins in the cell

Information on subcellular localization of proteins in the cell provides important clues about protein function and can be used to infer the function of predicted proteins. In addition, the subcellular localization of proteins with known function unravels where the corresponding biological processes take place and how they are connected among each other (Shen and Burger 2007). Thirty-one (31) contigs were separated into 12 subcellular structures according to their predicted functions (Figure 4). Putative transcripts in the *B. xylophilus* cDNA library represented genes expressed in the plasma membrane (P. CBR-ARF-6 protein and A8WMU9), cytoplasm (putative uncharacterized protein, A8XHC9; 60S ribosomal protein L3, RL3; ATP-dependent helicase DDX48, A8P212; heat shock protein 90, A4UU63; serine/threonine protein phosphatase, A8PJS7; ps4b-prov protein, A8QBR0; proteasome subunit alpha type, A8QCC8; 60S ribosomal protein L11, A8NKQ0; 26S

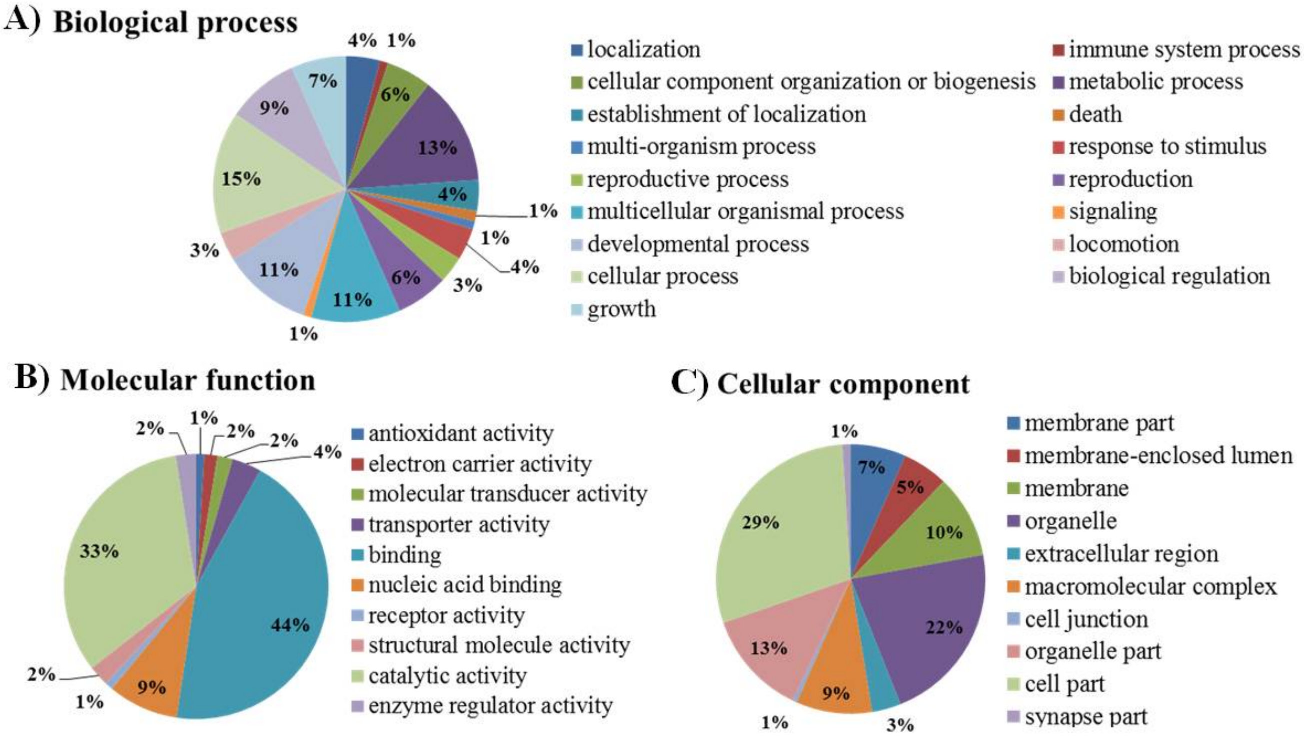

Figure 3. Percentage representation of gene ontology (GO) mappings for *Bursaphelenchus xylophilus* proteins. Biological process (A); Molecular function (B); Cellular component (C). Individual GO categories can have multiple mappings. Percentages shown represent the total categories annotated (not the total sequences annotated under each component).

proteasome regulatory subunit rpn11, A8PAL6; 40S ribosomal protein S3, A8NXR7; and elongation factor 1-alpha, A8PJ17), cytoskeleton (tubulin alpha chain, A8PYK7; RAS-like GTP-binding protein RhoA, A8PJ63; serine/threonine protein phosphatase, A8PJS7; beta tubulin isotype 1, A2TF56; and putative actin, A3QP75), microtubule cytoskeleton (serine/threonine protein phosphatase, A8PJS7), endoplasmic reticulum (heat shock protein 70, Q8MUA7 and probable transport protein Sec61 alpha subunit, A8Q009), golgi (AGAP011363-PA, A0NFB8; CBR-ARF-6 protein, A8WMU9; chromosome 2 SCAF14705, Q4S8K4; glycoprotein 25L2, A8PU74; and Rab fragment, Q6GXB2), intracellular transport vesicles (glycoprotein 25L2, A8PU74), nucleus (putative uncharacterized protein, A8XHC9; putative uncharacterized protein, P90904; 60S ribosomal protein L3, RL3; serine/threonine protein phosphatase, A8PJS7; proteasome subunit alpha type, A8QCC8; nucleolar protein K01G5.5, A8Q703; and uncharacterized protein uev-1, O45495), nucleolus (60S ribosomal protein L3, RL3 and nucleolar protein K01G5.5, A8Q703), other nuclear structures (nucleolar protein K01G5.5, A8Q703), mitochondrion (NADH-ubiquinone oxidoreductase 75-kDa subunit, Q86S77; CBR-RPS-18 protein, A8Y099; pyruvate carboxylase 1, PYC1; probable calcium-binding mitochondrial carrier F55A11.4, CMC2; 60S ribosomal protein L11, A8NKQ0; nucleolar protein K01G5.5, A8Q703; and putative uncharacterized protein, Q9BL46), and the endosome (Vps4b-prov protein, A8QBR0) (Table 2). The appropriate localization of proteins in subcellular structures and compartments plays an important role in their functional integrity. However, determination of subcellular localization by experimental means is not practical for all proteins due to time and cost constraints (Guda 2006). Our results may assist in inferring the function of proteins predicted from the *B. xylophilus* cDNA library.

Enzymatic classification based on EC number

EC numbers were extracted by the EC numbers method from UniProt database hits using the BLASTPGP algorithm. EC numbers were classified for 133 (8.3%) of 1,599 contigs. For the *B. xylophilus* data set, 133 contigs with EC numbers were further annotated, with 44 assigned as oxidoreductases, 35 as transferases, 33 as hydrolases, 9 as lyases, 4 as isomerases, and 8 as ligases (Table 3).

Classification of membrane proteins based on the number of transmembrane domains

The TMpred algorithm predicts membrane-spanning regions and their orientation. The prediction is made

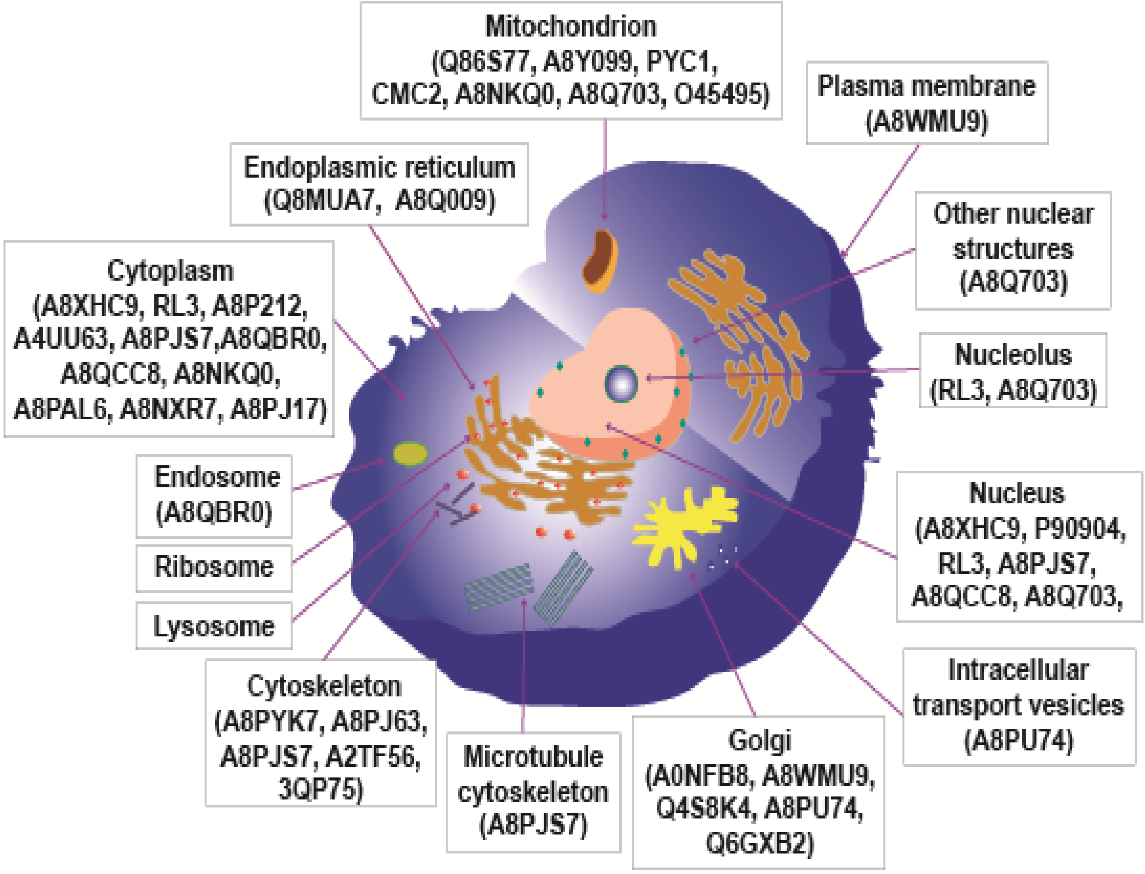

Figure 4. Putative localization of *Bursaphelenchus xylophilus* proteins in the cell. *B. xylophilus* subcellular structures and components were depicted by a single animal cell. The respective thicknesses and size of subcellular components were not accurate. Putative *B. xylophilus* proteins were replaced by annotation of the corresponding protein in the UniProt database.

Table 3. Classification of selected EC numbers.

EC number	EC description	Sorting
1	Oxidoreductases	44
1.1	Acting on the CH-OH group of donors	15
1.1.1	With NAD or NADP as acceptor	15
1.1.1.1	Alcohol dehydrogenase	3
1.1.1.2	Alcohol dehydrogenase (NADP)	3
1.1.1.37	Malate dehydrogenase	3
1.2	Acting on the aldehyde or oxo group of donors	6
1.2.1	With NAD or NADP as acceptor	3
1.2.4	With a disulfide as acceptor	3
1.2.4.1	Pyruvate dehydrogenase (lipoamide)	3
1.3	Acting on the CH-CH group of donors	5
1.3.1	With NAD or NADP as acceptor	3
1.4	Acting on the CH-NH_2 group of donors	3
1.6	Acting on NADH or NADPH	5
1.6.5	With a quinone or similar compound as acceptor	3
1.6.5.3	$NADH_2$ dehydrogenase (ubiquinone)	3
1.13	Acting on single donors with incorporation of molecular oxygen (oxygenases)	3
1.13.11	With incorporation of two atoms of oxygen	3
1.14	Acting on paired donors, with incorporation or reduction of molecular oxygen	4

Table 3. Contd.

2	Transferases	35
2.1	Transferring one-carbon groups	5
2.1.1	Methyltransferases	4
2.3	Acyltransferases	3
2.3.1	Transferring groups other than amino-acyl groups	3
2.4	Glycosyltransferases	4
2.5	Transferring alkyl or aryl groups, other than methyl groups	8
2.7	Transferring phosphorus-containing groups	12
2.7.7	Nucleotidyltransferases	3
3	Hydrolases	33
3.1	Acting on ester bonds	10
3.1.3	Phosphoric monoester hydrolases	7
3.1.3.16	Phosphoprotein phosphatase	6
3.4	Acting on peptide bonds (peptidases)	12
3.4.25	Threonine endopeptidases	4
3.4.25.1	Proteasome endopeptidase complex	4
3.5	Acting on carbon-nitrogen bonds, other than peptide bonds	3
3.5.1	In linear amides	3
3.6	Acting on acid anhydrides	6
3.6.1	In phosphorus-containing anhydrides	4
3.6.1.3	Adenosinetriphosphatase	3
4	Lyases	9
4.1	Carbon-carbon lyases	3
4.2	Carbon-oxygen lyases	5
4.2.1	Hydro-lyases	4
5	Isomerases	4
6	Ligases	8
6.1	Forming carbon-oxygen bonds	3
6.1.1	Ligases forming aminoacyl-tRNA and related compounds	3

EC numbers were extracted from UniProt database hits using the EC numbers method. Six representative EC classes were classified from *B. xylophilus* proteins. Each EC class is illustrated by a representative member. The number of genes associated with each EC class is indicated in the sorting column. The selected EC numbers have at least three hits.

Table 4. Classification of membrane proteins based on the number of transmembrane domains.

Number of domains	Number of proteins	% Protein
≥1 domain	481	30.90
1 domains	390	25.00
2 domains	62	4.00
3 domains	26	1.70
4 domains	3	0.20
none	1,078	69.10

The TMpred algorithm made a prediction of transmembrane helices in protein sequences.

using a combination of several weight matrices for scoring (Hofmann and Stoffel, 1993). Transmembrane regions of 1,599 clusters were grouped. For the *B. xylophilus* data set, 1,599 contigs with transmembrane regions could be further annotated, with 481 (30.9%) assigned to the 'above one' domain and 1,078 (69.1%) assigned as 'none' (Table 4). In the future, the long-ranged sequencing of insert cDNA will be performed to determine complete transmembrane regions of the full-length protein.

Table 5. Taxonomic classification.

Superkingdom (sorting)	Kingdom (sorting)	Phylum (sorting)	Class (sorting)
Eukaryota (665)	Metazoa (648)	Nematoda (564)	Chromadorea (563)
Bacteria (7)	Fungi (6)	Chordata (46)	Insecta (29)
	Viridiplantae (6)	Arthropoda (31)	Mammalia (20)
		Ascomycota (6)	Actinopterygii (14)
		Proteobacteria (5)	Amphibia (9)
		Streptophyta (5)	Saccharomycetes (5)
		Cnidaria (4)	Anthozoa (4)
		Platyhelminthes (3)	Aves (3)
		Apicomplexa (2)	Liliopsida (3)
		Bacteroidetes (1)	Aconoidasida (2)
		Chlorophyta (1)	Alphaproteobacteria (2)
		Firmicutes (1)	Gammaproteobacteria (2)
			Trematoda (2)
			Bacilli (1)
			Chilopoda (1)
			Chlorophyceae (1)
			Deltaproteobacteria (1)
			Enoplea (1)
			Eurotiomycetes (1)
			Malacostraca (1)
			Sphingobacteria (1)
			Spirotrichea (1)
			Turbellaria (1)

Taxonomy information was extracted by BLASTPGP hits from the UniProt database using the Taxon algorithm.

Taxonomic classification

The taxon algorithm extracted taxonomy information from BLASTPGP similarity hits using the UniProt database. Of 1,559 contigs, 672 (43.1%) were classified into taxonomy. For the *B. xylophilus* data set, the 672 contigs with taxonomy classification were further annotated, with seven assigned to bacteria and 665 assigned to eukaryota (Table 5). The number of kingdoms representing Metazoa was 648 (96.4%) of 672 contigs. The number of phylum representing Nematoda was 564 (83.9%) of 672 contigs. The number of classes representing Chromadorea was 563 (83.8%) of 672 contigs.

The class Chromadorea contains the *B. xylophilus*, which belongs to the order Tylenchida. Our results demonstrate that most genes classified in Chromadorea are highly homologous to the genes presented by Lilley and Parkinson and colleagues (Parkinson et al., 2004; Lilley et al., 2005).

Conflict of Interests

The author(s) have not declared any conflict of interests.

ACKNOWLEDGEMENTS

This project was supported by a grant from the National Academy of Agricultural Science (PJ008541), RDA, Korea. This study was supported by the 2014 Postdoctoral Fellowship Program (to H.Y.C. and I.C.C.) of the National Academy of Agricultural Science, RDA, Korea.

REFERENCES

Abelleira A, Picoaga A, Mansilla JP, Aguin O (2011). Detection of *Bursaphelenchus xylophilus*, causal agent of pine wilt disease on *Pinus pinaster* in Northwestern Spain. Plant Dis. 95:776.

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

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, *et al*. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 29:37-40.

Ashburner M, Lewis SE (2002). On ontologies for biologists: the Gene Ontology- uncoupling the web. Novartis Found Symp. 247: 66-80.

Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh LS, Ledley RS, Janda JF, Pfeiffer F, Mewes HW, Tsugita A, Wu C (2000). The protein information resource (PIR). Nucleic Acids Res. 28:41-44.

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

Donald PA, Stamps WT, Linit MJ (2003). Pine wilt disease in APSnet plant disease lessons. The American Phytopathological Society, St. Paul, MN, USA.

Dubreuil G, Magliano M, Deleury E, Abad P, Rosso MN (2007). Transcriptome analysis of root knot nematode functions induced in the early stages of parasitism. N. Phytol. 176:426-436.

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

Fan J, Kang L, Sun J (2007). Role of host volatiles in mate location by the Japanese pine sawyer, *Monochamus alternatus* Hope (Coleoptera: Cerambycidae). Environ. Entomol. 36:58-63.

Fielding NJ, Evans HF (1996). The pine wood nematode *Bursaphelenchus xylophilus* (Steiner and Buhrer) Nickle (*B. lignicolus* Mamiya and Kiyohara): an assessment of the current position. For. 69:35-46.

Guda C (2006). pTARGET: a web server for predicting protein subcellular localization. Nucleic Acids Res. 34:W210-213.

Henikoff S, Henikoff JG, Pietrokovski S (1999). Blocksþ: a nonredundant database of protein alignment blocks derived from multiple compilations. Bioinform. 15:471-479.

Hofmann K, Stoffel W (1993). TMbase-A database of membrane spanning proteins segments. Biol. Chem. Hoppe-Seyler 374:166.

Kang JS, Lee H, Moon IS, Lee Y, Koh YH, Je YH, Lim KJ, Lee SH (2009). Construction and characterization of subtractive stage-specific expressed sequence tag (EST) libraries of the pinewood nematode *Bursaphelenchus xylophilus*. Genom. 1:70-77.

Kang MJ, Kim YH, Hahn BS (2010). Expressed sequence tag analysis generated from a normalized full-length cDNA library of the root-knot nematode (*Meloidogyne incognita*). Genes Genom. 32:553-562.

Kikuchi T, Aikawa T, Kosaka H, Pritchard L, Ogura N, Jones JT (2007). Expressed sequence tag (EST) analysis of the pine wood nematode *Bursaphelenchus xylophilus* and *B. mucronatus*. Mol. Biochem. Parasitol. 155:9-17.

Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, Takanashi T, Tsai IJ, Assefa SA, Cock PJ, Otto TD, Hunt M, Reid AJ, Sanchez-Flores A, Tsuchihara K, Yokoi T, Larsson MC, Miwa J, Maule AG, Sahashi N, Jones JT, Berriman M (2011). Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*. PLoS Pathog. 7:e1002219.

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005). The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res. 1:D71-74.

Lilley CJ, Atkinson HJ, Urwin PE (2005). Molecular aspects of cyst nematodes. Mol. Plant Pathol. 6:577-588.

Linit MJ (1988). Nemtaode-vector relationships in the pine wilt disease system. J. Nematol. 20:227-235.

Lupas A, Van Dyke M, Stock J (1991). Predicting coiled coils from protein sequences. Sci. 24:1162-4116.

Martin DM, Berriman M, Barton GJ (2004). GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinform. 5:178.

Maruyama K, Sugano S (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138:171-174.

Menon R, Garg G, Gasser RB, Ranganathan S (2012). TranSeqAnnotator: large-scale analysis of transcriptomic data. BMC Bioinformatics 13(Suppl 17):S24.

Mota M, Vieira P (2008). Pine wilt disease: a worldwide threat to forest ecosystems. Springer XVIII, 406 p., ISBN: 978-1-4020-8454-6

Mota MM, Braasch H, Bravo MA, Penas AC, Burgermeister W, Metge K , Sousa E (1999). First report of *Bursaphelenchus xylophilus* in Portugal and in Europe. Nematology 1: 727-734.

Myers RF (1988). Pathogenesis in pine wilt caused by pinewood nematode, *Bursaphelenchus xylophilus*. J. Nematol. 20:236-244.

Nagaraj SH, Gasser RB, Ranganathan S (2008). Needles in the EST Haystack: Large-Scale Identification and Analysis of Excretory-Secretory (ES) Proteins in Parasitic Nematodes Using Expressed Sequence Tags (ESTs). PLoS Negl. Trop. Dis. 2:e301.

Oh JH, Kim YS, Kim NS (2003). An improved method for constructing a full-length enriched cDNA library using small amounts of total RNA as a starting material. Exp. Mol. Med. 35:586-590.

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

Rensing SA, Lang D, Reski R (2003). *In silico* prediction of UTR repeats using clustered EST data. Proceed. German Conference Bioinform. 117-122.

Rosso MN, Jones JT, Abad P (2008). RNAi and functional genomics in plant parasitic nematodes. Annu. Rev. Phytopathol. 47:207-232.

Sakai M, Yamasaki T (1990). (+)-Juniperol and (+)-pimaral: Attractants for the cerambycid beetle, Monochamus alternatus Hope. J. Chem. Eco. 16:3383-3392.

Scholl EH, Thorne JL, McCarter JP, Bird DM (2003). Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biol. 4:R39.

Shen YQ, Burger G (2007). 'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. BMC Bioinform. 8:420.

Shibata E (1987). Oviposition schedules, survivorship curves, and mortality factors within trees of two cerambycid beetles (Coleoptera: Cerambycidae), the Japanese pine sawyer, *Monochamus alternatus* hope, and sugi bark borer, *Semanotus japonicus* lacordaire. Res. Population Ecol. 29:347-367.

Sriwati R, Takemoto S, Futai K (2007). Cohabitation of the pine wood nematode, *Bursaphelenchus xylophilus*, and fungal species in pine trees inoculated with *B. xylophilus*. Nematol. 9:77-86.

Steiner G, Buhrer EM (1934). *Aphelenchoides xylophilus* n. sp. a nematode associated with bluestain and other fungi in timber. J. Agric. Res. 48:949-951.

Sultana T, Kim J, Lee SH, Han H, Kim S, Min GS, Nadler SA, Park JK (2013). Comparative analysis of complete mitochondrial genome sequences confirms independent origins of plant-parasitic nematodes. BMC Evol. Biol. 13:12.

Tae H, Ryu D, Sureshchandra S, Choi JH (2012). ESTclean: A Cleaning Tool for Next-Gen Transcriptome Shotgun Sequencing. BMC Bioinform. 13:247.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29:22-28.

Vieira J, Messing J (1987). Production of single-stranded plasmid DNA. Methods Enzymol. 153:3-11.

Wootton JC, Federhen S (1993). Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17:149-163.

Xie G, Chain PS, Lo CC, Liu KL, Gans J, Merritt J, Qi F (2010). Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. Mol. Oral Microbiol. 25:391-405.

Yan X, Cheng XY, Wang YS, Luo J, Mao ZC, Ferris VR, Xie BY (2012). Comparative transcriptomics of two pathogenic pinewood nematodes yields insights into parasitic adaptation to life on pine hosts. Gene 505:81-90.

Yang F, Xu B, Zhao S, Li J, Yang Y, Tang X, Wang F, Peng M, Huang Z (2012). De novo sequencing and analysis of the termite mushroom (*Termitomyces albuminosus*) transcriptome to discover putative genes involved in bioactive component biosynthesis. J. Biosci. Bioeng. 4:228-231.