Universidade de Aveiro Departamento de Química
Ano 2012

Daniel Ferreira Martins

Identification of horizontal gene transfer events in *B. xylophilus*

Universidade de Aveiro Departamento de Química
Ano 2012

Daniel Ferreira Martins

Identification of horizontal gene transfer events in *B. xylophilus*

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biotecnologia Molecular, realizada sob a orientação científica da Doutora Conceição Egas, diretora da Unidade de Serviços Avançados do Biocant e do Professor Doutor Jorge Saraiva investigador auxiliar do Departamento de Química da Universidade de Aveiro

Projeto financiado pela Autoridade Florestal Nacional: "O nemátode da madeira do pinheiro (NMP), *Bursaphelenchus xylophilus*"

Projecto financiado pela FCT "Transcriptómica e proteómica no estudo da base molecular da patogenicidade de *Bursaphelenchus xylophilus".*
PTDC/AGR-CFL/098916/2008
FCOMP-01-0124-FEDER-008794

Dedico este trabalho aos meus pais que me deram a maior das riquezas, o amor.

"Se um dia tiver que escolher entre o mundo e o amor lembre-se: se escolher o mundo, ficará sem o amor, mas se escolher o amor, com ele conquistará o mundo."
Albert Einstein

o júri

presidente

Prof. Doutor João Manuel da Costa e Araújo Pereira Coutinho
Professor associado com agregação no Departamento de Química da Universidade de Aveiro

Doutora Joana Sá Cardoso
Pós-Doutorada no IMAR, Departamento de Ciências da Vida, Universidade de Coimbra

Doutora Conceição Egas (Orientadora)
Diretora da Unidade de Serviços Avançados do Biocant

Prof. Doutor Jorge Manuel Alexandre Saraiva (Co-orientador)
Investigador auxiliar no Departamento de Química da Universidade de Aveiro

agradecimentos

Ao Biocant por me ter concedido a oportunidade de desenvolver este trabalho.
À minha orientadora, Doutora Conceição Egas pelo acompanhamento e dedicação prestadas, assim como pela motivação e conhecimentos transmitidos.
Às colegas Cristina, Paula, Maria José, Susana e Joana da Unidade de Serviços Avançados e ao Felipe e Miguel da Unidade de Bioinformática por todo o apoio prestado na realização deste trabalho, mas sobretudo pelo acolhimento, companheirismo e ânimo que sempre me transmitiram no período em que estive no Biocant.
À comunidade Biocant em geral que enriqueceu os meus dias e me fez sentir "parte da família".
Aos meus amigos de curso que ao longo destes anos na academia me acompanharam nos bons momentos e nos mais complicados, dando assim sentido ao caminho que fomos percorrendo em conjunto.
À minha família que sempre me apoiou.
Aos projetos paralelos que fui desenvolvendo ao longo deste ano, que me foram complementando e enriquecendo enquanto pessoa e que ao mesmo tempo me mostraram que é possível fazer sempre mais alguma coisa com a nossa vida!
À EAPJ-CM, aos escuteiros, aos amigos e conhecidos que me ajudam a ser quem sou.
Last but not least à minha namorada que me põe um sorriso na cara sempre que preciso, mesmo quando o trabalho parece mais complicado, e assim me vai dando força e motivação para seguir em frente.

palavras-chave

Bursaphelenchus xylophilus, doença da murchidão do pinheiro, nemátodo da madeira do pinheiro, transcriptoma, pirosequenciação, transferência horizontal de genes, analise filogenética, pipeline, diversidade microbiana

resumo

A doença da murchidão do pinheiro foi identificada pela primeira vez no Japão em 1905 e foi associada ao nemátodo *Bursaphelenchus xylophilus* em 1972. A devastação provocada por esta doença, assim como as perdas económicas a ela associadas e o impacto na fileira florestal levaram a Organização Europeia para a Proteção das Plantas a declarar o nemátodo da madeira do pinheiro como uma praga. Atualmente já existe informação extensa sobre a morfologia, ciclo de vida, associação com um vetor e potenciais hospedeiros para este organismo. No entanto, o mecanismo molecular da doença é pouco conhecido, mas um dos fatores apontados, passa pela aquisição de novas funções por incorporação horizontal de genes.

O transcriptoma de *B. xylophilus* foi sequenciado e anotado recentemente no Biocant, gerando um grande volume de dados. O objetivo do nosso trabalho foi identificar novos genes incorporados no genoma do nemátodo por transferência horizontal. Para levar a cabo esta tarefa, foi estabelecido um *pipeline* de análise de dados contendo um conjunto sequencial de filtros usados para remover genes com origem em nemátodos e reter os de origem bacteriana ou fúngica. Os transcritos de *B. xylophilus* anotados foram também filtrados de acordo com as funções dos genes e pelo E-value. Uma última filtragem foi realizada tendo por base a composição da comunidade microbiana associada ao nemátodo, obtida por pirosequenciação. O *pipeline* definido gerou 21 candidatos que foram validados através de análise filogenética com proteínas homólogas de bactérias, fungos, nemátodos e outros organismos e ainda com uma posterior identificação do gene no genoma do nemátodo, disponibilizado no decurso deste trabalho. Esta abordagem identificou três genes incorporados por transferência horizontal: uma β-1,3-endoglucanase, uma álcool desidrogenase, e um gene que pertence à família das desidrogenases/redutases de cadeia curta. Os primeiros dois genes já foram descritos em *B. xylophilus* e *C. elegans*, respetivamente, o que valida a nossa abordagem. O papel que o gene recém-identificado desempenha no nemátodo e na doença terá de ser estudado futuramente, na expectativa deste fornecer novas informações sobre o mecanismo da doença e novos alvos elegíveis para o controlo da doença.

keywords

Bursaphelenchus xylophilus, pine wilt disease, pine wood nematode, transcriptome, pyrosequencing, horizontal genes transfer, phylogenetic analysis, pipeline, microbial diversity

abstract

Pine wilt disease was first identified in Japan in 1905 and its association with the nematode *Bursaphelenchus xylophilus* dates back to 1972. The devastating nature of this disease and the big economical losses it generates along with the ecological impact led the European Plant Protection Organization to consider its causal agent as an european quarantine pest. Until now, information regarding the nature of *B. xylophilus* with reference to its morphological characters, life cycle, vector association and potential host have been gathered and documented. However, little is known on the molecular mechanisms of the disease. According to current knowledge one of the factors that may play a role in the disease is the acquisition of new gene functions through horizontal gene transfer (HGT).

The transcriptome of *B. xylophilus* was recently sequenced and annotated at Biocant, providing a large amount of genomics information. The goal of our work was to identify new genes incorporated through HGT in these transcripts. To accomplish our objective we established a data analysis pipeline composed of several filters to discard genes of nematodal origin and select only those of bacterial or fungal origin. *B. xylophilus* transcripts were screened based on the information held in their annotation: E-value, organism origin and existence of a specific function. A final screen was performed by removing all hits matching the microbial community associated with the nematode, as determined by barcoded pyrosequencing. The pipeline outlined 21 candidates that were compared to homologous sequences from bacteria, fungi, nematodes and other organisms through phylogenetic analysis. This last step confirmed the presence of three genes resulting from HGT a β-1,3- endoglucanase, an alcohol dehydrogenase and a short chain dehydrogenase/reductase. The first two genes were already described for *B. xylophilus* and *C. elegans*, respectively, validating our approach. The role of the newly identified HGT gene in the nematode and disease will be studied in the future, expecting to provide new information on the disease mechanism and contribute to the identification of new targets eligible for disease control.

Table of contents

Abbreviations

Ala……Alanine
BLAST……Basic Local Alignment Search Tool
BLASTn…….....… Nucleotide Basic Local Alignment Search Tool
BLASTp…….....…… Protein BLAST using protein query
BLASTX……...…… Protein BLAST using translated nucleotide query
bp……..…… Base pair
DNA……..… Deoxyribonucleic acid
cDNA……...……..… Complementary DNA
CAZymes……..... Carbohydrate active enzymes
CCD…….……..…… Charge-coupled device
D4S…….……...… Dispersal 4th larva stage
dNTPs……Deoxyribonucleotides
dsDNA…….. Double stranded DNA
EDTA……..….. Ethylenediamine tetraacetic acid
EST……...… Expressed sequence tag
EPPO…….…… European Plant Protection Organization
E-value……. Expect value
GHF…….. Glycosil hydrolase family
GC content……..... Guanine-cytosine content
Gly……..... Glycine
GO……Gene ontology
HAD…….. Haloacid Dehalogenase
HGT…….. Horizontal gene transfer
ITS II…….... Internal Transcribed Spacer 2
LSU…….....Large ribosomal subunit
mRNA…….…..… Messenger RNA
NGS…….…… Next generation sequencing
NCBI……..… National Center for Biotechnology Information
nBFNO…….. Non bacteria, non fungi nor nematode organisms
nt……Nucleotide

ORF…….....…Open reading frame

OTU…….……..Operational Taxonomic Units

PWN…….....Pine wood nematode

PWD……..... Pine wilt disease

PCR…….….. Polymerase Chain Reaction

PGPS…… Pine-grown propagative mixed stage

RDP…….……..Ribosomal Database Project II

ROS…….…..…Reactive oxygen species

RNA…….……. Ribonucleic acid

RNAi…… RNA interference

rRNA…… Ribosomal RNA

TAE……... Tris Acetate EDTA

USA……..……United States of America

vap…….… Venom allergen protein

List of tables

List of figures

Chapter I - Introduction

1-Nematodes

1.1-Overall view

Among the animal kingdom, *nematoda* is the largest phylum, not just for its extended number of individuals but also due to wide range of species it comprises. Their lifestyle is also very diverse and distinct. Nematodes can be parasites living in association with plants or animals, after infecting either one of them or live as free-living organisms[1-4]. They are also known for causing numerous human diseases and destroying livestock[1,5-6]. We can find nematodes in almost every known environment; in aquatic, marine and fresh water habitats; terrestrial habitats, soil or sedimentary environments and extreme environments, such as polar ice[7]. The classification of the *nematoda* phylum is based on their morphological and ecological traits. The traits used for this classification are the buccal and pharyngeal structures, along with the cuticle, lip region, intestine, reproductive system, sense organs and tail[1-4,8].

The recent development of molecular techniques, such as genome and transcriptome sequencing, along with bioinformatics allow us to obtain a much larger understanding of this phylum. By looking at these organisms on a genomic level, information on their evolution, distribution, behavior and preferences for feeding and inhabitation is possible and more easily accessible. With this type of analysis the factors that differentiate the diverse types of nematodes will become visible, allowing a better comprehension of these organisms. The first nematode to have his genetic code sequenced was *Caenorhabditis elegans* and following this case, several others have been sequenced[9].

1.2-Plant nematodes

Even though plant parasitic nematodes represent a minority within the *Nematoda* phylum, their consequences, when taken into account socio-economical factors, are dramatic[10]. These organisms are capable of destroying crops and threaten lumber related industries as well as other plant dependent industries[1,10]. This is a plague that is currently hard to fight due to the well developed parasitic mechanisms[10-11]. These nematodes have highly efficient parasitic machinery, resulting from their biology and physiological mechanisms. For example, they have special physiognomic structures that aid in plant penetration, root cell modification and food withdrawal, such as the stylet, responsible for

the wall perforation and the esophageal glands, responsible for producing secretions that modify the response of the plant cells to invasion[10-13].

The parasitic mechanism of the nematodes can be classified according to the way they approach and infect the plant. This categorization is mainly based on the position and progression of the nematode in the host. Therefore some nematodes are migratory ectoparasites remaining on the external part of roots and have limited interaction with its host, others are migratory endoparasites and cause damage to the host as they move through it. There is still a third class, the sedentary endoparasitic nematodes, which comprise the cyst nematodes and root knot nematodes that possess a very complex biotrophic interaction with their host[9,11,14].

1.3-*Bursaphelenchus xylophilus*

Among the plant parasitic nematodes *B. xylophilus* has been the subject of big focus and study, due to the economical losses it causes. While most of the nematodal species belonging to the *Bursaphelenchus* genera are exclusively fungal feeders, being transmitted only to dead or dying trees, *B. xylophilus* is known for its ability to feed on living plant cells, as well[14-15]. *B. xylophilus is* a migratory endoparasite[11,14] and its taxonomy dates back to 1970 when *Nickle*[16], based on the morphological characters and particularly the typical bursa, decided to transfer *Aphelenchoides xylophilus* to the genus *Bursaphelenchus spp.* [14,16]. Nowadays the nematode *B. xylophilus* follows the taxonomic classification presented in figure 1.

Kingdom: Animalia
Phylum: Nematoda
Class: Chromadorea
Order: Tylenchida
Family: Aphelenchoididae
Genus: *Bursaphelenchus*
Species: *B. xylophilus*

Figure 1 Taxonomic classification of the nematode *B. xylophilus*, adapted from National Center for Biotechnology Information (NCBI).

B. xylophilus was first identified in 1929 in the United States of America (USA) in a piece of wood in the process of house construction. This discovery was accomplished

due to the presence of bluish streaks on a piece of wood. After analysis, the blue regions were recognized as blue-stain fungi[17]. Analysis showed that besides the fungus, another organism was present. Scientists found that associated with the blue fungus were nematodes that after further analysis were proven to be *Aphelenchoides xylophilus*. The piece of wood where the nematode was found belonged to the longleaf pine, *Pinus palustrus*[17]. Since then this nematode has been referenced by several authors in other countries outside the USA.

1.3.1-Physiology

The knowledge of *B. xylophilus* morphological characters dates back to 1934, when *Steiner, and Buhrer*[17], dedicated their time studying this organism. Plant infecting nematodes share several physiological constituents (figure 2).

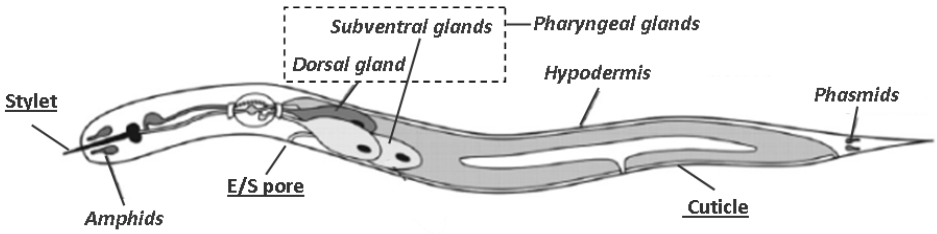


Figure 2 - Morphological constitution of a typical plant-parasitic nematode
(figure adapted from *Haegeman et al.*[11])
Morphological characters are underlined and glands are in italic.

B. xylophilus, like other plant-parasitic nematodes, has an external superficial multilayered cuticle that functions as an exoskeleton and a barrier to the entry of substances. The cuticle is a dynamic structure with a surface coat composed of proteins, carbohydrates and lipids[18]. Bacterial cells can be found attached to this exterior surface of the nematode body[19-20].

The nematode possesses glands that produce secretions that are useful for development and survival. The secretions used by the nematode for its physiological processes are produced in more than one gland spread throughout the body. These secretory organs are the amphids, the pharyngeal glands, the hypodermis, and the phasmids (figure 2)[11].

The two anterior sensilla, the amphids, are used for chemoreception which makes them important in the entire parasitism process as they help the nematode locate its host. The amphids are open to the surrounding environment and therefore their secretions are

directly exposed to the host playing an important role in the parasitic process[11]. The pharyngeal glands produce most of the nematode secretions, also known as effectors, that will participate in the parasitic mechanism. These effectors, are expelled by the stylet to the host cells[11,13]. In nematodes belonging to the Tylenchida order, the pharyngeal glands are composed of two subventral glands and one dorsal gland (figure 2). The functioning of the glands in nematodes from other orders is slightly different[11]. The hypodermis, the third type of gland, is responsible for the secretion of material to form the cuticle as well as the synthesis of proteins important for the parasitism process[11,18]. On the nematodes posterior end are the phasmids that have a role in the nematodes sensitivity and also produce secretions[11].

1.3.2-Life cycle

B. xylophilus has two development cycles, the propagative cycle and the dispersal cycle which are schematically represented in figure 3. These two cycles share in common four stages, the initial egg stage, the final adult form and two larval stages (L_1 and L_2) the remaining two development stages are specific for each cycle.

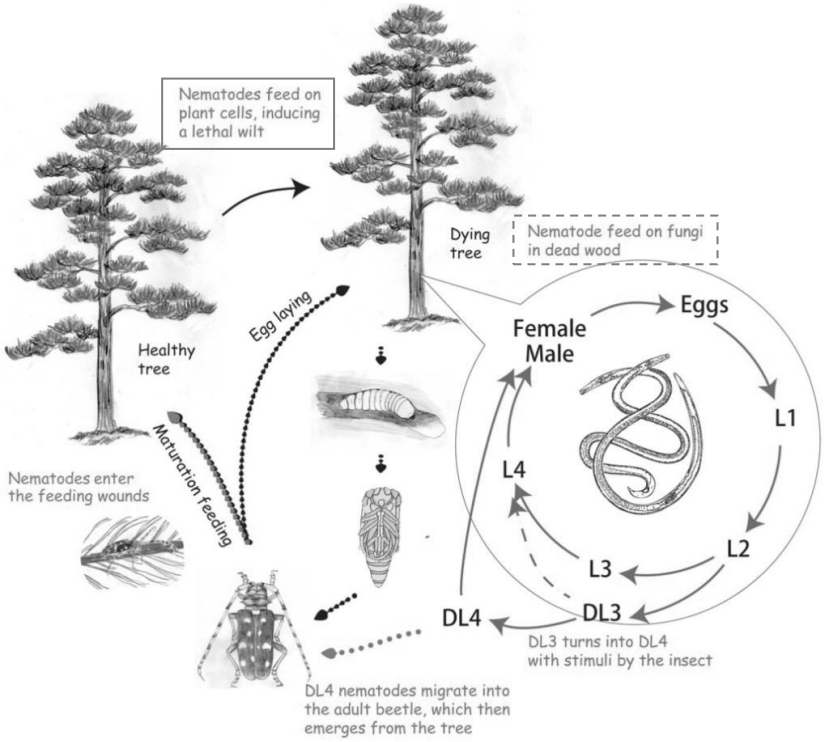


Figure 3 – *B. xylophilus* life cycle and feeding phases, adapted from *Kikuchi et al.*[9].
Life cycles are presented inside the circle. Propagative cycle: Eggs, L1, L2, L3, L4, Adult Female/Male; Dispersal cycle: Eggs, L1, L2, DL3, DL4, Adult Female/Male. Phytophagous phase (filled line rectangular) occurs when the nematode feeds on plant cells and the mycophagous phase (dashed line rectangular) occurs when the nematode feeds on fungi.

The first stage of the life cycle happens within the egg. The second stage, L_1, initiates right after the egg hatches and soon after initiating this phase, L_1 molts to the third stage L_2. Once the nematode reaches this stage it can evolve in two distinct ways. The nematode can either change into a fourth stage larvae and continue in the propagative cycle or evolve to a non-feeding dispersal stage and follow in the dispersal cycle. If it remains in the propagative cycle, the fourth stage larvae will eventually reach the adult stage, and after reproduction initiate a new cycle. The propagative cycle takes place in the sapwood under favorable conditions of temperature, moisture and nutrient accessibility[21]. In some situations the conditions for survival become unfavorable (lack of nutrients, adverse temperature and moisture conditions) and if the beetle pupae (nematode vector) is present, the nematode evolves into a pre-dauer larvae, the dispersal cycle. In this cycle, the third stage pre-dauer larvae aggregates on the wall of the beetle's pupal chamber and then molts to the dauer larvae stage. The dauer larvae enter the respiratory system of the beetle, being incorporated in its trachea, where it settles beneath the elytra[14,21-22]. Once the insect emerges and feeds in a conifer tree it transmits the dauer larvae to the host and in a period of aproximately 48h the dauer larvae evolves to the nematode's adult stage.

In figure 3 the development cycles of *B. xylophilus* are combined with his two feeding phases: the phytophagous phase and the mycophagous phase.

In the phytophagous phase *B. xylophilus* feeds himself on plants. This phase occurs when the nematode, still in its dauer stage is introduced in young pine trees by the beetle upon feeding. The emerging beetles migrate to young pine trees to feed on the bark of twigs, creating a feeding wound that the nematodes use to enter the tree. Once the nematode is introduced it evolves to the adult form and starts feeding on epithelial cells and resin ducts, quikly multiplying himself[21-22].

In the mycophagous phase *B. xylophilus* feeds on fungi (blue stain and other). This phase takes place in freshly cut softwood and in dead or dying conifer trees that have been invaded by opportunist fungus. This phase normally happens when the vector insect is in its oviposition transmission. The female bettle can lay her eggs in a recently killed or dying tree. The fourth stage dauer juvenile that are present in the insects trachea leave it to enter the tree through the oviposition slits created by the female beetle. Immediatley after entering the wood the dauer nematodes moult in to their adult stage, feed themselves on the available fungi and begin laying their eggs, and the cycle restarts[14,22].

Besides the morphological characteristics, *Steiner and Buhrer*[17] found that the *B. xylophilus* is associated with insects, using them as their carriers[14,17,23]. Thus the insects associated with *B. xylophilus* must have a larval development stage that as we saw has a partaking in the nematode life cycle. Among the possible insects *Monochamus spp.* was pointed as the most important vector of nematodes. This species is also known as the long-horn beetle and has long been identified as an inhabitant of conifer forests worldwide[23-24]. Among the *Monochamus* species, *Monochamus alternatus* and *Monochamus galloprovincialis* are the most important vector species[14]. In Portugal two species of *Monochamus* have been identified and documented as being vectors of *B. xylophilus*: *M. galloprovincialis*[23,25-26] and *M. suto*[25,27].

2-Pine Wilt Disease

The pine wilt disease (PWD) was first reported in 1905 in Japan[22]. The association of *B. xylophilus* as the causal agent of the PWD dates back to 1972[28] and in 1981 *Nickle et al*[29] classified *B. xylophilus* as the most important forest pest, causing serious monetary losses[29]. It is a devastating disease that if the environmental conditions are favorable kills pine trees in short time span[30].

2.1-Symptomatology

B. xylophilus, the pine wood nematode (PWN), infects the pine tree, but the visible signs of infection do not appear right away. Environmental conditions influence the progression of the infection and the density of nematode population. The physiological water status of the pine tree is also relevant to the development of the disease[14,24]. The temperature affects both *B. xylophilus* and the vector that transports it. In cooler temperatures, the vector has a shorter life span and his feeding habits are altered, feeding less. Summarily, the inhibitory effect of nematode due to temperature is based on three processes: reduced vector longetivity; reduced transmission efficiency and delayed nematode transmission[14,31]. As we see temperature affects PWD with regard on the biology of *B. xylophilus* and *Monochamus* and the biology of the interactions between themselves and between them and the host (pine tree)[31].

It is possible to monitor the presence and progression of the disease through the oleoresin flow that can be followed via wounds made on the trunk of trees[14]. According to *Xie and Zhao*[20], the PWD symtptoms can be divided into 6 stages (table 1). PWD stages go from the point where there are no visible signs of infection in the pine tree, until a stage where the pine tree is dead and all the visible signs have revealed themselves. In stage 1, even though the nematode has already entered the tree, there are no detectable signs of infection. However, throughout the stages the visible symptoms of the disease manifest themselves. The symptoms revealed by the tree as consequence of the infection with the PWN are mainly due to the inefficient water supply provoked by the destruction of the resin ducts by *B. xylophilus*[11,14]. The severity of the symptoms is influenced by the time of year in which the infection occurs. Infections that take place in the summer, lead to a rapid death of the pine tree (40-60 days) while infections in spring take longer to develop. Infections in autumn and winter may result in no development of symptoms[14,31].

Table 1- Symptoms of pine wilt disease, adapted from *Xie and Zhao*[20].

Stage	Symptom
1	Pine tree has a healthy aspect and resin flow is normal.
2	Resin secretions are reduced or ceased and the needles near the inoculated site are wilted or no longer shining.
3	Resin secretion ceased completely. Part of the needles on twigs near the inoculated twig turned yellowish or brown.
4	Start appearing yellow needles on 1-year-old twigs that were not inoculated.
5	Part of the needles on the top leader turned grayish or yellowish. No green needles were observed in the whole pine tree.
6	Most of the needles on the top leader turned brown and wilted and all the other twigs turned brown and the tree died completely.

2.2-Disease mechanisms

In order to develop and reproduce in the tree the nematode needs to overcome the plants defence mechanisms and create a survival interaction[14]. It is the nematode head that takes up the plants defence responses with the help of the stylet, that besides creating a passing route between the plant cells, also secretes efectors that soften and degrade the cell wall[7,11,24]. After entering the tree the nematode starts feeding on parenchymal cells and moves rapidly from the inoculation point to woody tissues, via the resin canals of xylem and cortex, feeding on their epithelial cells[7,11,14,32].

The movement of nematodes along the resin canal leads to a reduction in oleoresin exudate flow as the nematodes feed and reproduce inside these canals[14,24]. After two to three days following the invasion of PWN in the pine tree, the vascular system is compromised. The nematode increases the production of pine volatile substances for self defence[14,24,32]. These volatile subtances, such as ethylene, are exuded from the tree parenchyma and help cut the water columns under tension, leading to dehydration of tracheids and cavitation with xylem blockage[14,32]. The lack of a normal water flow throughout the pine tree leads to the worse and more visible symptoms, discoloration of pine needles and tree death.

2.3-Tree host

Even though the disease is called pine wilt disease (PWD), not all species of *Pinus spp.* are susceptible of infection by *B. xylophilus.* The pine species that have been identified as susceptible of dying of PWD are *P. bungeana, P. desinflora, P. luchuensis, P. massoniana, P. thunbergii, P. nigra, P. sylvestris and P pinaster*[22]. A major drawback is that species that are predisposed to infection by this parasite die in a short period[22].

Throughout time, some tree species have co-evolved with *B. xylophilus,* thanks to their natural and high tolerance to this pest. Amongst these pine species are the loblolly pine (*P. banhsiana),* the eastern white pine (*P. strobus*) and the table mountain pine (*P. pungens*)[14]. Interestingly, trees in North America, where the disease was first reported as being caused by *B. xylophilus,* have developed tolerance to this pathogen[9]. In tolerant trees the multiplication of the nematode population is prevented so that by the end of the infection the number of nematodes in the tree is lower than the number of nematodes that infect the plant[14,33]. It is thought that resistance arises from the expression of resistance genes and a hypersensitive response. A hypersensitive response towards the infection with PWN has already been reported in resistant trees[34].

Even pine trees that are not resistant to the PWN have their own natural defense mechanisms against this and other parasites. Plants produce toxins to protect themselves against pathogens. Some of these synthesized compounds have a broad spectrum of action so that the plant can have a defense mechanism against a wide number of pathogens. Phytoalexins, isoflavonoids or terpenoids are some of the anti-pathogenic compounds and some of these are known as being nematicidal[11]. Another class of compounds produced by

plants and used in their defense mechanism are reactive oxygen species (ROS). Histopathological studies have shown that these compounds are synthesized in response to nematode invasion and might also activate other signaling pathways that may lead to the strengthening of cell walls[11].

2.4-Pathogenic interactions

Even before a major increase of the nematode population within the tree, there seems to be an extensive death of pine cells, suggesting that the nematode may produce phytotoxins that may lead to this situation[14,35]. Two phytotoxins 8-hydroxycarbotanacetone and 10-hydroxyverbenone, have been identified and characterized in trees infested with *B. xylophilus*[14]. More recently it was found that the disease process might be influenced by a group of bacteria that is associated with the PWN and that favor the pathogenic mechanism, probably by the release of toxins[14,19-20,35-36]. It has also been suggested that PWD is the result of a complex infectious process of the nematode and the bacteria associated with it[37-39]. It is important to keep in mind that the bacteria alone cannot cause PWD[14,36,38].

Many studies have been conducted to determine which bacterial strains can be found associated with *B. xylophilus* and the dynamic of such association. *Xie and Zhao*[20], analyzed chips from twigs that were not inoculated, in different stages of the disease and found that in the first stages, bacterial colonies were absent, but with the progression of the disease, these became more representative and diverse. Recent published investigations on this matter reveal that PWD development decreases when nematode surface is sterilized[22,40] and that the bacterial colonies associated with *B. xylophilus* differ between geographic regions[19,40-41]. Thus it is possible to find bacteria belonging to the genera *Bacillus* in Japan, *Pseudomonas* in China, both are present in Korean samples, and according to the studies in Portuguese isolates bacteria belonging to the genera *Pantoea, Klebsiella* and *Serratia* are the most common[19-20,40-41] .

2.5-Disease progression

Tree diseases, like PWD cannot be confined to geographical or geopolitical boundaries and therefore, throughout time, many important trees on a worldwide scale have been destroyed due to pests[42]. The timber trade was responsible for the spread of *B.*

xylophilus to Japan, China, Taiwan and Korea, and later on to Europe.[30,43] The consequences of this epidemic spread are so devastating that by 2004 a large area of pine forests in these regions had been destroyed[43]. In 1999 the PWN was identified on the maritime pine, *Pinus pinaster* near Setúbal, Portugal. This was the first time the virulent species of *Bursaphelenchus* was reported in Europe. However, over the years, this pest has spread to pine trees in other locations of the country affecting negatively the economical entities that rely on healthy pine trees to generate products with economical significance such as lumber, resin, pulp and pine seeds[30,44].

Due to the big losses caused by PWN, *B. xylophilus* is now considered in the European Union as a European Plant Protection Organization (EPPO) quarantine pest, legislated by the Directive 77/93 EEC. This directive comprises the nematode *B. xylophilus* and it's vector *Monochamus spp*.

Among the nematodal species of the *Bursaphelenchus spp.* genera only *B. xylophilus* is widely known as having pathogenic properties in normal environmental conditions, causing PWD[14]. However we can find several reports mentioning other species of the *Bursaphelenchus spp.* genera with pathogenic properties, but only in laboratory assays; none of them have been reported as pathogenic in natural conditions[14,36,45].

3-Molecular mechanisms of parasitism

Little is known regarding the molecular mechanisms of the interaction of *B. xylophilus* with its host. On the other hand, some progresses have been accomplished in this field with the creation of expressed sequence tags (EST) libraries and genome sequencing projects that allowed the discovery of genes coding for proteins that play a role in the molecular mechanism of the disease[9,46-47]. Events partaking in the parasitic mechanism have been characterized independently, but no further interdependent relations have been established among them. At least three pathways have been proposed for the pine wood nematode parasitism process; effectors produced in *B. xylophilus*' glands which are expelled through the stylet[11,48], bacteria associated with the nematode[5,19-20,40-41] and acquisition of genes by HGT from bacteria and fungus, providing *B. xylophilus* with the tools to infect the host[9,15,47,49-50]. In order to develop solutions that will help solve or at least control the spread of *B. xylophilus*, it is necessary to know which factors trigger the

expression of each gene and the gene networks that contribute to the pathogenetic mechanism of the PWN.

3.1-Genomics and transcriptomics

Researchers are working on the sequencing and assembly of the genome of several organisms, such as human, mouse and other organisms with the main goal of knowing the sequence of genes and their organization. The transcriptome provides the information for understanding when, where and how each gene is expressed. The knowledge of *B. xylophilus* genome sequence would be of much interest due to the vast amount of information that would be available regarding its biology. This information would be of utmost use to identify the mechanisms beneath the pathogenic ability of *B. xylophilus,* and provide information on potential targets to control the spreading of the disease (enzyme pathways or essential genes)[14]. The gathering of genomic and transcriptome information is only possible due to the development of next generation sequencing technologies (NGS).

3.2-Next generation sequencing technologies

Pyrosequencing is a next generation sequencing (NGS) method, a high throughput technology that allows dramatic increases in cost-effective sequencing based on real-time DNA synthesis that is monitored through bioluminescence. NGS technologies have been used for genome sequencing, resequencing as well as metagenomics. Sequencing services carried out by the Advanced Services Unit at Biocant rely on the Roche 454 Sequencer, the first next-generation sequencing technology released. The 454 sequencing is based on emulsion PCR, a highly efficient *in vitro* DNA amplification step. In this PCR method, individual DNA fragments obtained by DNA shearing are attached onto beads by adapters, and afterwards incorporated in separate emulsion droplets. These droplets act as individual amplification reactors, producing approximately 10^7 clonal copies of the template fragment attached to the bead. Each template containing bead is subsequently transferred into a well (each well contains only one bead) on a picotiter plate. The template attached to the bead is sequenced in the Roche 454 by solid-phase pyrosequencing technology, based on three enzymatic reactions that produce light (figure 4). In these reactions, dNTPs are added one at a time to the immobilized template on the bead and in the case of incorporation, a chemiluminescent signal proportional to the number of incorporated dNTPs is released and detected by a CCD (charge-coupled device) camera[51-53].

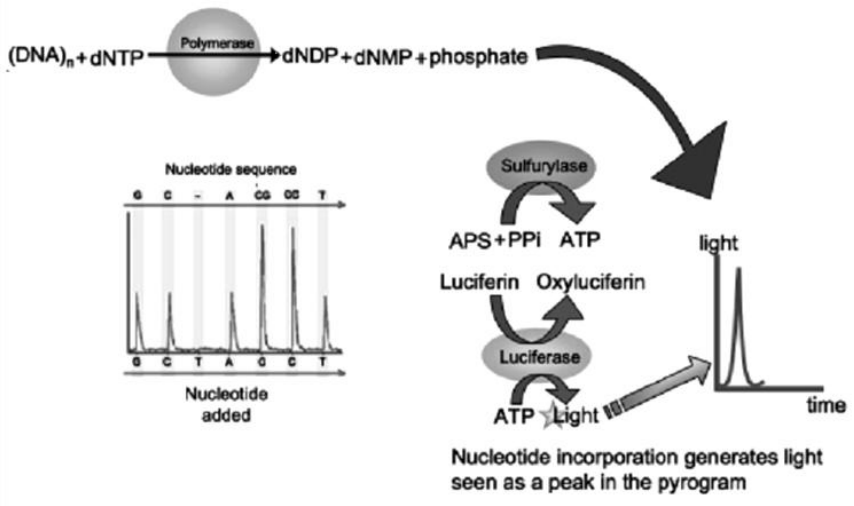


Figure 4 - Schematic view of the pyrosequencing reaction and the enzymatic reactions involved, adapted from *Petrosino et al.*[54].

3.3-Genome of *B. xylophilus*

Recent progresses on the genome of *B. xylophilus* have been accomplished by *Kikuchi et al.*[9]. This research team conducted a study to look for evidence and mechanistic triggers used by *B. xylophilus* through the analysis of the genome sequence. According to their results the genome of *B. xylophilus* is 74.5 Mb long and through assembly they obtained 6 pairs of nuclear chromosomes. *B. xylophilus'* genome size is smaller than *C. elegans* and some of the other published nematodal genomes, except *M. hapla.* However the GC content was 40.4%, higher than the other sequenced nematodes except *P. pacificus*. Mitochondrial genome shows gene content and organization similar to *C. elegans.* They found 18 074 protein coding genes, less than *C. elegans* (20 416) and *M. incognita* (19 212), but bigger than *M. hapla* (14 420), though the average protein length is similar. However, *B. xylophilus* displays the largest mean exon size (289 bp), the smallest average number of exons per gene (4.5) and the smallest mean intron size (153 bp). A comparison of *B. xylophilus* protein set with those of other nine nematodes revealed that the PWN genome is relatively conserved over the long divergence from other plant parasitic nematodes.

In their study, *Kikuchi et al.*[9], reported important information about *B. xylophilus* biology and genes involved in development, chemoreception and neuropeptides. A thorough description of the RNA interference (RNAi) pathway was also documented, enlightening the use of this mechanism as a tool to study the nematode biology and

pathogenic mechanism. Their quest for carbohydrate active enzymes (CAZymes) found cellulases (11 GHF 45), pectinases (15 PL3), expansins (8), that can act upon plant cell wall and 1,3-glucanases (6 GHF 16), chitinases (9 GHF 18, 2 GHF 19, 7 GHF 20) that participate on the hydrolysis of the fungal cell wall. These enzymes were compared with those of other five nematodes and GHF 45 as well as GHF 16 were found to be present only in *B. xylophilus*. The analysis of the genes encoding both proteins revealed their high similarity to those of fungi and bacteria, respectively, suggesting that they might have been acquired by HGT. These authors also found 581 peptidase genes, the largest number found in any characterized nematode genome. Peptidases in nematodes partake critical roles in physiological processes including embryogenesis and cuticle remodeling during larval development and also in the parasitic process, such as tissue penetration, digestion of host tissue and evasion of the host immune response. The identified proteins belong to the peptidase families involved in extracellular digestion and lysosomal activity. One of the identified endopeptidase families, aspartic-type endopeptidase, also seems to have been acquired by HGT from ascomycete fungi. On their analysis to the effectors, secreted proteins that mediate parasitic interactions, they found three other proteins, encoded by genes, potentially acquired by HGT. These were 6-phosphogluconolactonase from bacteria, a protein containing a cystatin like domain, from gamma proteobacteria and HAD-super family hydrolase from bacteria.

3.4-Transcriptome

The information contained in the genome is used to build and maintain cells and is transcribed into messenger RNA (mRNA). The collection of all mRNA in a given cell is called the transcriptome[55]. The transcriptome is composed only by the genes that encode proteins therefore it is different from the genome. When taking up a genetic study through this method, only the genes that are being expressed at that moment are taken into account[55].

To obtain information from the transcriptome of *B. xylophilus* regarding its pathogenic ability, *Kikuchi et al.*[15] cloned and analyzed 5 000 EST. This project revealed the presence of genes in *B. xylophilus* that might have been acquired by HGT. Therefore, in 2007, *Kikuchi et al.*[46] took up a new EST project, but this time on a much larger scale. The generated library had a total of 13 327 ESTs (transcripts) from *B. xylophilus*. In this

study, the research team included another nematode species, *B. mucronatus*, with close relation to *B. xylophilus*, but lacking the ability to parasitize plants. For this non-pathogenic nematode species, the number of generated EST sequences (3 193) was slightly smaller than the amount for the PWN. The team found genes encoding for proteins with structural and metabolic functions in the nematode, such as actin, myosin and collagen that belong to the cytoskeleton and cytochrome C oxidase, ADP/ATP translocase and elongation factor proteins that participate in the energetic and metabolic processes. Genes potentially acquired via HGT, with no prior reference, were also identified in this assay as encoding for chitinases and expansins.

B. xylophilus life cycle has two stages which comprise different morphologic and physiologic characteristics, therefore, it is also expected that genes expressed in each one of these phases will be different. In order to check for such differences in gene expression throughout the PWN life cycle, *Kang et al.*[56] developed a study elucidating gene expression at specific stages of *B. xylophilus* development. These researchers conducted their experiments with two subtractive *B. xylophilus* cDNA libraries, to study the dispersal 4th larva stage (D4S) and the pine-grown propagative mixed stage (PGPS). They found in their PGPS library a group of proteins, the allergen-like proteins, which have also been found in other plant parasitic nematodes and are suggested to play an important role against the host. In their study they found vap1 and vap2, two proteins rich in cystein, that were known to be secreted from the esophagus gland cells, during the parasitism process. Vap2 was found to be over expressed in the PWN propagative phase when growing in pine trees, leading the research team to suggest that this protein plays a crucial role in *B. xylophilus* parasitic process. This type of analysis could also be of great interest to expose the molecular basis between the host-nematode interactions in each step of the development.

Another input into the molecular knowledge of *B. xylophilus* was the transcriptome pyrosequencing at the Advanced Service Unit at Biocant. Using 454 pyrosequencing technology, this group sequenced the transcriptomes of seven PWN isolates from four distinct geographical locations: Portugal, China, Japan and United States of America. Although the aim was to discover differences in the genome related to geography, the study produced 16 297 transcripts, corresponding to 10 776 transcripts with conserved protein domains as identified through InterPro and 7 969 transcripts matching the Gene

Ontology terms. The transcriptome and its corresponding protein information were organized in a searchable database, where targeted searches can be made on each individual population. Additionally, by assembling together the set of transcripts from all populations, the group generated a digital transcriptome of the species *B. xylophilus*. Genes characteristic of pathogenicity in this nematode were identified in the transcriptome, namely cellulases, chitinases, expansins or venom allergen proteins (unpublished results). Transcripts encoding proteins with bacterial and fungal origins were also found in this transcriptome. This evidence instigated, in this unit, the quest for proteins eventually incorporated in the genome of *B. xylophilus* by HGT.

3.5-Horizontal gene transfer

One of the aspects of the molecular mechanism of parasitism by *B. xylophilus* is the presence of new features that participate and influence the parasitic interaction of the nematode with the pine tree. These new features are acquired by HGT, also known as lateral gene transfer. This is a mechanism that explains the transmission of genetic material between individuals of different species and different organisms. Through HGT the genetic information is not necessarily transmitted by reproduction or vertical inheritance. It has been demonstrated that HGT can occur between cells, viruses, plasmids and also between closely-related bacterial strains, as well as transfer between domains. HGT aids organisms acquire genes with new functions and therefore an important mechanism in shaping genomes and generating the ability of certain organisms to acquire 'tools' that will help them take up new processes[57-58]. A gene incorporated by HGT in a foreign genome may suffer changes in its size and number of introns[50].

3.5.1-Horizontal gene transfer in nematodes

The transcriptome analysis of some nematodes such as *M. hapla* and *M. incognita*, belonging to the Tylenchida order, found quite a few cell-wall-degrading enzymes that are of much use in the parasitic process. Among these enzymes are cellulases[50,59], xylanases[60], other members of the glicosylhydrolase family[11,61], polygalacturonase[11], pectate lyases[62] as well as proteins that break the non-covalent bonds between plant cell walls[63]. The presence of such enzymes in these nematodes, along with the genes that encode them, and the absence of these in other free living nematodes, suggest that these genes were acquired by HGT, most probably from bacteria or unicellular organisms where very similar genes can

be found[7,63]. *Mayer et al.*[50] demonstrated that nematodes from different genera, in their case, *Koerneria* and *Pristionchus*, had genes encoding for the same enzyme, acquired by HGT that did not precede from the same donor[50].

3.5.2-Horizontal gene transfer in *B. xylophilus*

Horizontal gene transfer (HGT) is an hypothesis that might explain how the nematode *B. xylophilus* could have acquired the necessary tools to become a parasite of conifer trees. As we have seen until this moment, various genes encoding proteins have been suggested to be acquired and incorporated in *B. xylophilus* genome by HGT. *B. xylophilus* has incorporated genes that code for cell wall degrading enzymes and effectors as well as proteins involved in physiological processes. The proteins that have been identified in *B. xylophilus* as potentially being incorporated by HGT are resumed in table 2.

Table 2 - Proteins identified in *B. xylophilus* and proposed to have been incorporated in its genome by HGT.

Acquired protein	Origin	Reference
GHF 45	Ascomycete fungi	*Kikuchi et al.*[9,15]
GHF 16	Gammaproteobacteria	*Kikuchi et al.*[9,47]
Expansin-like	Bacteria	*Kikuchi et al.*[49]
Pectate lyase	Bacteria	*Kikuchi et al.*[62]
Aspartic-type endopeptidases	Ascomycete fungi	*Kikuchi et al.*[9]
6-phosphogluconolactonase	Bacteria (firmicutes)	*Kikuchi et al.*[9]
Protein with cystatin like domain	Gammaproteobacteria	*Kikuchi et al.*[9]
HAD-super family hidrolase	Bacteria (firmicutes)	*Kikuchi et al.*[9]

The PWN ability to secrete cellulases may assist the pathogenic process as well as its migration within the host, since cellulose is the major component of plant cell walls[14]. *Kikuchi et al.*[15] found among their transcripts a sequence with similarity to cellulases (Bx-eng-1). The identified transcript was 762 bp long, and possessed an ORF of 672 bp. The SignalP program deduced a Bx-ENG-1 protein sequence that contains in its N-terminal a predicted 15 amino acid signal that presumably will be found in the mature peptide. Further EST analysis, identified two other transcripts (Bx-ENG-2 and Bx-ENG-3) that could also encode proteins similar to GHF 45 cellulases. These last two identified transcripts also enclosed the predicted peptide signal at the N-terminus, yet these transcripts were slightly bigger than the first one, however this difference did not influence

the sequence alignment. BLASTX software analysis showed that the three transcripts from *B. xylophilus* were highly similar to GHF 45 cellulases from fungus.

The ability of *B. xylophilus* to secrete β-1,3-glucanases is not relevant to its pathogenic mechanism and consequent phytophagous phase, however, these proteins play an important role in its mycophagous phase of development. β-1,3-glucan is the major constituent of the fungal cell wall and β-1,3-glucanases cleave the β-1,3-D-glucosidic bonds between the monomers[9,47]. During an EST project on *B. xylophilus, Kikuchi et al.*[47], found a transcript (Bx-lam-16A) encoding for a protein BxLAM16A with similarity to β-1,3-glucanases. The identified transcript was 803 bp long, containing a 753 bp ORF. The predicted amino acid enclosed a hydrophobic peptide signal to be cleaved between Ala^{15} and Gly^{16}, originating a mature peptide with an expected molecular mass of 26 441.73 Da and a predicted pI of 4.7. A BLASTp search found that the deduced BxLAM16A peptide sequence was most similar to GHF 16 β-1,3-glucanases from bacteria.

Expansins are, like GHF 45, another type of proteins that could play a part in *B. xylophilus* pathogenic mechanism because despite not degrading the plant cell wall, these proteins are thought to lose non-covalent bonds that hold cells together, easing the enzymatic attack[49]. *Kikuchi et al.*[49] in an EST analysis on *B. xylophilus* found a transcript that could encode for a expansin-like protein. The detected coding region was 593 bp long and contained a 220 bp ORF. The protein encoded by this gene is 171 amino acid long with a predicted molecular mass of 17.602 kDa. SignalP program predicted a signal peptide near the N-terminal of the predicted protein sequence. Searches with BLASTp and BLASTX revealed that the predicted expansin-like peptide sequence is similar to those found in bacteria but could also be identified in potato cyst nematodes (*G. rostochiensis*).

Pectin, like cellulose, is another component of the plant cell wall. Pectin is a structural heteropolysaccharide that can be degraded by pectate lyases. These enzymes can be found in *B. xylophilus* and are produced in the esophageal gland cells being secreted through the stylet. Pectate lyases can also be found in the non-pathogenic nematode *B. mucronatus*, therefore suggesting that this plant cell wall degrading enzyme is used for migration in both living and dead plant tissues[62]. *Kikuchi et al*[62] constructed an EST library and searched the generated transcripts through BLASTn and BLASTX programs to detect similar DNA and protein sequences in databases. Searches came up with a transcript (Bx-pel-1) encoding for a protein similar to pectate lyase. The identified transcript sequence

was 848 bp long with a 252 bp ORF. SignalP program detected a 18 amino acid signal near the N-terminus of the deduced BxPEL1 polypeptide. The predicted mature protein had a molecular mass of 25 kDa and a theoretical pI of 8.45. Further EST analysis found another sequence (BxPEL2) that can also encode a protein alike to a pectate lyase from the polysaccharide lyase family. Both sequences, BxPEL1 and BxPEL2, shared 64% amino acid similarity. BLAST search revealed that the genes encoding the pectate lyase incorporated in *B. xylophilus* genome were more similar to the gene sequences found in bacteria.

The last four proteins presented in table 2 (Aspartic-type endopeptidases, 6-phosphogluconolactonase, protein with cystatin like domain and HAD-super family hidrolase) were identified by *Kikuchi et al.*[9] in their research on the genome of *B. xylophilus*, but no further information regarding the molecular organization of the genomic sequences encoding these proteins has been published until the moment.

3.5.3-Horizontal gene transfer analysis

One of the simplest criteria to suspect that a gene has been incorporated by HGT is by analyzing the amino acid sequence of the protein it encodes. However, the similarity of a bacterial protein sequence with the sequence of a nematode protein is not enough to prove it was incorporated in the nematode's genome by HGT. In order to prove so, it is essential to testify incongruence of that gene with the nematode phylogeny[64]. Initial approaches to identify HGT cases were based on biochemical and immunological assays. With the development of bioinformatics tools, the quest for HGT candidates became faster and more efficient[11]. The quest for HGT candidates can be resumed according to figure 5.

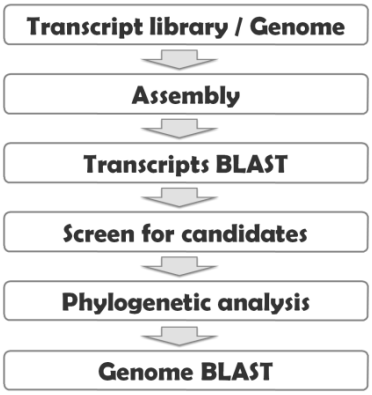


Figure 5 - Outline of the process to identify HGT candidates.

Next generation sequencing is the technology currently in use to put together transcript libraries. Transcript sequences are clustered according to the similarity of their sequence, generating a final consensus sequence, the transcript. The transcript sequence codes for a protein and in order to find the corresponding protein, a BLASTX search on a non-redundant database, like UniBank or GenBank, is performed. This search generates a database formed by proteins and the identification of the organism where it is expressed. BLASTX search also returns an expected value (E-value) that quantifies the level of similarity of the transcript sequence, with the sequence that is found in the database. At this point a number of criteria need to be established to trim down the number of candidates. Amongst the options that can be used to eliminate unviable candidates are BLAST hit E-value, transcript sequences with no identified protein in database, proteins encoded by organisms with no relation to the nematode, making it impossible to transfer genes (for example, humans), or proteins with an unspecified function (hypothetical proteins). Criteria used in HGT analysis have to be carefully chosen so that unviable candidates can be screened leaving only those with a bacterial or fungal origin, an E-value$\leq 10^{-6}$ and a specified protein function. Therefore the candidates that emerge from the screening process are analyzed according to their evolutionary relationships, phylogenetics. Phylogenetic analysis is the method performed to determine or estimate these evolutionary relationships that are frequently represented as branches in tree-like diagrams, called phylogenetic trees[65]. The sequences submitted to phylogenetic analysis are first subjected to multiple sequence alignment which generates gap scores. This information will further on be used to build a phylogenetic tree to determine the origin of the suspected protein. After identifying a gene incorporated by HGT it is crucial to prove it is viable and functional. Biochemical tests should be performed to determine protein activity and assess the process evolution from the original organism.

In *Kikuchi et al.*[47] the sequence of β-1,3-glucanases from *B. xylophilus* was aligned using CLUSTALX (ver. 1.81) against β-1,3-glucanase like proteins from bacteria, fungi and animals. BLOSUM (BLOck SUbstitution Matrix) 30 was used as the protein weight matrix. BLOSUM 30 is a series of matrices that are used to correct the sequence changes between long and closely related species. Gap-opening and gap-extension penalties had to be tested to determine the factors that would result in a better alignment. Little manual corrections needed to be performed on the aligned sequence, based on the crystal structure

of endo-β-1,3-1,4-glucanases from *Bacillus licheniformis* and *Bacillus macerans*. Furthermore phylogenetic trees were constructed based on the alignment with neighbor joining and maximum parsimony methods through PAUP* v. 4.0b10. This methodology created trees with similar topology. In *Kikuchi et al.*[15], the phylogenetic analysis of GHF 45 cellulase protein sequences, was performed using CLUSTALW and SOAP v1.1 software to align the sequences. To carry out the alignment, the signal peptides and the N-terminal extensions of the Bx-ENG-2 and Bx-ENG-3 peptides were removed, yet leading to no effect on the overall alignment. For the alignment a series of 35 parameters were defined to filter out ambiguous alignment sites (gap opening and extension penalties and 100% conservation across alignment). A phylogenetic tree was constructed based on the maximum likelihood analysis from the aligned sequences, just like the previous case. These two reports of *Kikuchi et al.* used the same software to align their sequences, but used different tools to fine tune the alignments.

Kikuchi et al.[9,49] and *Richards et al.*[58] did not share much detail on the phylogenetic sequence analysis, only that it is accomplished with MUSCLE. However, *Richards et al.*[58] mentioned the use of GBlocks to sample conserved regions in the alignment. As seen in references, alignments performed with MUSCLE took up less steps than those conducted with CLUSTAL. In fact, MUSCLE offers improvements in scalability with comparable accuracy[66]. CLUSTAL and MUSCLE are the main multiple sequence alignment softwares used in the referenced searches, yet, others are available (table 3). Though MUSCLE is quit a viable option for most alignment tasks[66].

Table 3 - Summary of the multiple sequence alignment programs, adapted from *Edgar and Batzoglou, 2006* [66].

Program	Advantages	Cautions
CLUSTALW	Uses less memory than other programs	Less accurate or scalable than modern programs
DIALIGN	Attempts to distinguish between alignable and non-alignable regions	Less accurate than CLUSTALW on global benchmarks
MAFFT, MUSCLE	Faster and more accurate than CLUSTALW; good trade-off of accuracy and computational cost. Options to run even faster, with lower average accuracy, for high-throughput applications.	For very large data sets (say, more than 1000 sequences) select time- and memory-saving options
PROBCONS	Highest accuracy score on several benchmarks	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)
ProDA	Does not assume global alignability; allows repeated, shuffled and absent domains.	High computational cost and less accurate than CLUSTALW on global benchmarks
T-COFFEE	High accuracy and the ability to incorporate heterogeneous types of information	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)

4-Challenges to the academic community

The molecular mechanism behind the parasitism ability of *B. xylophilus* is not well known, however the economical losses that this pest generates on a worldwide scale are tremendous. Some mechanisms and factors that lead to the development and spread of the PWD have already been identified but a more profound knowledge of the biology of this organism will help develop methods to combat and restrain this disease.

4.1-Purpose of this work

As we saw, HGT is one of the proposed mechanisms by the scientific community that led *B. xylophilus* to acquire the ability to cause PWD. Some genes present in *B. xylophilus* have already been found to most likely be acquired by HGT from bacteria and fungi, but almost certainly more genes have been incorporated in the genome of the PWN through this route. The transcriptome of *B. xylophilus* prepared at Biocant contains hints of potential HGT candidates, therefore it is important to find all these candidates as well as their phylogenetic origin, to identify new parasitic pathways and discover new targets for the design of new strategies to control the disease.

Chapter II – Materials & Methods

1-Pipeline development

1.1-Background

The transcriptome of *B. xylophilus* was sequenced at the Advanced Services Unit at Biocant, using the 454 pyrosequencing technology (454 Life Sciences, Roche, Branford, USA). Results were generated from seven PWN isolates (table 4) cultured on *Botrytis cinerea* and maintained in laboratory conditions, originating from four distinct geographical regions: Portugal (4 isolates), China (1 isolate), Japan (1 isolate) and USA (1 isolate).

Table 4 - Summary of the samples used, regarding their origin and the concentration of the extracted genomic DNA.

Origin	Region	Sample reference	Concentration (ng/µL)
Portugal	Santiago do Cacém	**Pt15**	153.3
	Alcácer do Sal	**Pt17**	181.5
	Santa Comba Dão	**Pt19**	177.9
	Tábua	**Pt21**	151
China	-	**ChJs**	182.5
Japan	-	**J10**	141.9
USA	-	**USA618**	180.8

(-) Not applicable

The results were assembled and annotated according to *Bettencourt et al.*[67] (the methodology is resumed in figure 6) and organized in a searchable database for posterior analysis.

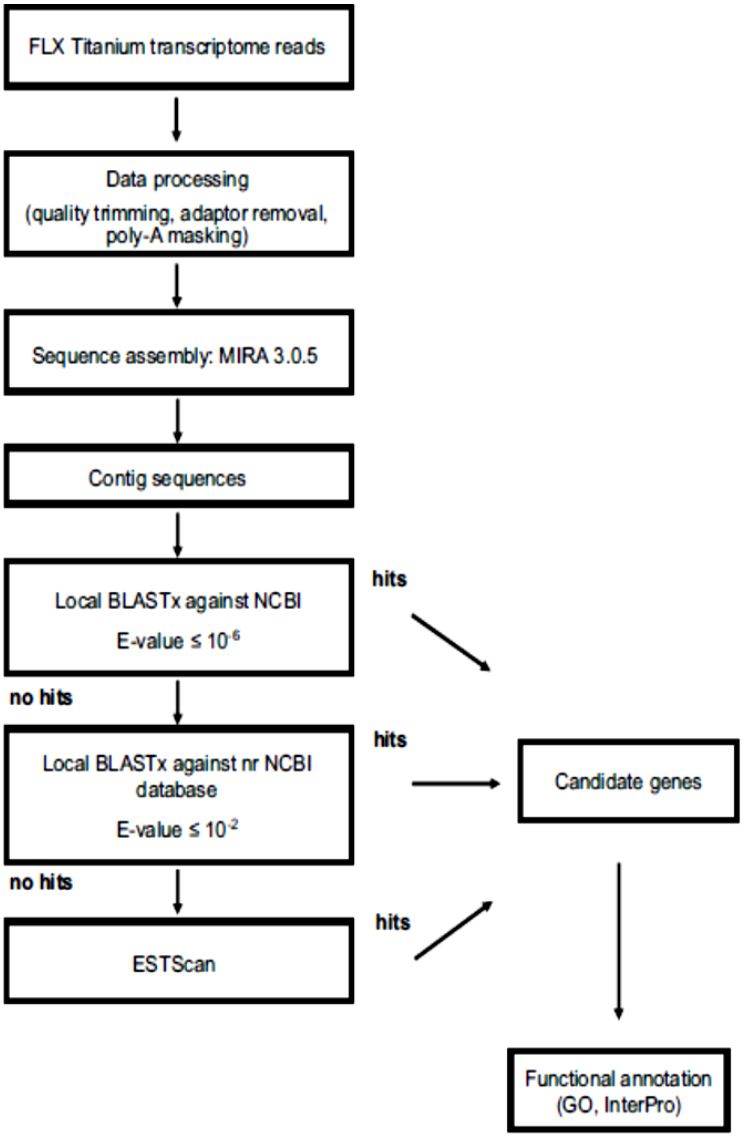


Figure 6 – Flowchart representing the data processing pipeline for *de novo* transcriptome assembly and annotation adapted from *Bettencourt et al.*[67]

1.2-Screening using information held in the database

The transcripts were placed in the database along with their BLASTX, InterPro and gene ontology description. BLASTX description contained the expected protein, the organism from where it originated and the corresponding E-value. Based on this information the transcripts were screened using the parameters E-value, recognized protein function and taxonomy, according to figure 7, until the appropriate pipeline was met.

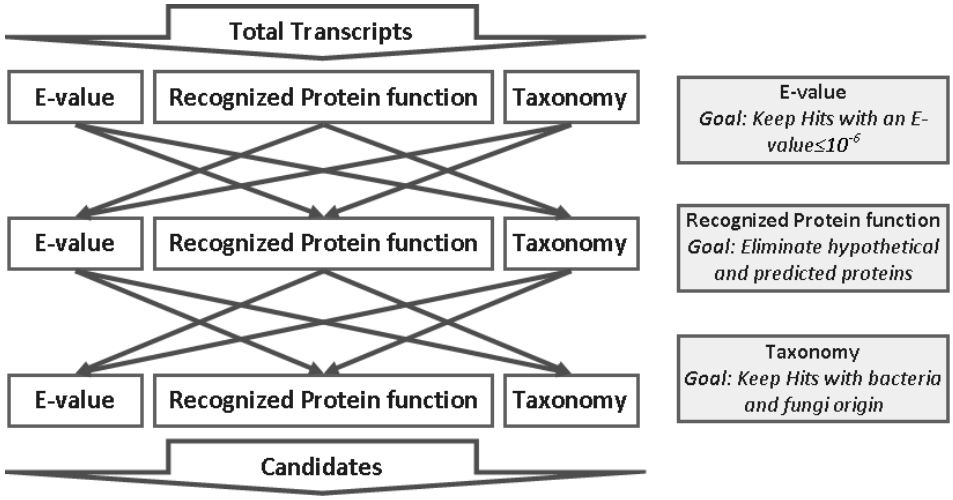


Figure 7 – Ordering trials of the chosen parameters in the development of the final pipeline.

1.3-*B. xylophilus* microbial community

The composition of the microbial community living with the nematode was incorporated in the pipeline and used as a filter to the candidates resulting from the screening process presented in section 1.2.

1.4-Phylogenetic analysis

The candidates resulting from section 1.3 were individually submitted to phylogenetic analysis according to the outline presented in figure 8, to determine their origin.

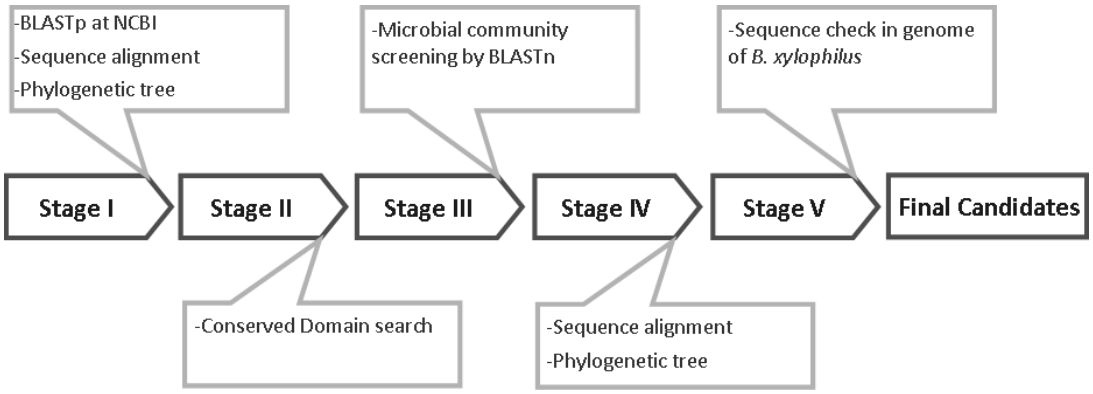


Figure 8 – Outline of the phylogenetic analysis.

All sequence alignments were performed using MUSCLE[68-69] and the phylogenetic trees were built in MEGA 5 using the Maximum Likelihood method[70]. All BLAST and conserved domain searches were conducted at the National Center for Biotechnology Information website *http://blast.ncbi.nlm.nih.gov/* and *http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml*, respectively. Final candidates were checked in the genome of *B. xylophilus* at *http://www.ncbi.nlm.nih.gov/genome?term=bursaphelenchus*.

In stage I the transcript was submitted to BLASTp search against bacteria, in first place, fungi, in second, nematodes in third and in fourth place other organisms excluding the last three tested classes. Three to four sequences, resulting from the search, containing the highest score and E-value were downloaded. These sequences were afterwards aligned according to the parameters presented in table 5a) and their phylogenetic tree was built following the parameters presented in table 5b). Transcripts that appeared in the phylogenetic tree grouped with nematodes were discarded, remaining those that grouped with bacteria and fungi. In stage II the transcript was tested for the existence of a conserved domain and only transcripts containing a conserved domain were maintained. In stage III the transcript was screened with the results of the surrounding microbial community by BLASTn search. This step was introduced to ensure the transcript sequence was not homologous to the genera identified as belonging to the microbial community surrounding the PWN. In stage IV the transcript was aligned with the sequences downloaded in stage I, according to the parameters of table 6a) and the phylogenetic tree built according to the parameters in table 6b). In stage V, the remaining transcripts were checked in the genome of *B. xylophilus*.

Table 5 - Parameters used for phylogenetic analysis in Stage I

Parameters used for sequence alignment in MUSCLE a) and phylogenetic analysis in MEGA 5 b).The presented conditions were applied in Stage I of the phylogenetic analysis.

a) Parameters for sequence alignment

Gap penalties	
Gap open	-2.9
Gap extend	0
Hydrophobicity multiplier	1.2
Memory/Iterations	
Max memory in MB	931
Max iterations	1000
More Advanced Options	
Clustering Method (Iteration 1,2)	UPGMB
Clustering Method (Other iterations)	UPGMB
Min Diag Length (lambda)	24

b) Parameters for phylogenetic tree

Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	100
Substitution Model	
Substitutions Type	Amino acid
Model/Method	JTT with Freqs. (+F) model
Rates and Patterns	
Rates among Sites	Uniform rates
Data Subset to Use	
Gaps/Missing Data Treatment	Use all sites
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically

Table 6 - Parameters used for phylogenetic analysis in Stage IV

Parameters used for sequence alignment in MUSCLE a) and phylogenetic analysis in MEGA 5 b). The presented conditions were applied in Stage IV of the phylogenetic analysis.

a) **Parameters for sequence alignment**

Gap penalties	
Gap open	-2.9
Gap extend	0
Hydrophobicity multiplier	1.2
Memory/Iterations	
Max memory in MB	931
Max iterations	5000
More Advanced Options	
Clustering Method (Iteration 1,2)	UPGMB
Clustering Method (Other iterations)	UPGMB
Min Diag Length (lambda)	24

b) **Parameters for phylogenetic tree**

Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	1000
Substitution Model	
Substitutions Type	Amino acid
Model/Method	JTT with Freqs. (+F) model
Rates and Patterns	
Rates among Sites	Uniform rates
Data Subset to Use	
Gaps/Missing Data Treatment	Use all sites
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically

2-Analysis of the microbial community composition

The bacterial community structure and composition was characterized by sequencing the V6 hypervariable region and the fungal community was characterized by sequencing the ITS II and D2 region by 454 sequencing, using barcoded (six-base) fusion primers with the Roche-454 A Titanium sequencing adapters. The amplification reactions were performed in a final volume of 25 µL containing 2 µl of DNA, 0.2 µM of each PCR primer (table 7), 0.2 mM dNTPs, 0.5 µl of 50X Advantage2 Polymerase Mix, 1.5 µl of DMSO 5% (v/v) and 2.5 µl of 10X Advantage® 2 SA PCR Buffer (resumed in table 8). The samples were amplified according to the program presented in table 9 in MyCycler Thermal Cycler (Bio-Rad Laboratories, Hercules, California, USA). The amplicons V6, D2 and ITS II were amplified for all samples presented in table 4.

Table 7 - Primers used in PCR reactions.

Primer	Target	Sequence (5'-3')
V6_Forward	V6 hypervariable region of bacterial rDNA	ATGCAACGCGAAGAACCT
V6_Reverse		TAGCGATTCCGACTTCA
D2_Forward	large ribosomal subunit	AAGMACTTTGRAAAGAGAG
D2_Reverse		GGTCCGTGTTTCAAGACG
ITSII_Forward	internal transcribed spacer 2 region	GCATCGATGAAGAACGC
ITSII_Reverse		CCTCCGCTTATTGATATGC

Table 8 - PCR mix used for rDNA amplification.

Reagent	Volume	Supplier
DNA	2 µl	a)
Forward Primer (0.2 µM)	1 µl	Integrated DNA Technologies, Inc (Leuven, Belgium)
Reverse Primer (0.2 µM)	1 µl	Integrated DNA Technologies, Inc (Leuven, Belgium)
dNTPs (0.2 µM)	0.5 µl	Bioron GmbH, Ludwigshafen, Germany
50X Advantage2 Polymerase Mix	0.5 µl	Clontech, Mountain View, CA, USA
DMSO 5% (v/v)	1.5 µl	Roche Diagnostics GmbH, Mannheim, Germany
10X Advantage® 2 SA PCR Buffer	2.5 µl	Clontech, Mountain View, CA, USA
H_2O	16 µl	-
Total reaction volume	**25 µl**	

(-) Not applicable; a) samples presented in table 4.

Table 9 - PCR programs used to amplify the ribosomal regions.

	V6			D2 / ITS II		
Initial denaturation	95ºC	3’		94ºC	3’	
Denaturation	95ºC	30’’	30x	94ºC	30’’	30x
Annealing	52ºC	40’’		50ºC	40’’	
Extension	68ºC	40’’		68ºC	40’’	
Final extension	68ºC	10’		68ºC	10’	

(’) stands for minutes; (’’) stands for seconds; (30x) means that the cycle: denaturation, annealing and extension, was repeated thirty times, following the presented conditions.

After the PCR reactions, the samples were visualized in a 1% (w/v) agarose gel stained with ethidium bromide to check the quality of the amplicon. Electrophoresis was performed in 1x Tris-acetate-EDTA (TAE) buffer at 90V for 30 min and gel images were digitally captured using Gel Doc™ XR+ System (Bio-Rad Laboratories, California, USA).

The amplified DNA sequences were afterwards purified using the most suited method based on the analysis of the electrophoresis gel. Therefore, some were purified with Agencourt® AMPure® XP beads (Agencourt, Beckman Coulter, USA) and others were excised from the agarose gel, being purified with High Pure PCR Product Purification Kit (Roche Diagnostics GmbH, Mannheim, Germany).

Once the purification step was accomplished the nucleic acid quantity was assessed by fluorometry with the Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen, Life Technologies, Carlsbad, California, USA) via the Gemini™ EM Fluorescence Microplate Reader (Molecular Devices, Sunnyvale CA, USA), pooled at equimolar concentrations and sequenced in the A direction with GS 454 FLX Titanium chemistry, according to

manufacturer's instructions (Roche, 454 Life Sciences, Brandford, CT, USA). The sequencing data (fasta files) retrieved from the FLX 454 System was processed using an automatic annotation pipeline implemented in Bioinformatics Unit at Biocant. This pipeline was written in python and combines several widely used software packages for sequence and phylogenetics analysis, thus allowing a rapid analysis of rDNA projects. The raw pyrosequencing reads were treated according to the schematic presented in figure 9.

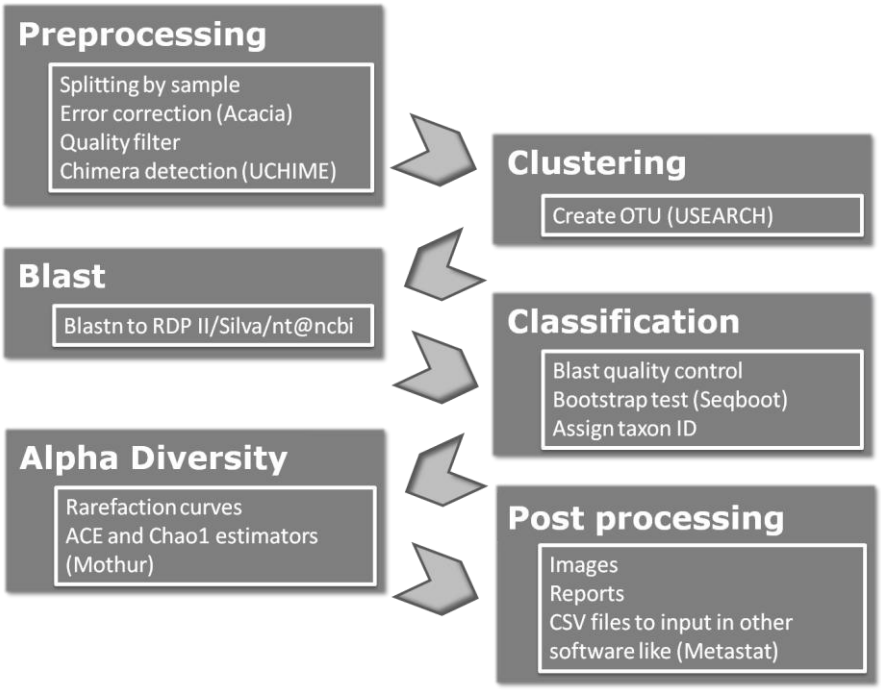


Figure 9 - Flowchart of the bioinformatic methods used to analyze the pyrosequencing results.

In a first step, the sequencing reads were assigned to the appropriated samples and rDNA specific region (ITS II, D2 or V6) based on the respective barcode. In this step, in order to minimize effects of random sequencing errors, we eliminated (i) sequence reads with <120 bp, (ii) sequences that contained more than two undetermined nucleotides (nt), (iii) sequences with more than 50% of low complexity regions made by DustMasker[71], (iv) cut the sequences by reverse primer if sequencing reached the B adaptor, (v) chimera sequences identified by UChime[72]. In a further step of our pipeline the sequences were grouped according to their phylogenetic distance of 3%, which corresponds to species-level threshold, creating the Operational Taxonomic Units (OTU). This operation is made by USearch[73] and automatically creates the consensus obtained by sequence overlapping. Richness of population (rarefaction curves) and the diversity indices (Chao1 and ACE) were calculated using Mothur package[74]. In the next stage, consensus sequences for each

OTU were submitted to BLAST against curated databases which allowed the taxonomic annotation. According to the type of microorganisms a different database was searched for, the Ribosomal Database Project II (RDP) database for prokaryotes and nt@ncbi or SILVA for eukaryotes. After BLAST results, the best hits were selected and subjected to another quality control. All sequences with an alignment of less than 40% were rejected as well as those with an E-value greater than $1e^{-50}$. A bootstrap test was applied for all sequences that passed the previous quality check, using 100 replicates in seqBoot from the Phylip package[75]. The identification process was rather complex because it was necessary to correct the E-value scores, walk through the taxonomy path and identify the least common taxonomy level in the bootstrap process. Only the sequences with a bootstrap greater than 70% are reported. The taxonomic assignment of the OTUs was completed with the attribution of the NCBI taxonomy identification number, which allows building the complete taxonomy of all identified organisms. Finally, for each taxon identified in the sample, the total number of sequences summed up. Therefore, a file with all the abundance of all the identified organisms within a sample was generated, allowing the posterior population statistics analysis. An image output was also generated using the software Phylostat 1.0 beta (unpublished).

3-Gene annotation

3.1-MAKER

The candidate HGT gene was annotated at the Bioinformatics unit at Biocant using the genome annotation pipeline, MAKER. MAKER is a portable and easily configurable genome annotation pipeline which has the ability to identify repeats, align EST's and proteins with a genome, make gene predictions and automatically integrate these data into protein-coding gene annotations, possessing evidence–based quality indices. The annotation was performed according to the procedure presented by *Cantarel et al.*[76] and resumed in figure 10.

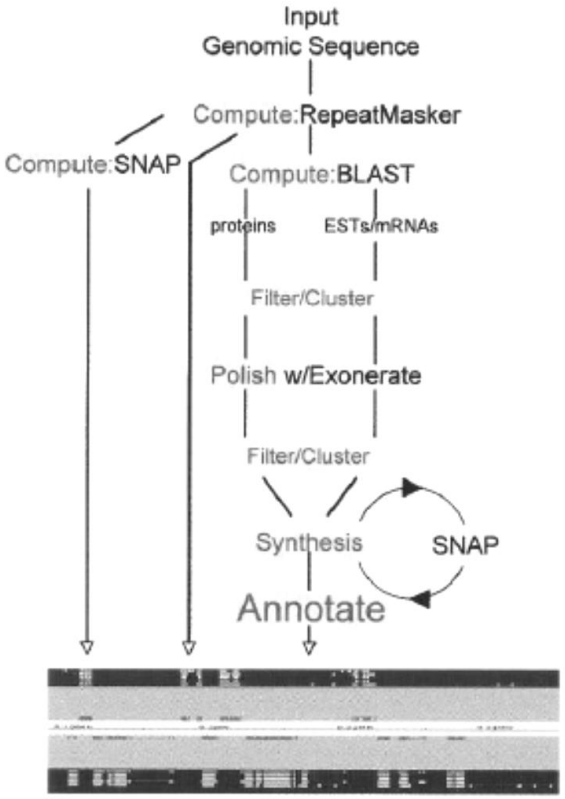


Figure 10 – Flowchart of the MAKER pipeline, adapted from *Cantarel et al.*[72]

The MAKER pipeline can be divided into five steps presented in figure 10. In step 1, the compute phase, a battery of sequence analysis programs was run with the purpose of identifying and masking repeats as well as assembling protein EST and mRNA alignments that were used in the MAKER's gene annotation process. Step 2 comprises the filtering/clustering stage. This phase consisted in identifying and removing marginal predictions as well as sequence alignments based on their scores, percentage of identity, etc. Step 3 is called the polish phase, BLAST hits were realigned using another alignment algorithm in order to obtain superior precision at exon boundaries. Step 4, termed synthesis, collected the clustering information from step 2 (ESTs) and the polished data from step 3 (aligned proteins) and generated evidence used for the annotation held in step 5. In this step MAKER post-processed the synthesis-generated SNAP predictions, recombining them with evidence, generating complete annotations.[76]

3.2-Blast2GO®

Blast2GO® is a universal GO annotation, visualization and statistics framework, designed to allow automatic and high throughput sequence annotation along with the

integration of functionality for annotation-based data mining[77]. The proceedings implied in the method are resumed in figure 11 according to *Conesa et al.*[77].

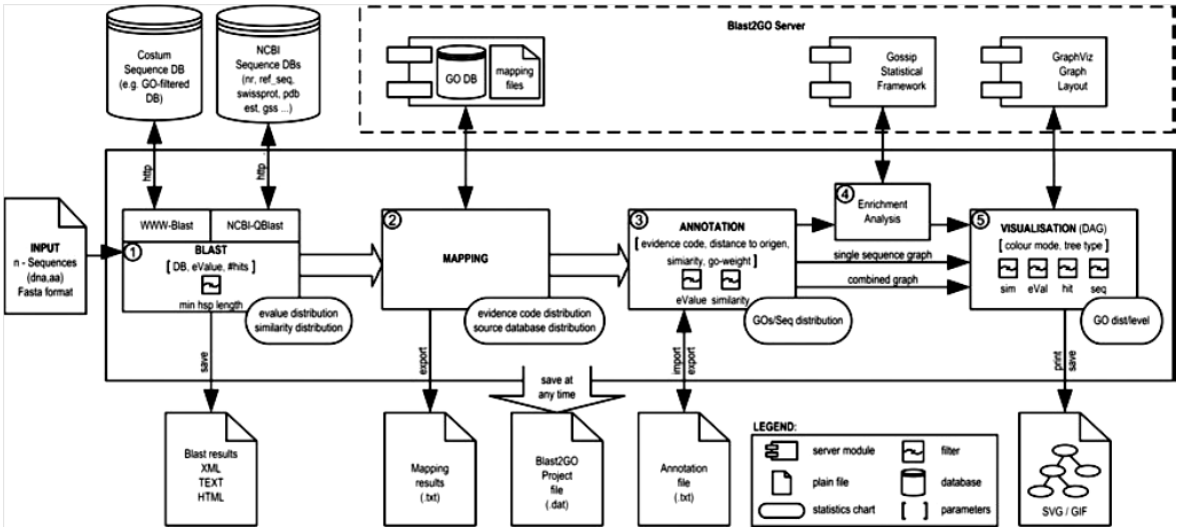


Figure 11 – Blast2GO® application overview, adapted from *Conesa, A. et al.*[77].
The figure shows schematically a typical run of Blast2GO®. Used symbols are described in the embedded legend. Numbered circles denote the major application steps. From the left to the right these are (1) Blasting: a group of selected sequences is blasted against either the NCBI or custom databases, (2) Mapping: GO terms are mapped on the BLAST results using annotation files provided by the GO Consortium that are downloaded on a monthly basis at the Blast2GO® server, (3) Annotation: sequences are annotated using an annotation rule that takes parameters provided by the user, (4) Statistical analysis: optionally, analysis of GO term distribution differences between groups of sequences can be performed and (5) Visualization: annotation and statistics results can be visualized on the GO DAG. At each of these steps, different charts are available to evaluate the progress of the analysis and data can be saved and exported in different formats.

Blast2GO® is a user friendly tool that allows monitoring and interaction at different steps of the analysis generating a visual output. Blast2GO® uses BLAST to locate homologs of the introduced fasta files. The program extracts GO terms for each one of the obtained hits by mapping to existent annotation associations. An annotation rule assigns GO terms to the query sequence and the resulting annotation can be visualized in a graph.[77]

Chapter III – Results & Discussion

Horizontal gene transfer (HGT) is a mechanism for organisms to acquire new genes and consequently new functions. In the case of *B. xylophilus,* at least eight genes have been described as resulting from HGT. Our aim is to identify new uncharacterized HGT candidates among the transcripts of *B. xylophilus* obtained at the Advanced Services Unit at Biocant. This laboratory sequenced the transcriptome of seven nematode isolates from four geographic locations and assembled all reads together to create an *in silico* transcriptome of *B. xylophilus*. After gene annotation the laboratory developed a database to organize transcript information and search for gene functions. The following sections explain the methodology employed in screening and testing of the new HGT candidates.

1-Establishment of filters to select horizontal gene transfer genes

Pipeline development was an attempt-error process, since it was necessary to tryout several layouts to achieve the most appropriate outline retrieving the most accurate results. The variables, taxonomic origin of the organism and existence of a specific function as well as E-value, were conjugated in different manners in order to minimize the errors and the duplication of steps. At the end of the pipeline we expected to find candidates with a bacterial or fungal origin that possessed a significant E-value and a specific protein function, to deduce the role of the gene in the nematode and infer on the participation in the pathogenic mechanisms of *B. xylophilus*.

The incorporation of genes acquired by HGT from bacteria and fungi in the genome of *B. xylophilus* is one of the pathways that are believed to have a role in the development and progression of PWD[14]. By assimilating genes that naturally wouldn't be present, the PWN acquires abilities that could help him survive in different conditions and develop new feeding habits. It has been shown that the genome of *B. xylophilus* integrates at least eight genes from bacterial[9,49,62] and fungal[9,15] origin, therefore, in the quest for new genes acquired by HGT, the organism origin, is a logical and useful filter to apply. In the transcript annotation, we found proteins with bacterial and fungal origin, proteins belonging to nematodes and a group of proteins with non-bacterial, non-fungal, nor nematode origin (nBFNO). Transcript annotation is accompanied by E-value, another of the chosen filtering parameters, which indicates the statistical significance of the BLAST annotation[78]. Through E-value, one can determine the number of matches of a given pairwise alignment expected to be found by chance in a particular database. Thus, the

lower the E-value, the higher the chance of a given match being unique and more similar to the sequence of analysis, rising the significance of the hit[79]. E-values that tend to zero are more suitable to detect reliable results, therefore in our analysis we established E-value $\leq 10^{-6}$ as the cutoff point[67]. A different parameter established was the existence of a particular function since it was commonly found among the transcripts the existence of hypothetical or predicted proteins. This designation is attributed whenever there is no experimental evidence regarding the translation of such gene and the protein hasn't been experimentally characterized[80-81]. Some of these hypothetical proteins are thought to originate from the transcription of pseudogenes, DNA sequences that look like functional genes that however have no proven purpose[80,82]. Even though hypothetical proteins may eventually be proteins with new uncharacterized roles, an *in silico* analysis on its own does not achieve conclusions, without the *in vitro* or *in vivo* investigation[80]. Another denomination found amongst our transcripts was conserved hypothetical proteins. This designation comprehends a large portion of genes that could be found in sequenced genomes from a transversal range of phylogenetic lines and that haven't been functionally or chemically characterized[81]. In consequence, maintaining in the final results candidates with such classifying hits would difficult the task of proving the reason of its incorporation in the genome of the PWN and its role in the nematodes survival. By relying our study on proteins of known function, understanding the outcome of the incorporation of such gene in the PWN is likely to be more successful.

Our essay started off with 6 751 transcripts that were classified according to the taxonomic origin of the hits which classified the candidate. This step besides providing a bottom line for the quest for HGT candidates also supplied overall information regarding the taxonomic representativity and statistical significance of our transcripts, summarized in table 10.

Table 10 - Summary of taxonomic affiliation of our transcripts

Hits	[1]Total	[2]% Total	[3]E-value≤10^{-6}	[4]% E-value
Nematodes	3 915	58%	3 481	89%
nBFNO	1 977	29.3%	1 229	63%
Bacteria and Fungi	496	7.3%	198	54.5%
Non-classified	363	5.4%	-	-

The presented results are based on the classification of the first BLAST hit; Total number of hits analyzed 6 751; [1]contains the total number of hits belonging to the defined taxonomic group; [2]shows the percentage of the taxonomic representation in the total of transcripts; [3]contains the hits with an E-value$\leq 10^{-6}$, considered acceptable; [4]represents the percentage of hits with an acceptable E-value, among the selected group. nBFNO - Non bacteria, non fungi nor nematode organisms; non-classified – transcripts that do not possess a classifying hit; (-) not applicable.

We performed this analysis based exclusively on the first BLAST hit result, since it is the most representative of all, holds the highest BLAST score and the highest E-value. The majority of the transcripts (58%) were classified as possessing a nematode origin, and the transcripts that did not have a classifying hit, represented the minority (5.4%). The percentage of nematode hits containing an acceptable E-value was high (89%). *B. xylophilus* belongs to the nematoda phylum, therefore it was naturally expected that the bulk of the transcripts also fit in the same phylum due to the phylogenetic evolution. However, and according to prior published reports[9,15,46-47,49,63] it is also possible that some of the genes belonging to the genome of *B. xylophilus* may have been acquired from other phyla, by HGT, thus the existence of 7.3% of the transcripts belonging to bacteria and fungi is acceptable and are the searching basis for candidates. The number of hits designated as belonging to bacteria and fungi with an acceptable E-value represent only 54.5% of the total of hits in this group but still represent the greater part of the results.

Looking further into the results of the interest group, bacteria and fungi, it was possible to determine the representativity of each domain and the number of proteins with a specified function, these results are presented in table 11.

Table 11 - Summary of bacteria and fungi representativity and protein specificity of the transcripts.

Hits	[1]Total		[2]E-value≤10^{-6}		[3]Hypothetical Protein		[4]Acceptable candidates	
Bacteria	204	3.1%	82	40%	87	43%	49	24%
Fungi	292	4.3%	114	39%	174	60%	47	16%

The presented results are based on the classification of the first BLAST hit; [1]total number of hits belonging to the defined taxonomic group / percentage of the taxonomic representation in the total of transcripts; [2]hits with an E-value≤10^{-6}, considered acceptable / percentage of hits with an acceptable E-value, among the total transcripts of the selected group; [3]hits identified as coding for hypothetical proteins / percentage of hits coding for hypothetical proteins among the total transcripts of the taxonomic group; [4]transcripts containing an acceptable E-value (E-value≤10^{-6}) and simultaneously a specified protein function / percentage of candidates that hold an acceptable E-value and detain a specified protein function.

The percentage of transcripts identified as being closer to bacteria is 3.1% of the total of candidates, among these only 40% hold an acceptable E-value, yet only [1]57% of the total of transcripts with bacterial origin possessed a protein with a specified function, this gives us a total of 24% bacterial transcripts relevant for the study. The percentage 4.3% of fungal transcripts is slightly higher than those with bacterial origin however the outcome of

[1] 100% *(total)* – 43% *(Hypothetical Proteins)* = 57% *(Proteins with a specified function)*

viable candidates from this domain is only 16%, less than the amount available from bacteria.

2-Development of the pipeline

The pipeline was developed based on attempt-error approach. The established screening parameters were applied to the total 6 751 transcripts in different arrangements until the best approach was met. Seven different outlines (presented in appendix I) were tested to set up the final pipeline, which results from the analysis of the weaknesses and strengths of the several attempts, resumed in figure 12.

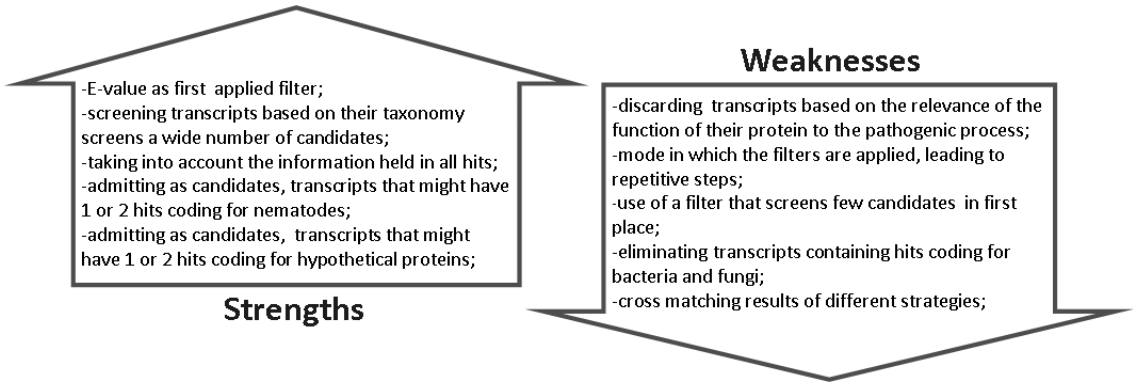


Figure 12 - Summary of weaknesses and strengths of the tested pipeline strategies.

The several developed attempts to achieve the final pipeline resulted as an evolving process in an effort to remove an identified flaw in the prior strategy. These flaws are presented as weaknesses in figure 12 and the positive aspects of the strategies are presented as strengths. Based on this schematization of filters and filter arrangement we realized that a pipeline to screen transcripts in a quest to find HGT candidates, should comprise a first step with the ability of screening a significant amount of candidates; must eliminate all transcripts containing hits coding for nBFNO organisms; shouldn't consider as valid transcripts that possess three or more hits coding for nematodes and should disregard transcripts with hypothetical proteins. Based on these criteria we designed the following flowchart, presented in figure 13.

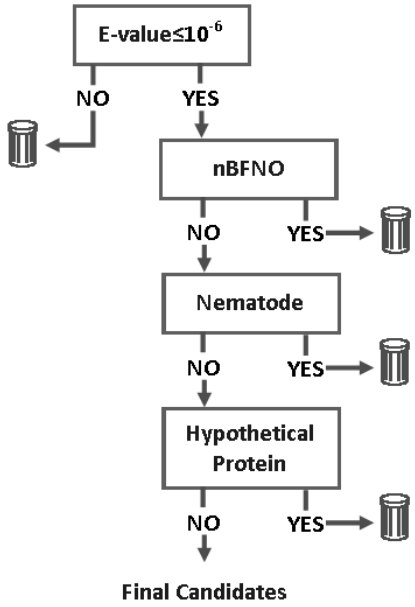


Figure 13 – Flowchart of an ideal pipeline.
This pipeline was determined based on the observation of prior attempts and hints presented in figure 12.

We adopted E-value as the first filter, not only for its ability of screening a large amount of transcripts but also because if we screened out all the transcripts containing an inappropriate E-value we were keeping the ones with a relevant hit classification. This means that, when screening nBFNO organisms, transcripts containing at least one hit coding for these organisms was sufficient to eliminate that transcript. This justifies our second step in the flowchart. The order of the following two steps was determined based on the results of two strategies. The final strategy is presented in figure 14.

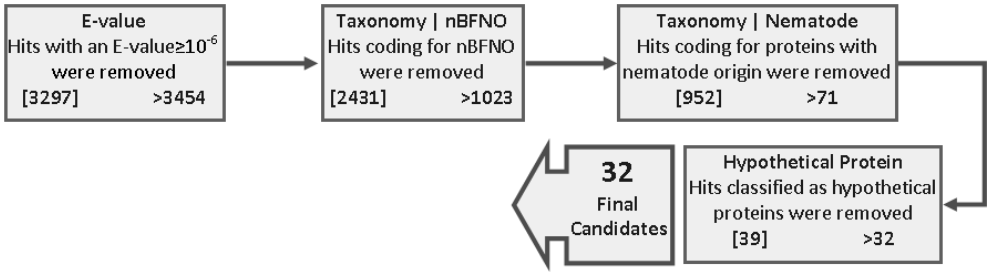


Figure 14 - Outline of the final pipeline (Strategy VII).
Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

Through the observation of the diagram we can conclude that with our approach we were able to obtain 32 final candidates, with the desired requirements, from the initial set of 6 751 transcripts.

3-*B. xylophilus* microbial community analysis

B. xylophilus has a microbial community in its surrounding cuticle and maybe in its digestive tract, and it is possible that some of the genetic material of these microorganisms might have been sequenced together with the nematode and could be present in the transcripts. This circumstance would adulterate our results, because we might have been considering as HGT candidates, genes that in fact belong to bacteria living associated with the PWN; from the fungus where the nematode was cultured or from other fungi contaminating the culture. The nematode samples used to sequence the transcriptome were disinfected, but this step may not ensure the death of the microbial community nor guarantee the complete removal of the microbial community from the surface. Analysis of the composition of the microbial community surrounding the samples could help distinguish the transcripts belonging to *B. xylophilus* from those belonging to bacteria and fungi without *in vitro* testing. Using next generation DNA sequencing technologies, the *in silico* identification of the microbes present in the community is now possible[54,71,83]. At the Advanced Services Unit at Biocant, this assay was conducted using the Roche 454 pyrosequencing technology. Via sequencing the PCR products of phylogenetically conserved sequences we can define the composition of the microbial community. Genes from ribosomal RNA (rRNA) are conserved in all organisms and can be particularly informative for identifying the genus and in some cases the species of microorganisms[83]. In this biodiversity study we amplified three different regions of the rRNA. In order to identify bacterial genera we used the V6 hypervariable region of the 16S rRNA with approximately 50-70 bp, generating an amplicon of approximately 450 bp[71,83-85] (figure 15).

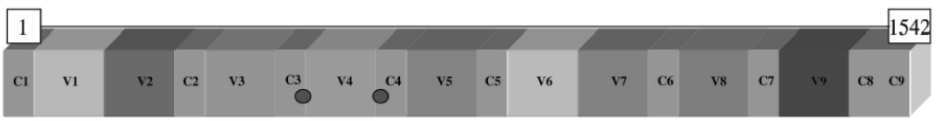


Figure 15 - Conserved and hypervariable regions in the 16S bacterial rRNA, adapted from *Petrosino et al.*[54].
Conserved regions (C1-C9) and the hypervariable regions (V1-V9)

For the identification of fungal individuals, two sequencing regions from the rRNA have been used as illustrated in figure 16. The D2 region from the large ribosomal subunit (LSU) generates an amplicon with about 350 bp and is more suited for the identification of yeast while the internal transcribed spacer 2 region (ITS II), which generates an amplicon

with approximately 450 bp long is more indicated for the recognition of filamentous fungi[86-88].

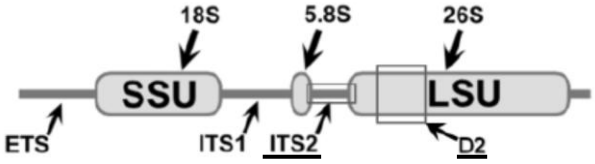


Figure 16 - Representation of the fungal rRNA operon, adapted from *Accugenix, Inc*.[89]
Underlined regions were used in the metagenomic analysis.

3.1-rDNA amplification and sequencing

PCR amplification procedures were conducted using aliquots of 50 ng/µL of genomic DNA that was extracted at the Advanced Services Unit at Biocant by a colleague in a prior research and were preserved at -4ºC. The genomic DNA was isolated from the seven samples of *B. xylophilus*, the same used to sequence the transcriptome, which originated from four distinct geographical locations according to table 12.

Table 12 - Summary of the samples used, regarding their origin.

Origin	Region	Sample reference
Portugal	Santiago do Cacém	**Pt15**
	Alcácer do Sal	**Pt17**
	Santa Comba Dão	**Pt19**
	Tábua	**Pt21**
China	-	**ChJs**
Japan	-	**J10**
USA	-	**USA618**

(-) Not applicable

PCR products were analyzed by electrophoresis in agarose gel and pooled according to figure 17 to be purified. Our goal was to obtain high quality PCR products for pyrosequencing in order to classify the microbial community.

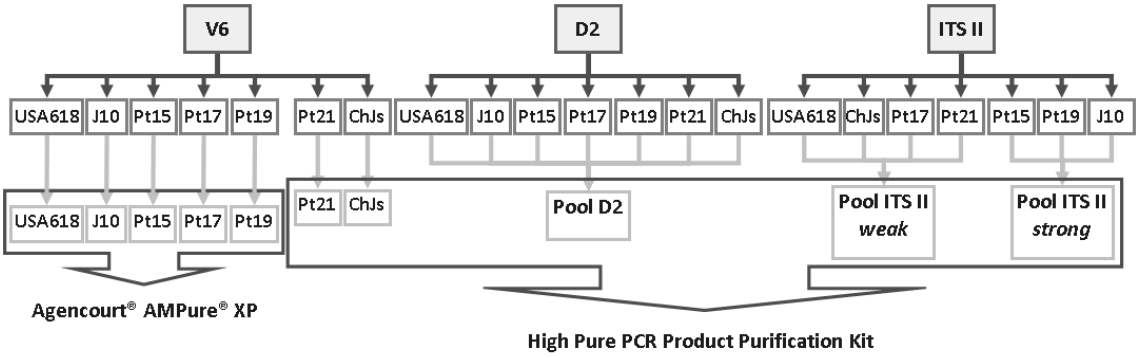


Figure 17 - PCR products pooling scheme.

The purification method was chosen in order to be the most suitable for each situation, therefore, some amplicons were purified using Agencourt® AMPure® XP beads and others by band excision from agarose gel using High Pure PCR Product Purification Kit. PCR products that generated well defined and intense bands where purified individually by Agencourt® AMPure® XP. However most of the samples needed to be gathered in pools, according to their distinctiveness and were cut out from the agarose gel.

All purified PCR products were visualized by agarose gel electrophoresis to verify the size, quality of the amplicon and presence of unincorporated primers (figure 18).

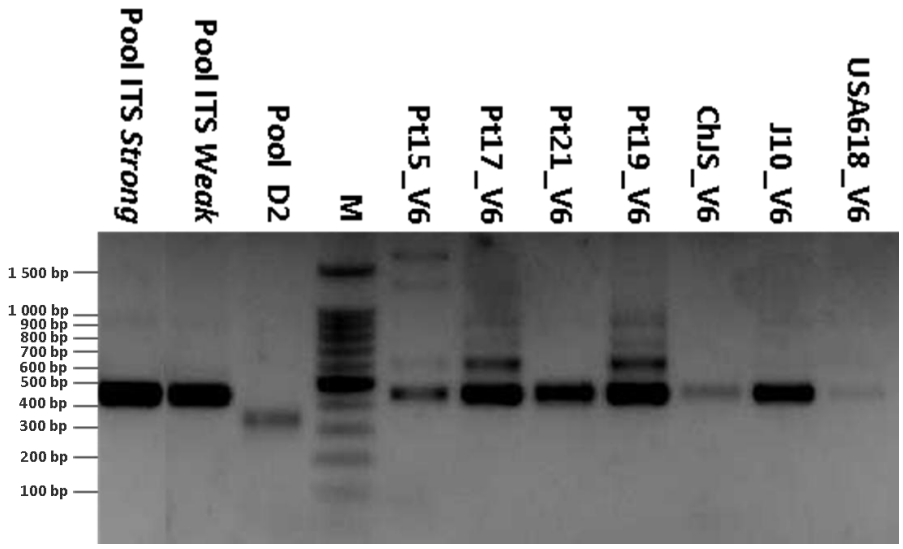


Figure 18 - Agarose gel electrophoresis of the PCR products after the purification step to assess quality. 1% agarose gel stained with ethidium bromide 1% (w/v); electrophoresis was performed in 1x Tris-acetate-EDTA (TAE) buffer at 90V for 30 min and gel images were digitally captured using Gel Doc™ XR+ System (Bio-Rad Laboratories, California, USA).

3.2-Microbial community composition

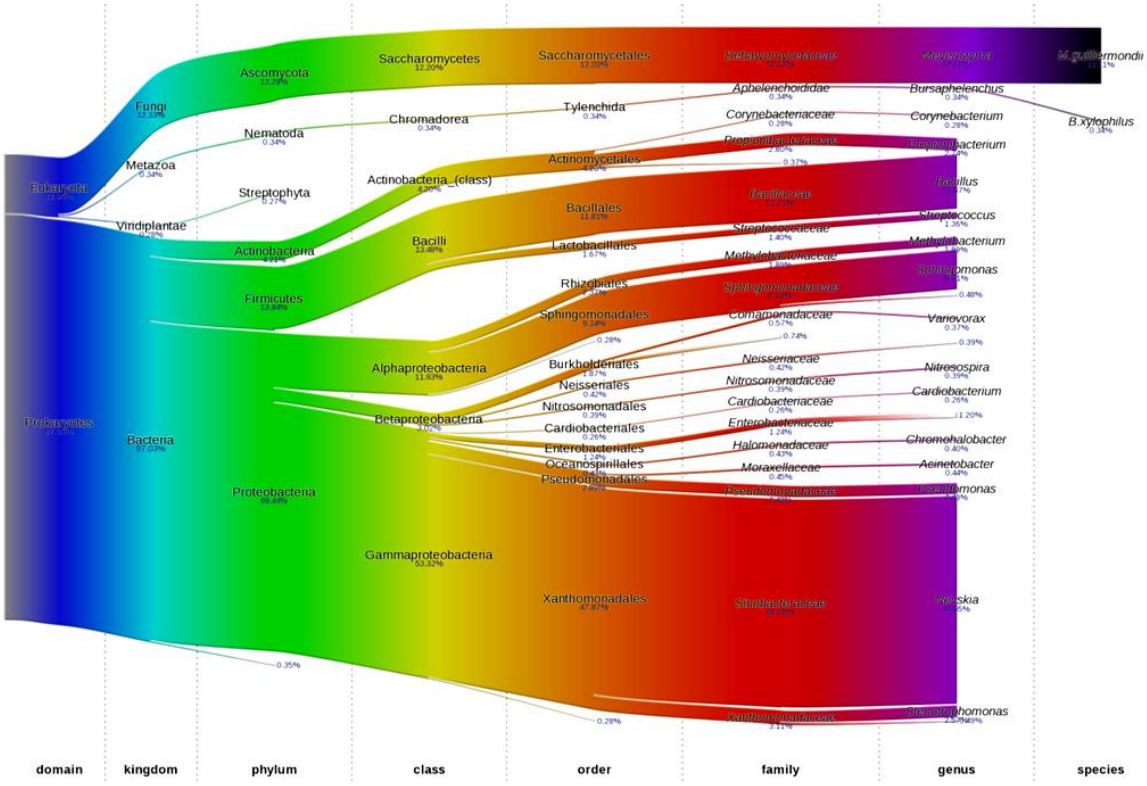


Figure 19 – Visual representation of the biodiversity results.

The result of the analysis of the pooled samples is summarized in figure 19. The figure corresponds only to results with a representativity of transcripts in the biodiversity results >0.26%. Microbial communities present in our PWN samples were divided in 2 domains, 4 kingdoms, 11 phyla, 16 classes, 33 orders, 70 families and 96 genera. These rDNA regions were only able to discriminate a total of 6 species, only 6.3% of the total number of recognized genus. In the biodiversity results we found two well represented kingdoms, bacteria, which embraces the majority of the identified genus and fungi that is only represented by *M. guillermondi*. One of the other identified kingdoms was the metazoa that holds our subject of study *B. xylophilus*. This subject was identified in the biodiversity assay because the primers used to amplify the ITS II region of rDNA in fungi are coincident for the same region in the PWN (appendix II). Viridiplantae is the other identified kingdom, but in this case, no genus was recognized as belonging to it. In this biodiversity report there are two intriguing results due to their high predominance in the sum of the results, the *Meyerozyma* and *Nevskia* genus, because none of both have been reported as belonging to the microbial community surrounding the PWN. However, understanding why there is such a great occurrence of microorganisms belonging to these

two genera is not a goal of this thesis. According to figure 19 we uncover that the bacteria kingdom holds the biggest representation of microbial individuals in our microbial community. Published studies that describe the presence of microbial communities living in close relationship with the PWN report the occurrence of bacterial strains, some of which can also be found in our samples, thus validating our results[19-20,40-41]. However, our biodiversity study, besides identifying bacterial strains that have already been reported as living with *B. xylophilus*, also finds different genus (table 13).

Table 13 – Bacterial community surrounding *B. xylophilus*.

Xie and Zhao 2008[20]	Roriz, Santos et al. 2011[19]	Vicente, Nascimento et al. 2011[40]		Proença, Francisco et al. 2010[41]	This study
[1]China	[1]Portugal	[1]Portugal		[1]Portugal	[1]Portugal, China, Japan, USA
		wild	*lab*		
Pantoea	*Pantoea*	*Pantoea*	-	*Pantoea*	-
Pseudomonas	-	*Pseudomonas*	*Pseudomonas*	*Pseudomonas*	*Pseudomonas*
Sphingomonas	-	-	-	-	*Sphingomonas*
-	*Citrobacter*	-	-	-	-
-	*Terribacillus*	-	-	-	-
-	*Klebsiella*	-	*Klebsiella*	*Klebsiella*	-
-	*Enterobacter*	*Enterobacter*	*Enterobacter*	-	*Enterobacter*
-	*Bacillus*	-	-	-	*Bacillus*
-	*Paenibacillus*	-	-	-	-
-	*Escherichia*	-	-	-	-
-	-	*Serratia*	*Serratia*	*Serratia*	*Serratia*
-	-	*Erwinia*	-	*Erwinia*	-
-	-	*Rahnella*	*Rahnella*	*Rahnella*	-
-	-	*Ewingella*	*Ewingella*	*Ewingella*	-
-	-	*Burkholderia*	-	*Burkholderia*	*Burkholderia*
-	-	*Staphylococcus*	*Staphylococcus*	-	*Staphylococcus*
-	-	-	*Acinetobacter*	-	*Acinetobacter*
-	-	-	*Comamonas*	-	-
-	-	-	*Herbaspirillum*	-	-
-	-	-	*Enterococcus*	-	-
-	-	-	-	*Curtobacterium*	*Curtobacterium*
-	-	-	-	*Yersinia*	-
-	-	-	-	*Hafnia*	-
-	-	-	-	*Cronobacter*	-
-	-	-	-	*Janthinobacterium*	-
-	-	-	-	*Luteibacter*	-
-	-	-	-	-	*Propionibacterium*
-	-	-	-	-	*Streptococcus*
-	-	-	-	-	*Methylobacterium*
-	-	-	-	-	*Nevskia*
-	-	-	-	-	*Stenotrophomonas*

Chart containing the bacterial genus identified in bibliography as living with *B. xylophilus* as well as most represented bacterial genus in our samples; [1]Origin of the PWN samples; Shadded lines respresent the bacterial genus that are common to the majority of the presented studies; (-) Not applicable.

The conditions in which the studies indicated in table 13 were conducted are summarized in table 17 of appendix II. This comparative approach showed us that there are bacterial strains like *Pseudomonas, Enterobacter, Serratia, Burkholderia* and *Staphylococcus* that were identified in most of the studies presented in table 13, including ours, but on the other hand, there are strains like *Pantoea, Klebsiella, Rahnella* and *Ewingella* that are reported in most of the other studies, but seem to be absent in our samples. Mainly reported as associated with the PWN are bacteria belonging to the genus *Pantoea*, *Xanthomonas* and *Pseudomonas*, still only this last genus was discovered in our biodiversity assay[19,36,90]. Some studies have stated that the bacterial representation varies geographically and that some genus can be considered as characteristic for samples collected in specific regions. In Japanese samples *Bacillus* is the predominant genus. On the other hand bacteria belonging to *Pseudomonas* is the main representation in Chinese samples and in Korean samples both genus are present[36,91-92]. Nonetheless, no bacterial genus has yet been identified as characteristic of the Portuguese PWN community and in accordance with table 13 none of the presented genus can be pointed out as distinctive of the Portuguese population since none of them appeared in all study cases. The bacterial strains that have geographically been attributed to Japan and China isolates were found among our results as expected, since we possessed samples collected from these two locations. According to table 13 *Pseudomonas* is not only found in Chinese samples but also in some Portuguese samples.

The goal of this biodiversity study was to verify if the transcriptome library built from our PWN samples possessed transcripts belonging to the microbial community. These results could then be used as a screening step in the pipeline to ensure that the transcripts suggested as candidates originate from the genome of *B. xylophilus* and not from a microorganism belonging to the surrounding microbial community. The microbial genera discovered in the microbiome were cross matched with the transcript library and these results are presented in table 14.

Table 14 - Genus of microorganisms identified in the microbial community and their presence in the transcriptome

		Genus from the microbiome identified in the transcriptome	[1]%	Number of transcripts from the transcriptome identified as belonging to the microbiome	[2]%
Total genus identified in the microbial community	**96**	35	36%	279	4.1%
Genus with higher representativity	**13**	6	6.3%	133	2%

[1] Represents the percentage of genus from the microbial community identified in the transcriptome in the total number of genus from the microbial community (95); [2] Represents the fraction of transcripts belonging to the transcriptome identified as being part of the microbial community in the total number of transcripts from the transcriptome (6 751).

Cross matching the results of the microbiome with the information from the transcriptome pointed out that 36% of the microbial genera identified in the biodiversity study appeared in the transcriptome and that these implicate 279 transcripts, 4.1% of the total. The genus containing a higher representativity in the biodiversity assay held 2% of the sum of transcripts in the transcriptome. The distribution of transcripts belonging to the transcript library among the several microbial genera is presented in figure 20.

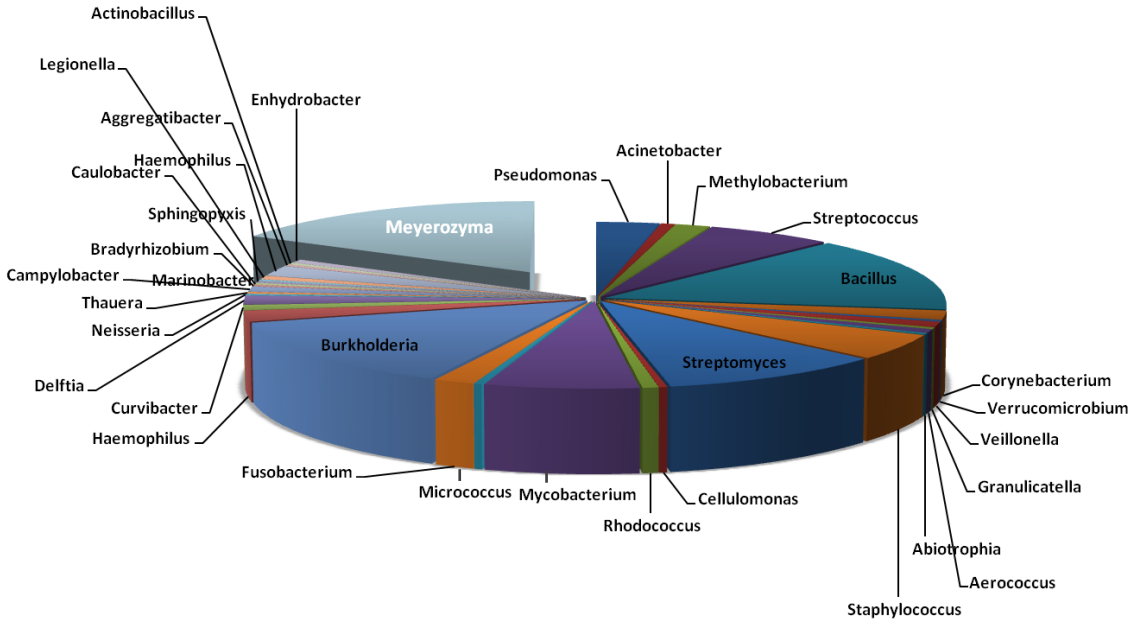


Figure 20 - Representation of the genus belonging to the microbiome that were identified in the transcriptome and their abundance.

The separated slice (*Meyerozyma*) belongs to the fungi kingdom, while all the other slices belong to bacteria.

Figure 20 shows that the fungi *Meyerozyma guillermondi* besides representing a meaningful portion of the microbial community also occupies a significant fraction of the transcripts identified in the transcriptome as belonging to the microbiome. Among the genera belonging to the bacteria, *Burkholderia*, *Staphylococcus*, *Bacillus*, *Streptococcus*, *Pseudomonas*, *Methylobacterium Mycobacterium*, *Streptomyces*, and *Fusobacterium* have

more transcripts in the transcript library. It is curious to notice that not all bacterial genera with higher representativity in the microbiome (figure 19) were the ones with most identified transcripts in the transcriptome (figure 20), Mycobacterium, Streptomyces and Fusobacterium fell into this category.

This analysis led to the knowledge that the transcriptome assembled at the Advanced Services Unit at Biocant contains transcripts that are not originated from *B. xylophilus.* Consequently, we checked if the final candidates of the pipeline withhold any of these transcripts belonging to the microbiome. This screening process is represented in figure 21.

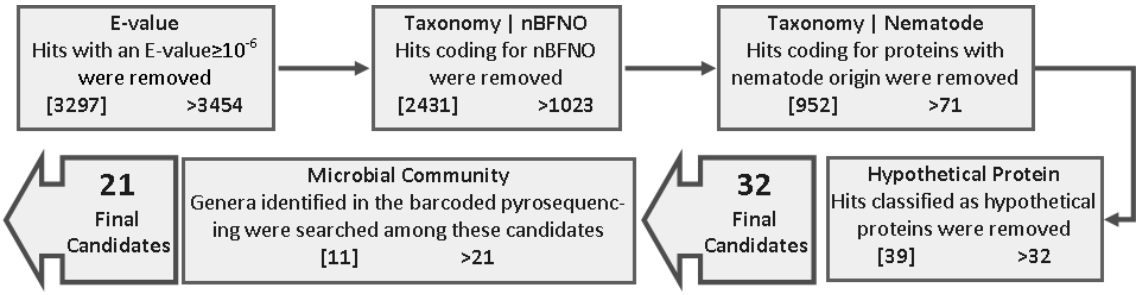


Figure 21 – Final pipeline with the microbiome results applied as a filter.
Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

From the observation of figure 21 we concluded that performing this microbiome step was a good choice because by solely using the other applied filters we were unable to screen transcripts that were in the samples but were not incorporated in the genome of the PWN. We saw that among the final 32 candidates of the pipeline were still 11 candidates that belonged to the microbial community living with the PWN, which left us with 21 candidates, presented in table 15.

Table 15 - Transcripts candidate to HGT event and corresponding protein function.

[1]Candidate reference	[2]Expected protein function
All_gs454_005644	beta-1,3-endoglucanase
All_gs454_005676	beta-1,3-endoglucanase
All_gs454_005732	beta-1,3-endoglucanase
All_gs454_005733	beta-1,3-endoglucanase
All_gs454_003220	alcohol dehydrogenase 1
All_gs454_005442	alcohol dehydrogenase 1
All_gs454_002803	alcohol dehydrogenase
All_gs454_007597	Tor1p
All_gs454_001719	M13 family peptidase
All_gs454_003991	pyrroline-5-carboxylate reductase
All_gs454_005127	beta-galactosidase (Lactase)
All_gs454_005696	*indeterminate*
All_gs454_007949	DNA translocase FtsK
All_gs454_008421	RNA polymerase II largest subunit B220
All_gs454_008990	Mucin-associated surface protein
All_gs454_009220	Ribosomal protein L19
All_gs454_011000	40S ribosomal protein S9-B
All_gs454_011202	Endonuclease/exonuclease/phosphatase family protein
All_gs454_007978	cyanate hydratase family protein
All_gs454_008739	enoyl-CoA hydratase
All_gs454_010630	short-chain dehydrogenase/reductase SDR

[1] Name of the transcript during the study; [2] Protein function attributed to the transcript based on the classification of the first hit; Shaded lines correspond to proteins already reported as HGT.

Amongst the final candidates of the adopted pipeline were two proteins that had already been reported as incorporated by HGT in the genome of nematodes. These proteins were identified as HGT proteins in two different nematodes, beta-1,3-endoglucanase, was identified in *B. xylophilus* and alcohol dehydrogenase was identified in *C. elegans*[47,93]. Their phylogenetic origin were also different, the gene responsible for the synthesis of beta-1,3-endoglucanase was originated from bacteria, while the gene responsible for coding alcohol dehydrogenase was descendant from fungi[47,93]. The presence of these proteins in the final candidate group indicated that the approach we used to screen the transcripts was appropriate. The two identified HGT candidates took up seven of the transcripts. The remaining fourteen transcripts followed into further analysis to determine their origin. The two identified and published HGT genes were used as a control group in the phylogenetic analysis, to determine the suitability of the adopted methodology.

4-Phylogenetic analysis

Phylogenetic analysis was the last step in the process of determining which transcripts were HGT candidates. While in prior stages of candidate screening, transcripts were analyzed in group according to the established parameters, in this phase they were explored individually. In order to determine which candidates had been incorporated by HGT in the genome of the PWN, a work flow was adopted to guide the phylogenetic analysis. The result of this approach is shown in figure 22.

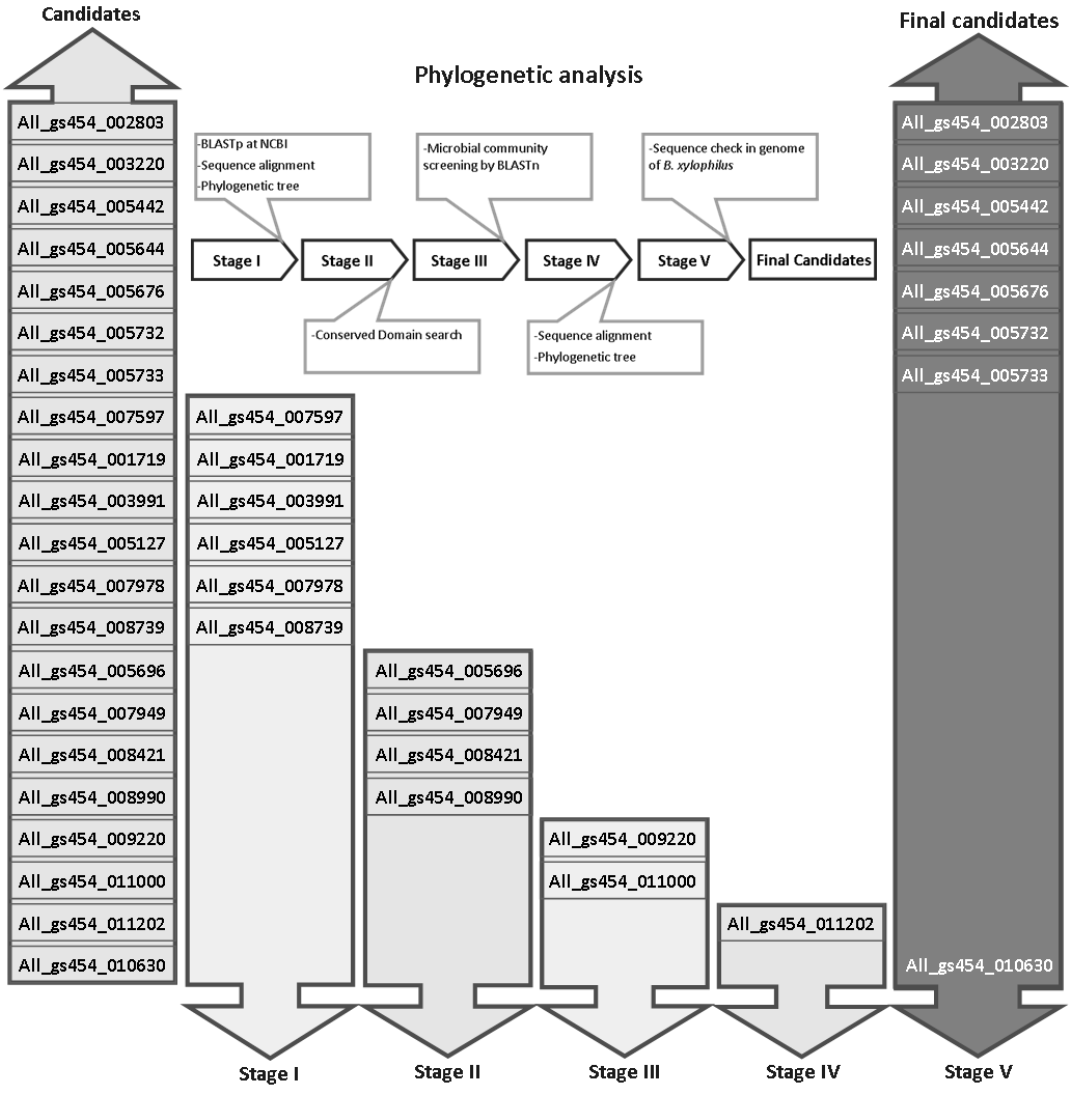


Figure 22 - Representation of the results of the phylogenetic analysis in each phase.
Arrows that illustrate stages 1 to 4 contain the transcripts that were eliminated in that step. Stage 5 is the confirmation step, therefore contains the final candidates of our study.

Figure 22 indicates that the work flow established for the phylogenetic analysis continued screening candidates and that by the end of it, only one of the unreported candidates was left. A deeper look into each stage will help understand the final results.

In stage I the transcript sequence was subjected to BLASTp against bacteria, in first place, fungi, in second, nematodes, in third and in fourth place other organisms excluding the last three tested classes. For each of the four BLAST searches we downloaded three to four sequences containing the highest scores and E-value. The sequences were afterwards aligned using MUSCLE algorithm in MEGA 5, manually adjusted and submitted to phylogenetic tree construction using Maximum Likelihood, also in MEGA 5. Figure 23 exemplifies the expected outcome generated by the phylogenetic trees.

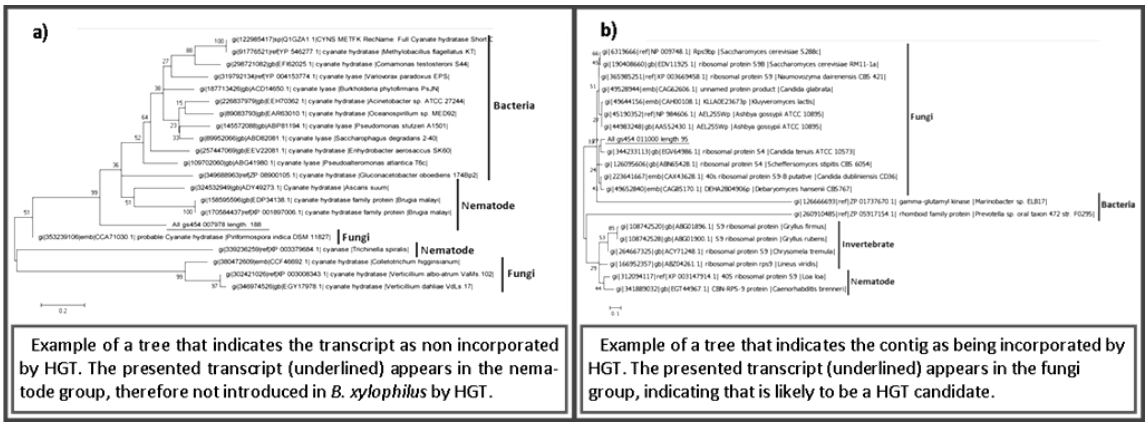


Figure 23 - Examples of phylogenetic trees in the analysis of HGT candidates.
a) tree exemplifying a non HGT candidate. The tested candidate has a bigger proximity to the nematode taxon; b) tree exemplifying a HGT candidate. The tested candidate has a bigger proximity to the fungi taxon.

Our goal was to detect genes incorporated in the genome of *B. xylophilus* that have been acquired by HGT, originating from bacteria or fungi. A transcript sequence that appears in the phylogenetic tree closer to the nematoda phylum undoubtedly clusters in this phylum and was not incorporated exogenously. Stage I eliminated six candidates. In stage II we checked, in the remaining candidates, the presence of a conserved domain. The existence of a conserved domain was essential in order to determine the role of that gene in the biology of the PWN. In this stage four candidates were deleted for not possessing a conserved domain. In stage III, the remaining candidates were screened by cross reference, through BLASTn search, with the genera identified in the barcoded pyrosequencing. Results with an identity ≥97% indicated that the transcript originated in the same organism. This stage eliminated two transcripts, leaving us with two more, besides the other identified and reported HGT candidates. In stage IV we performed a new sequence

alignment, using 5 000 iterations, instead of the initial 1 000 iterations used in stage I and a new phylogenetic tree with 1 000 bootstraps as an alternative to the 100 bootstraps used in stage I. These improvements applied to the sequence alignment and phylogenetic tree showed that the transcript All_gs454_011202 was taxonomically more similar to the nematode taxon, thus leaving us with one last candidate All_gs454_010630. Stage V confirmed the presence of the sequence of the last transcript by BLAST search in the genome of *B. xylophilus*, recently turned accessible.

Even though not referred, the sequences of the genes already identified and reported as HGT were subjected to all stages of the phylogenetic analysis. The fact that these sequences passed in all proposed filters revealed that the adopted methodology was appropriate to discover genes incorporated by HGT. Figure 24 illustrates the combination of the phylogenetic analysis to the pipeline and the three final HGT candidates.

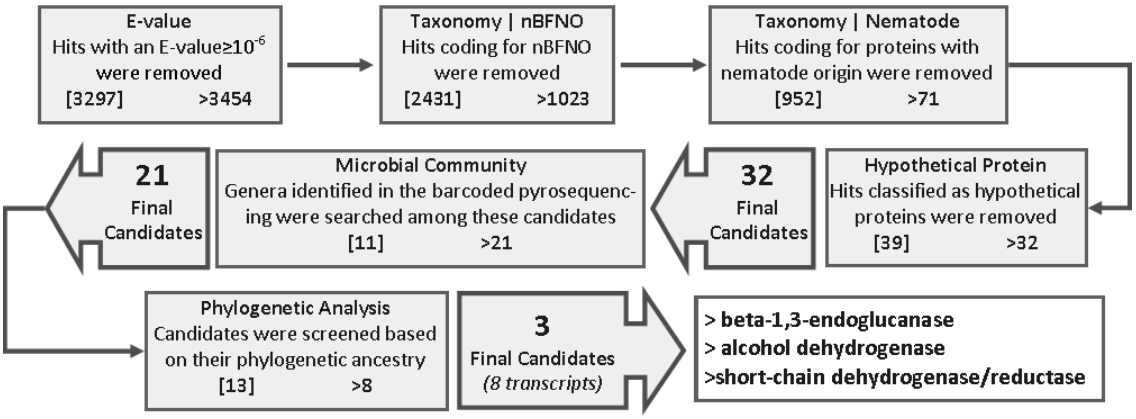


Figure 24 - Final pipeline with the phylogenetic analysis applied as a filter and the three final candidates.
Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

5-Horizontal gene transfer candidate validation

Through the adopted methodology we demonstrated the existence of three HGT candidates in the transcriptome of a laboratory maintained culture of *B. xylophilus* grown on the fungus *Botrytis cinerea*. Two of the final three candidates had already been reported as genes incorporated by HGT in the genome of nematodes, leaving us with one new uncharacterized gene. Phylogenetic analysis was the core element to settle on the phylogenetic origin of the candidate gene. For that reason, we present the phylogenetic trees of the three candidate genes incorporated in the genome of *B. xylophilus* by HGT. In

the presented phylogenetic trees we noticed that the candidate HGT genes did not appear clustered with nematodes, which points out their low homology to the genes in this phylum.

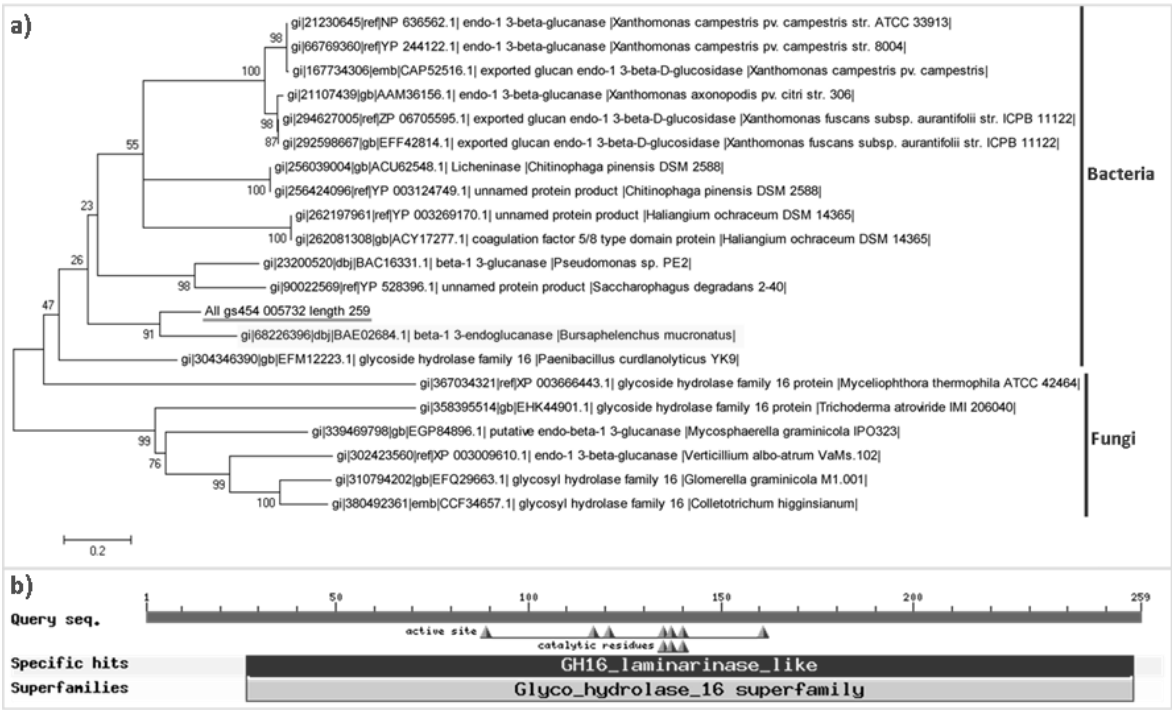


Figure 25 – Phylogenetic tree of the beta-1,3-endoglucanase candidate.
a) phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (1 000 bootstraps) and the JTT-F substitution model; b) conserved domain search from NCBI; the candidate transcript is underlined.

From the phylogenetic tree presented in figure 25 a) we saw that the transcript All_gs454_005732, that codes for beta-1,3-endoglucanase, appears in the Bacteria cluster inside a separated branch with *B. mucronatus*. *Kikuchi et al.*[47] reported that beta-1,3-endoglucanase was incorporated in the genome of both nematodes from the *Bursaphelenchus* genus, *B. xylophilus* and *B. mucronatus*. To build the phylogenetic tree of this transcript we used the sequences retrieved by BLASTp. No sequence from the nematoda phylum was presented in the phylogenetic tree because BLAST only retrieved nematode sequences with a very low score and without the conserved domain presented in figure 25 b). In the phylogenetic tree published in the paper of *Kikuchi et al.*[47] nematode sequences were also absent.

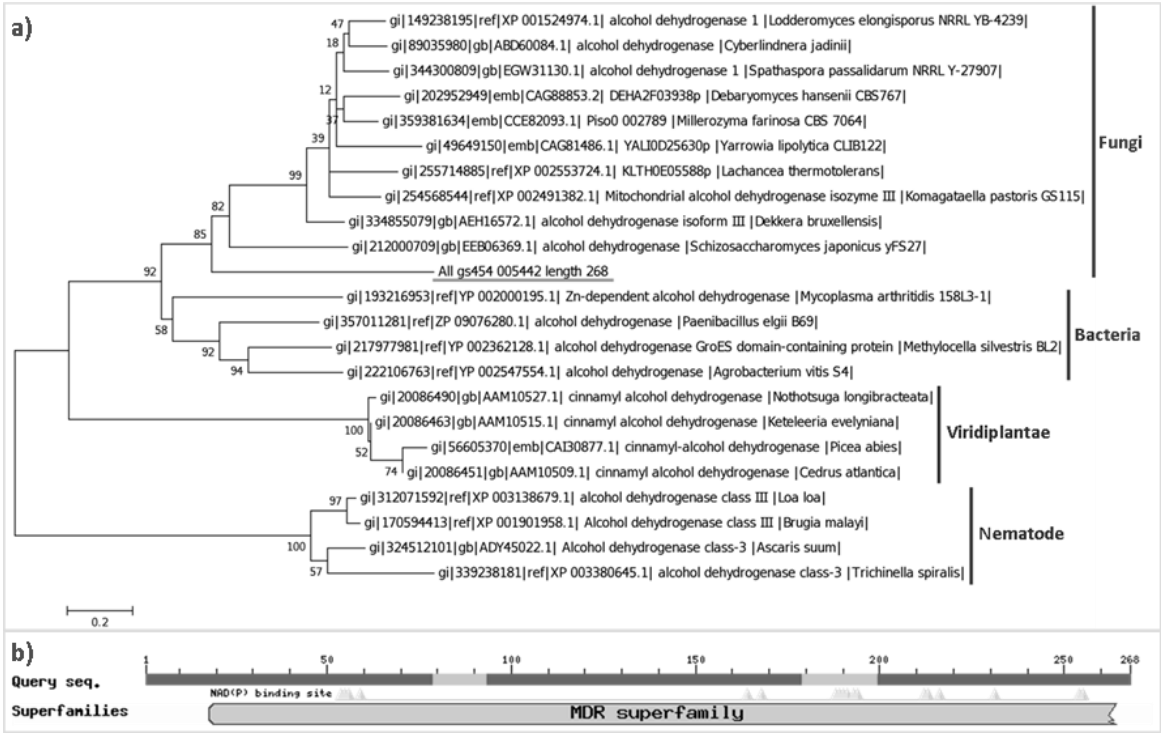


Figure 26 - Phylogenetic tree of the alcohol dehydrogenase candidate.
a) phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (1 000 bootstraps) and the JTT-F substitution model; b) conserved domain search from NCBI; the candidate transcript is underlined.

Parkinson et al.[93] developed a Java/Pearl-based application to identify genes with unexpected patterns of phylogenetic affinity, which led them to the discovery of a alcohol dehydrogenase gene incorporated in the genome of *C. elegans* by HGT. By analyzing the presented phylogenetic tree (figure 26a), we concluded that the gene clustered in fungi.

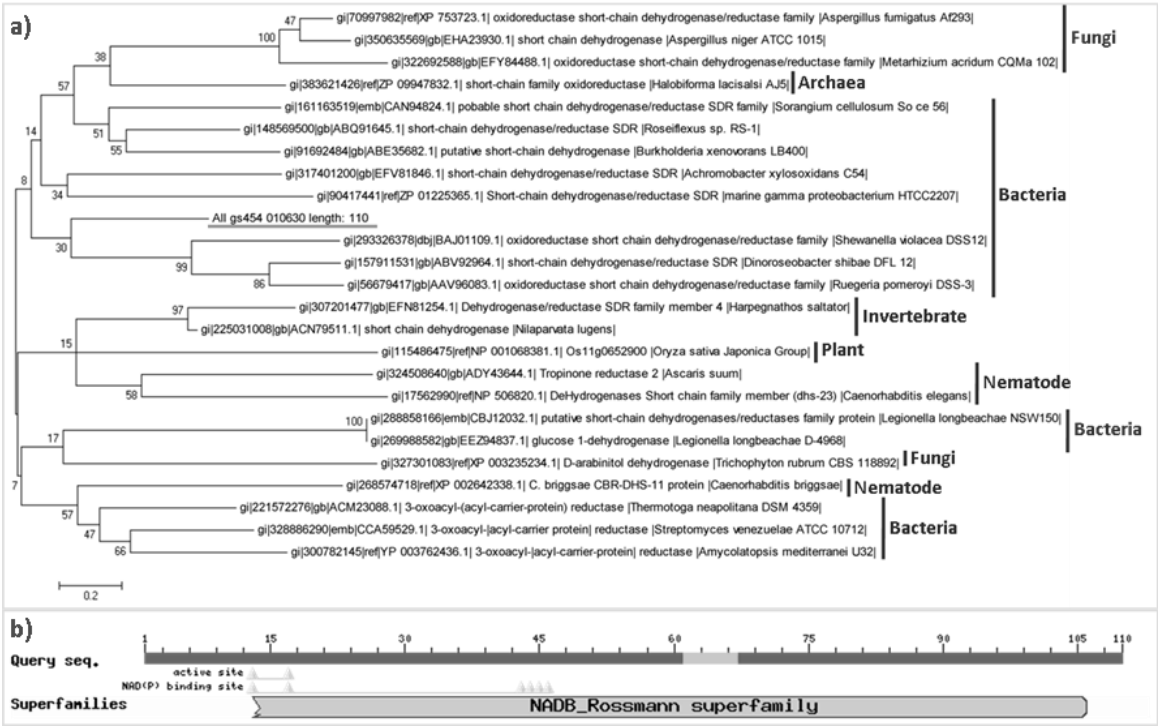
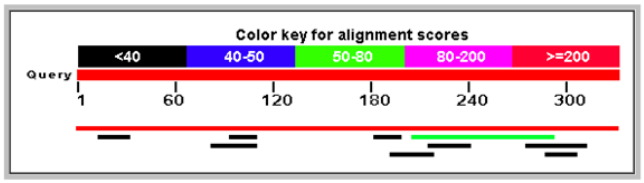


Figure 27 - Phylogenetic tree of the short-chain dehydrogenase/reductase candidate.
a) phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (1 000 bootstraps) and the JTT-F substitution model; b) conserved domain search from NCBI; the candidate transcript is underlined.

The phylogenetic tree presented in figure 27a) corresponds to the last HGT candidate, a Short-chain dehydrogenase/reductase (further on referred to as HGT gene), clustered within bacterial proteins. The next step was to understand its role in the pathogenic process and survival of the PWN. In the phylogenetic analysis we checked all candidates for the presence of a conserved functional domain and only those possessing one passed. The conserved domain search for the HGT gene is presented in figure 27b) and revealed a Rossman-fold NAD(P)(+)-binding protein. Proteins containing this domain are involved in numerous processes and metabolic pathways such as glycolysis. Proteins with this domain generally contain a second one, conferring the protein a specificity for substrate binding, or enzymatic activity. Yet, no further domain was presented in this transcript.




```
 Score =  596 bits (660),  Expect = 6e-170
 Identities = 330/330 (100%), Gaps = 0/330 (0%)
 Strand=Plus/Minus

Query  1      GGGGTGGCCCGCAAAATTCAACCCGCAATTCCCATGTACGCGATCAGCAAAGCCGCCGTA  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  20593  GGGGTGGCCCGCAAAATTCAACCCGCAATTCCCATGTACGCGATCAGCAAAGCCGCCGTA
20534

Query  61     GAGCATTACGCTCGGCACGCCGCCTTCGAATACGCCGCGGATGGGATTCGCGTGAACTGC  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  20533  GAGCATTACGCTCGGCACGCCGCCTTCGAATACGCCGCGGATGGGATTCGCGTGAACTGC
20474

Query  121    GTTGCGCCCGGAATCATCGAGAGTGCGTTCCATGAACGTGGGGCCAATCCGCATCCGGAC  180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  20473  GTTGCGCCCGGAATCATCGAGAGTGCGTTCCATGAACGTGGGGCCAATCCGCATCCGGAC
20414

Query  181    GCCGCCAAAGCGGCGGCCGCGATGGTGCCGCTGAAGAGGATGGGAAAGCCGGAAGAAGCG  240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  20413  GCCGCCAAAGCGGCGGCCGCGATGGTGCCGCTGAAGAGGATGGGAAAGCCGGAAGAAGCG
20354

Query  241    GCGGCTATGATGGTGTTTGTCGCCAGTGACAAGGCCAGCTACATCACTGGGGAATTCTTC  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  20353  GCGGCTATGATGGTGTTTGTCGCCAGTGACAAGGCCAGCTACATCACTGGGGAATTCTTC
20294

Query  301    GGAGTCAACGGTGGACTCATGATAAAACCA  330
              ||||||||||||||||||||||||||||||
Sbjct  20293  GGAGTCAACGGTGGACTCATGATAAAACCA  20264
```

Locus: emb|CADV01009067.1|
Organism: Bursaphelenchus xylophilus
Definition: Bursaphelenchus xylophilus, WGS project CADV00000000 data, strain Ka4C1 (inbred line), scaffold01226.19, whole genome shotgun sequence
Length: 29986 bp

Figure 28 – BLAST search result of HGT gene against the genome of *B. xylophilus*.

The transcript sequence was searched in the genome of *B. xylophilus* to confirm its origin. Figure 28 presents the BLASTn result and the sequence alignment. The transcript sequence had an identity of 100% with part of the locus emb|CADV01009067.1| belonging to *B. xylophilus* (figure 28). Yet, with the information gathered so far it was not possible to determine the role of this gene in the biology of the PWN. Annotating the gene could shed a light, given that this process would provide the whole gene sequence.

6-Short chain dehydrogenase annotation

Since we were able to identify the transcript sequence in the genome of *B. xylophilus*, annotating the gene seemed like a viable solution to better characterize the HGT gene. Our transcript was only part of a larger gene, therefore, annotating the region of

DNA were it was found, could help us determine the complete sequence of the gene and consequently establish its role the PWN.

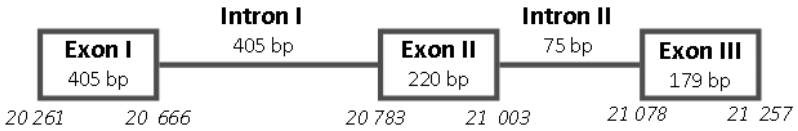


Figure 29 - Short-chain dehydrogenase/reductase gene identification.
Boxes represent the exons and the lines represent the introns. The numbers indicated below the exons/introns indicate their size in bp. The numbers in italic indicate the positions in the locus emb|CADV01009067.1| belonging to *B. xylophilus.* The gene has a total size of 807 bp.

This step provided additional information regarding the assembly of the HGT gene. We found that the gene was 807 bp long, from position 20261 to 21257 of the genome, had 3 exons and 2 introns, presented in figure 29. The transcript sequence fitted entirely in exon I, the biggest of the three and slightly bigger than the average exon (288.9 bp), reported by *Kikuchi et al.*[9]. The gene sequence was analyzed with Blast2GO®, an all in one tool for functional annotation of (novel) sequences and the analysis of annotation data[77]. Unfortunately, no conclusive result was recovered. Blast2GO®, was unable to attribute a metabolic pathway to our gene and the indicated enzyme (EC:1.1.1.0) after extensive search in online databases revealed to be inexistent. In the quest to determine the function of the candidate gene, searches were conducted in EXPASY, PDB, Procognate, Pfam, KEGG, Rebase, DBGet and Brenda, but none was conclusive. At this moment we are unable to infer the role of this gene in the biology of the PWN.

Chapter IV - Conclusion

Pine wilt disease (PWD) is a devastating disease that kills pine trees in a short time span. This condition is caused by *Bursaphelenchus xylophilus*, also known as the pine wood nematode (PWN). The molecular mechanism of the disease is unknown, however the finding of genes unusual to nematodes in *B. xylophilus*, acquired through horizontal gene transfer (HGT), has pointed out this mechanism as an hypothesis to explain how *B. xylophilus* could have acquired the necessary tools to become a parasite of conifer trees. Currently HGT genes identified in *B. xylophilus* code for cell wall degrading enzymes, expansins, pectate lyases and other genes, like the beta-1,3-endoglucanase found in this study. However, these identified and characterized genes alone are not sufficient to explain the development and progression of the disease. The goal of our study was to find in the transcriptome of *B. xylophilus* new uncharacterized genes incorporated by HGT in its genome.

The transcripts were placed in a searchable database and screened to encounter the HGT candidate genes. No previous report shed lights on which parameters should be taken into account when searching for HGT candidates amongst a transcript library nor the best methodology to screen data to encounter candidates. These studies only reported the approach used to prove the horizontal transfer of such gene as well as the methods to verify its functionality. Therefore our first objective was to develop a viable data pipeline that could be used to encounter HGT candidates. The final pipeline could be divided in three phases: first, where the candidates were filtered based on the information contained in the database; second, using the results from the diversity of the microbial community associated with the nematode and third, through phylogenetic analysis, which confirmed the origin of the candidate. As a result of the pipeline we obtained three HGT candidates: beta-1,3-endoglucanase, alcohol dehydrogenase and short-chain dehydrogenase/reductase. The first two candidates have already been reported, which proved that our approach was appropriate to screen HGT genes. The last candidate had never been reported and even though we can't determine its role in the development of *B. xylophilus*, at this point, we demonstrated its presence in the genome of *B. xylophilus* as well as its bacterial origin.

The biodiversity assay not only identified the microbial community present in the nematodes surrounding but also revealed itself as extremely useful to screen HGT candidates. The bacterial genera identified in this trial which contain samples from four distinct geographical origins match the results reported in the bibliography.

Future work

The results accomplished in this study open paths to new studies that will help understand the mechanisms behind PWD and hopefully prevent its progression, reducing its socio-economical impact. Among these suggestions are:

I. Silencing of the short-chain dehydrogenase/reductase gene in *B. xylophilus* so that by observing the resulting phenotype, its role can be determined;
II. Sequencing the transcriptome of *B. xylophilus* grown in pine trees to determine which genes are being expressed and search for HGT candidates;
III. Develop a software application, adopting all steps of the proposed pipeline to aid in the quest for new HGT candidates, using transcripts;

Bibliography

1.Meldal, B. H. M.; Debenham, N. J.; De Ley, P.; De Ley, I. T.; Vanfleteren, J. R.; Vierstraete, A. R.; Bert, W.; Borgonie, G.; Moens, T.; Tyler, P. A.; Austen, M. C.; Blaxter, M. L.; Rogers, A. D.; Lambshead, P. J. D., An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. *Molecular Phylogenetics and Evolution* **2007,** *42* (3), 622-636.
2.Williamson, V. M.; Hussey, R. S., Nematode Pathogenesis and Resistance in Plants. *The Plant Cell Online* **1996,** *8* (10), 1735-1745.
3.Williamson, V. M.; Kumar, A., Nematode resistance in plants: the battle underground. *Trends in Genetics* **2006,** *22* (7), 396-403.
4.Dieterich, C.; Clifton, S. W.; Schuster, L. N.; Chinwalla, A.; Delehaunty, K.; Dinkelacker, I.; Fulton, L.; Fulton, R.; Godfrey, J.; Minx, P.; Mitreva, M.; Roeseler, W.; Tian, H.; Witte, H.; Yang, S.-P.; Wilson, R. K.; Sommer, R. J., The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet* **2008,** *40* (10), 1193-1198.
5.Bird, D. M.; Opperman, C. H.; Davies, K. G., Interactions between bacteria and plant-parasitic nematodes: now and then. *International Journal for Parasitology* **2003,** *33* (11), 1269-1276.
6.Ghedin, E.; Wang, S.; Spiro, D.; Caler, E.; Zhao, Q.; Crabtree, J.; Allen, J. E.; Delcher, A. L.; Guiliano, D. B.; Miranda-Saavedra, D.; Angiuoli, S. V.; Creasy, T.; Amedeo, P.; Haas, B.; El-Sayed, N. M.; Wortman, J. R.; Feldblyum, T.; Tallon, L.; Schatz, M.; Shumway, M.; Koo, H.; Salzberg, S. L.; Schobel, S.; Pertea, M.; Pop, M.; White, O.; Barton, G. J.; Carlow, C. K. S.; Crawford, M. J.; Daub, J.; Dimmic, M. W.; Estes, C. F.; Foster, J. M.; Ganatra, M.; Gregory, W. F.; Johnson, N. M.; Jin, J.; Komuniecki, R.; Korf, I.; Kumar, S.; Laney, S.; Li, B.-W.; Li, W.; Lindblom, T. H.; Lustigman, S.; Ma, D.; Maina, C. V.; Martin, D. M. A.; McCarter, J. P.; McReynolds, L.; Mitreva, M.; Nutman, T. B.; Parkinson, J.; Peregrín-Alvarez, J. M.; Poole, C.; Ren, Q.; Saunders, L.; Sluder, A. E.; Smith, K.; Stanke, M.; Unnasch, T. R.; Ware, J.; Wei, A. D.; Weil, G.; Williams, D. J.; Zhang, Y.; Williams, S. A.; Fraser-Liggett, C.; Slatko, B.; Blaxter, M. L.; Scott, A. L., Draft Genome of the Filarial Nematode Parasite Brugia malayi. *Science* **2007,** *317* (5845), 1756-1760.
7.Dieterich, C.; Sommer, R. J., How to become a parasite – lessons from the genomes of nematodes. *Trends in Genetics* **2009,** *25* (5), 203-209.
8.Dorris, M.; De Ley, P.; Blaxter, M. L., Molecular analysis of nematode diversity and the evolution of parasitism. *Parasitology Today* **1999,** *15* (5), 188-193.
9.Kikuchi, T.; Cotton, J. A.; Dalzell, J. J.; Hasegawa, K.; Kanzaki, N.; McVeigh, P.; Takanashi, T.; Tsai, I. J.; Assefa, S. A.; Cock, P. J. A.; Otto, T. D.; Hunt, M.; Reid, A. J.; Sanchez-Flores, A.; Tsuchihara, K.; Yokoi, T.; Larsson, M. C.; Miwa, J.; Maule, A. G.; Sahashi, N.; Jones, J. T.; Berriman, M., Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen *Bursaphelenchus xylophilus*. *PLoS Pathog* **2011,** *7* (9), e1002219.
10.Gheysen, G.; Fenoll, C., Gene expression in nematode feeding sites. *Annual Review of Phytopathology* **2002,** *40*, 191-219.
11.Haegeman, A.; Mantelin, S.; Jones, J. T.; Gheysen, G., Functional roles of effectors of plant-parasitic nematodes. *Gene* **2012,** *492* (1), 19-31.
12.Davis, E. L.; Hussey, R. S.; Baum, T. J.; Bakker, J.; Schots, A., Nematode parasitism genes. *Annual Review of Phytopathology* **2000,** *38*, 365-396.
13.Davis, E. L.; Hussey, R. S.; Mitchum, M. G.; Baum, T. J., Parasitism proteins in nematode–plant interactions. *Current Opinion in Plant Biology* **2008,** *11* (4), 360-366.
14.Jones, J. T.; Moens, M.; Mota, M.; Hongmei, L.; Kikuchi, T., *Bursaphelenchus xylophilus*: opportunities in comparative genomics and molecular host–parasite interactions. *Mol. Plant Pathol.* **2008,** *9* (3), 357-368.
15.Kikuchi, T.; Jones, J. T.; Aikawa, T.; Kosaka, H.; Ogura, N., A family of glycosyl hydrolase family 45 cellulases from the pine wood nematode *Bursaphelenchus xylophilus*. *Febs Letters* **2004,** *572* (1-3), 201-205.
16.Nickle, W. R., A taxonomic review of the genera of the *Aphelenchoidea* (Fnchs, 1937) Thorne, 1949 (Nematoda: *Tylenchida*). *J. Nematol.* **1970,** (2), 375-392.
17.Steiner, G.; Buhrer, E. M., *Aphelenchoides xylophilus*, n. sp., a nematode associated with blue-stain and other fungi in timber. *Journal of Agricultural Research* **1934,** *48*, 0949-0951.
18.Spiegel, Y.; McClure, M. A., The surface-coat of plant-parasitic nematodes - chemical-composition, origin, and biological role - a review. *J. Nematol.* **1995,** *27* (2), 127-134.
19.Roriz, M.; Santos, C.; Vasconcelos, M. W., Population dynamics of bacteria associated with different strains of the pine wood nematode *Bursaphelenchus xylophilus* after inoculation in maritime pine (*Pinus pinaster*). *Exp. Parasitol.* **2011,** *128* (4), 357-364.
20.Xie, L. Q.; Zhao, B. G., Post-inoculation population dynamics of *Bursaphelenchus xylophilus* and associated bacteria in pine wilt disease on Pinus thunbergii. *J. Phytopathol.* **2008,** *156* (7-8), 385-389.

21.Forestry, N. A. S. P. How to Identify and Manage Pine Wilt Disease and Treat Wood Products Infested by the Pinewood Nematodes. http://www.na.fs.fed.us/spfo/pubs/howtos/ht_pinewilt/pinewilt.htm (accessed 4-1-2012).
22.Li, H. Identification and pathogenicity of *Bursaphelenchus* species (Nematoda: Parasitaphelenchidae). Thesis submitted in fulfillment of the requirements for the degree of Doctor (PhD) in applied biological sciences, 2008.
23.Sousa, E.; Naves, P.; Bonifácio, L.; Bravo, M. A.; Penas, A. C.; Pires, J.; Serrão, M., Preliminary survey for insects associated with *Bursaphelenchus xylophilus* in Portugal. *EPPO Bulletin* **2002,** *32* (3), 499.
24.Suzuki, K., Pine wilt disease - A threat to pine forests in Europe. In *Pinewood Nematode, Bursaphelenchus Xylophilus*, Mota, M.; Vieira, P., Eds. 2004; Vol. 1, pp 25-30.
25.Sousa, E.; Bravo, M. A.; Pires, J.; Naves, P.; Penas, A. C.; Bonifacio, L.; Mota, M. M., *Bursaphelenchus xylophilus* (Nematoda; Aphelenchoididae) associated with *Monochamus galloprovincialis* (Coleoptera; Cerambycidae) in Portugal. *Nematology* **2001,** *3*, 89-91.
26.ZUZARTE, A., Contribuição para o conhecimento dos Cleridae, Buprestidae e Cerambycidae de Portugal (Insecta, Coleoptera. Descrição de duas novas espécies de Vesperus Latreille (Co. Cerambycidae). *Boletim Sociedade Portuguesa de Entomologia* **1985,** (1), 95-103.
27.Paulino de Oliveira, M., *Catalogue des insectes du Portugal : Coléoptères*. Imprensa da Universidade: Coimbra, 1882.
28.Mamiya, Y.; Kiyohara, T., Description of bursaphelenchus-lignicolus n-sp (nematoda-aphelenchoididae) from pine wood and histopathology of nematode-infested trees. *Nematologica* **1972,** *18* (1), 120-&.
29.Nickle, W. R.; Golden, A. M.; Mamiya, Y.; Wergin, W. P., On the taxonomy and morphology of the pine wood nematode, *Bursaphelenchus xylophilus* (steiner and buhrer 1934) Nickle 1970. *J. Nematol.* **1981,** *13* (3), 385-392.
30.Mota, M. M.; Braasch, H.; Bravo, M. A.; Penas, A. C.; Burgermeister, W.; Metge, K.; Sousa, E., First report of *Bursaphelenchus xylophilus* in Portugal and in Europe. *Nematology* **1999,** *1* (7/8), 727-734.
31.Jikumaru, S.; Togashi, K., Temperature effects on the transmission of *Bursaphelenchus xylophilus* (Nemata : *Aphelenchoididae*) by *Monochamus alternatus* (Coleoptera : *Cerambycidae*). *J. Nematol.* **2000,** *32* (1), 110-116.
32.Kuroda, K., Mechanism of cavitation development in the pine wilt disease. *European Journal of Forest Pathology* **1991,** *21* (2), 82-89.
33.Trudgill, D. L., Resistance to and tolerance of plant parasitic nematodes in plants. *Annual Review of Phytopathology* **1991,** *29*, 167-192.
34.Myers, R. F., Pathogenesis in pine wilt caused by pinewood nematode, *Bursaphelenchus xylophilus*. *J. Nematol.* **1988,** *20* (2), 236-244.
35.Oku, H.; Shiraishi, T.; Ouchi, S.; Kurozumi, S.; Ohta, H., Pine wilt toxin, the metabolite of a bacterium associated with a nematode. *Naturwissenschaften* **1980,** *67* (4), 198-199.
36.Han, Z. M.; Hong, Y. D.; Zhao, B. G., A study on pathogenicity of bacteria carried by pine wood nematodes. *J. Phytopathol.* **2003,** *151* (11-12), 683-689.
37.Guo, D.; Zhao, B.; Gao, R., Observation of the site of pinewood nematode where bacteria are carried with SEM and TEM. *J Nanjing For Univ* **2000,** (4), 69–71.
38.Zhao, B.; Wang, H.; Han, S.; Han, Z., Distribution and pathogenicity of bacteria species carried by *Bursaphelenchus xylophilus* in China. *Nematology* **2003,** (6), 899–906.
39.Zhao, B. G.; Lin, F., Mutualistic symbiosis between *Bursaphelenchus xylophilus* and bacteria of the genus *Pseudomonas*. *Forest Pathology* **2005,** *35* (5), 339-345.
40.Vicente, C. S. L.; Nascimento, F.; Espada, M.; Mota, M.; Oliveira, S., Bacteria associated with the pinewood nematode Bursaphelenchus xylophilus collected in Portugal. *Antonie Van Leeuwenhoek* **2011,** *100* (3), 477-481.
41.Proença, D. N.; Francisco, R.; Santos, C. V.; Lopes, A.; Fonseca, L.; Abrantes, I. M. O.; Morais, P. V., Diversity of Bacteria Associated with *Bursaphelenchus xylophilus* and Other Nematodes Isolated from *Pinus pinaster* Trees with Pine Wilt Disease. *PLoS One* **2010,** *5* (12), e15191.
42.Dwinell, L. D., The pinewood nematode: Regulation and mitigation. *Annual Review of Phytopathology* **1997,** *35*, 153-166.
43.Sriwati, R.; Takemoto, S.; Futai, K., Cohabitation of the pine wood nematode, *Bursaphelenchus xylophilus*, and fungal species in pine trees inoculated with *B. xylophilus*. *Nematology* **2007,** *9*, 77-86.
44.Nacional, A. F. Pragas e Doenças. http://www.afn.min-agricultura.pt/portal/pragas-doencas (accessed 25-06-12).

45.Ye, W. M.; Giblin-Davis, R. M.; Braasch, H.; Morris, K.; Thomas, W. K., Phylogenetic relationships among *Bursaphelenchus* species (Nematoda : Parasitaphelenchidae) inferred from nuclear ribosomal and mitochondrial DNA sequence data. *Molecular Phylogenetics and Evolution* **2007,** *43* (3), 1185-1197.
46.Kikuchi, T.; Aikawa, T.; Kosaka, H.; Pritchard, L.; Ogura, N.; Jones, J. T., Expressed sequence tag (EST) analysis of the pine wood nematode *Bursaphelenchus xylophilus* and *B. mucronatus*. *Molecular and Biochemical Parasitology* **2007,** *155* (1), 9-17.
47.Kikuchi, T.; Shibuya, H.; Jones, J. T., Molecular and biochemical characterization of an endo-β-1,3-glucanase from the pinewood nematode *Bursaphelenchus xylophilus* acquired by horizontal gene transfer from bacteria. *Biochem. J.* **2005,** *389*, 117-125.
48.Pritchard, L.; Birch, P., A systems biology perspective on plant–microbe interactions: Biochemical and structural targets of pathogen effectors. *Plant Science* **2011,** *180* (4), 584-603.
49.Kikuchi, T.; Li, H. M.; Karim, N.; Kennedy, M. W.; Moens, M.; Jones, J. T., Identification of putative expansin-like genes from the pine wood nematode, *Bursaphelenchus xylophilus*, and evolution of the expansin gene family within the Nematoda. *Nematology* **2009,** *11*, 355-364.
50.Mayer, W. E.; Schuster, L. N.; Bartelmes, G.; Dieterich, C.; Sommer, R. J., Horizontal gene transfer of microbial cellulases into nematode genomes is associated with functional assimilation and gene turnover. *Bmc Evolutionary Biology* **2011,** *11* (13), 1471-2148.
51.Ahmadian, A.; Ehn, M.; Hober, S., Pyrosequencing: History, biochemistry and future. *Clin. Chim. Acta* **2006,** *363* (1-2), 83-94.
52.Ronaghi, M., Pyrosequencing sheds light on DNA sequencing. *Genome Research* **2001,** *11* (1), 3-11.
53.Morozova, O.; Marra, M. A., Applications of next-generation sequencing technologies in functional genomics. *Genomics* **2008,** *92* (5), 255-264.
54.Petrosino, J. F.; Highlander, S.; Luna, R. A.; Gibbs, R. A.; Versalovic, J., Metagenomic Pyrosequencing and Microbial Identification. *Clin Chem* **2009,** *55* (5), 856-866.
55.Institute, N. H. G. R. Transcriptome. http://www.genome.gov/13014330 (accessed 9-1-2012).
56.Kang, J. S.; Lee, H.; Moon, I. S.; Lee, Y.; Koh, Y. H.; Je, Y. H.; Lim, K. J.; Lee, S. H., Construction and characterization of subtractive stage-specific expressed sequence tag (EST) libraries of the pinewood nematode *Bursaphelenchus xylophilus*. *Genomics* **2009,** *94* (1), 70-77.
57.Hotopp, J. C. D.; Clark, M. E.; Oliveira, D.; Foster, J. M.; Fischer, P.; Torres, M. C.; Giebel, J. D.; Kumar, N.; Ishmael, N.; Wang, S. L.; Ingram, J.; Nene, R. V.; Shepard, J.; Tomkins, J.; Richards, S.; Spiro, D. J.; Ghedin, E.; Slatko, B. E.; Tettelin, H.; Werren, J. H., Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **2007,** *317* (5845), 1753-1756.
58.Richards, T. A.; Soanes, D. M.; Jones, M. D. M.; Vasieva, O.; Leonard, G.; Paszkiewicz, K.; Foster, P. G.; Hall, N.; Talbot, N. J., Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proceedings of the National Academy of Sciences of the United States of America* **2011,** *108* (37), 15258-15263.
59.Smant, G.; Stokkermans, J. P. W. G.; Yan, Y.; de Boer, J. M.; Baum, T. J.; Wang, X.; Hussey, R. S.; Gommers, F. J.; Henrissat, B.; Davis, E. L.; Helder, J.; Schots, A.; Bakker, J., Endogenous cellulases in animals: Isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proceedings of the National Academy of Sciences* **1998,** *95* (9), 4906-4911.
60.Haegeman, A.; Vanholme, B.; Gheysen, G., Characterization of a putative endoxylanase in the migratory plant-parasitic nematode *Radopholus similis*. *Mol. Plant Pathol.* **2009,** *10* (3), 389-401.
61.Opperman, C. H.; Bird, D. M.; Williamson, V. M.; Rokhsar, D. S.; Burke, M.; Cohn, J.; Cromer, J.; Diener, S.; Gajan, J.; Graham, S.; Houfek, T. D.; Liu, Q.; Mitros, T.; Schaff, J.; Schaffer, R.; Scholl, E.; Sosinski, B. R.; Thomas, V. P.; Windham, E., Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. *Proceedings of the National Academy of Sciences* **2008**.
62.Kikuchi, T.; Shibuya, H.; Aikawa, T.; Jones, J. T., Cloning and characterization of pectate lyases expressed in the esophageal gland of the pine wood nematode *Bursaphelenchus xylophilus*. *Mol. Plant-Microbe Interact.* **2006,** *19* (3), 280-287.
63.Jones, J. T.; Furlanetto, C.; Kikuchi, T., Horizontal gene transfer from bacteria and fungi as a driving force in the evolution of plant parasitism in nematodes. *Nematology* **2005,** *7*, 641-646.
64.Scholl, E.; Thorne, J.; McCarter, J.; Bird, D. M., Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. *Genome Biology* **2003,** *4* (6), R39.
65.Baxevanis, A. D.; Ouellette, B. F. F., *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Second ed.; John Wiley & Sons, Inc.: 2001.
66.Edgar, R. C.; Batzoglou, S., Multiple sequence alignment. *Current Opinion in Structural Biology* **2006,** *16* (3), 368-373.

67.Bettencourt, R.; Pinheiro, M.; Egas, C.; Gomes, P.; Afonso, M.; Shank, T.; Santos, R., High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus*. *BMC Genomics* **2010,** *11* (1), 559.
68.Edgar, R., MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **2004,** *5* (1), 113.
69.Edgar, R. C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **2004,** *32* (5), 1792-1797.
70.Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S., MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **2011,** *28* (10), 2731-2739.
71.Sogin, M. L.; Morrison, H. G.; Huber, J. A.; Welch, D. M.; Huse, S. M.; Neal, P. R.; Arrieta, J. M.; Herndl, G. J., Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* **2006,** *103* (32), 12115-12120.
72.Edgar, R. C.; Haas, B. J.; Clemente, J. C.; Quince, C.; Knight, R., UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **2011,** *27* (16), 2194-200.
73.Edgar, R. C., Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**.
74.Schloss, P. D., A High-Throughput DNA Sequence Aligner for Microbial Ecology Studies. *PLoS One* **2009,** *4* (12).
75.Felsenstein, J., PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics 5* **1989**, 164-166.
76.Cantarel, B. L.; Korf, I.; Robb, S. M. C.; Parra, G.; Ross, E.; Moore, B.; Holt, C.; Sánchez A., A.; Yandell, M., MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **2008,** *18* (1), 188-196.
77.Conesa, A.; Götz, S.; Garcia-Gomez, J. M.; Terol, J.; Talon, M.; Robles, M., Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **September, 2005,** *21*, 3674-3676.
78.EV, K.; MY, G., *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Kluwer Academic: Boston, 2003.
79.McEntyre, J.; Ostell, J. The NCBI Handbook [Internet]. http://www.ncbi.nlm.nih.gov/books/NBK21101/.
80.Desler, C.; Suravajhala, P.; Sanderhoff, M.; Rasmussen, M.; Rasmussen, L., In Silico screening for functional candidates amongst hypothetical proteins. *BMC Bioinformatics* **2009,** *10* (1), 289.
81.Lubec, G.; Afjehi-Sadat, L.; Yang, J.-W.; John, J. P. P., Searching for hypothetical proteins: Theory and practice based upon original data and literature. *Progress in Neurobiology* **2005,** *77* (1–2), 90-127.
82.Pitman, S. D. Pseudogenes And other Forms of 'Junk' DNA. http://www.detectingdesign.com/pseudogenes.html#Pseudogenes.
83.Kysela, D. T.; Palacios, C.; Sogin, M. L., Serial analysis of V6 ribosomal sequence tags (SARST-V6): a method for efficient, high-throughput analysis of microbial community composition. *Environmental Microbiology* **2005,** *7* (3), 356-364.
84.Huse, S. M.; Dethlefsen, L.; Huber, J. A.; Welch, D. M.; Relman, D. A.; Sogin, M. L., Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing. *PLoS Genet* **2008,** *4* (11), e1000255.
85.Heuer, H.; Hartung, K.; Wieland, G.; Kramer, I.; Smalla, K., Polynucleotide Probes That Target a Hypervariable Region of 16S rRNA Genes To Identify Bacterial Isolates Corresponding to Bands of Community Fingerprints. *Applied and Environmental Microbiology* **1999,** *65* (3), 1045-1049.
86.Seifert, K. A., Progress towards DNA barcoding of fungi. *Mol. Ecol. Resour.* **2009,** *9*, 83-89.
87.Müller, T.; Philippi, N.; Dandekar, T.; Schultz, J.; Wolf, M., Distinguishing species. *RNA* **2007,** *13* (9), 1469-1472.
88.Nilsson, R. H.; Kristiansson, E.; Ryberg, M.; Hallenberg, N.; Larsson, K.; shy; Henrik, Intraspecific ITS Variability in the Kingdom Fungi as Expressed in the International Sequence Databases and Its Implications for Molecular Species Identification. *Evolutionary Bioinformatics* **2008,** *4*, 193-201.
89.Accugenix, I. D2 vs. ITS2: Direct Comparison of Two Sequencing Targets and Their Ability to Differentiate Fungal Species. http://www.accugenix.com/microbial-identification-bacteria-fungus-knowledge-center/micro-id-basics/ribosomal-fungal/.
90.Higgins, D. F.; Harmey, M. A.; Jones, D. L., Pathogenicity related gene expression in *Bursaphelenchus xylophilus*. In *Sustainability of pine forests in relation to pine wilt and decline*, Proceedings of International Symposium: Tokyo, Japan, October, 1998; pp 27-28.
91.Zhao, B.; Li, R., The role of bacteria associated with the pinewood nematode in pathogenicity and toxin–production related to pine wilt. *Pine wilt disease.* **2008**, pp 250–259.

92.Wang, Z.; Wang, C. Y.; Fang, Z. M.; Zhang, D. L.; Liu, L.; Lee, M. R.; Li, Z.; Li, J. J.; Sung, C. K., Advances in research of pathogenic mechanism of pine wilt disease. *African Journal of Microbiology Research* **2010,** *4* (6), 437-442.
93.Parkinson, J.; Blaxter, M., SimiTri—visualizing similarity relationships for groups of sequences. *Bioinformatics* **2003,** *19* (3), 390-395.

Appendix

Appendix I – Pipeline development

In this section we present the various attempts to define the data analysis pipeline.

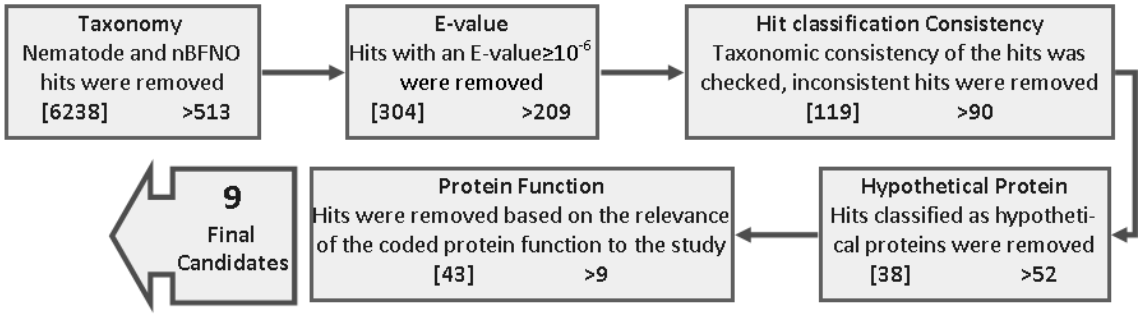


Figure 30 - Strategy I

Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

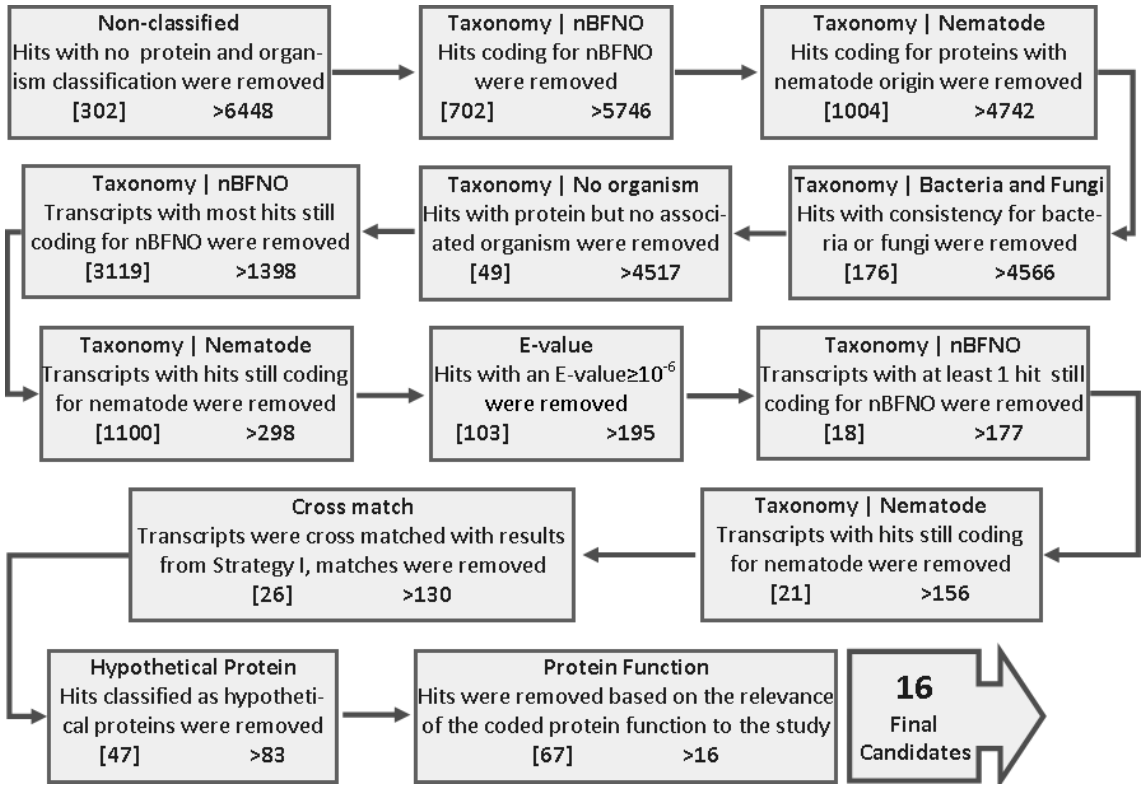


Figure 31 - Strategy II

Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

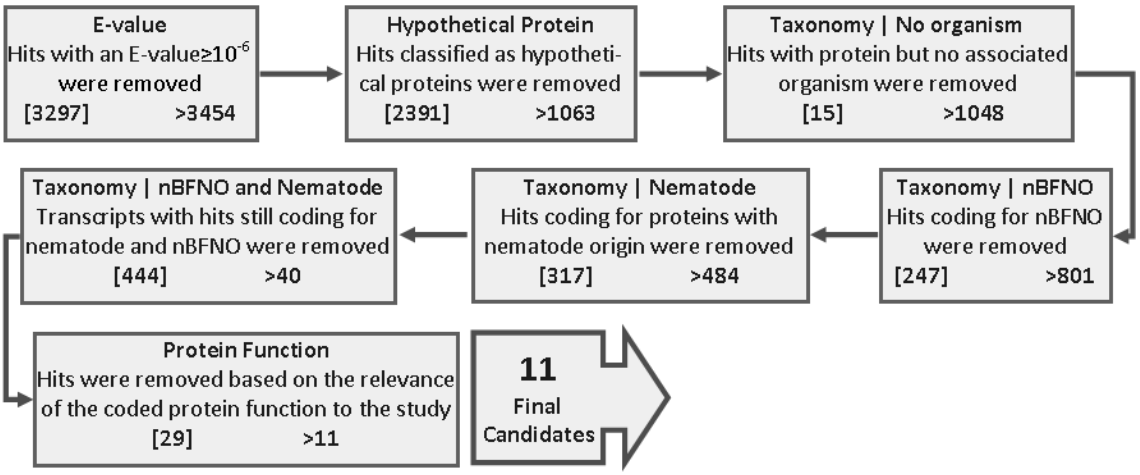


Figure 32 - Strategy III

Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

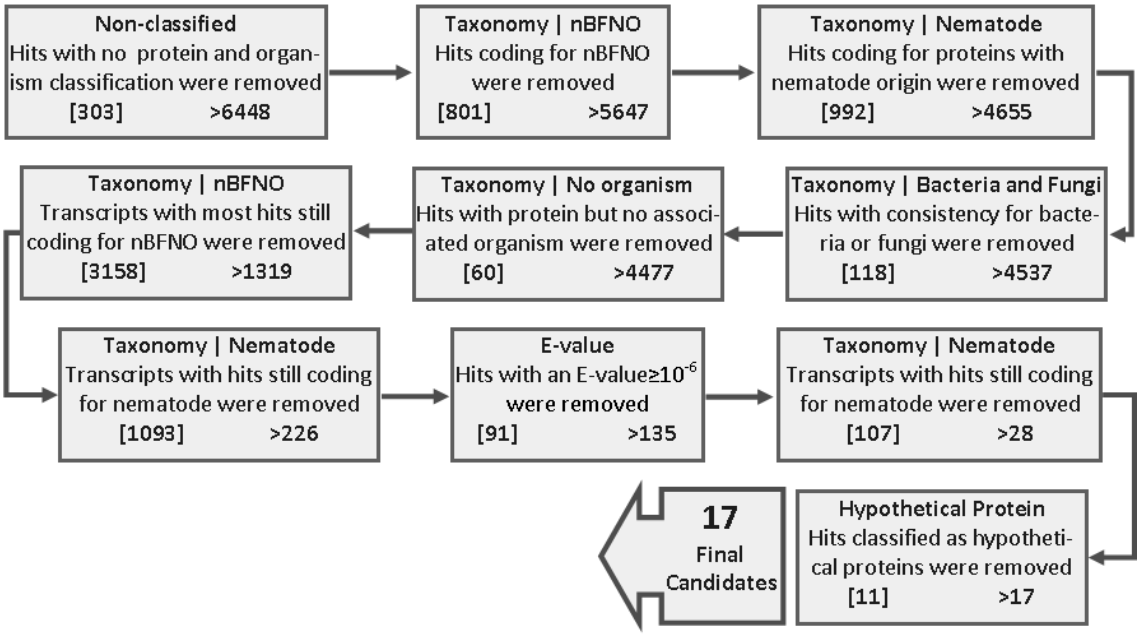


Figure 33 - Strategy IV

Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

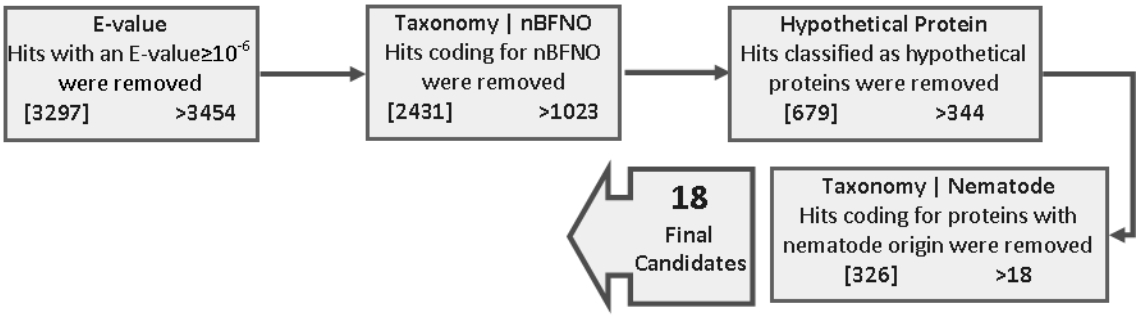


Figure 34 - Strategy V

Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

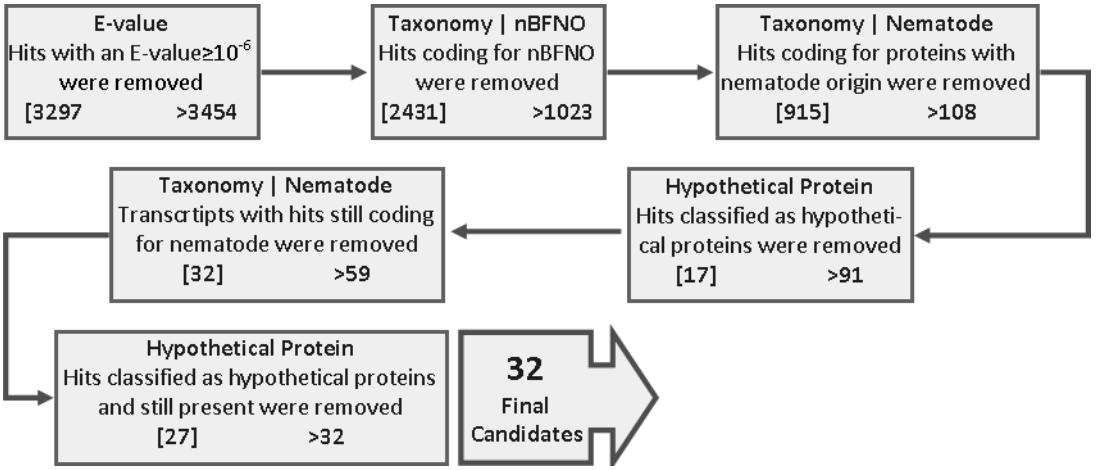


Figure 35 - Strategy VI

Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

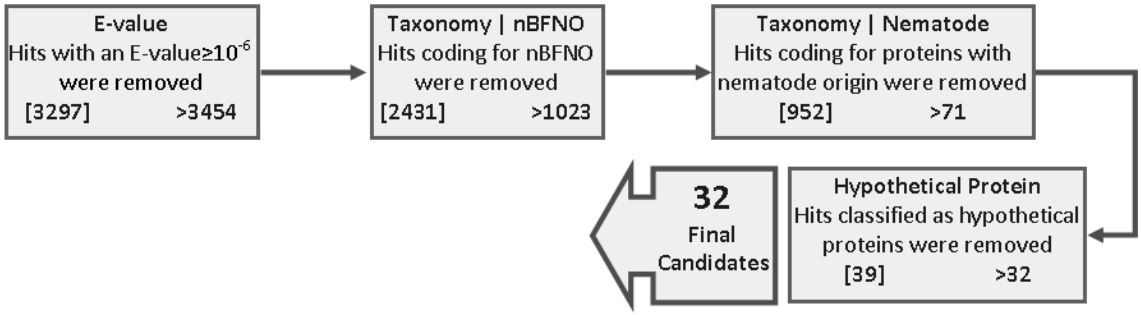


Figure 36 - Strategy VII

Numbers in straight brackets [ ] represent hits that have been removed in that step while the numbers that are preceded by > represent the hits that are maintained and pass to the next step.

Appendix II – Microbial community associated with *B. xylophilus*

ITS II Primer and *B. xylophilus*

In this section we present the reason why *B. xylophilus* was identified in the microbial diversity study.

Table 16 - ITS II forward and reverse primer.

Primer	Target	Sequence (5'-3')
ITSII_Forward	internal transcribed spacer 2 region	GCATCGATGAAGAACGC
ITSII_Reverse		CCTCCGCTTATTGATATGC

>gi|149368663|gb|EF446952.1| Bursaphelenchus xylophilus isolate BXCZZ internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence

CCCGATCCTAATGCACATTTATTCGTGCTCGTCACGATGATGCGATTGGTGACTTCGGTTGCCGCGCATGATG
GCGGTTCGATTCGCGTCGTTCCGCCTACTGATGGTTCGCATGGAAGCCGAGAGGCGACCGTGCAACGGTGAAG
TCTGGGTTTCTACGTGCTGTTGTTGAGTTGGCGTTTTACCGTGCCGACAGATGAGACCAGCCAGCTGCTTGCC
GATTCGTTCTGGCGAGCGTAGGATTGAAAAGCCCGAGAGGCTGCCCTGACAAAACATTCATTTTACATTTATT
TTGTTGGAAAAGAGCTTTAAGTTACTCCGGTGGATCACTTGGCTCGCGGGTCGATGAAGAACGCAGTGAATTG
CGATAATAAGTACGAATTACAGATATTATGAGTACCATGTTTTTGAATGCATATTGCGCTCTTGGGCTTTGCT
CTTGAGCATATTCGATTCAGGGTGTGTTTTTAAACTCGAGCAGAAACGCCGACTTGTTTTTTTCAAGTTTCTG
CACGTTGTGACAGTCGTCTCGCATTGTTCGCGCAATGTTAGGCACCATCTGTTTTACGCGGTTTGTTCCGCGA
CCAATATCTTCTACGCACTGTTTGTCCGTGCGGGGCGAGAGGGCTTCGTGCTCGATTGTCGTGCGCGGCTAAA
CCGTTTGGTGATGTTGTTTCAACGGCGCGGCCGTCAGGGACGTTCGGATGAGAATGTTTGGAGTCCTGGCTGC
GGTTTGTTGAGCTTCGTCGTGAAGCCTTGCGGGCAGTGTTGTCGGAATTGGTTGAAACCACCTGAGTTGGGTA
TGACTACCTGCTGAACTTAAGCATATCAGTAAGCAGAGGAAA

Bacterial diversity studies

Table 17 - Studies developed to identify the bacterial genus accompanying *B. xylophilus*.

	Xie and Zhao, 2008	Roriz, Santos et al. 2011		Vicente, Nascimento et al. 2011		Proença, Francisco et al. 2010	This study
	China	Portugal		Portugal		Portugal	Portugal, China, Japan, USA
				wild	*lab*		
PWN Source	*15 year old dead Japanese black pine in Nanjing, China (Apr.02)*	*2 isolates from Setúbal, 1 from central region and 1 avirulent isolate from Japan*		*5 symtomatic trees from Avô, Coimbra*	*21 lab cultures collected between 05-08 from 11 locations in Portugal*	*3 different areas Alcacer do Sal – Grândola(1) Coimbra(2)*	Lab maintained cultures
Pine species	*Pinus thunbergii*	-		*Pinus pinaster*		*Pinus pinaster*	-
Culture conditions	*Mycelia mat of Botrytis cinerea Pers. on potato-dextrose-agar medium. 10 days | 28ºC | dark*	*Botrytis cinerea Pers. mycelium on barley seeds 7days | 26ºC | dark*		-	*Botrytis cinerea on barley seeds 4ºC*	-	-
Re-insertion	*6/7 year old pine trees*	*1 year old pine trees*		-	-	-	-
Duration	-	*14 days*		-	-	-	-
Pine species	*Pinus thunbergii*	*Pinus pinaster*		-	-	-	-
Monitoring strategies	*-resin flow -color of needles*	*-resin flow -visual analysis*		-	-	-	-
Bacterial source	*Inoculated pine wood*	*Inoculated pine wood*	*PWN surface*	*PWN surface*	*PWN surface*	*Inoculated pine wood*	*Entire PWN*
Disinfection conditions	*75% ethanol*	*75% ethanol*	*3% H_2O_2 | 5min*	*3% H_2O_2 | 3min*		-	-
DNA obtaining technique	*Wood chips placed in NB medium 28ºC | 3 days*	*Wood chips placed in NB medium 28ºC | 3 days*	*Trail analysis | 1PWN placed on NA medium 26ºC*	*Trail analysis TSA (trypticase soy agar) NA (nutrient agar) LA (luria agar) PSDm (Pseudomonas complex médium) 28ºC | 1 week*		*Trail analysis Wood chips placed in R2A medium 25ºC | 3 days*	*PWN was chopped*
Identification technique	*API auto-identidfication*	*Purification and sequencing by Macrogen Korea*		-	-	*ABI 310 DNA Sequencer*	*Roche 454*

(-) not applicable

Appendix III – Phylogenetic analysis

Stage I

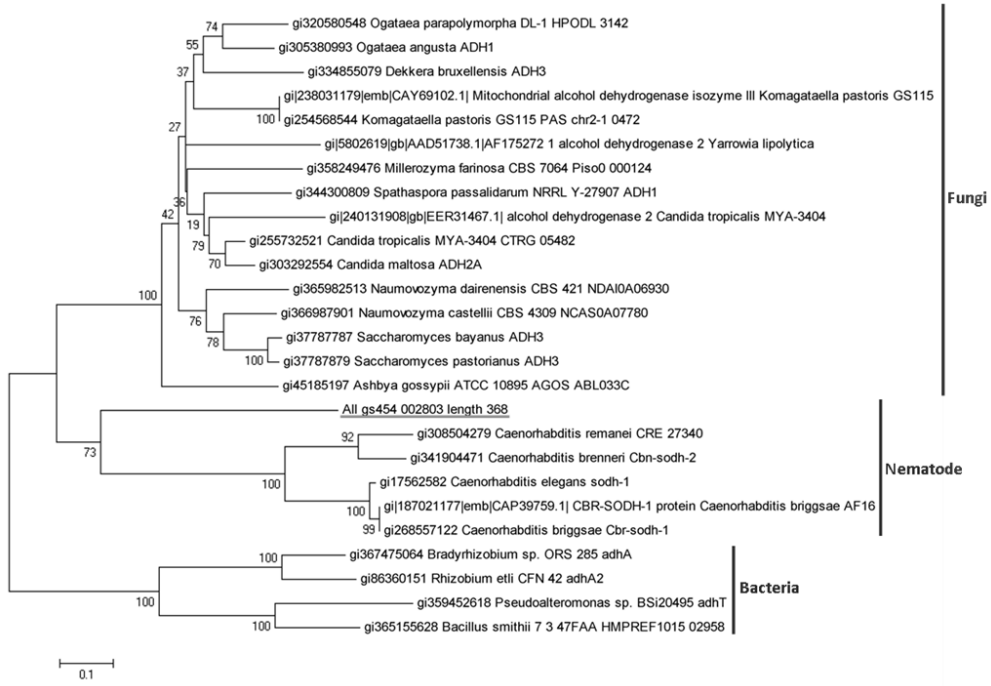


Figure 37 - Phylogenetic tree of the transcript All_gs454_002803 (alcohol dehydrogenase) with *Caenorhabditis* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes.

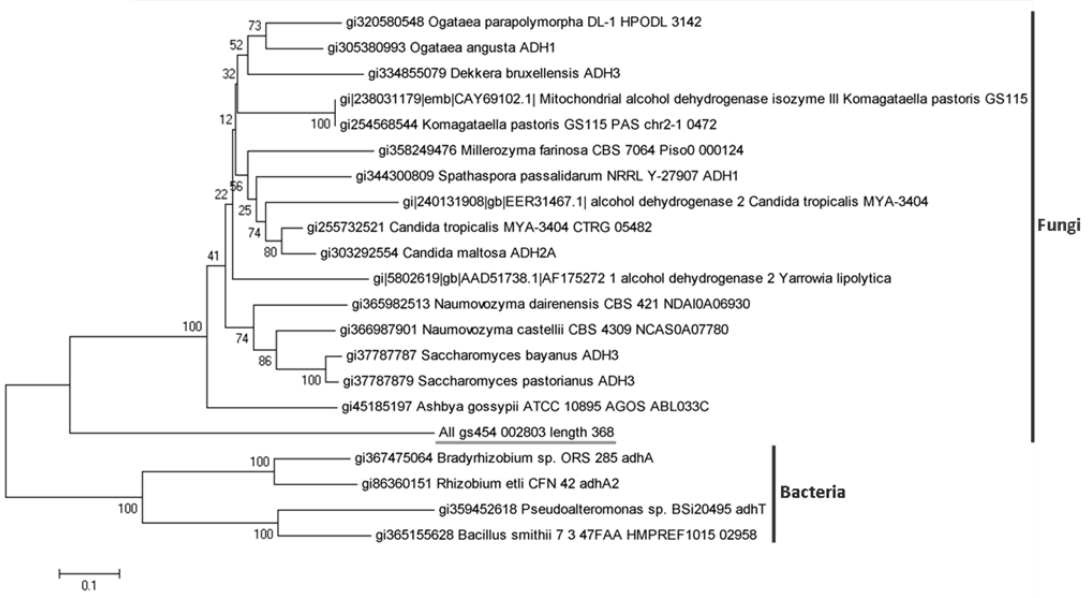


Figure 38 - Phylogenetic tree of the transcript All_gs454_002803 (alcohol dehydrogenase) without *Caenorhabditis* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within fungi, therefore, valid.

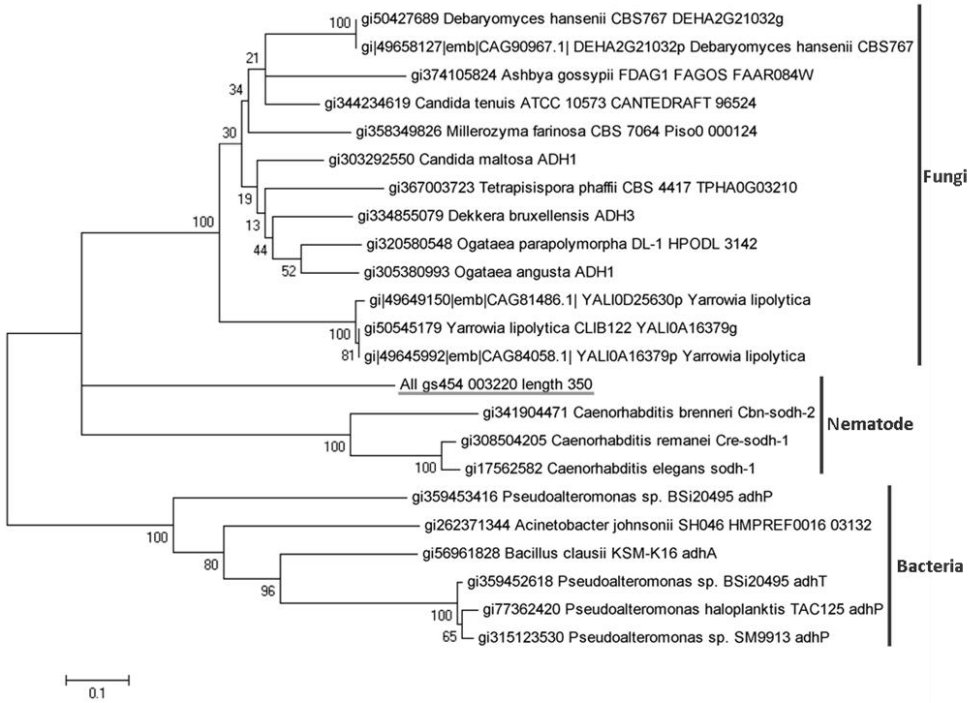


Figure 39- Phylogenetic tree of the transcript All_gs454_003220 (alcohol dehydrogenase) with *Caenorhabditis* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes.

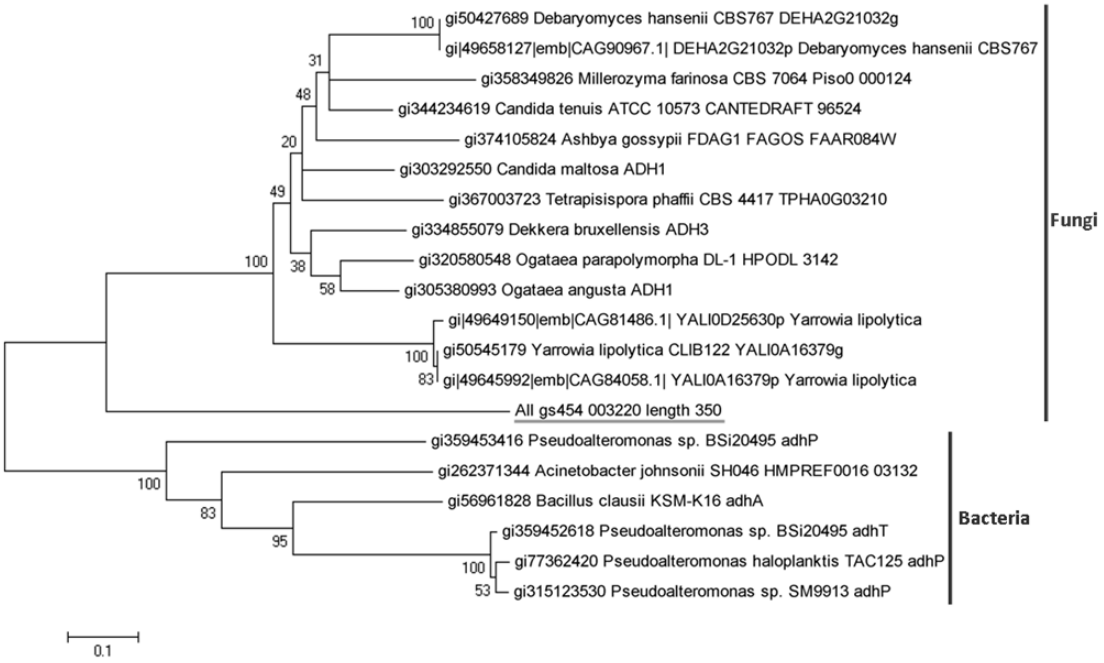


Figure 40 - Phylogenetic tree of the transcript All_gs454_003220 (alcohol dehydrogenase) without *Caenorhabditis* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within fungi, therefore, valid.

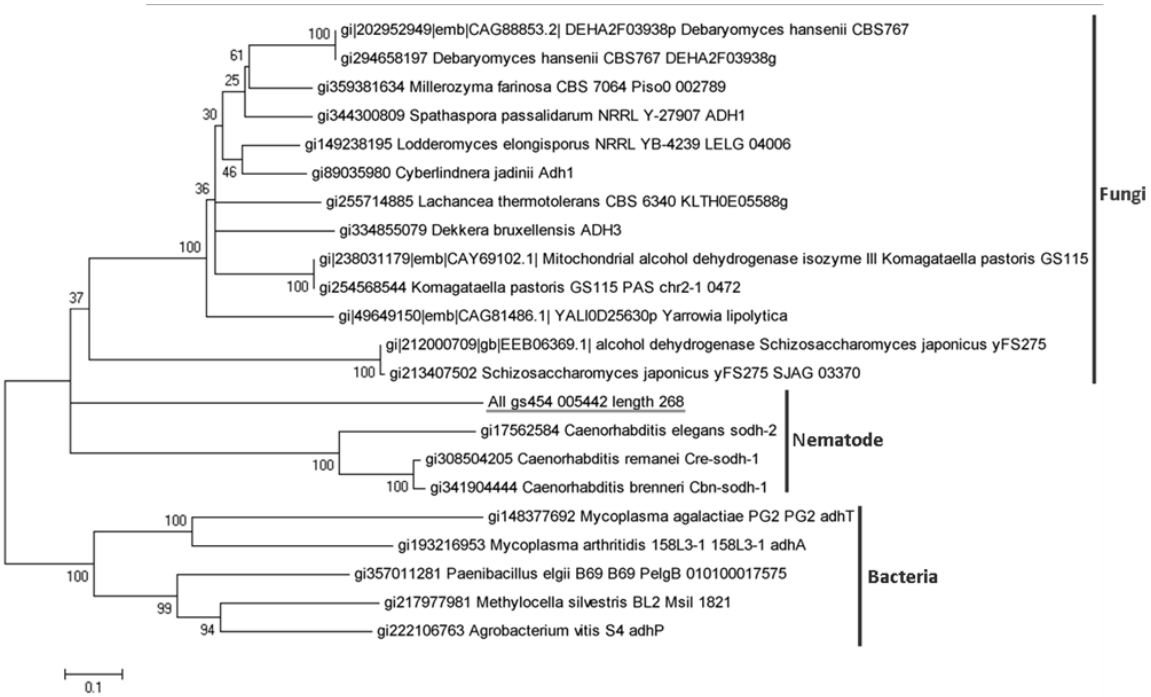


Figure 41 - Phylogenetic tree of the transcript All_gs454_005442 (alcohol dehydrogenase) with *Caenorhabditis* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes.

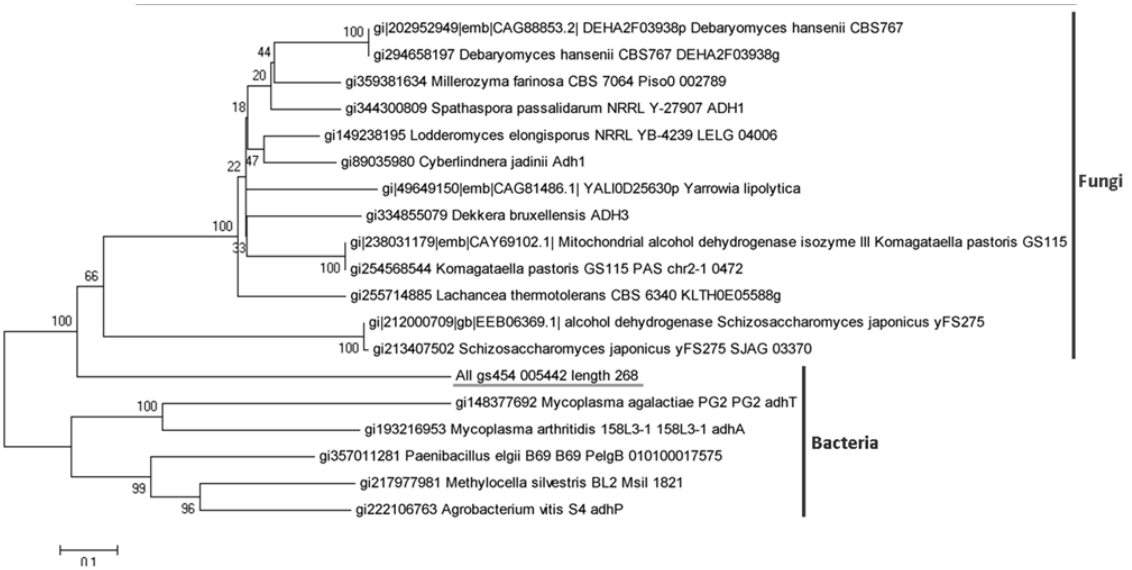


Figure 42 - Phylogenetic tree of the transcript All_gs454_005442 (alcohol dehydrogenase) without *Caenorhabditis* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within bacteria, therefore, valid.

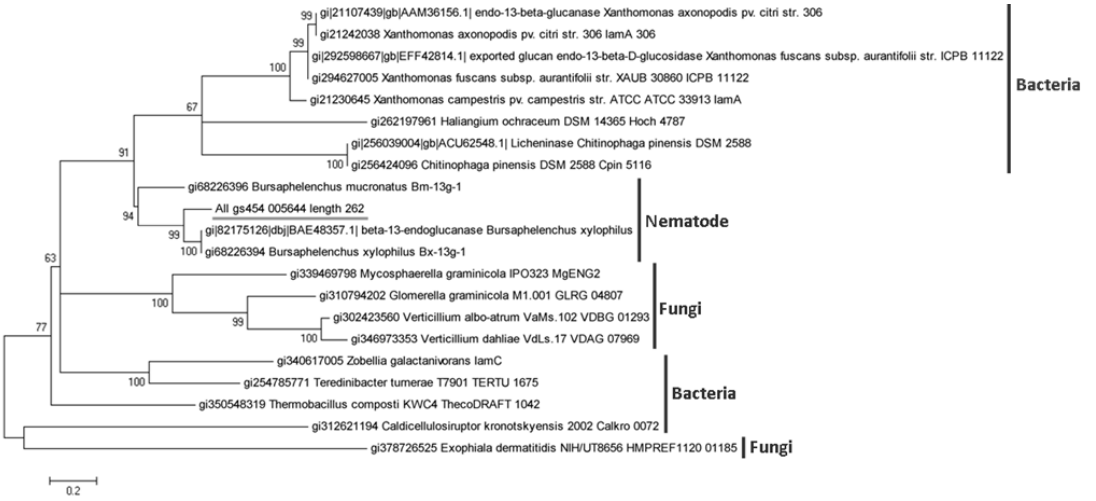


Figure 43 - Phylogenetic tree of the transcript All_gs454_005644 (beta-1,3-endoglucanase) with *Bursaphelenchus* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes.

Figure 44 - Phylogenetic tree of the transcript All_gs454_005676 (beta-1,3-endoglucanase) with *Bursaphelenchus* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes.

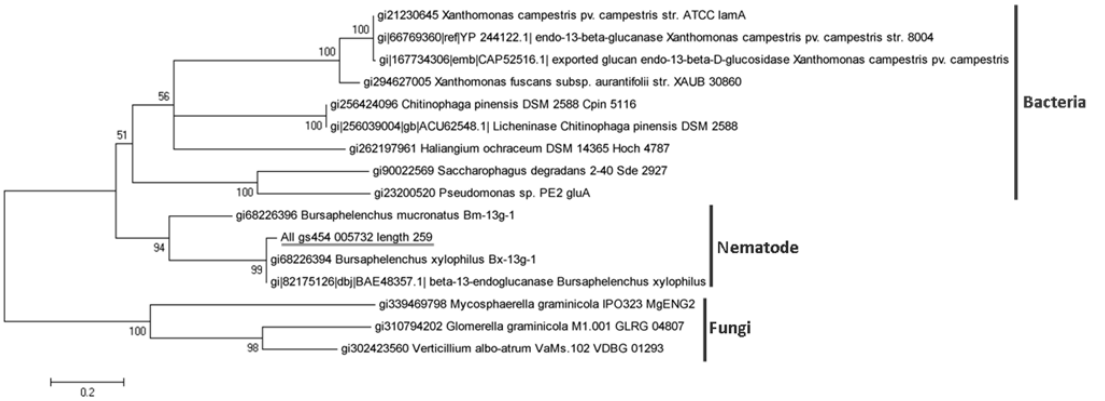


Figure 45 - Phylogenetic tree of the transcript All_gs454_005732 (beta-1,3-endoglucanase) with *Bursaphelenchus* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes.

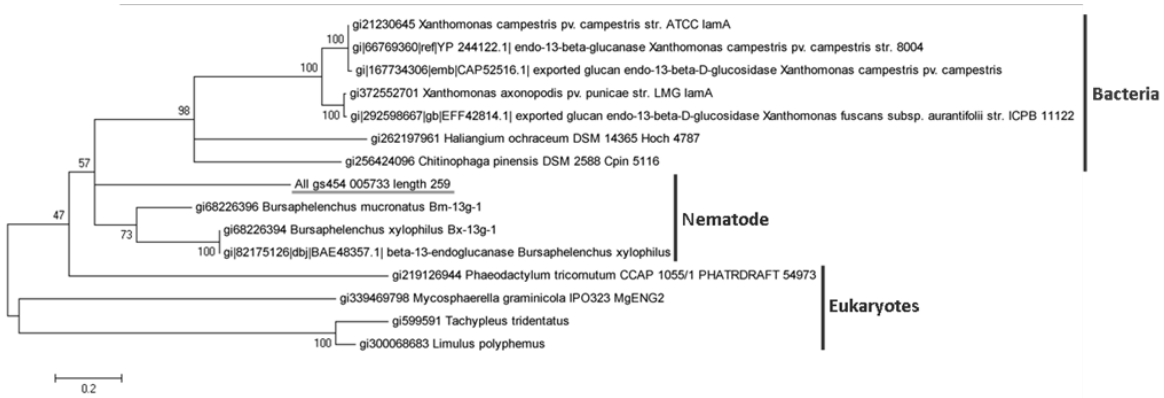


Figure 46 - Phylogenetic tree of the transcript All_gs454_005733 (beta-1,3-endoglucanase) with *Bursaphelenchus* nematodes.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes.

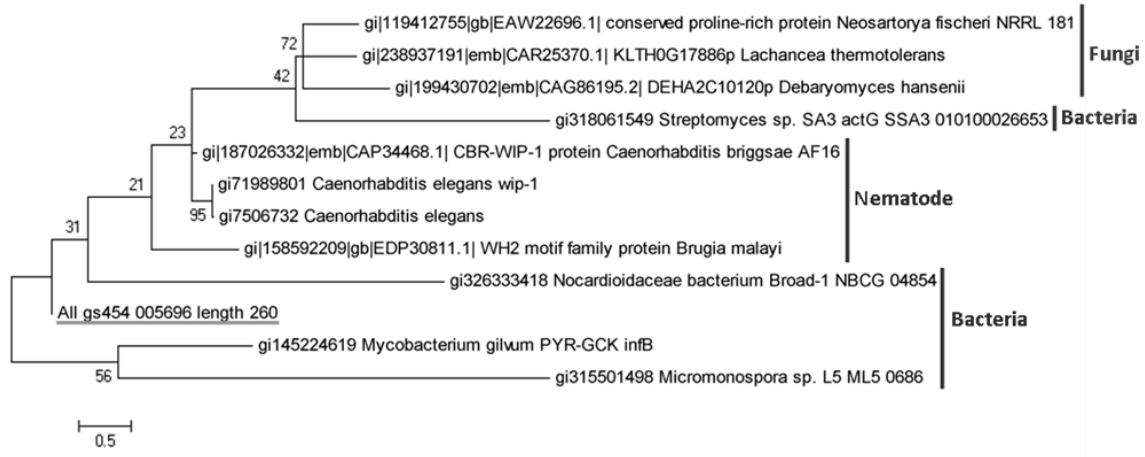


Figure 47 - Phylogenetic tree of the transcript All_gs454_005696.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within bacteria, therefore, valid.

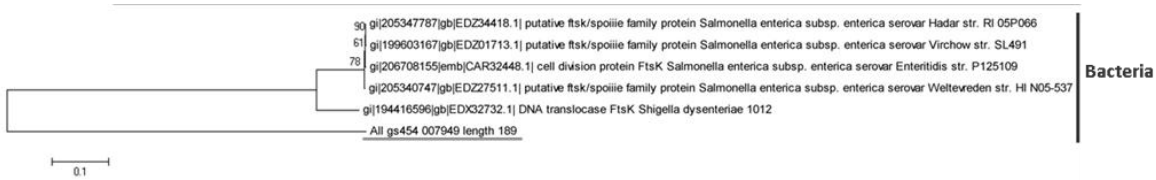


Figure 48 - Phylogenetic tree of the transcript All_gs454_007949.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within bacteria, therefore, valid.

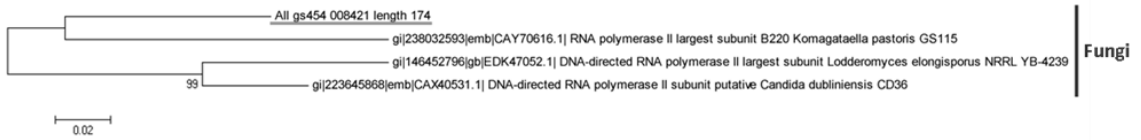


Figure 49 - Phylogenetic tree of the transcript All_gs454_008421.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within fungi, therefore, valid.

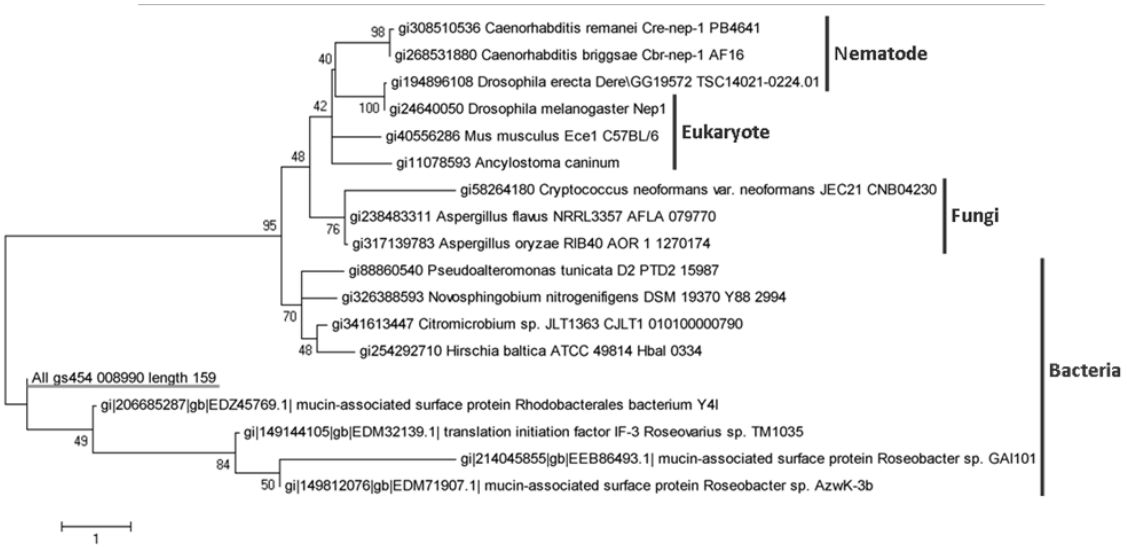


Figure 50 - Phylogenetic tree of the transcript All_gs454_008990.
The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within bacteria, therefore, valid.

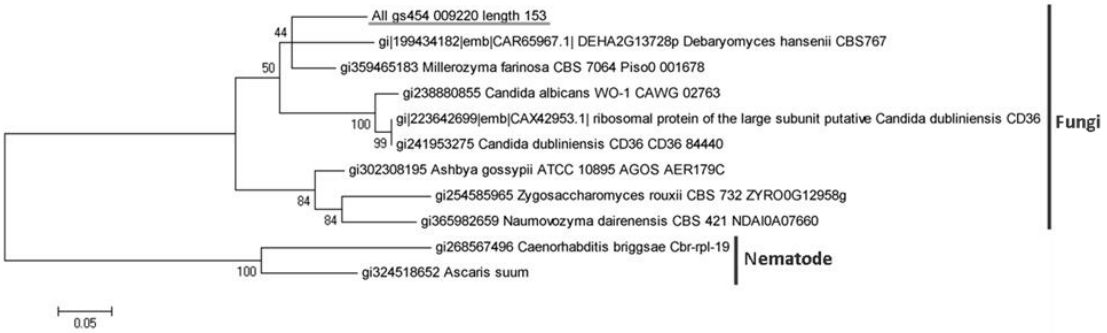


Figure 51 - Phylogenetic tree of the transcript All_gs454_009220.
The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within fungi, therefore, valid.

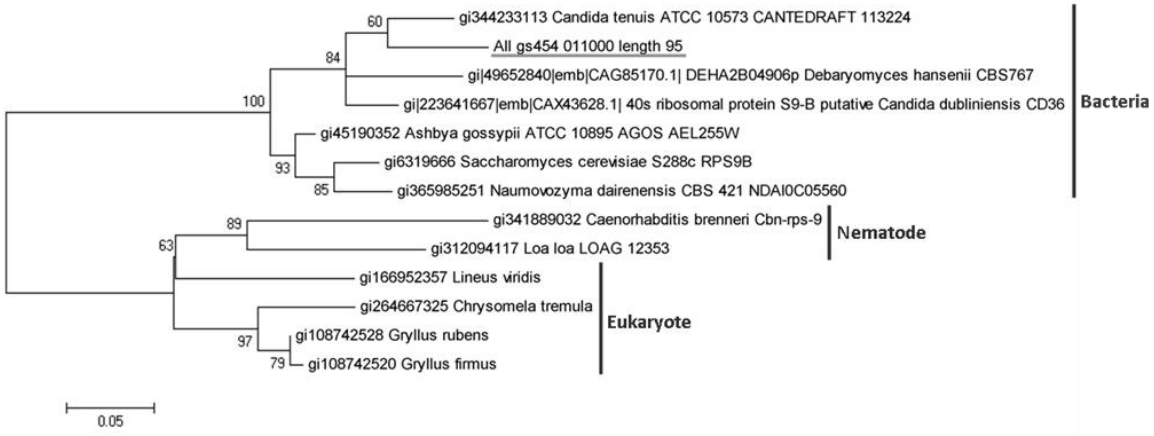


Figure 52 - Phylogenetic tree of the transcript All_gs454_011000.
The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within bacteria, therefore, valid.

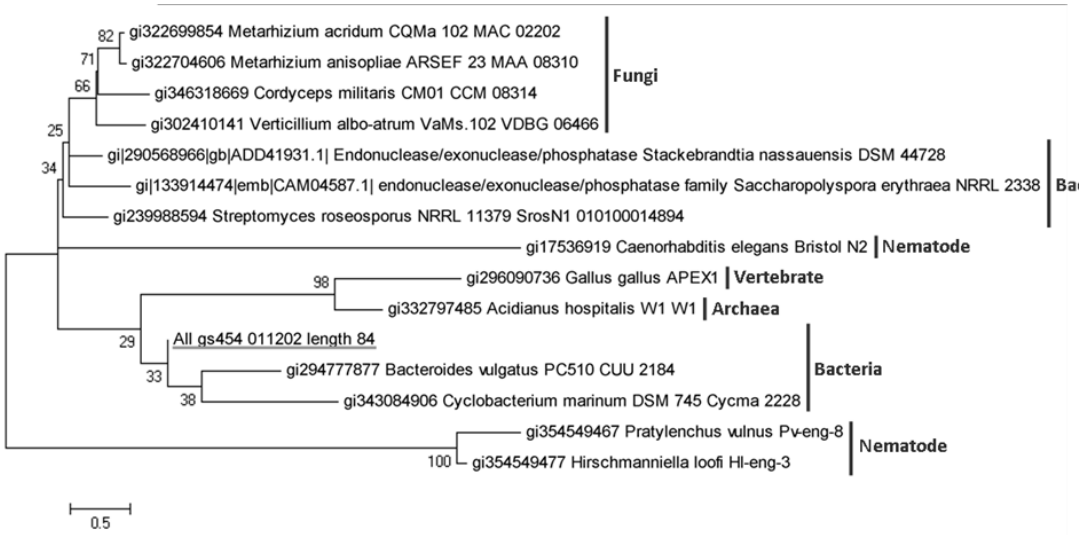


Figure 53 - Phylogenetic tree of the transcript All_gs454_011202.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within bacteria, therefore, valid.

Figure 54 - Phylogenetic tree of the transcript All_gs454_010630.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within bacteria, therefore, valid.

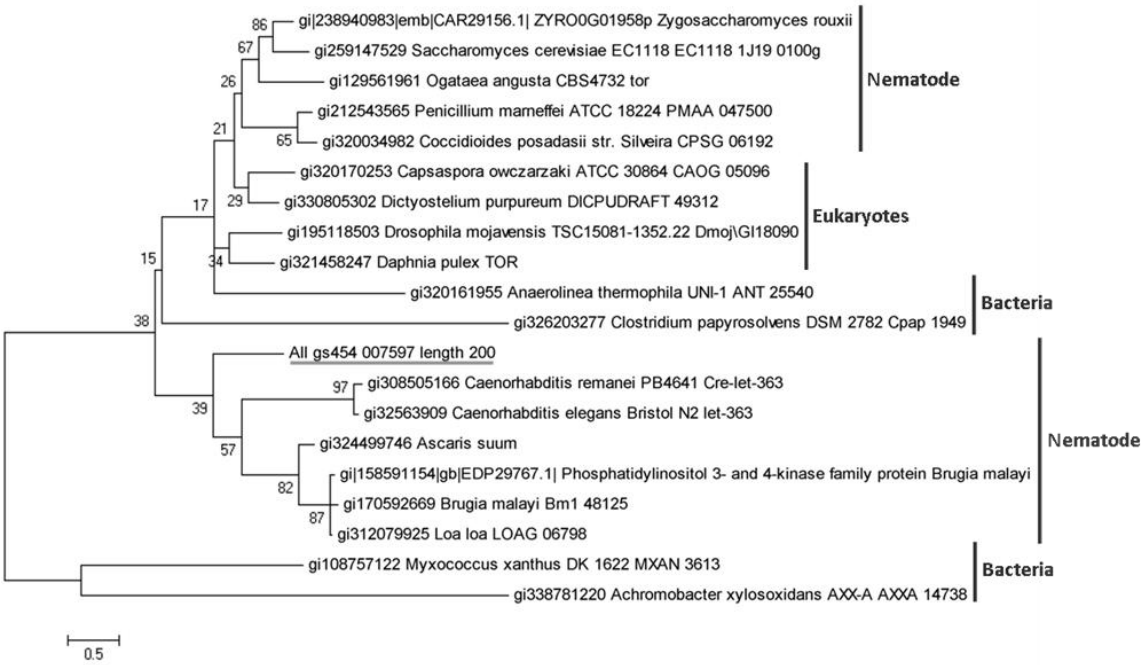


Figure 55 - Phylogenetic tree of the transcript All_gs454_007597.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes, therefore, invalid.

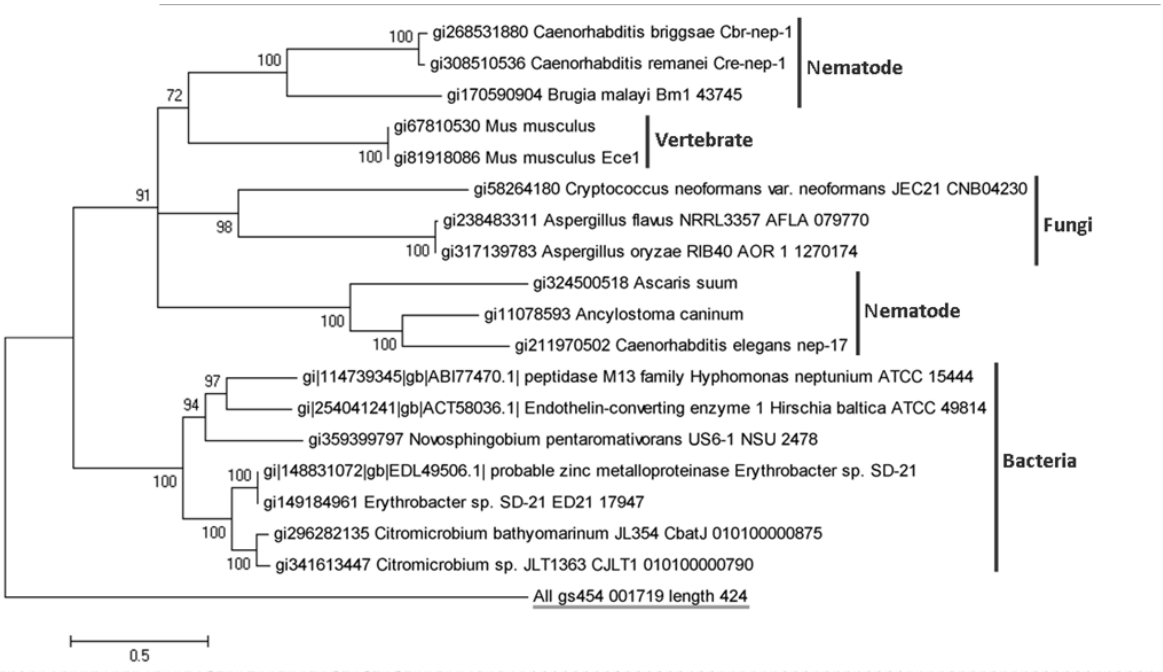


Figure 56 - Phylogenetic tree of the transcript All_gs454_001719.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate does not group with the presented taxons, therefore, invalid.

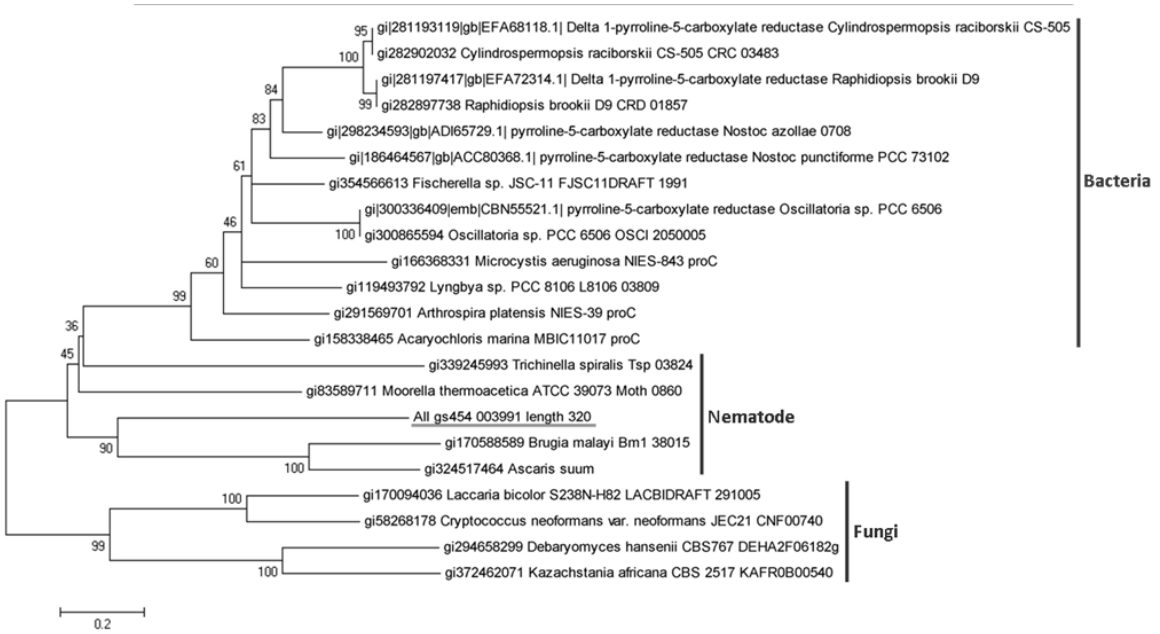


Figure 57 - Phylogenetic tree of the transcript All_gs454_003991.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes, therefore, invalid.

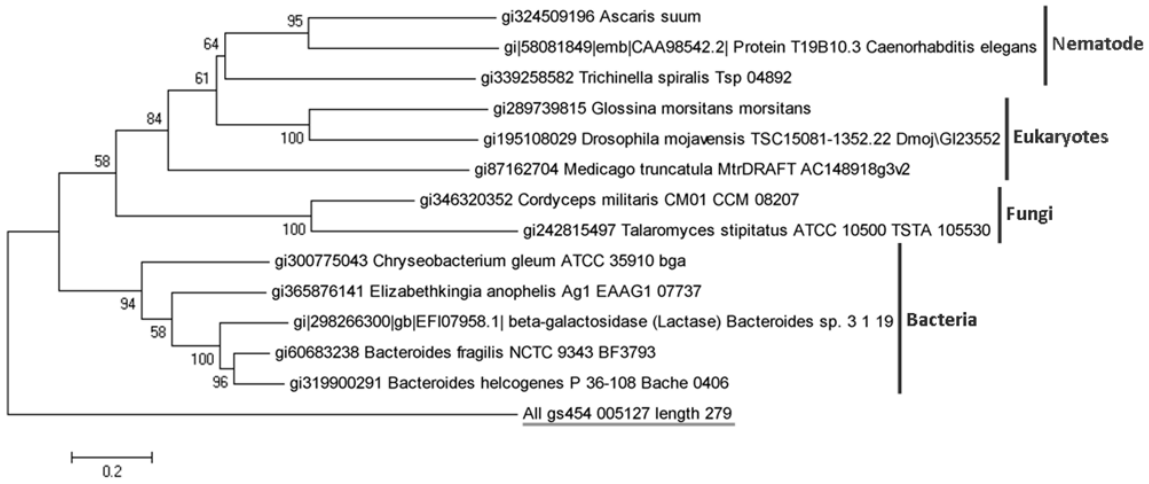


Figure 58 - Phylogenetic tree of the transcript All_gs454_005127.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate does not group with the presented taxons, therefore, invalid.

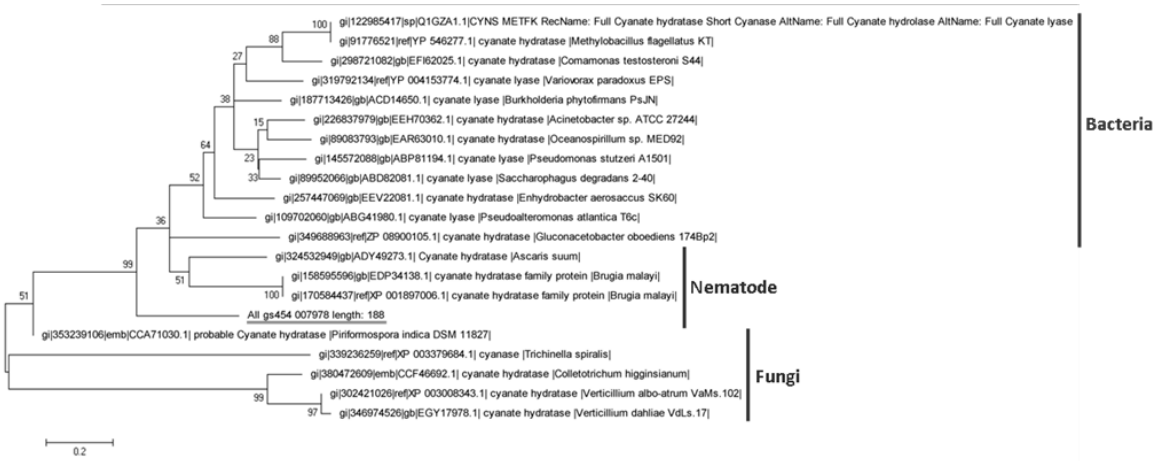


Figure 59 - Phylogenetic tree of the transcript All_gs454_007978.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes, therefore, invalid.

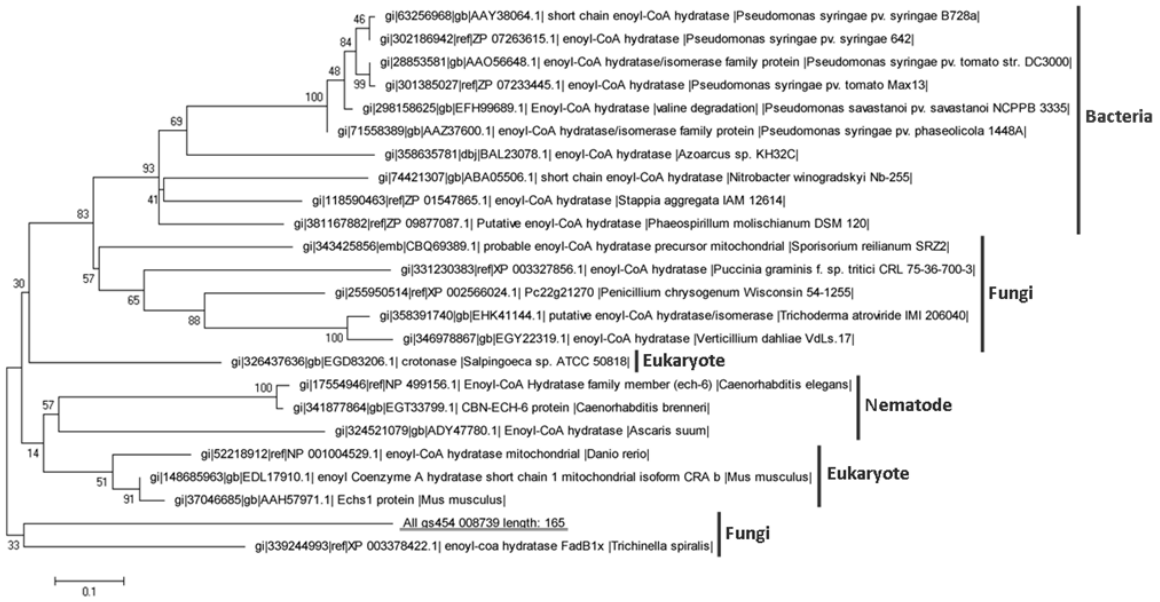
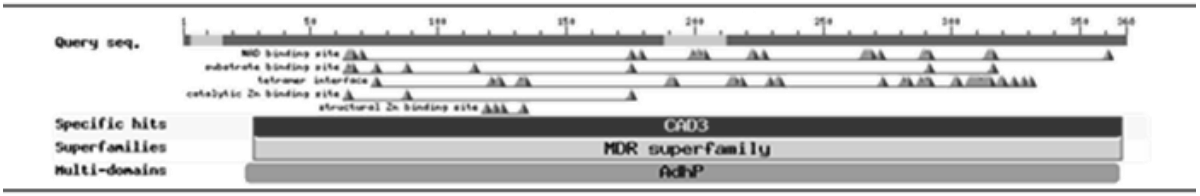


Figure 60 - Phylogenetic tree of the transcript All_gs454_008739.

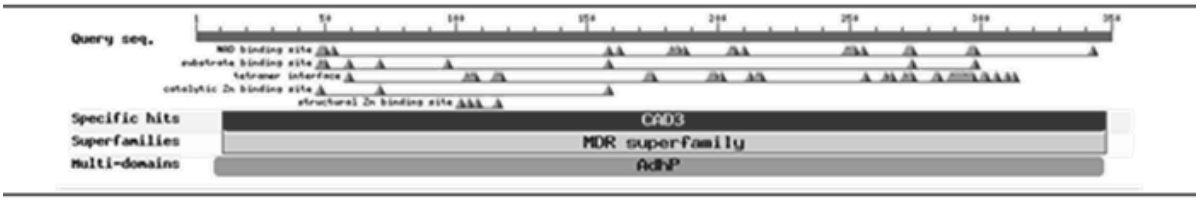
The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (100 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes, therefore, invalid.

Stage II

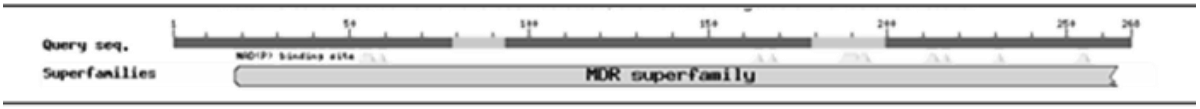
All_gs454_002803



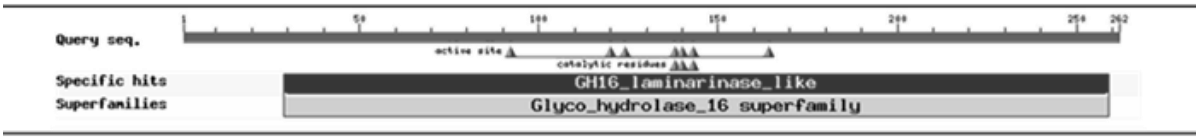
All_gs454_003220



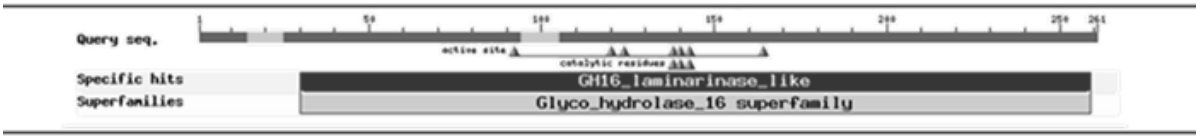
All_gs454_005442



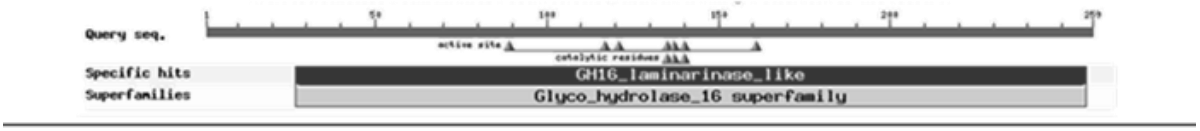
All_gs454_005644



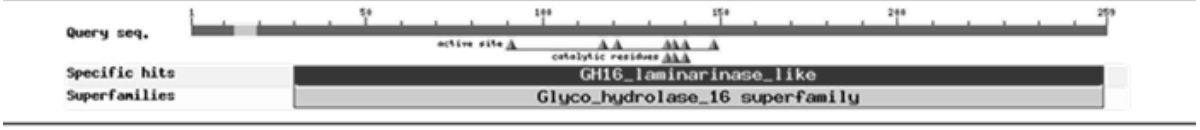
All_gs454_005676



All_gs454_005732



All_gs454_005733



All_gs454_005696

No conserved domain

All_gs454_007949

No conserved domain

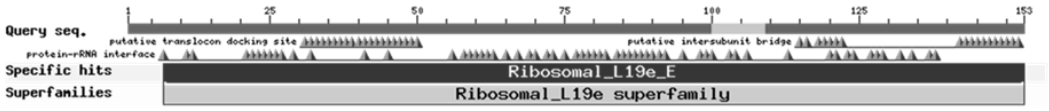
All_gs454_008421

No conserved domain

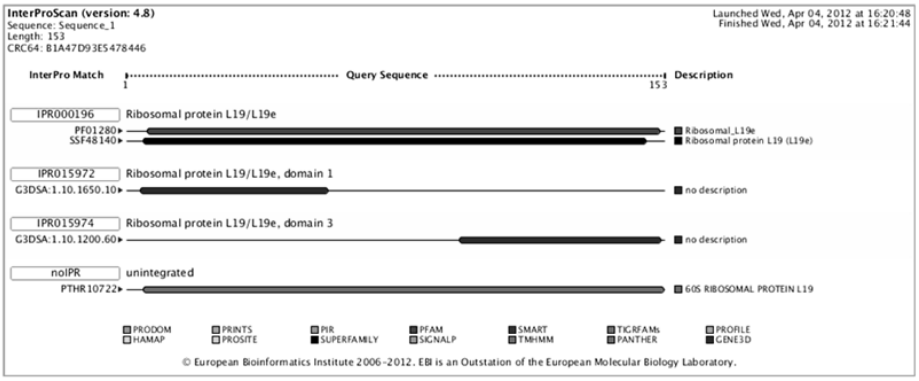
All_gs454_008990

No conserved domain

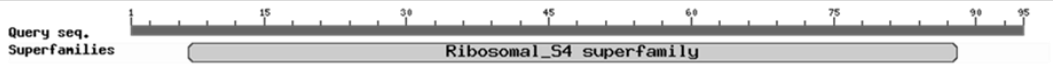
All_gs454_009220



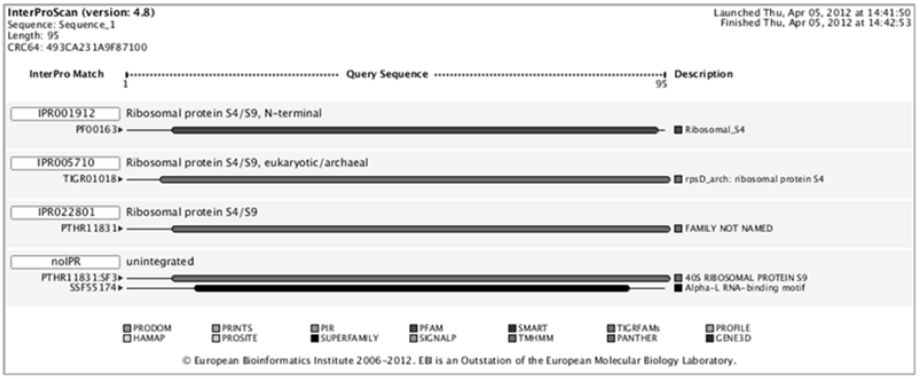
Ribosomal protein L19e, eukaryotic. L19e is found in the large ribosomal subunit of eukaryotes and archaea. L19e is distinct from the ribosomal subunit L19, which is found in prokaryotes. It consists of two small globular domains connected by an extended segment. It is located toward the surface of the large subunit, with one exposed end involved in forming the intersubunit bridge with the small subunit. The other exposed end is involved in forming the translocon binding site, along with L22, L23, L24, L29, and L31e subunits.



All_gs454_011000

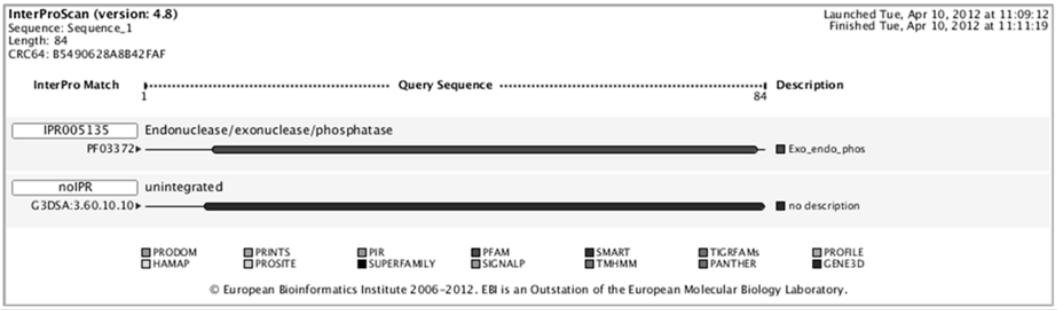


Ribosomal protein S4/S9 N-terminal domain; This family includes small ribosomal subunit S9 from prokaryotes and S16 from metazoans. This domain is predicted to bind to ribosomal RNA. This domain is composed of four helices in the known structure. However the domain is discontinuous in sequence and the alignment for this family contains only the first three helices.

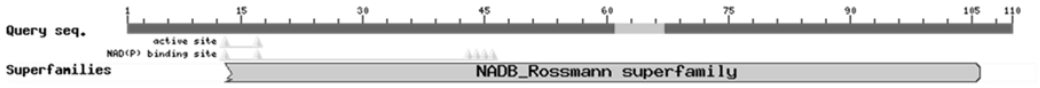


All_gs454_011202

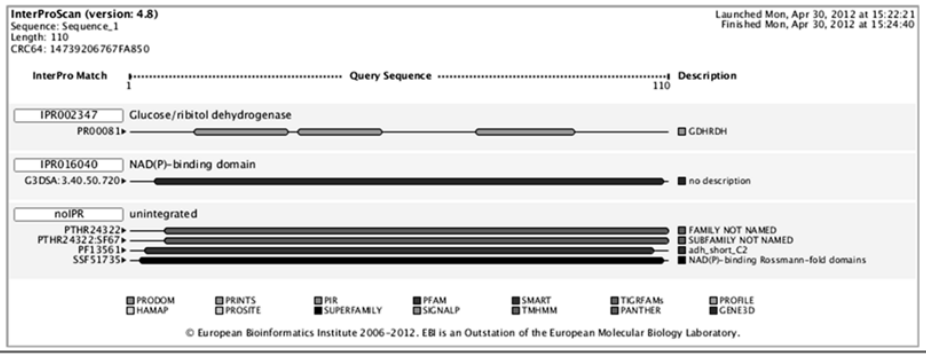
Exonuclease-Endonuclease-Phosphatase (EEP) domain superfamily; This large superfamily includes the catalytic domain (exonuclease/endonuclease/phosphatase or EEP domain) of a diverse set of proteins including the ExoIII family of apurinic/apyrimidinic (AP) endonucleases, inositol polyphosphate 5-phosphatases (INPP5), neutral sphingomyelinases (nSMases), deadenylases (such as the vertebrate circadian-clock regulated nocturnin), bacterial cytolethal distending toxin B (CdtB), deoxyribonuclease 1 (DNase1), the endonuclease domain of the non-LTR retrotransposon LINE-1, and related domains. These diverse enzymes share a common catalytic mechanism of cleaving phosphodiester bonds; their substrates range from nucleic acids to phospholipids and perhaps proteins.



All_gs454_010630



Rossmann-fold NAD(P)(+)-binding proteins; A large family of proteins that share a Rossmann-fold NAD(P)H/NAD(P)(+) binding (NADB) domain. The NADB domain is found in numerous dehydrogenases of metabolic pathways such as glycolysis, and many other redox enzymes. NAD binding involves numerous hydrogen-bonds and van der Waals contacts, in particular H-bonding of residues in a turn between the first strand and the subsequent helix of the Rossmann-fold topology. Characteristically, this turn exhibits a consensus binding pattern similar to GXGXXG, in which the first 2 glycines participate in NAD(P)-binding, and the third facilitates close packing of the helix to the beta-strand. Typically, proteins in this family contain a second domain in addition to the NADB domain, which is responsible for specifically binding a substrate and catalyzing a particular enzymatic reaction.



Stage III

The following table is the report of the BLASTn search performed for each candidate against the genera identified in the microbial community of *B. xylophilus.*

(×) indicates that the BLASTn search returned no significant results;

x) indicates that the BLASTn search returned a result and this is presented below the table.

	All_gs454_009220	All_gs454_011000	All_gs454_011202	All_gs454_010630
Abiotrophia	×	×	×	×
Acetivibrio	×	×	×	×
Achromobacter	×	×	×	×
Acidisoma	×	×	×	×
Acidovorax	×	×	×	×
Acinetobacter	×	×	×	×
Actinobacillus	×	×	×	×
Actinobaculum	×	×	×	×
Actinomyces	×	×	×	×
Actinoplanes	×	×	×	×
Aerococcus	×	×	×	×
Aeromonas	×	×	×	×
Aggregatibacter	×	×	×	×
Agreia	×	×	×	×
Agrobacterium	a)	×	×	×
Agromyces	×	×	×	×
Aquabacterium	×	×	×	×
Arenimonas	×	×	×	×
Arthrobacter	×	×	×	×
Asticcacaulis	×	×	×	×
Azohydromonas	×	×	×	×
Bacillus	×	×	×	×
Bacteriovorax	×	×	×	×
Bdellovibrio	×	×	×	×
Blastococcus	×	×	×	×
Bosea	×	×	×	×
botrytis cinerea	×	×	×	×
Brachybacterium	×	×	×	×
Bradyrhizobium	×	×	×	×
Brevundimonas	×	×	×	×
Brochothrix	×	×	×	×
Burkholderia	×	d)	×	×
Bursaphelenchus	×	×	×	×
Butyricicoccus	×	×	×	×

	All_gs454_009220	All_gs454_011000	All_gs454_011202	All_gs454_010630
Campylobacter	×	×	×	×
Cardiobacterium	×	×	×	×
Caulobacter	×	×	×	×
Cellulomonas	×	×	×	×
Chromohalobacter	×	×	×	×
Citrobacter	×	×	×	×
Corynebacterium	×	×	×	×
Curtobacterium	×	×	×	×
Curvibacter	×	×	×	×
Dechloromonas	×	×	×	×
Delftia	×	×	×	×
Dermacoccus	×	×	×	×
Desulfovibrio	×	×	×	×
Dolosigranulum	×	×	×	×
Enhydrobacter	×	×	×	×
Enterococcus	×	×	×	×
Fusobacterium	×	×	×	×
Gemmata	×	×	×	×
Geothrix	×	×	×	×
Granulicatella	×	×	×	×
Haemophilus	×	×	×	×
Halomonas	×	×	×	×
Herminiimonas	×	×	×	×
Hydrogenophilus	×	×	×	×
Hyphomicrobium	×	×	×	×
Janibacter	×	×	×	×
Janthinobacterium	×	×	×	×
Kingella	×	×	×	×
Kocuria	×	×	×	×
Kuraishia	×	×	×	×
Lactococcus	×	×	×	×
Legionella	×	×	×	×
Leifsonia	×	×	×	×
Leptotrichia	×	×	×	×
Leuconostoc	×	×	×	×
Luteimonas	×	×	×	×
Marinobacter	×	×	×	×
Marmoricola	×	×	×	×
Massilia	×	×	×	×
Methylibium	×	×	×	×
Methylobacterium	×	×	×	×

	All_gs454_009220	All_gs454_011000	All_gs454_011202	All_gs454_010630
Methyloversatilis	×	×	×	×
Meyerozyma	b)	e)	×	×
Microbacterium	×	×	×	×
Micrococcus	×	×	×	×
Microlunatus	×	×	×	×
Mycobacterium	×	×	×	×
Neisseria	×	×	×	×
Nevskia	×	×	×	×
Nitrosospira	×	×	×	×
Nitrospira	×	×	×	×
Nocardioides	×	×	×	×
Novosphingobium	×	×	×	×
Paenibacillus	×	×	×	×
Pantoea	×	×	×	×
Paracoccus	×	×	×	×
Pelomonas	×	×	×	×
Phascolarctobacterium	×	×	×	×
Phenylobacterium	×	×	×	×
Phycicoccus	×	×	×	×
Phyllobacterium	×	×	×	×
Pirellula	×	×	×	×
Planctomyces	×	×	×	×
Pleomorphomonas	×	×	×	×
Promicromonospora	×	×	×	×
Propionibacterium	×	×	×	×
Propionivibrio	×	×	×	×
Prosthecomicrobium	×	×	×	×
Pseudoclavibacter	×	×	×	×
Pseudolabrys	×	×	×	×
Pseudomonas	×	×	×	×
Psychrobacter	×	×	×	×
Quadrisphaera	×	×	×	×
Rheinheimera	×	×	×	×
Rhizobium	×	f)	×	×
Rhodococcus	×	×	×	×
Roseomonas	×	×	×	×
Rothia	×	×	×	×
Rudaea	×	×	×	×
Selenomonas	×	×	×	×
Serratia	×	×	×	×
Silanimonas	×	×	×	×

	All_gs454_009220	All_gs454_011000	All_gs454_011202	All_gs454_010630
Skermanella	×	×	×	×
Sphingobium	×	×	×	×
Sphingomonas	×	×	×	×
Sphingopyxis	×	×	×	×
Sporichthya	×	×	×	×
Staphylococcus	×	g)	×	×
Stenotrophomonas	×	×	×	×
Streptococcus	×	×	×	×
Streptomyces	c)	h)	×	×
Tepidimonas	×	×	×	×
Terrabacter	×	×	×	×
Tessaracoccus	×	×	×	×
Thauera	×	×	×	×
Thermus	×	×	×	×
Thiohalophilus	×	×	×	×
Turicella	×	×	×	×
Uliginosibacterium	×	×	×	×
Variovorax	×	×	×	×
Veillonella	×	×	×	×
Verrucomicrobium	×	×	×	×
Vibrio	×	×	×	×
Weissella	×	×	×	×
Yaniella	×	×	×	×
Zoogloea	×	×	×	×

```
>emb|CR382139.2| Debaryomyces hansenii CBS767 chromosome G complete sequence
Length=2051050

 Features in this part of subject sequence:
   DEHA2G13728p

 Score =  499 bits (270),  Expect = 7e-140
 Identities = 389/448 (87%), Gaps = 2/448 (0%)
 Strand=Plus/Minus

Query  12       GGCTAATCTTAGAACTCAGAAAAGACTTGCTGCCAGTGTTGTGGGTGTTGGTAAGAGAAA  71
                 ||||||||||||||||| || |||||||| || |||||||| ||||||||||||||||||
Sbjct  1136107  GGCTAATCTTAGAACTCAAAAGAGACTTGCAGCTAGTGTTGTTGGTGTTGGTAAGAGAAA  1136048

Query  72       GATTTGGATGGATCCAAATGAGACCAACGAAATTGCCAACGCCAACTCACGTCAAGCCAT  131
                 ||| ||| | ||||| || ||  |||  |||||  |||| |||||||| |||||||| ||
Sbjct  1136047  GATCTGGTTAGATCCTAACGAAGCCACTGAAATCTCCAATGCCAACTCTCGTCAAGCTAT  1135988

Query  132      CAGAAAATTATACAGAAATGGTACTATTGTGAAGAAGCCTAATGTGATCCACTCCAGATC  191
                 |||||| |||||||||| ||||| ||||| ||||||||   |||  ||||||| ||| |
Sbjct  1135987  CAGAAAGTTATACAGAAACGGTACCATTGTCAAGAAGCCAGTTGTTGTCCACTCTAGAGC  1135928

Query  192      CAGAGCTAGAGCTTTGGCTGAATCCAAGAGAGCTGGTAGACACACCGGTTACGGTAAGAG  251
                 |||||| |||||||||   ||||||||||||| ||||||||||  ||||||||||||||
Sbjct  1135927  CAGAGCAAGAGCTTTGTTAGAATCCAAGAGAGCCGGTAGACACATGGGTTACGGTAAGAG  1135868

Query  252      AAAGGGTACCAAGGATGCCCGTATGTCTTCTCAAGTTGTGTGGATGAGAAGATTGAGAGT  311
                 |||||||||||||| || ||||||||| |||||||| | ||||||||||||||||||||
Sbjct  1135867  AAAGGGTACCAAGGACGCTCGTATGTCTGCTCAAGTTTTATGGATGAGAAGATTGAGAGT  1135808

Query  312      GTTGAGAAGATTGTTGGCCAAGTACAGAGATGCTGGTAAGATTGACAAGCACTTGTACCA  371
                  || |||||||| ||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  1135807  CTTAAGAAGATTATTGGCTAAGTACAGAGATGCTGGTAAGATTGACAAGCACTTGTACCA  1135748

Query  372      CACCTTGTACAAGTCTGCCAAGGGTAACACTTTCAAGCACAAGAGATCATTAGTCGAGCA  431
                 | | || || || ||||||||||||||| ||||||||||||||||||| |||||||| ||
Sbjct  1135747  CTCATTATATAAATCTGCCAAGGGTAACGCTTTCAAGCACAAGAGATCTTTAGTCGAACA  1135688

Query  432      CATCAT-CACCGCCAAGGCTGAGGCTTT  458
                 |||||| ||  |||||||| || |||||
Sbjct  1135687  CATCATTCAA-GCCAAGGCCGAAGCTTT  1135661
```

Figure 61 - a) Alignment of the transcript All_gs454_009220 with the Agrobacterium genus.

```
>ref|XM_001483596.1| G Meyerozyma guilliermondii ATCC 6260 hypothetical protein (PGUG_04375)
partial mRNA
Length=498

 GENE ID: 5125558 PGUG_04375 | hypothetical protein
[Pichia guilliermondii ATCC 6260]

 Score =  704 bits (381),  Expect = 0.0
 Identities = 381/381 (100%), Gaps = 0/381 (0%)
 Strand=Plus/Plus

Query  79   ATGGATCCAAATGAGACCAACGAAATTGCCAACGCCAACTCACGTCAAGCCATCAGAAAA  138
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    ATGGATCCAAATGAGACCAACGAAATTGCCAACGCCAACTCACGTCAAGCCATCAGAAAA  60

Query  139  TTATACAGAAATGGTACTATTGTGAAGAAGCCTAATGTGATCCACTCCAGATCCAGAGCT  198
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TTATACAGAAATGGTACTATTGTGAAGAAGCCTAATGTGATCCACTCCAGATCCAGAGCT  120

Query  199  AGAGCTTTGGCTGAATCCAAGAGAGCTGGTAGACACACCGGTTACGGTAAGAGAAAGGGT  258
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  AGAGCTTTGGCTGAATCCAAGAGAGCTGGTAGACACACCGGTTACGGTAAGAGAAAGGGT  180

Query  259  ACCAAGGATGCCCGTATGTCTTCTCAAGTTGTGTGGATGAGAAGATTGAGAGTGTTGAGA  318
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  ACCAAGGATGCCCGTATGTCTTCTCAAGTTGTGTGGATGAGAAGATTGAGAGTGTTGAGA  240

Query  319  AGATTGTTGGCCAAGTACAGAGATGCTGGTAAGATTGACAAGCACTTGTACCACACCTTG  378
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  AGATTGTTGGCCAAGTACAGAGATGCTGGTAAGATTGACAAGCACTTGTACCACACCTTG  300

Query  379  TACAAGTCTGCCAAGGGTAACACTTTCAAGCACAAGAGATCATTAGTCGAGCACATCATC  438
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  TACAAGTCTGCCAAGGGTAACACTTTCAAGCACAAGAGATCATTAGTCGAGCACATCATC  360

Query  439  ACCGCCAAGGCTGAGGCTTTG  459
            |||||||||||||||||||||
Sbjct  361  ACCGCCAAGGCTGAGGCTTTG  381
```

Figure 62 - b) Alignment of the transcript All_gs454_009220 with the Meyerozyma genus.

```
>emb|CU928179.1| D Zygosaccharomyces rouxii strain CBS732 chromosome G complete
sequence
Length=1865392

 Features in this part of subject sequence:
   ZYRO0G12958p

 Score =  241 bits (130),  Expect = 1e-61
 Identities = 345/450 (77%), Gaps = 10/450 (2%)
 Strand=Plus/Minus

Query  12       GGCTAATCTT-AGAACTCAGAAAAGACTTGCTGCCAGTGTTGTGGGTGTTGGTAAGAGAA  70
                ||||||  |||  | |||||  || |||||  || ||    |||  | ||||  ||||||||||
Sbjct  1034679  GGCTAA-CTTGCGTACTCAAAAGAGACTCGCCGCTTCTGTCATCGGTGCTGGTAAGAGAA  1034621

Query  71       AGATTTGGATGGATCCAAATGAGACCAACGAAATTGC-CAACGCCAACTCACGTCAAGCC  129
                || | ||| | || || || || || |||   ||  | || ||| || |||||  |   | |||
Sbjct  1034620  AGGTCTGGTTAGACCCTAACGAAACCTCTGAGTTGGCTCAA-GCTAACTCCAGAAACGCC  1034562

Query  130      ATCAGAAAATT-ATACAGAAATGGTACTATTGTGAAGAA-GCCTAATGTGATCCACTCCA  187
                ||||||||||  |  || |||||||| |||||||||| |||   | || | |||||||
Sbjct  1034561  ATCAGAAAATTGGTTAAG-AATGGTACCATTGTGAAGAAGGCCGTCT-TGGTGCACTCCA  1034504

Query  188      GATCCAGAGCTAGAGCTTTGGCTGAATCCAAGAGAGCTGGTAGACACACCGGTTACGGTA  247
                |||| ||| |||||   |  ||||  || ||||||  |||| | ||||||||||||||||
Sbjct  1034503  GATCTAGAACTAGAAGATACGCTGCTTCTAAGAGAAGTGGTCGTCACACCGGTTACGGTA  1034444

Query  248      AGAGAAAGGGTACCAAGGATGCCCGTATGTCTTCTCAAGTTGTGTGGATGAGAAGATTGA  307
                |||||||||||||||||| || ||| || | ||| | || || ||||| |||||||||
Sbjct  1034443  AGAGAAAGGGTACCAAGGAAGCTCGTTTGCCCTCTAAGGTCGTCTGGATCAGAAGATTGC  1034384

Query  308      GAGTGTTGAGAAGATTGTTGGCCAAGTACAGAGATGCTGGTAAGATTGACAAGCACTTGT  367
                | || |||||||||| | ||| | || ||| | || ||||||||||| ||||  |||||||
Sbjct  1034383  GTGTTTTGAGAAGACTATTGTCAAAATACCGTGACGCTGGTAAGATCGACAGACACTTGT  1034324

Query  368      ACCACACCTTGTACAAGTCTGCCAAGGGTAACACTTTCAAGCACAAGAGATCATTAGTCG  427
                |||||  ||||||| |    |||||||||||||| |||||||||||||| | || ||||
Sbjct  1034323  ACCACGTCTTGTACCACGAAGCCAAGGGTAACACCTTCAAGCACAAGAGAGCTTTGGTCG  1034264

Query  428      AGCACATCAT-CACCGCCAAGGCTGAGGCT  456
                | ||||||||  ||  || |||||||| |||
Sbjct  1034263  AACACATCATTCAA-GCTAAGGCTGATGCT  1034235
```

Figure 63 - c) Alignment of the transcript All_gs454_009220 with the Streptomyces genus.

```
>emb|CR382134.2| D Debaryomyces hansenii CBS767 chromosome B complete sequence
Length=1344482

 Features in this part of subject sequence:
   DEHA2B04906p

 Score =  285 bits (154),  Expect = 1e-75
 Identities = 232/271 (86%), Gaps = 0/271 (0%)
 Strand=Plus/Plus

Query  14      GTGCCCCAAGAACTTACTCTAAGACCTACTCTGTGCCAAAGCAGCCATATGAGTCTGCCC  73
               ||||||||||||||||||| ||||| |||||||| |||||||| ||||  |||||||| |
Sbjct  399864  GTGCCCCAAGAACTTACTCCAAGACTTACTCTGTCCCAAAGCAACCATTCGAGTCTGCTC  399923

Query  74      GTTTGGACAGTGAATTGAAGTTGGCTGGTGAACTTGGATTGAAGAACAAGAGAGAAATCT  133
               |||| |||  |||||| |||||| ||||||||    || || ||||||||||||||||||
Sbjct  399924  GTTTAGACGCTGAATTAAAGTTAGCTGGTGAATACGGTTTAAAGAACAAGAGAGAAATCT  399983

Query  134     ACAGAATTCAATTCCAATTGTCCAAGATCAGAAGAGCCGCTCGTGACTTGTTGACCAGAG  193
               ||||||||   |||||||||||| ||||| |||||||| |||||||| || ||||||||||
Sbjct  399984  ACAGAATTGGTTTCCAATTGTCTAAGATTAGAAGAGCTGCTCGTGATTTATTGACCAGAG  400043

Query  194     ACGAGAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTGATCAGAAGATTGGTGAAGG  253
               | || ||||||||||||||||||||||||||||||||||| || |||||||| ||||
Sbjct  400044  ATGAAAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTAATTAGAAGATTAGTGAGAA  400103

Query  254     TTGGTGTGTTGCCAGAGGACAAGATGAAGTT  284
                ||||||  ||  |  || ||||||||||||||
Sbjct  400104  CTGGTGTCTTATCCGAAGACAAGATGAAGTT  400134
```

Figure 64 - d) Alignment of the transcript All_gs454_011000 with the Burkholderia genus.

```
>ref|XM_001484261.1| G Meyerozyma guilliermondii ATCC 6260 40S ribosomal protein S9-B
(PGUG_03692) partial mRNA
Length=576

 GENE ID: 5126298 PGUG_03692 | 40S ribosomal protein S9-B
[Pichia guilliermondii ATCC 6260]

 Score =  497 bits (269),  Expect = 3e-141
 Identities = 271/272 (99%), Gaps = 0/272 (0%)
 Strand=Plus/Minus

Query  14   GTGCCCCAAGAACTTACTCTAAGACCTACTCTGTGCCAAAGCAGCCATATGAGTCTGCCC  73
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  572  GTGCCCCAAGAACTTACTCTAAGACCTACTCTGTGCCAAAGCAGCCATATGAGTCTGCCC  513

Query  74   GTTTGGACAGTGAATTGAAGTTGGCTGGTGAACTTGGATTGAAGAACAAGAGAGAAATCT  133
            ||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  512  GTTTGGACAGTGAATTGAAGTTGGCTGGTGAATTTGGATTGAAGAACAAGAGAGAAATCT  453

Query  134  ACAGAATTCAATTCCAATTGTCCAAGATCAGAAGAGCCGCTCGTGACTTGTTGACCAGAG  193
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  452  ACAGAATTCAATTCCAATTGTCCAAGATCAGAAGAGCCGCTCGTGACTTGTTGACCAGAG  393

Query  194  ACGAGAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTGATCAGAAGATTGGTGAAGG  253
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  392  ACGAGAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTGATCAGAAGATTGGTGAAGG  333

Query  254  TTGGTGTGTTGCCAGAGGACAAGATGAAGTTG  285
            ||||||||||||||||||||||||||||||||
Sbjct  332  TTGGTGTGTTGCCAGAGGACAAGATGAAGTTG  301
```

Figure 65 - e) Alignment of the transcript All_gs454_011000 with the Meyerozyma genus.

```
>emb|CR380959.2| D Candida glabrata strain CBS138 chromosome M complete sequence
Length=1402899

 Features in this part of subject sequence:
   unnamed protein product

 Score =  300 bits (162),  Expect = 4e-80
 Identities = 236/272 (87%), Gaps = 4/272 (1%)
 Strand=Plus/Plus

Query  16      GCCCCAAGAACTTACTCTAAGACCTACTCTGTGCCAAAGCAG-CCATATGAGTCTGCCCG  74
               ||||||||||||||||||||||||||||||    |||||| || || || || ||| | ||
Sbjct  673626  GCCCCAAGAACTTACTCTAAGACCTACTCTACCCCAAAG-AGACCTTACGAATCTTCTCG  673684

Query  75      TTTGGACAGTGAATTGAAGTTGGCTGGTGAACTTGGATTGAAGAACAAGAGAGAAATCTA  134
               |||||||  |||||||||| | ||||||||| | || ||||||||||||||||||||| ||
Sbjct  673685  TTTGGACGCTGAATTGAAGCTTGCTGGTGAATTCGGTTTGAAGAACAAGAGAGAAATTTA  673744

Query  135     CAGAATTCAATTCCAATTGTCCAAGATCAGAAGAGCCGCTCGTGA-CTTGTTGACCAGAG  193
               |||||||   |||||||||||||||||||||||| ||| | || ||| ||||||||||
Sbjct  673745  CAGAATTTCTTTCCAATTGTCCAAGATCAGAAGAGCTGCTAGAGATCTT-TTGACCAGAG  673803

Query  194     ACGAGAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTGATCAGAAGATTGGTGAAGG  253
               |||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  673804  ACGAAAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTGATCAGAAGATTGGTGAGAA  673863

Query  254     TTGGTGTGTTGCCAGAGGACAAGATGAAGTTG  285
               | ||||| ||| | || ||||||| |||||||
Sbjct  673864  TCGGTGTCTTGTCCGAAGACAAGAAGAAGTTG  673895
```

Figure 66 - f) Alignment of the transcript All_gs454_011000 with the Rhizobium genus.

```
>emb|FO082059.1| D Millerozyma farinosa CBS 7064 chromosome A complete sequence
Length=1055225

 Features in this part of subject sequence:
   Piso0_000336

 Score =  276 bits (149),  Expect = 9e-73
 Identities = 232/273 (85%), Gaps = 2/273 (1%)
 Strand=Plus/Minus

Query  14      GTGCCCCAAGAACTTACTCTAAGACCTACTCTGTGCCAAAGCAG-CCATATGAGTCTGCC  72
               |||||||||||||||||||||||||| ||| |||| |||||| || || |  ||||| ||
Sbjct  580542  GTGCCCCAAGAACTTACTCTAAGACTTACACTGTCCCAAAG-AGACCTTTCGAGTCCGCT  580484

Query  73      CGTTTGGACAGTGAATTGAAGTTGGCTGGTGAACTTGGATTGAAGAACAAGAGAGAAATC  132
               ||||||||    ||||||||||||||||||||    || ||||||||||||||| ||||||
Sbjct  580483  CGTTTGGATGCCGAATTGAAGTTGGCTGGTGAATACGGTTTGAAGAACAAGAGGGAAATC  580424

Query  133     TACAGAATTCAATTCCAATTGTCCAAGATCAGAAGAGCCGCTCGTGACTTGTTGACCAGA  192
               |||||||||   | ||||||||| ||||| |||||||||||| ||||||||||||| |||
Sbjct  580423  TACAGAATTGGTTACCAATTGTCTAAGATTAGAAGAGCCGCCCGTGACTTGTTGACTAGA  580364

Query  193     GACGAGAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTGATCAGAAGATTGGTGAAG  252
               || |||||||||||||||||||||||||||||||| || |||||||||||||||||| |
Sbjct  580363  GATGAGAAGGACCCAAAGAGATTGTTCGAAGGTAACGCCTTGATCAGAAGATTGGTCAGA  580304

Query  253     GTTGGTGTGTTGCCAGAGGACAAGATGAAGTTG  285
               | ||||||| ||| | || ||||  ||||| |||
Sbjct  580303  ATTGGTGTCTTGTCTGAAGACAGAATGAAATTG  580271
```

Figure 67 - g) Alignment of the transcript All_gs454_011000 with the Staphylococcus genus.

```
>emb|CR382134.2| Debaryomyces hansenii CBS767 chromosome B complete sequence
Length=1344482

 Features in this part of subject sequence:
   DEHA2B04906p

 Score =  285 bits (154),  Expect = 3e-75
 Identities = 232/271 (86%), Gaps = 0/271 (0%)
 Strand=Plus/Plus

Query  14      GTGCCCCAAGAACTTACTCTAAGACCTACTCTGTGCCAAAGCAGCCATATGAGTCTGCCC  73
               ||||||||||||||||||| ||||| |||||||| |||||||| ||||  |||||||| |
Sbjct  399864  GTGCCCCAAGAACTTACTCCAAGACTTACTCTGTCCCAAAGCAACCATTCGAGTCTGCTC  399923

Query  74      GTTTGGACAGTGAATTGAAGTTGGCTGGTGAACTTGGATTGAAGAACAAGAGAGAAATCT  133
               |||| |||  |||||| ||||| |||||||||    || || ||||||||||||||||||
Sbjct  399924  GTTTAGACGCTGAATTAAAGTTAGCTGGTGAATACGGTTTAAAGAACAAGAGAGAAATCT  399983

Query  134     ACAGAATTCAATTCCAATTGTCCAAGATCAGAAGAGCCGCTCGTGACTTGTTGACCAGAG  193
               ||||||||   |||||||||||| ||||| |||||||| |||||||| || ||||||||||
Sbjct  399984  ACAGAATTGGTTTCCAATTGTCTAAGATTAGAAGAGCTGCTCGTGATTTATTGACCAGAG  400043

Query  194     ACGAGAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTGATCAGAAGATTGGTGAAGG  253
               | || ||||||||||||||||||||||||||||||||||| || |||||||| ||||
Sbjct  400044  ATGAAAAGGACCCAAAGAGATTGTTCGAAGGTAATGCTTTAATTAGAAGATTAGTGAGAA  400103

Query  254     TTGGTGTGTTGCCAGAGGACAAGATGAAGTT  284
               |||||| ||  | || ||||||||||||||
Sbjct  400104  CTGGTGTCTTATCCGAAGACAAGATGAAGTT  400134
```

Figure 68 - h) Alignment of the transcript All_gs454_011000 with the Streptomyces genus.

Stage IV

Figure 69 - Phylogenetic tree of the transcript All_gs454_011202.

The transcript is underlined; Phylogenetic tree generated by maximum likelihood analysis in MEGA 5, using the bootstrap method phylogeny test (1 000 bootstraps) and the JTT-F substitution model; the tree indicates that the candidate groups within the nematodes, therefore, invalid.

HGT gene report

All_gs454_010630	Short chain dehydrogenase/reductase

AA

>All_gs454_010630 length: 110

GVARKIQPAIPMYAISKAAVEHYARHAAFEYAADGIRVNCVAPGIIESAFHERGANPHPDAAKAAAAMVPLKRMGKPEEAAAMMVFVASDKASYITGEFFGVN
GGLMIKP

ORF

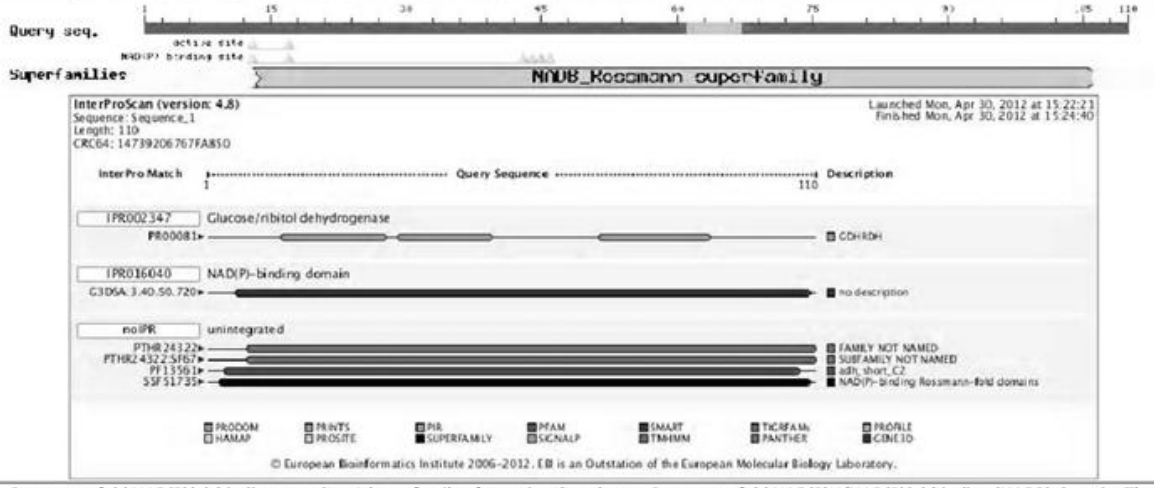
>isotig14747 All_gs454_010630 length: 330

GGGGTGGCCCGCAAAATTCAaCCCGCAATTCCCATGTACGCGATCAGCAaAGCCGCCGTAGAGCATTACGCTCGGCACGCCGCCTTCGAATACGCcGCGgat
GGGATTCGCGTGAACTGCGTTGCGCCCGGAATCATCgaGAGTGCGTTCCAtGAACGTGGGgCCAATCCGCAtCcGGACGCCGCCAAAGCGGCGGCCGCGAT
GGTGCCGCTGAAGAGGATGGGAAAGCCGGAAGAAGCGGCGGCTATGATGGTGTTTGTCGCCAGTGACAaGGCCAGCTACaTCACTGGGGAATtcTtCGGA
GTCAACGGTGGACTCATGATAAAACCA

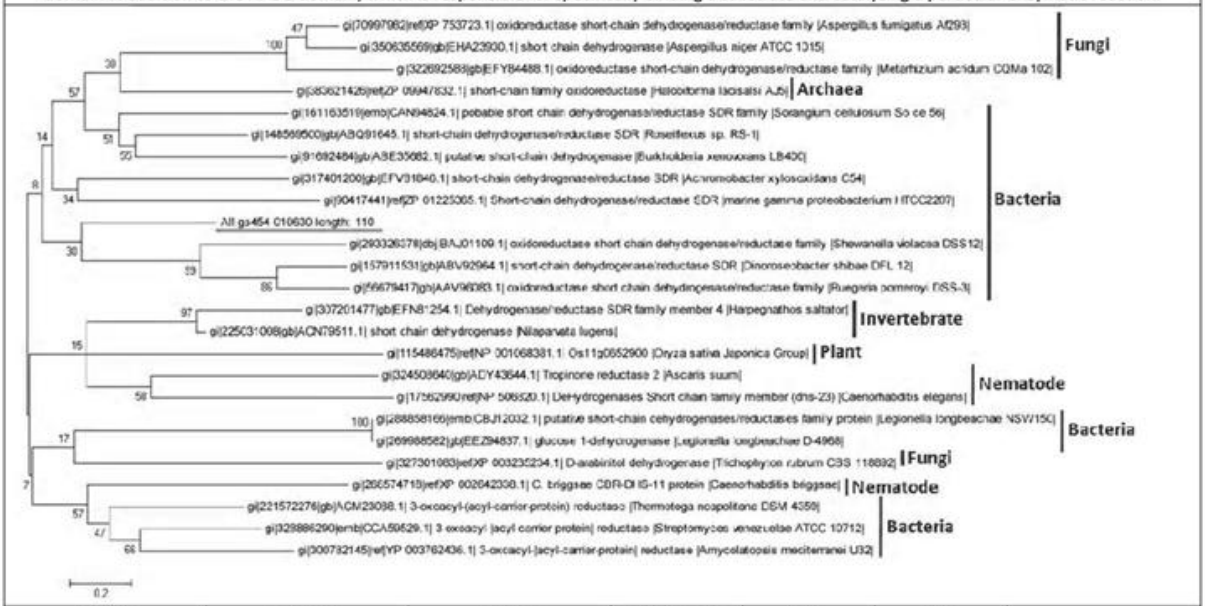
Contig

>isotig14747 length: 371

GGGGTGGCCCGCAAAATTCAaCCCGCAATTCCCATGTACGCGATCAGCAaAGCCGCCGTAGAGCATTACGCTCGGCACGCCGCCTTCGAATACGCcGCGgat
GGGATTCGCGTGAACTGCGTTGCGCCCGGAATCATCgaGAGTGCGTTCCAtGAACGTGGGgCCAATCCGCAtCcGGACGCCGCCAAAGCGGCGGCCGCGAT
GGTGCCGCTGAAGAGGATGGGAAAGCCGGAAGAAGCGGCGGCTATGATGGTGTTTGTCGCCAGTGACAaGGCCAGCTACaTCACTGGGGAATtcTtCGGA
GTCAACGGTGGACTCATGATAAAACCATAATTTAATTACcTTATTCTtcATTGCAAATATTACttttt



Rossmann-fold NAD(P)(+)-binding proteins; A large family of proteins that share a Rossmann-fold NAD(P)H/NAD(P)(+) binding (NADB) domain. The NADB domain is found in numerous dehydrogenases of metabolic pathways such as glycolysis, and many other redox enzymes. NAD binding involves numerous hydrogen-bonds and van der Waals contacts, in particular H-bonding of residues in a turn between the first strand and the subsequent helix of the Rossmann-fold topology. Characteristically, this turn exhibits a consensus binding pattern similar to GXGXXG, in which the first 2 glycines participate in NAD(P)-binding, and the third facilitates close packing of the helix to the beta-strand. Typically, proteins in this family contain a second domain in addition to the NADB domain, which is responsible for specifically binding a substrate and catalyzing a particular enzymatic reaction.



BLAST		Metagen		Genome		Allignment	CDD	Tree	
✓		✓		✓					

Appendix IV –Blast2GO®

In this section we present the result of the annotation performed in the Blast2GO® program.

nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
1	snap_masked-CADV010...	fabg_thema ame: full=3-oxoacyl- reductase ame: full=3-ketoacyl-acyl carrier protein reductase ame: full=beta-ketoacyl-acyl carrier protein reductase ame: full=beta-ketoacyl-acp	268	10	2,2E-40	51.8%	4	F:oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor; C:cytoplasmic part; P:metabolic process; F:nucleotide binding	EC:1.1.1.0	IPR002198; IPR002347; IPR016040; IPR020904; PTHR24322 (PANTHER), PTHR24322:SF76 (PANTHER), PF13561 (PFAM), seg (SEG), SSF51735 (SUPERFAMILY)