ABSTRACT

CHARLES, LAUREN ELIZABETH. Phylogenetic studies of *Pasteuria penetrans* looking at the evolutionary history of housekeeping genes and collagen-like motif sequences**.** (Under the direction of Charles Opperman)

Pasteuria penetrans, a gram-positive, endospore-forming eubacterium, is an obligate parasite of the root-knot nematodes, *Meloidogyne* spp. These nematodes are common root pests of economically important crop plants, making *P. penetrans* a potential biocontrol agent. Currently, the genome of the bacterium is being sequenced from spores and the data recovered utilized to better understand its mechanisms of parasitism.

Phylogenic analysis, at the protein level, has been done on a total of thirty-three bacterial species using the concatenation of forty housekeeping genes, two subsets of these involving twenty-eight taxa and twenty-seven genes, with and without indels removed, and each gene individually. Through the application of maximum likelihood, maximum parsimony, and Bayesian methods on these datasets, *P. penetrans* is found to cluster tightly within the class *Bacilli* of the gram-positive, low G+C content eubacteria with a high level of confidence. The exact placement of the bacterium as ancestral to *Bacillus* spp. has been resolved. Surprisingly, it is more closely related to the saprophytic extremophile *Bacillus haladurans* and *B. subtilis* than to the parasitic *B. anthracis* and *B. cerus*. These findings facilitate the phylogenetic context of *Pasteuria* to be exploited through further research in comparative genomics towards the development of biocontrol strategies.

Collagen-like genes, recently discovered in bacteria, are predicted to be virulence factors in *Bacillus anthracis* and *B. cereus,* close relatives to *Pasteuria penetrans*. These genes are identified by a GXY triple helix repeat motif that has only been found in the parasitic *Bacillus spp.*, namely the above two species and *B. thuringenisis*. They have,

however, been studied in depth in higher eukaryotes, such as humans, leading to assumptions about the content and structural stability of the triple helix GXY-repeats. As genomes are sequenced and *in silico* analyses deem possible, the true nature of collagen-like repeats have come to light along with their presence in other types of organisms, ranging from fungi to bacteria. Taking this approach to study *Pasteuria penetrans* for the possible role in virulence of collagen-like genes, more information has been revealed about the true characteristics and evolution of these motifs in bacteria, vertebrates, invertebrates, and fungi.

The percentage of each amino acid in the X and Y position of the GXY-repeat motif of 85 vertebrate, 228 invertebrate, 17 fungus, and 76 bacterial sequences, including 21 *Pasteuria penetrans* and 45 *Bacilli*, were examined. In the X-position, the following amino acids were preferred by each group: proline and then alanine for bacteria; proline and then threonine, alanine, and cysteine for *P. penetrans*; proline and then alanine and glutamine for invertebrates and vertebrates; and surprisingly, glycine and alanine with proline as a second choice for fungi. In the Y-position were the following results: bacteria utilize threonine then glycine and alanine; *P. penetrans* uses alanine, glycine, proline, and then threonine; Invertebrates contain proline followed by alanine; vertebrates utilize proline then alanine and arginine; and fungi use glycine, proline then serine and alanine.

Maximum Parsimony Analyses were performed on the collagen genes due to the heterogeneity in sites seen in the evolution across kingdoms. *P. penetrans* sequences were first paired with each of the other groups, including the other bacteria, invertebrates, vertebrates, and fungi. Then, a full phylogenetic tree was assembled from 15 *P. penetrans*, 40 other bacteria, 61 invertebrates, 16 vertebrates, and 17 fungi GXY-repeat motifs. These results strengthened the similarity of Xaa and Yaa amino acid usage within groups, except

for the fungi which paired closest to invertebrates. The tree also shows a pattern of evolution from fungi, invertebrates, vertebrates, to *P. penetrans* and other bacteria, mainly *Bacilli,* which cluster adjacent to one another.

PHYLOGENETIC STUDIES OF *PASTEURIA PENETRANS* LOOKING AT THE EVOLUTIONARY HISTORY OF HOUSEKEEPING GENES AND COLLAGEN-LIKE MOTIF SEQUENCES.

By
LAUREN ELIZABETH CHARLES

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

PLANT PATHOLOGY

Raleigh

2005

APPROVED BY:

Charles Opperman
Chair of Advisory Committee

David Bird
Advisory Committee

Ignazio Carbone
Advisory Committee

BIOGRAPHY

Lauren Charles was born and raised on a rural farm in the small town of Easton, CT. Graduating high school second in her class, she accepted a full scholarship to Boston College. She received an undergraduate degree in Mathematics and a minor in Environmental Studies, graduating Magnum Cum Laude. After spending a year in Richmond, VA taking Biology classes and working as a veterinarian technician, she decided to move to North Carolina and explore biological research. Here, she began working as a research technician under Charles Opperman in the Plant Pathology Department. Her first experiences with laboratory and greenhouse work focused on the discovery of resistance genes in *Medicago trunculata* for root-knot nematodes, *Meloidogyne javanica*. She was then recruited into a master's program at NCSU. Jumping into the new fields of Genomics and Bioinformatics, she was able to combine the knowledge gained from plant host-parasite interactions with her mathematical background. Her research concentrated on *Pasteuria penetrans*, a bacterial parasite of root-knot nematodes, *Meloidogyne arenaria*. As the sequencing of the genome progressed, she was able to utilize the data available to do a deep phylogeny on the organism and then focus on collagen-like repeat motifs. The results of these research projects are found in the following pages.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

REVIEW

NEMATODES IMPACT IN AGRICULTURE AND CURRENT CONTROL

Root-knot nematodes of the *Meloidogyne* spp. are among the most prevalent economical crop pests causing greater than $50 billion per year in economic losses (Sasser and Freckman, 1987). Due to the polyphagous nature of these parasites, they can feed on a wide range of hosts including tomatoes, peanuts, tobacco, peppers, watermelons, and cotton. They form root-galls while often injuring plant tissue which blocks water and nutrient flow in the plant leading to stunted growth, leaf chlorosis, impaired fruit production, and secondary infection. The most effective control method for root-knot nematodes is the application of nematicides, including the fumigant methyl bromide (Hague and Growen, 1987). Currently, the alternatives to soil applied chemicals are crop rotation, bare fallow, use of certified nematode-free planting material and resistant varieties, along with occasional use of parasitic or predatory fungi and nematodes (Akhtar, 1997). Recent discoveries of an obligate, soil-borne, bacterial pathogen of *Meloidogyne* spp., *Pasteuria penetrans*, bring hope for a safer, more effective way of biocontrol (Akhtar and Malik, 2000**;** Chen and Dickson, 1998; Stirling, 1988).

Root-knot nematodes are obligate, sedentary endoparasites of plants. Their life cycle begins when they hatch from the soil, in a mobile J2 stage, and migrate to the plant roots. Here, they penetrate the root in the zone of elongation and travel intercellularly until they find a suitable feeding site in the developing vascular region. The initiation of feeding through the stylet leads to a loss in mobility, secretion of chemicals from esophageal glands hypothesized to cause giant-cell formation, and three molts resulting in sexual differentiation. Males, which retain the veriform shape, travel back out of the root, while the pyriform females remain to feed on the multinucleate giant-cells. The increased metabolic activities

from the four to six induced giant-cells provide her with enough nutrients to reproduce parthenogenetically. After laying 500-1,000 eggs in an external egg sac, the female nematode dies and the eggs are then free to hatch spontaneously in the soil (Davis, 2003a; Davis, 2003b; and Trudgill, 1991).

To control plant pathogenic nematodes, the use of toxic fumagant and nonfumagant nematicides are the most effective treatments available but inevitably lead to environmental problems and residues. Besides being extremely costly, there is often re-infestation of the soil after harvest, and contamination of ground water. There can also be residues left on fruits, vegetables, and ground surfaces which have extremely toxic effects on the birds and mammals that come in contact. This has consequently led to the suspension of certain chemicals on the national and state level often based on soil texture and organic matter content, depth to the water table, and proximity to drinking water wells.

In the USA and some other countries, the common fumigant nematicides, such as dibromochloropropane (DBCP), dichloropropene/dichloropropane mixture (D-D), and ethylene dibromide (EBE), cause serious groundwater contamination and, therefore, are banned from use. Nonfumigant organophosphate and carbamate chemicals, such as aldicarb and carbofuran, are also regionally restricted in the USA. When these chemicals decompose, they often leach into highly biologically active soils causing problems. States, such as California and Florida, restrict aldicarb usage to low rainfall months to cut down on the toxic effects. There has also been a suspension of the use of 1-3 dichloropropene in CA in 1990 which further reduces the choices for chemical nematode control. This year, 2005, the use of methyl bromide has also been prohibited from use due to its toxicity and contribution to the depletion of the atmospheric ozone layer. Alternatively, granular and systemic nematicides

are thought to be better for the environment since they are only applied to areas of high root density and, therefore, rich nematode concentration. Here under the canopy, there is restricted rainfall and less movement of the toxic chemicals through the soil but the harmful environmental effects are still evident (Duncan, 1991).

With these restrictions and biological problems relating to chemical nematicides, there is an increase in cultural practices, such as crop rotation and bare fallow, use of certified nematode-free planting material and resistant varieties. Besides the limited effectiveness, there are inherent problems with each of these practices. In areas such as dry lands, a narrow range of crop plants are often able to be grown in the region and the soil is inherently low in nutrients. Bare fallow is not possible here and crop rotation, as a general rule, is extremely limited due to the selection of agricultural plants being good hosts for many of the nematode genera. Flooding paddy rice followed by vegetable crops (Thames and Stoner, 1953) or tobacco (de Guiran, 1970) is found to be effective, but again not practical in most situations. Resistant crop varieties are few in number and often due to single-gene resistance such as in tomato which is quickly overcome in the field. Use of nematode-free plants and planting material is inefficient in areas of high soil infestation. There are some predatory or parasitic fungi and nematodes, such as *Pochonia chlamydosporia* (Kerry, 2001) and *Arthrobotrys* spp. (Stirling and Smith, 1998), that have provided evidence for successful biological control. *Bacillus thuringiensis*, a common bacterial insecticide, produce toxic proteins that have recently been discovered to also target nematodes, including *Rotylenchulus reniformis* and *Meloidogyne* spp., showing potential use as a nematicide (Chen *et al.,* 2000; Wei *et al*., 2003; Zuckerman *et al*., 1993; http://www.nal.usda.gov/bic/BTTOX/BT-patents/05378460.html). The problem with these

biocontrol agents is that they are not fully exploited and rarely used in the field (Duncan, 1991; and Akhtar, 1997), although the use of molecular techniques provides future potential (Morton *et al.*, 2004).

PASTEURIA PENETRANS POTENTIAL BIOCONTROL

Pasteuria penetrans, an obligate parasite of the root-knot nematode *Meloidogyne* spp., has the potential to provide crop protection from these pests in an effective, environmentally safe, and controlled way. Often found associated with nematode suppressive soils, these gram-positive bacteria have the potential to become a commercially available biocontrol agent (Akhtar and Malik, 2000**;** Chen and Dickson, 1998; Stirling, 1988). Having a high tolerance to adverse environmental conditions, high and low temperatures, humidity changes, desiccation, storage, and most pesticides (Mankau and Prasad, 1972; Sayre and Starr, 1989; Stirling *et al.*, 1986), these nematode hyperparasites have an advantage over proposed biocontrol fungi in addition to a long shelf life. The bacteria, currently unexplainable, are highly selective to individual populations within species of nematodes in a field to all nematodes within a species as shown through attachment assays, and easily detected in the soil (Carneiro *et al.*, 1999; Channer and Gowen, 1992; Davies and Danks, 1992; Davies *et al.*, 2001; Stirling, 1985). Through studying the biology, life cycle and virulence factors of *P. penetrans*, the tritrophic relationship between the bacterium, nematode and plant can be better understood and exploited.

Pasteuria penetrans lifecycle is intimately linked to that of its host. The highly durable bacterial endospores reside in the soil awaiting the chance encounter of a juvenile nematode traveling to its host root system. As the nematode passes by, the bacterium

adheres to the cuticle of the plant parasite which can often end up encumbered with endospores before reaching its host. This attachment of spores alone can reduce the penetration of nematodes into the roots (Stirling, 1988). Once the nematode has entered the plant and initiated feeding, the *P. penetrans* spores begin to germinate and produce rhizoids throughout the infected females (personal communication with Keith Davies). As the females grow and take in nutrients, the rhizoids produce bacterial rods which rapidly undergo exponential growth. The reproductive tract and egg production of the nematode is thereby inhibited by the bacterium and instead its cavity becomes filled with bacterial microcolonies (Sayre and Starr, 1985). Sporogenesis is soon triggered in *P. penetrans,* similar to *Bacilli* spp., and involves the initiation of a metal-ion sensitive, phosphor-relay pathway (Phillips and Strauch, 2002). Once the bacteria fully develop into endospores, the female nematodes die and the individual cadaver can release up to $2x10^6$ spores into the soil (Chen and Dickson, 1998; Davies *et al*., 1988; Hewlett and Dickson, 1993; Sayre and Starr, 1985). After about 3 years (Oostendorp *et al.*, 1991; Chen and Dickenson, 1998), this cycle naturally suppresses the growth of the nematodes resulting in an estimated 10^4 to 10^5 endospores per gram of soil (Stirling, 1988). The plant parasites are replaced with the endospores of the nematode hyperparasite, *Pasteuria penetrans,* in an environmentally controlled manner since the bacterium can only actively proliferate inside *Meloidogyne* spp.

To be able to exploit the tritophic relationship between the bacterium, nematode and plant in the field, the biology and chemistry of the bacterium must be fully understood. The obligate nature of the host-parasite interaction makes it difficult to study *Pasteuria penetrans* on its own. Some attempts have been made in traditional laboratory and field experiments to grasp the mechanisms of parasitism by the nematode hyperparasite. The two major

categories that are under investigation for *P. penetrans* potential in biocontrol are the ability to mass produce the fastidious bacteria as well as expanding its limited host range to enable the attachment to and germination into a broader array of phytoparasitic nematodes (Oka *et al*, 1997).

Attempts to mass produce *Pasteuria penetrans* without its obligate *Meloidogyne* host have been futile. Although it is possible to view the bacteria on laboratory media, the ability to keep the cultures active, increase the rate of growth and identify growth factors has been fruitless (Bishop and Ellar, 1991; Riese *et al.*, 1998; Williams *et al.*, 1989). Exponential growth followed by sporulation must be controlled to successfully culture and mass produce the bacterium *in vitro*. The closest to achieving this were Bishop and Ellar in their 1991 manuscript which proposed two different types of media triggering each of the two growth phases respectively. These media, while triggering the desired vegetative growth and sporulation phase, were only feasible on a small scale. Since the bacteria lives in the pseudocoelom of the nematode, studies that model the physical and chemical makeup of this fluid may be central to the development of an *in vitro* culture medium. Although there have been attempts to define a growth medium from *in silico* analysis of obligate bacterial parasites before such as for *Xylella fastidiosa* (Lemos *et al.*, 2003), this approach has not yet been looked at yet for *P. penetrans*. Using genomic sequences to predict nutritional requirements for steady proliferation is a new but possibly more effective method to obtain results for fastidious, obligate parasites such as seen here. Once vegetative growth is possible *in vitro*, the transition steps between vegetative growth and sporulation need to be understood so that this step can also be manipulated, allowing for mass distribution of inactive endospores into nematode infested soils.

To enable the controlled expansion of *Pasteuria penetrans* host range, experiments have been conducted to identify factors effecting the attachment and germination in the nematode. There have been attachment assays and attempts to find a link between host specificity of *P. penetrans* and the *Meloidogyne* cuticle without avail. Although there has been one successful attachment assay using three different *P. penetrans* populations each with unique spore binding properties that could successfully identify *M. arenaria, M. incognita,* and *M. hapla* (Orui and Ozawa, 1999), all other reported experiments showed conflicting results. Despite the fact that that the nematodes reproduce parthenogenetically and the bacteria clonally, there are still large differences observed between each individual's surface composition for attachment. Due to this heterogeneity in nematode cuticles and spore binding properties of the bacteria, no connection has been consistently found between the two. Attempts to link spore attachment to nematode phylogeny have strengthened this claim of respective heterogeneity between the spore attachment and cuticle (Davies *et al.*, 2001), along with immunological studies using monoclonal antibodies (Davies and Redden, 1997) and baiting experiments on *Meloidogyne* (Davies *et al.*, 1994). In a study done by Oostendorp, Dickson, and Mitchell (1990), it was discovered that isolates of *Pasteuria* could adapt to host nematodes by propagation on the specific population, increasing in successful attachment and germination each time. This biological adaptation of concomitant heterogeneity between the host and parasite is similar to the Red Queen Hypothesis or arms race where one creature must adapt quickly to just survive from generation to generation. Without the link between recognition, attachment, and germination, expanding the host range of the potential biocontrol remains a challenge.

This inability to predict attachment of *P. penetrans* spores to a nematode relays the urgency to explore other means of studying this interaction to fully understand the first step in the infection process. When examining the potential for use as a biocontrol agent in the field, this heterogeneity in attachment surfaces of both the host and parasite can be misleading. If incompatible subgroups of bacteria and nematodes are required to interact, the ability to suppress the soil will be lower than expected (Tzortzakakis *et al.*, 1997). It is also believed that inefficient suppression may reflect other factors such as soil conditions, relating to texture and chemical make-up, rather than just the effect of the parasitic interaction (Mateille *et al*., 2002). If this recognition step could be exploited so that there was a universal attachment and germination of spores, the chance for *P. penetrans* use as a biocontrol agent would drastically increase.

A GENOMIC APPROACH TO STUDYING *P. PENETRANS*

In an effort to resolve issues such as those presented above, Genomics, the study of an organism through DNA sequences, and Bioinformatics, the use of mathematical techniques and models to solve biological problems, were developed. The tools created enable the research of an obligate parasite through DNA or protein sequences when laboratory and field experiments are limited. Then, with this information, evolutionary relationships and histories can be inferred providing snapshots into the past. *Pasteuria penetrans* is a prime candidate for this type of research. It requires the presence of the host, *Meloidogyne* spp., in order to proliferate, and ambiguity is faced when attempting to identify factors for this interaction and for proliferation. The genetic information gained through

sequencing and comparative genomics can be used to understand the determinants of its host range to expand its capabilities as a biocontrol agent.

To be able to exploit these methods, pure DNA from the bacteria must first be extracted and then sequenced in the laboratory. Once there is at least a six-fold coverage of the genome, the series of cloned DNA sequences or contigs are matched together. Although there still may be DNA or genes, often non-functional, which have not been sequenced or discovered through the current techniques, the genome is still considered complete at this time. With fastidious, obligate parasites, such as *Pasteuria penetrans*, preparing the DNA to be sequenced can often be the hardest step in the process.

Multilocus sequence typing (MLST) (Maiden *et al*., 1998), which has been shown to be extremely useful in characterizing pathogenic strains (Urwin and Maiden, 2003), has now replaced the use of multilocus enzyme electrophoresis (MLEE). MLST exploits the variation that is slowly accumulating in housekeeping genes within populations and is selectively neutral for characterizing bacteria. The use of enzymes through MLEE has never been exploited before in the characterization of *Pasteuria penetrans,* due to its obligate parasitic nature. In this case, it is extremely difficult to distinguish *Pasteuria* enzymes from host enzymes. However, the ability to extract and amplify only *Pasteuria* DNA from endospores makes MLST a feasible approach to characterizing different groups of the bacteria along with providing a pathway to infer the organism's closest relatives along with its phylogenetic history.

The next step in this type of approach is to identify the taxonomic groups closest to the organism using the genomic sequences along with past knowledge from morphological and biochemical studies. Often, a phylogenetic analysis (i.e., a systematic way to determine

the ancestral relationships among known species using methods such as parsimony, maximum likelihood, and Bayesian analysis) is performed to find these relationships. This knowledge gained will lead to comparative studies with close relatives, using similar techniques, and eventually, testable inferences regarding topics of interest about the organism in question, such as virulence, growth and transition factors.

TAXONOMIC HISTORY

When examining the history of *Pasteuria penetrans,* there were many changes seen in the relationship between taxonomy and morphology leading to an unclear resolution of the bacteria's exact placement in the tree of life. There have been many attempts to classify this genus of bacteria since first described by Metchnikoff in 1888 on *Daphnia*, a waterflea, as *Pasteuria ramosa*, in honor of Louis Pasteur. In 1906, Cobb studied the morphology of this parasite on the nematode *Dorylaimus bulbiferous* and declared it a protozoan. This change was later accepted and renamed by Thorne in 1940 as *Duboscqia penetrans*. Since then, electron microscope techniques have shown the bacteria to be more *Bacillus*-like rather than a protozoan and hence, it was renamed again to *B. penetrans* (Mankau, 1975). In 1985, there was a rediscovery of Metchnikoff's work suggesting similarities to the original *P. ramosa,* which led to a reversion of the genus back to *Pasteuria* by Sayre and Starr. Based on its morphological features as gram-positive, mycelial and endospore-forming bacteria, it was classified as an *Actinomycetales* (Sayre and Starr, 1989). Within the *Pasteuria* genus, based on the host range, life cycle and morphology, three different species of hyperparasites of the phytoparasitic nematodes have been proposed, namely *P. penetrans* on *Meloidogyne* spp., *P.*

thornei on *Pratylenchus* spp., and *P. nishizawae* on *Heterodera* and *Globodera* spp. (Chen and Dickson, 1998).

With the recent development of automated sequencing techniques, there have been several subsequent attempts to utilize single genes to re-evaluate the classification of *Pastueria penetrans* and define its closest relatives. Studies using 16S rDNA and rRNA showed *Pasteuria* spp. including *P. penetrans,* to be members of the *Clostridium-Bacillus-Streptococcus* branch of gram-positive eubacteria (Anderson *et al*., 1999, Atibalentja *et al.*, 2000; Preston *et al*., 2003). This has recently been verified through the examination of *spo0A* sequence, placing *Pasteuria* with members of the supergenus *Bacillus* (Trotter and Bishop, 2003). However, phylogenetic analytical procedures relying on only one gene at a time to classify a species have been found to be inaccurate in many cases while the use of multi-locus approaches have been shown to be more reliable (Doyle, 1992; Fox *et al.*, 1992; Pamilo and Nei, 1988; Rokas *et al.*, 2003a,b; Scholl *et al.*, 2003; Young, 2001). Therefore, the exact placement of *P. penetrans* in the tree of life showing which organism is its closest relative is still left unresolved through morphological, biochemical, and single-gene analyses. A more robust approach to phylogenetic analysis of the bacteria would provide stronger evidence towards *P. penetrans* evolutionary history.

To resolve the placement of *Pasteuria penetrans* among bacteria, a deep phylogenetic analysis was performed on forty housekeeping genes separately and concatenated together along with a subset of these twenty-seven genes. Various phylogenetic programs were used (Bayesian, maximum likelihood and maximum parsimony) to further strengthen the validity of the drawn conclusions. This analysis, procedures, and outcomes can be found in Chapter 2 of this manuscript. The information acquired will be utilized to gain more knowledge about

the obligate parasite through comparisons of its closest relatives and available DNA sequences.

CLOSEST RELATIVES OF THE *BACILLI* CLASS

Through morphological and biological studies, *Pasteuria penetrans* is clearly among the *Bacilli* class. From this information, biological characteristics of the bacterium can be proposed. This may include specifics about the genome, such as size, G+C content and placement of genes, conserved pathways like sporogenesis along with potential factors in virulence and pathogenicity. With this information, hypotheses relating to attachment, germination, and factors for mass production, including triggers for sporulation, can help unravel the necessary steps for successful biocontrol.

The genome of *Pasteuria* might be expected to have many similar characteristics to that of the *Bacilli*, especially the pathogenetic bacteria. Most of the fully sequenced *Bacillus* genomes, including *Bacillus subtilis* and *B. halodurans,* have about 4.2 Kb and have an average G+C content of 44% (Kunst *et al.,* 1997; Takami *et al.,* 2000). Bacteria which have evolved the ability to parasitize other organisms tend to utilize their host for a number of metabolic activities. This often leads to a loss in gene function and DNA, resulting in a reduction of genome size (Klasson and Andersson, 2004; Moran, 2002). Due to this minimization of pathogenetic bacteria's DNA and its similarities to *Bacilli*, *Pasteuria* genome size is most likely between 2.5-4Mb. It would not be surprising if the average G+C content and amino acid usage should also parallel these bacteria. These factors are exploited in a phylogenetic analysis used to determine the closeness of the bacteria on an evolutionary scale, as seen in Chapter 2 of this thesis.

The minimum genome of pathogenetic bacteria is comprised of the genes necessary for virulence, pathogenicity and basic cell functions. Many parasitic bacteria contain pathogenicity islands which are genetic elements made up of clustered chromosomal genes with a slightly different G+C composition that encode for virulence determinants and genetic flexibility (Hacker *et al.,* 1997). Pathogenicity islands have been recently discovered in gram-positive bacteria, including mammalian pathogens *Staphylococcus aureus* (Yarwood *et al.,* 2002) and *Bacillus anthracis* (Read *et al.,* 2003), and phytopathogenic *Stremtomyces* spp. (Kers *et al.,* 2004). Such genetic markers may hold the key to *Pasteuria penetrans* host preference and ability to infect certain nematode populations.

The *Bacilli* class contains a diverse collection of gram-positive, rod-shaped, aerobic, bacteria most of which undergo endospore formation upon deprivation of an essential nutrient (Errington, 1993). These endospore-forming bacteria include *Alicyclobacillus, Amphibacillus, Bacillus, Clostridium, Desulfotomaculum, Sporohalobacter, Sporolactobacillus, Sporosarcina, Sulfobacillus, Syntrophosphora,* and *Thermoactinomyces* (Berkeley and Ali, 1994). This process of endospore formation called sporogenesis, which the bacteria undergo while inside the female nematode, is an essential component for survival and mass production. Sporogenesis is a highly conserved biological process in bacteria and it is believed that *Pasteuria* and *Bacillus thuringiensis* follow a very similar sporogenesis pathway (Chen *et al.*, 1997). The resulting endospore is the survival stage of the bacterium and allows it to endure harsh conditions in the soil until the chance encounter of its host root-knot nematode. If the underlying process and key factors of sporulation could be identified in *Pasteuria* with the help of comparative genomics then the utilization of the bacteria as a biocontrol agent could be possible.

Contributing to the attachment and germination, these bacterial endospores have an outermost layer compromised of two distinct parts, a proteinaceous coat and a peripheral layer seen in pathogenic spores called the exosporuim. In a proteomic analysis of the spore coats of *Bacillus subtilis* and *Bacillus anthracis*, a core group of proteins is shared between species coding for mostly morphological characteristics necessary for spore resistance and survival (Lai *et al.*, 2003). The spore coat constituents and polysaccharide components that make up the outer surface contain the most variety between organisms, such as noted between *B. anthracis* and *B. subtilis* (Read *et al.,* 2003). These outer layer differences may be key factors in the identification and host specificity seen in different organisms including the close relative *Pasteuria penetrans*.

Serving as the primary permeability barrier, the exosporuim is a prominent, heavily glycosylated, loose-fitting, balloon-like layer containing spore surface antigens (Steichen *et al.*, 2003). Integral to spore survival, germination and disease, this proteinaceous layer is an essential component for pathogenesis (Sylvestre *et al.*, 2002). The antigens present on this surface of *Pasteuria penetrans* have been characterized by monoclonal antibodies in attempt to link them to attachment and recognition of host nematodes without avail (Davies and Redden, 1997). However, since then a link has been identified between attachment, infection, and an immunodominant antigen, BclA, on *Bacillus anthracis* spore coat, found by unique antispore monoclonal antibodies (Steichen *et al.,* 2003; Sylvestre *et al.*, 2002, 2003). BclA, along with several other glycoproteins identified in *B. thuringiensis* and *B. cereus* (Garcia-Patrone and Tandecarz, 1995; Charlton *et al.,* 1999), are only found in spores or sporulating cells and are structural components of exosporium filaments (Sylvestre *et al.*, 2002, 2003). These glycosylated proteins contain internal collagen-like regions of GXY

repeats of different lengths, which have been shown to be responsible for variation in filament length in *Bacillus anthracis* (Sylvestre *et al.*, 2003). These collagen-like genes, also found in *Pasteuria penetrans,* may be a principle factor in host specification, attachment and germination. A closer look at these genes, seen in Chapter 3 and discussed below, may bring scientific research one step closer to utilizing *P. penetrans* as a biocontrol agent.

COLLAGEN-LIKE SEQUENCES

The majority of the life-cycle of *Pasteuria penetrans* is spent as an endospore in the soil prior to nematode attachment. The first stage of the infection process begins with the exosporium of the bacterium adhering to the nematode cuticle. The exosporium forms wing-like structures containing hair-like filaments that provide a solid attachment to the nematode. These filamentous proteins contain triple helices, similar to those found in collagen proteins and, moreover, found in *Bacillus anthracis, B. thuringiensis* and *B. cereus* as a factor in endospore attachment (Sylvestre *et al.*, 2002, 2003; Garcia-Patrone and Tandecarz, 1995; Charlton *et al.*, 1999).

Collagens are filamentous proteins characterized by contiguous GXY-triplet motifs forming triple-helices where G represents glycine and X and Y are variable amino acids. They have many functions including being the major protein of connective tissue and the most abundant protein in animals (Bairati and Garrone, 1985). These collagens were characterized by their helical protein shape, a right-handed superhelix composed of three left-handed polyproline type II-like chains wound around their central axis. Glycine is present every third amino acid since its small, neutral nature enables the helical protein shape while still maintaining planar peptide bonds. They are stabilized by hydrogen bonds between the

backbone groups, the NH of glycine (donor) and the CO in the X-position of another chain (acceptor) (Beck and Brodsky, 1998; Xu *et al.,* 2002). The OH of hydroxyproline can also serve as the donor in addition to water-mediated interactions aiding in the stabilization. A high content of imino acids in the X and Y, namely proline and hydroxyproline, not only help in hydrogen bonding but create steric repulsion due to the pyrrolidone rings in these residues which force the extended helical form.

Requiring the post-translational hydroxylation of proline to stabilize the triple helix shape, which other organisms such as bacteria are unable to execute due to an absence of proline hydoxylases, collagens were thought to be exclusive to the animal kingdom. To this day, there is no evidence for their presence in protozoa, plants, or algae. However, recent discoveries of collagen-like genes in other organisms, namely the surface structures or spore components of bacteria, the fungi fimbriae, and other proteins within invertebrate and viral genomes, have lead to discussion of other means for stabilization of the helical shape (Bann *et al.*, 2000; Xu, *et al.*, 2002; Yang, *et al.*, 1997). This also suggests the presence of a common ancestor existing before the divergence of these organisms, such as between fungi and animals (Celerin *et al.,* 1996), or the possibility of horizontal gene transfer between organisms or kingdoms (Rasmussen *et al.*, 2003).

When looking at the composition of the X and Y groups, it has been found that the side chains of these residues have minor effects on the triple-helix stability and are mainly used for intermolecular interactions since they are always exposed to the solvent (Beck and Brodsky, 1998; Chan *et al*., 1997). This allows for a variety of functions the collagen-like proteins perform in organisms, most commonly directed towards these types of interactions. The Y-position does seem to have some impact on stability as shown in melting temperature

studies using a variety of GXY triplets, favoring Gly-Pro-Arg and Gly-Pro-Hyp (Yang *et al*., 1997). It is found through calorimetric studies that the major stabilizing factor in these molecules is through hydrogen bonding to the glycine backbone structure (Privalov, 1982). Cysteine residues are also thought to contribute to the triple-helix stabilization through covalent inter-chain bonding along with a conservation of polar/apolar nature in the residues (Beck and Brodsky, 1998). For example, it has been noted that if there is a negative amino acid in the X-position, such as glutamate or aspartate, there is usually a complimentary positive amino acid present in the Y-position, such as arginine or lysine, that can form stable trimers and vise-versa (Rasmussen *et al.*, 2003). Therefore, the X and Y-positions and their side chains seem to show less involvement in the stabilizing of the triple-helix shape which was previously thought to require the use of imino acids, as seen in vertebrates. This would allow for substitution in these sites contributing to different evolutionary forces acting on each type of organism providing ground works for phylogenetic analysis seen in Chapter 3 of this manuscript.

Current glimpses into the use of alternative residues in the GXY triplet motif through biochemical studies and protein models are shown in bacteria and invertebrates, but not yet in fungi. The most studied invertebrates are of the Phylum Annelida, including *Nereis sp.*, *Pheretima sieboldi* (earthworm) and *Riftia pachyptila* (hydrothermal vent worm). *Nereis sp.* and the earthworm were found to have significant amounts of alanine and serine, with lesser amounts of threonine and glutamic acid, in the Y-position and were resistant to proteolysis by clostridial collagenase, implying that the collagens were extremely stable (Goldstein and Adams, 1970; Waite *et al.*, 1980). The hydrothermal vent worms had a very low content of

proline and hydroxyproline, but instead use threonine in the Y-position creating a large amount of stability when glycosylated (Mann *et al*., 1996; Bann *et al.*, 2000).

Collagen-like triple-helix repeats have been found in only a limited number of microbrial genomes, including the gram-positive bacteria *Bacillus anthracis, B. cerus, B. thuringiensis*, and *Streptococcus pyogenes*. In all bacteria studied, proline was still favored in the X-position while threonine dominated the Y location allowing for direct hydrogen bonding to the backbone and hence, a strong collagen-like triple helix structure (Rasmussen *et al.*, 2003; Sylvestre *et al.*, 2003; and Xu *et al.*, 2002). There were the most variations in X and Y-positions found within these bacterial sequences. For example, *Bacillus anthracis* BclA protein contained some aspartic acid and threonine in the X-position (Sylvestre *et al.*, 2003) and the streptococcal Scl1 and Scl2 proteins had significant amounts of arginine, glutamic acid, aspartic acid and lysine in the X while arginine, aspartic acid and lysine were also found in Y-position (Xu *et al.*, 2002). There has also been notation of high amounts of glutamine, along with threonine, present in the Y-position of gram-positive bacteria when the majority of the X-position is occupied by proline. This same study also noted that if the sequences had greater than fifty-percent of the Y spot occupied by threonine, then the X was a charged amino acid such as alanine, serine, or proline (Rasmussen *et al.*, 2003).

Since *Pasteuria penetrans* has been found to be embedded in the *Bacilli* clade (Charles *et al.,* 2005) and these collagen-like repeats have only been found in pathenogenic *Bacilli*, these GXY-repeat sequences have been further studied in this paper for insight into their evolutionary history. While *Bacillus anthracis* is a true pathogen, and *B. cereus* and *B. thuringiensis* are facultative pathogens and *P. penetrans* is an obligate parasite, they all have collagens-like motifs. The genes containing these GXY-repeats have been noted in *B.*

anthracis and *B. cereus* as possible virulence factors in endospore attachment, which is the first stage in *P. penetrans* infection (Sylvestre *et al.*, 2003; Charlton *et al.*, 1999).

Since the collagens are very diverse and old molecules, there is a large amount of heterogeneity noted within and throughout the different kingdoms. The non-GXY domains are very different with less than 20% identity in most cases. This has lead to two major assumptions: 1.These genes have evolved many different uses, one possibly being for *Pasteuria penetrans* to attach to the nematode verses *Bacillus anthracis* attaching to its mammalian host; and 2. Phylogenetic analysis would be meaningless on such diverse proteins (Rasmussen *et al.*, 2003). Chapter 3 of this manuscript overcomes this obstacle by excluded those regions of dissimilarity focusing on only the triple-helix motif. By looking at these collagen-like genes and their evolutionary history, key factors affecting attachment and germination may be eluded leading to an understanding of host range and ultimately utilization as a biocontrol agent.

CONCLUSIONS

In attempt to reduce the amount of toxic chemical pesticide use, the tritrophic relationship between *Pasteuria penetrans*, *Meloidogyne* spp., and various crop plant species is being studied for utilization in biocontrol. There have been many attempts to study the fastidious bacteria, *P. penetrans*, in the laboratory without much success. Currently, *P. penetrans* genome is being sequenced providing a new source of information to be exploited. This thesis describes a bioinformatics approach to unraveling the secrets of attachment, germination, and proliferation.

A deep phylogenetic analysis of a variety of housekeeping genes was undertaken to resolve the bacteria's true placement in the tree of life. With this information, new hypotheses have been drawn regarding the biological processes utilized by the obligate parasite, *P. penetrans*, through comparisons with its closest relatives among the *Bacilli* class.

Utilizing this information, a closer look was taken at collagen-like genes encoding proteins presumed to be found on the exosporium or endospore surface. These filaments are believed to be virulence factors in attachment and germination of similar pathogenic bacteria. This is one of the first attempts to study *Pasteuria penetrans* biology through *in silico* analysis using comparative genomics. Although there is a large amount of heterogeneity in the spore surface and collagen-like repeat motifs, this analysis gives a broad idea as to which of these genes may have a role in pathogenicity and the origin of each of them. It proposes genes that may be horizontally transferred between species and even kingdoms.

In light of the information presented before, the fourth chapter summarizes the conclusions that can be drawn from the analyses and proposes additional hypothesis which can be studied in the same manner. The utilization of genetic information through mathematical models opens a whole new window of opportunity for biological discovery unavailable in the laboratory for obligate parasites. Through implementation of these tools, *Pasteuria penetrans* is well on its way to being fully exploited in the field for natural biocontrol against the economically devastating crop pathogen, root-knot nematodes of the *Meloidogyne* spp.

LITERATURE CITED

Akhtar, M. 1997. Current options in integrated management of plant-parasitic nematodes. Integrated Pest Management Reviews **2(4):** 187 – 197.

Akhtar, M. and Malik, A. 2000. Roles of organic soil amendments and soil organisms in the biological control of plant-parasitic nematodes: a review. Bioresource Technology **74:** 35-47.

Anderson, J.M., Preston, J.F., Dickson, D.W., Hewlett, T.E. and Maruniak, J.E. 1999. Phylogenetic analysis of *Pasteuria penetrans* by 16S rRNA gene cloning and sequencing. Journal of Nematology **31:** 319-325.

Atibalentja N., Noel G.R., and Domier L.L. 2000. Phylogenetic position of the North American isolate of Pasteuria that parasitizes the soybean cyst nematode, *Heterodera glycines*, as inferred from 16S rDNA sequence analysis. International Journal of Systematic and Evolutionary Microbiology **50:** 605-13.

Bairati, A. and Garrone, R. (Eds) 1985. Biology of invertebrate and lower vertebrate collagens. Plenum Press, New York.

Bann, J.G., Payton, D.H., and Bachinger, H.P. 2000. Sweet is stable: glycosylation stabilizes collagen. FEBS Letters **473:** 237-240.

Beck, K., and Brodsky, B. 1998. Supercoiled protein motifs: The collagen triple-helix and the α-helical coiled coil. Journal of Structural Biology **122:** 17-29.

Berkeley, R.C.W., and Ali, N. 1994. Classification and identification of endospore-forming bacteria. Pages 1-8 in: Fundamental and Applied Aspects of Bacterial Spores. G.W. Gould, A.D. Russell, and D.E.S. Stewart-Tull, eds. Blackwell Scientific Publishing, Oxford.

Bishop, A.H. and Ellar, D.J. 1991. Attempts to culture *Pasteuria penetrans in vitro*. Biocontrol Science and Technology **1:** 101-114.

Carneiro, R.M.D.G., Randig, O., Freitas, L.G., and Dickson, D.W. 1999. Attachment of endospores of *Pasteuria penetrans* to males and juveniles of *Meloidogyne* spp. Nematology **1(3):** 267-271.

Celerin, M., Ray, J.M., Schisler, N.J., Day, A.W., Stetler-Stevenson, W.G., and Laudenbach, **D.E. 1996.** Fungal fimbriae are composed of collagen. The EMBO Journal **15(17):** 4445-4453.

Chan, V.C., Ramshaw, J.A.M, Kirkpatrick, A., Beck, K., and Brodsky, B. 1997. Positional preferences of ionizable residues in Gly-X-Y triplets of the collagen triple-helix. Journal of Biological Chemistry **272:** 31441-31446.

Channer, A.G. and Gowen, S.R. 1992. Selection for increased host resistance and increased pathogen specificity in the *Meloidogyne – Pasteuria penetrans* interaction. Fundamental and Applied Nematology **15:** 331-339.

Charles, L., Carbone, I., Bird, D., Burke, M., Opperman, C., Davies, K., and Kerry, B. 2005. Phylogenetic Analysis of *Pasteuria penetrans* using multiple genetic loci. Journal of Bacteriology (submitted).

Charlton, S., Moir, A.J.G., Baillie, L., and Moir, A. 1999. Characterization of the exosporium of *Bacillus cereus*. Journal of Applied Microbiology **87:** 241-245.

Chen, Z.X., Dickson, D.W., Freitas, L.G., and Preston, J.F. 1997. Ultrastructure, Morphology, and Sporogenesis of *Pasteuria penetrans*. Phytopathology **87(3):** 273-283.

Chen, Z.X. and Dickson, D.W. 1998. Review of *Pasteuria penetrans*: biology, ecology, and biocontrol potential. Journal of Nematology **30(3):** 313-340.

Chen, J., G. S. Abawi, and B. M. Zuckerman. 2000. Efficacy of *Bacillus thuringiensis*, *Paecilomyces marquandii*, and *Streptomyces costaricanus* with and without organic amendment against *Meloidogyne hapla* infecting lettuce. Journal of Nematology **32:** 70-77.

Cobb, N.A. 1906. *Fungus Maladies* of the sugar cane, with notes on associated insects and nematodes, 2nd ed. Hawaiian Sugar Planters Assoc. bulletin no. 5, Hawaiian Sugar Planters Association, Honolulu.

Davies, K.G., Kerry, B.R., and Flynn, C.A. 1988. Observations on the pathogenicity of *Pasteuria penetrans* a parasite of root-knot nematodes. Annals of Applied Biology **112:** 491-501.

Davies, K.G. and Danks, C. 1992. Interspecific differences in the nematode surface coat between *Meloigogyne incognita* and *M. arenaria* related to the adhesion of the bacterium *Pasteuria penetrans*. Parasitology **105:** 475-480.

Davies, K.G., Fargette, M., Balla, G., Daudi, A., Duponnois, R., Gowen, S.R., Mateille, T., Phillips, M.S., Sawadogo, A. Trivino, C., Vouyoukalou, E., and Trudgill, D.L. 2001. Cuticle heterogeneity as exhibited by *Pasteuria* spore attachment is not linked to the phylogeny of parthenogenetic root-knot nematodes (*Meloidogyne* spp.). Parasitology **122 (1):** 111-120.

Davies, K.G. and Redden, M. 1997. Diversity and partial characterization of putative virulence determinants in *Pasteuria penetrans*, the hyperparasite of root-knot nematodes. Journal of Applied Microbiology **83(2):** 227-235.

Davies, K.G., Redden, M., and Pearson, T.K. 1994. Endospore heterogeneity in *Pasteuria penetrans* related to attachment to plant-parasitic nematodes. Letters in Applied Microbiology **19:** 370-373.

Davis, E.L. 2003a. PP 501 Lecture I: Nematology. NCSU Plant Pathology.

Davis, E.L. 2003b. PP 501 Lecture II: Root-knot and Cyst Nematodes. NCSU Plant Pathology.

de Guiran, G. 1970. Le problème *Meloidogyne* sur tabac à Madagascar. *Cah. ORSTOM*, Ser. Biol. **11:** 187-208.

Doyle, J.J. 1992. Gene trees and species tree: Molecular systematics as one-character taxonomy. Systematic Botany **17:** 144-163.

Duncan, L.W. 1991. Current options for nematode management. Annual Review of Phytopathology **29:** 469-490.

Errington, J. 1993. Bacillus subtilis sporulation: regulation of gene expression and control of morphogenesis. Microbiol. Rev. 57:1-33; and Priest, FG. 1993. Systemics and ecology of Bacillus, p. 3-16. In A.L. Sonenshein, J.A. Hoch, and R. Losick (ed.), Bacillus subtilis and other gram-positive bacteria. Biochemistry, physiology, and molecular genetics. American Society for Microbiology, Washington, D.C.

Fox, G.E., J.D. Wisotzkey, and P. Jurtshuk, Jr. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. International Journal of Systematic Bacteriology **42:** 166-170.

Garcia-Patrone, M., and Tandecarz, J.S. 1995. A glycoprotein multimer from *Bacillus thuringiensis* sporangia: Dissociation into subunits and sugar composition. Molecular and Cellular Biochemistry **145:** 29-37.

Goldstein, A. and Adams, E. 1970. Glycylhydroxyprolyl sequences in earthworm cuticle collagen: glycylhydroxyprolylserine. Journal of Biological Chemistry **245(20):** 5478-5483.

Hacker, J., Blium-Oehier, G., Muhldorfer, I., and Tschape, H. 1997. Pathogenicity idlands of virulent bacteria: structure, function and impact on microbial evolution. Molecular Microbiology **23(6):** 1089-1097.

Hague, N.G.M. and Growen, S.R. 1987. Chemical control of nematodes. Sydney, Australia: Academic.131-178.

Hewlett and Dickson, 1993. A centrifugation method for attaching endospores of *Pasteuria* spp. to nematodes. Journal of Nematology **25:** 785-788.

Kerry, B.R. 2001. Exploitation of the nematophagous fungus *Verticullium chlamydosporium* Goddard for the biological control of root-knot nematodes (*Meloidogyne* spp.). Butt, T.M., Jackson, C., and Magan, N. (Eds). Fungi as biocontrol agents: progress, problems and potential. Wallingford, UK, CABI Publishing: 155-168.

Kers, J.A., Cameron, K.D., Joshi, M.V., Bukhalid, R.A., Morello, J.E., Wach, M.J., Gibson, D.M., and Loria, R. 2004. A large, mobile pathogenicity island confers plant patholgenicity on *Streptomyces* species. Molecular Microbiology.

Klasson, L. Andersson, S.G.E. 2004. Evolution of minimal-gene-sets in host dependent bacteria. TRENDS in Microbiology **12(1):** 37-43.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., *et al.* 1997. The comlete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature **390:** 249-256.

Lai, E., Phadke, N.D., Kachman, M.T., Giorno, R., Vazquez, S., Vazquez, J.A., Maddock, J.R., and Driks, A. 2003. Proteomic analysis of the spore coats of *Bacillus subtilis* and *Bacillus anthracis*. Journal of Bacteriology **185(4):** 1443-1454.

Lemos, E.G., Alves, L.M., and Campanharo, J.C. 2003. Genomics-based design of defined growth media for the plant pathogen *Xylella fastidiosa*. FEMS Microbiology Letters **219(1):** 39-45.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. and Spratt, B.G. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences (USA) **95:** 3140-3145.

Mankau, R. 1975. *Bacillus penetrans* n. comb. causing a virulent disease of plant-parasitic nematodes. Journal of Invertebrate Pathology **26: 333-339.**

Mankau, R. and Prasad, N. 1972. Possibility and problems in the use of a sporozoan endoparasite for biological control of plant parasitic nematodes. Nematropica **2:** 7-8.

Mann, K., Mechling, D.E., Bächinger, H.P., Eckerskorn, C., Gaill, F., and Timpl, R. 1996. Glycosylated threonine but not 4-hydoxyproline dominates the triple helix stabilizing positions in the sequence of a hydrothermal vent worm cuticle collagen. J. Mol. Biol. **261:** 255-266.

Mateille, T., Trudgill, D.L., Trivino, C., Bala, G., Sawadogo, A., and Vouyoukalou, E. 2002. Multisite survey of soil interactions with infestation of root-knot nematodes (*Meloidogyne* spp.) by *Pasteuria penetrans*. Soil Biology and Biochemistry **34:** 1417-1424.

Metchnikoff, E. 1888. *Pastueria ramose*, un représentant des bactéries à divisions longitudinale. Ann. Inst. Paster (Paris) **2**: 165-170.

Moran, N.A. 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell **108:** 583-586.

Morton, C.O., Hirsch, P.R, and Kerry, B.R. 2004. Infection of plant-parasitic nematodes by nematophagous fungi – a review of the application of molecular biology to understand infection processes and to improve biological control. Nematology **6(2):** 161-170.

Oka, Y., Chet, I., Mor, M., and Spiegel, Y. 1997. A fungal parasite of *Meloidogyne javanica* eggs: Evaluation of its use to control the root-knot nematode. Biocontrol Science and Technology **7:** 489-497.

Oostendorp, M., Dickson, D.W., and Mitchell, D.J. 1990. Host range and ecology of isolates of *Pasteuria* spp. from the southeastern United States. Journal of Nematology **22(4):** 525-531.

Oostendorp, M., Dickson, D.W., and Mitchell, D.J. 1991. Population development of *Pasteuria penetrans* on *Meloidogyne arenaria*. Journal of Nematology **23:** 58-64.

Orui, Y. and Ozawa, H. 1999. Identification of three major *Meloidogyne* Species by stained spores of host-specific *Pasteuria penetrans* isolates. Applied Entomology and Zoology **34(2):** 195-203.

Pamilo, P., Nei, M. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution **5:** 568-583.

Phillips, Z.E., and Strauch, M.A. 2002. *Bacillus subtilis* sporulation and stationary phase gene expression. Cell. Mol. Life Science **59(3):** 392-402.

Privalov, P.L. 1982. Stability of proteins. Proteins which do not present a single cooperative system. Advances in Protein Chemistry **35**:1-104.

Preston, J.F., Dickson, D.W., Maruniak, J.E., Brito, J.A., Schmidt, L.M. and Giblin-Davis, R.M. 2003. *Pasteuria* spp., Sytematics and phylogeny of these bacterial parasites of phytopathogenic nematodes. Journal of Nematology **35:** 198-207.

Rasmussen, M., Jacobsson, M. and Björck, L.2003. Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. Journal of Biological Chemistry **278(34):** 32313-32316.

Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., *et al.* 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature **423:** 81-86.

Riese, R.W., Hackett, K.J., Sayre, R.M., and Huettel, R.N. 1998. Factors affecting cultivation of three isolates of *Pasteuria* spp. Journal of Nematology **20:** 657.

Rokas, A., Williams, B.L., King, N. and Carroll, S.B. 2003a. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425:** 798-804.

Rokas, A., King, N., Finnerty, J., and Carroll, S. 2003b. Conflicting phylogenetic signals at the base of the metazoan tree. Evolution and Development **5:** 346-359.

Sasser, J. N., and Freckman, D. W. 1987. A world perspective on nematology: The role of the society. Pages 7–14 in: Vistas on Nematology. J. A. Veech and D. W.Dickson, eds. Society of Nematology, Inc., Hyattsville, MD, U.S.A.

Sayre, R. M., and M. P. Starr. 1985. *Pasteuria penetrans* (ex Thorne 1940) nom. rev., comb. n., sp. n., a mycelial endospore-forming bacterium parasitic in plant-parasitic nematodes. Proceedings of the Helminthological Society of Washington **52:** 149-165.

Sayre, R. M., and M. P. Starr. 1989. Genus *Pasteuria* Metchnikoff, 1888. S.T. Williams (ed.), Bergey's Manual of Systematic Bacteriology **4:** 2601-2615.

Scholl, E.H., Thorne, J.L., McCarter, J.P., Bird, D.McK., 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biology **4:** R39.

Steichen, C., Chen, P., Kearney, J.F., and Turnbough, C.L. Jr. 2003. Identification of the immunodominant protein and other proteins of the *Bacillus anthracis* expsporium. Journal of Bacteriology **185(6):** 1903-1910.

Stirling, G.R. 1985. Host specificity of *Pasteuria penetrans* within the genus *Meloidgyne*. Nematologica **31:** 203-209.

Stirling, G.R., Bird, A.F., and Cakurs, A.B. 1986. Attachment of *Pasteuria penetrans* spores to the cuticles of root-knot nematodes. Revue de N'ematologie **9:** 251-260.

Stirling, G.R. 1988. Biological control of plant parasitic nematodes. Pp. 93-139 in G.O. Poinar and H.-B Jansson, eds. Diseases of nematodes, vol. 2. Boca Raton, FL: CRC Press.

Stirling, G.R. and Smith, L.J. 1998. Field tests of formulated products containing either *Verticillium chlamydosporium* or *Arthrobotrys dactyloides* for biological control of root-knot nematodes. Biological Control **11:** 231-239.

Slyvestre, P., Couture-Tosi, E., and Mock, M. 2002. A collagen-like surface glycoprotein is a structural component of the *Bacillus anthracis* exosporium. Molecular Microbiology **45(1):** 169-178.

Slyvestre, P., Couture-Tosi, E., and Mock, M. 2003. Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in the exosporium filament length. Journal of Bacteriology **185(5):** 1555-1563.

Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Cirama, C., Nakamura, Y., Ogasawara, N., Kuhara, S., and Horikoshi, K. 2000.

Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and gemonic sequence comparison with *Bacillus subtilis*. Nucleic Acids Research **28(21):** 4317-4331.

Thames, W.H. and Stoner, W.N. 1953. A preliminary trial of lowland culture rice in rotation with vegetable crops as a means of reducing root-knot nematode infestations in the Everglades. Plant Disease Reporter **37:** 187-192.

Thorne, G. 1940. *Duboscqia penetrans* n. sp. (Sporozoa: Microsporidia, Nosematidae), a parasite of the nematode *Pratylenchus pratensis* (de Man) Filipjev. Proc. Helminthol. Soc. Washington **7:** 51-53.

Trotter, J.R., Bishop, A.H. 2003. Phylogenetic analysis and confirmation of the endospore-forming nature of *Pasteuria penetrans* based on the spo0A gene. FEMS Microbiology Letters **29:** 249-256.

Trudgill, D.L. 1991. Resistance to and tolerance of plant parasitic nematodes in plants. Annual Review of Phytopathology **29:** 167-192.

Tzortzakakis, E.A., de R Channer, A.G., Gowen, S.R., and Ahmed, R. 1997. Studies on the potential use of *Pasteuria penetrans* as a biocontrol agent of root-knot nematodes (*Meloidogyne* spp.). Plant Pathology **46(1):** 44-55.

Urwin, R. and Maiden, M.C.J. 2003. Multi-locus sequences typing: a tool for global epidemiology. Trends in Microbiology **11:** 479-487.

Waite, J.H., Tanzer, M.L., and Merkel, J.R. 1980. *Nereis* cuticle collagen: proteolysis by marine vibrial and clostridial collagenases. Journal of Biological Chemistry **255(8):** 3596-3599.

Wei, J.Z., Hale, K., Carta, L., Platzer, E., Wong, C., Fang, S.C, and Aroian, R.V. 2003. *Bacillus thuringiensis* crystal proteins that target nematodes. Proceedings of the National Academy of Sciences (USA) **100(5):** 2760-2765.

Williams, A.B., Stirling, G.R., Hayward, A.C., and Perry, J. 1989. Properties and attempted culture of *Pasteuria penetrans*, a bacterial parasite of root-knot nematodes (*Meloidogyne javanica*). Journal of Applied Bacteriology **67:** 145-156.

Xu, Y., Keene, D.R., Bujnicki, J.M., Höök, M. and Lukomski, S. 2002. Streptococcal Scl1 and Scl2 proteins form collagen-like triple helices. Journal of Biological Chemistry **277(30):** 27312-27318.

Yarwood, J.M., McCormick, J.K., Paustian, M.L, Orwin, P.M, Kapur, V., and Schlievert, P.M. 2002. Characterization and expression analysis of *Staphylococcus aureus* pathogenicity island 3 – implications for the evolution of staphylococcal pathogenicity islands. Journal of Biological Chemistry **277(15):** 13138-13147.

Young, J.M. 2001. Implications of alternative classifications and horizontal gene transfer for bacterial taxonomy. International Journal of Systematic and Evolutionary Microbiology **51:** 945-953.

Yang, W., Chan, V.C., Kirkpatrick, A., Ramshaw, J.A.M., and Brodsky, B. 1997. Gly-Pro-Arg congers stability similar to Gly-Pro-Hyp in the collagen triple-helix of host-guest peptides. Journal of Biological Chemistry **272 (46):** 28837-28840.

Zuckerman, B.M., Dicklow, M.B., and Acosta, N. 1993. A strain of *Bacillus thuringiensis* for the control of plant-parasitic nematodes. Biocontrol Science and Technology **3:** 41-46.

Phylogenetic Analysis of *Pasteuria penetrans* using multiple genetic loci

Running title: *P. penetrans* phylogenetic analysis

Lauren Charles[1], Ignazio Carbone[2], Keith G. Davies[3], David Bird[1], Mark Burke[1], Brian R. Kerry[3], & Charles H. Opperman[1]*

[1]Center for the Biology of Nematode Parasitism
Department of Plant Pathology
North Carolina State University
Raleigh, NC 27606, USA

[2]Center for Integrated Fungal Research
Department of Plant Pathology
North Carolina State University
Raleigh, NC 27606, USA

[3]Nematode Interactions Unit
Rothamsted Research, Ltd.
Harpenden, Herts. AL5 2JQ, UK

*Correspondence should be addressed to C.H.O. (email: warthog@ncsu.edu; fax: 919.515.9500; phone: 919.515.6699)

Pasteuria penetrans is a gram-positive, endospore-forming eubacterium, apparently in the Bacillus-Clostridium clade. It is an obligate parasite of root-knot nematode (*Meloidogyne* spp.) and preferentially grows on the developing ovaries, inhibiting reproduction. Root-knot nematodes are devastating root pests of economically important crop plants and are difficult to control. Consequently, *P. penetrans* has long been recognized as a potential biocontrol agent for root-knot nematode, but the fastidious life cycle and the obligate nature of parasitism have inhibited progress on mass culture and deployment. We are currently sequencing the genome of the *Pasteuria* bacterium and have performed amino acid level analyses of a total of 33 bacterial species (including *P. penetrans*) using the concatenation of forty housekeeping genes, with and without insertions/deletions (indels) removed, and with each gene individually. Through the application of maximum likelihood, maximum parsimony and Bayesian methods on these datasets, *P. penetrans* was found to cluster tightly, with a high level of confidence, within the class *Bacilli* of the gram-positive, low G+C content eubacteria. Strikingly, our analyses placed *P. penetrans* as ancestral to *Bacillus* spp. Additionally, all analyses reveal that *P. penetrans* is surprisingly more closely related to the saprophytic extremophile *Bacillus haladurans* and *B. subtilis* than to the pathogenic *B. anthracis* and *B. cereus*. Collectively, these findings strongly imply that *P. penetrans* is an ancient member of the Bacillus group. We further suggest that *P. penetrans* may have evolved from an ancient symbiotic bacterial associate of nematodes, possibly as the root-knot nematode evolved to be a highly specialized parasite of plants.

Introduction

Pasteuria penetrans is an endospore-forming, gram-positive, obligate parasitic bacterium of root-knot nematode (RKN), *Meloidogyne* spp. RKN have a very broad host range, encompassing more than 2,000 plant species, and most cultivated crops are attacked by at least one species of *Meloidogyne* (33), causing economic losses greater than $50 billion per annum. The problem in the sub-tropics and tropics is particularly severe, and many developing nations are seriously impacted in both food security and economics by RKN. Mature female RKN release hundreds of eggs into a proteinaceous matrix on the surface of the root. Following a first molt in the egg, motile second-stage (J2) juveniles hatch in the soil and typically re-infect the same plant. RKN J2 destructively penetrate the root, preferentially in the zone of elongation or at the site of a lateral root emergence, and migrate intercellularly into the vascular cylinder, causing little or no injury. Once in the vascular cylinder, the nematode makes a commitment to establish a highly specialized feeding site, referred to as a giant cell. The relationship between RKN and its host is both intimate and complex and involves dramatic changes both in plant and nematode, leading to giant cell induction and gall formation. The *Meloidogyne* J2 is a non-feeding, developmentally arrested, long-lived dispersal stage, and can survive in the soil for weeks or even months on stored lipid reserves, and this is the nematode stage exposed to *P. penetrans* spores in the soil.

The life-cycle of this bacterium has co-evolved with its host and begins when J2 migrating through the soil become encumbered with endospores. The endospores do not germinate until the J2 has entered the plant root and established a feeding site. Sometime between the establishment of a feeding site and the second nematode molt, the endospore

germinates and produces rhizoids which extend throughout the developing nematode. The rhizoids eventually produce bacterial rods that undergo rapid exponential growth, resulting in degeneration of the nematode's reproductive tract and inhibition of egg production. Sporogenesis is triggered in a manner similar to other *Bacilli* and involves the initiation of a metal-ion sensitive, phospho-relay pathway (16) resulting in the production of endospores. Because of both its high efficacy and host specificity, *P. penetrans* represents a potentially ideal biological control agent of these economically important crop pests (37, 26, 5, 43). However, the obligate nature of the bacterium's life-style and its host specificity have rendered it difficult to develop into a commercial product. For these reasons, a genomic approach has recently been undertaken to help understand mechanisms of parasitism of *Pasteuria* spp. and the possible exploitation of their ecological niche.

There have been many attempts to classify this group of bacteria since first described as *Pasteuria ramosa* by Metchnikoff (23) in 1888 on *Daphnia*, a waterflea. In 1906, Cobb (6) studied the morphology of this parasite on the nematode *Dorylaimus bulbiferous* and claimed it should be placed among the protozoa. This change was later accepted and the bacterium renamed by Thorne (40) in 1940 as *Duboscqia penetrans*. Since then, electron microscope techniques have shown the bacteria to be more *Bacillus*-like rather than a protozoan and hence, it was renamed again, as *B. penetrans* (21). In 1985, there was a rediscovery of Metchnikoff's work, suggesting similarities to the original *P. ramosa,* which led to a reversion of the genus back to *Pasteuria* by Sayre and Starr (34). Because *P. penetrans* is a gram-positive, mycelial and endospore-forming bacterium, it was classified in the *Actinomycetales* (34).

The use of multilocus sequence typing (MLST) exploits the variation that is slowly accumulating in housekeeping genes within populations and offers a selectively neutral method for characterizing bacteria (20). This approach has been shown to be extremely useful in characterizing pathogenic strains (44). The use of enzymes in multilocus enzyme electrophoresis (MLEE) was never exploited to characterize *P. penetrans*, due in large part to its obligate parasitic nature and the difficulty in separating *Pasteuria* enzymes from host enzymes. However, the acquisition of genome sequence data from *P. penetrans* makes MLST a feasible approach and provides a method to infer the organism's closest relatives along with its phylogenetic history.

Automated sequencing techniques have been applied to utilize single genes to re-evaluate the classification of this bacterium. Studies using 16S rDNA and rRNA indicated that *Pasteuria* spp., including *P. penetrans,* belonged to the *Clostridium-Bacillus-Streptococcus* branch of gram-positive eubacteria (2, 3, 28). This has recently been supported through the use of the spo0A gene, placing *Pasteuria* with members of the supergenus *Bacillus* (42). However, phylogenetic analytical procedures relying on only one gene at a time to classify a species have been found to be inaccurate in many cases, while the use of multi-locus approaches have been shown to be more reliable (12, 14, 27, 30, 31, 35, 48).

To resolve the placement of *Pasteuria penetrans* among bacteria, a deep phylogenetic analysis was performed on 40 housekeeping genes separately and concatenated. Further analyses were performed with a subset of 27 of these genes. Various phylogenetic methods were used, including Bayesian, maximum likelihood and maximum parsimony, to further strengthen the validity of the analysis.

Methods

Bacterial species from a supertree by Daubin *et al.* (7) were searched in the NCBI database (45) to identify completely sequenced genomes. Completely sequenced genomes are necessary to ensure that the genes collected are orthologous. These species, whose phyla are collectively distributed throughout the superkingdom Bacteria, were chosen to avoid taxon bias in the resulting trees (4).

In previous single-gene phylogenetic analyses, *P. penetrans* has been demonstrated to fall among the *Bacilli* clade (2, 3, 42), but the specific relationships are still unresolved. Kobayashi *et al.* (18) in 2003 identified the genes essential to *Bacillus subtilis*. Since this list of genes contains independent, housekeeping genes that are known to be slow evolving, these genes provide the highest probability of giving the most accurate and unbiased tree when concatenated (30). These chosen genes were blasted using a tblastn query (1) and Blossum62 matrix against *P. penetrans* unpublished sequence data with restrictions of a maximum e-value of 1e-20 and minimum bit score of 100. Those genes recovered were then evaluated for evidence of horizontal gene transfer (HGT) based on G+C content, codon usage, amino acid usage and gene position (15); potential HGT candidates were discarded. The remaining genes were then blasted in NCBI against the bacterial species chosen using the above criteria and blast tools. These blast results were used as a guide to identify full gene sequences within the respective species genome. Those genes, 40 in total, which were found in the most species and contained the longest sequence identities, ranging from 50 to over 400 amino acids, were chosen for analysis (Table 1 and 2).

The multiple sequence alignments were done using both the global alignment program ClustalW (39) and the local alignment program DiAlign (24). The best alignment of the two was chosen and hand-eye adjustments were made in GeneDoc (25) until sequence identities were at least 30 percent. This was done by pruning any region violating position homology in the multiple alignment and also excluding species with a consistently low sequence identity. These adjustments were the basis for delimiting the start and stop positions of the gene sequences in each alignment (Table 2). Since the species have a high degree of diversity, the alignments were performed at the amino acid level.

Phylogenetic analyses were then conducted on these amino acid sequence alignments; see below for parameters used in each program. The first analysis was performed on each gene separately using maximum likelihood and maximum parsimony. Then, separate alignments were concatenated using SNAP Workbench (29) for a total of 6,036 amino acids and analyzed again using the same programs. Two different subsets of this concatenation, both consisting of the same 28 taxa and 27 genes, were analyzed using Bayesian and maximum likelihood methods, as well as maximum parsimony. These subsets were created to minimize the potential of phylogenetic artifacts, such as unequal taxon sampling bias, derived from the inclusion of species and genes with large amounts of missing data. By including only those species containing at least 27 genes, the taxon number was reduced to 28. The species that were removed included all those in the phyla Crenarchaeota and Euryarchaeota. After taxa were removed, the multiple sequence alignment of each gene was reexamined and kept if there were no regions in the alignment that disrupted positional homology (Table 2). The largest of these subsets

included all 27 full gene sequences totaling 4,236 amino acids. The other subset had all of the insertions/deletions (indels) removed using SNAP Workbench, leaving the most conservative dataset with a total of 3,032 amino acids.

The phylogenetic analyses for all datasets were performed as described below. In all cases, a strict consensus tree was inferred to avoid placing emphasis on any one topology. Gaps were treated as missing data or ambiguous characters, and no molecular clock was assumed. Maximum likelihood analysis using PAML 3.12 (47) inferred a single tree via stepwise addition. The discrete-gamma model and empirical amino acid substitution matrix of Jones *et al.* (17) were used to perform the analysis. All model parameters (e.g. amino acid substitution rates) were estimated empirically from the data. Maximum parsimony (MP) searches in MEGA 2.1 (19) were performed using close-neighbor-interchange (CNI), uniform weighting and 5,000 bootstrap replicates. The CNI initial trees were selected via random addition trees with 10 replications each and a search level of 3. Bayesian analysis using MrBayes (32) was performed to allow the most heterogeneity among sites while using similar starting parameter values as in the maximum likelihood analyses. We used gamma-distributed rates across sites with substitutions occurring according to a time-reversible model. A uniformly shaped prior was chosen, which included all topologies equally probable, *a priori,* and unconstrained branch lengths. The rate matrix for the prior was fixed to the Jones-Taylor-Thornton model for consistency with the ML analyses. One cold chain and three heated chains were used with a sample frequency set to 50. The Markov chain Monte Carlo analysis ran for 10,000 generations for each analysis due to the complexity of the protein dataset.

Results

All phylogenetic analyses performed on 40 housekeeping genes of 33 bacterial species consistently placed *Pasteuria penetrans* between *Bacillus halodurans* and *Staphylococcus aureus* within the gram-positive, low G-C content *Bacilli* with good support (Fig.1-3).

The initial maximum likelihood and maximum parsimony analyses performed on each gene separately gave varying results. For the majority of the trees, *P. penetrans* falls within the low G+C content, gram-positive clade, as expected. There were, however, seven instances where the placement of bacteria within the phylogeny was seemingly random (Table 1). For example, *P. penetrans* is shown to be most closely related to proteobacteria for the *groEL* gene, with a bootstrap value of 99 (Fig. 4) and to the high G+C content, gram-positive bacteria for the gene *eno*, with a bootstrap value of 90 (Fig. 5). There were also a few clades with strong support in one gene tree that were incongruent with clades supported by data in other gene loci. The genes with the most consistent phylogenetic placement of species, with some minor incongruencies within clades, coincide with the subset of genes chosen for the concatenation analyses. When examining the maximum ln likelihood of these individual gene trees (Table 3), it is apparent that trees inferred from alignments that are similar in length have log likelihood values similar in magnitude even though the overall tree topologies and the specific placement of species differ.

The inferred maximum likelihood tree from the concatenation of all 40 genes was concordant with 33 of the 40 individual gene trees (Fig 2). The tree for the combined data set resolved accepted clades of species within the same phylum. Each clade was

further split into smaller clades corroborating with widely accepted phylogenetic and taxonomic relationships (7, 45). The placement of *P. penetrans* within this phylogeny agrees with the initial hypothesis as being a member of the class *Bacilli* and resolves the exact placement of the bacterium between *B. halodurans* and *S. aureus*.

The next series of phylogenetic analyses were performed on 27 housekeeping genes and 28 bacterial species. These genes were concatenated and analyzed using maximum likelihood and maximum parsimony methods (Fig 3). The phylogenetic inferences from each of these methods were consistent with the placement of *Pasteuria* ancestral to *Bacillus* and the *Lactococcus/Streptococcus* clade. As shown in Fig. 3, maximum parsimony does not provide as much support of phylogenetic relationships in the inferred tree as do the parametric methods (Fig. 1 and 2).

The most conservative analysis was performed on the set of 27 genes with all indels removed. This dataset was then examined using Bayesian, maximum likelihood and maximum parsimony methods. All three phylogenies placed *P. penetrans* within the *Bacilli* clade, between *B. halodurans* and *S. aureus* (Fig 1). The posterior probability was 80 and the bootstrap value was 85. Importantly, the results from each of these analyses mirrored those of their full gene sequences and the initial dataset of 40 genes and 33 species.

Discussion

The results of the concatenated species tree analysis place *Pasteuria penetrans* among the *Bacilli* clade, between *B. halodurans* and *S. aureus* (Fig 1 and 2). This result

suggests that *Pasteuria* is ancestral to other *Bacillus* spp. Within this group, *P. penetrans* is most closely related to the saprophytic extremophile *B. haladurans* and its close saprophytic relative *B. subtilis,* than to the pathogenic species *B. anthracis* and *B. cereus*. Unfortunately, at the time of this analysis the genome of *B. thuringiensis* was not available. These results were independent of which phylogenetic method was used and whether the data set contained 40 or 27 genes with or without indels.

For most of the phylogenetic trees based on single genes, the inferred clades are also found in the concatenated tree with slight variations in exact clade positions. However, for some of the genes, especially those with low species counts (Table 2), the positioning of *P. penetrans* varies (Table 1, Fig.4 and 5). The incongruence observed here is a typical result for single-gene phylogenies (30, 31). This is due to a number of biological and methodical factors that are related to gene evolution and phylogenetic analyses, such as rate of evolution, variable sites, gene size, taxon bias, base composition and number of parsimony-informative sites. A contributing factor to the differences seen in these single-gene phylogenies, which was also why these particular genes were selected, is the fact that they are mostly unlinked (Table 1). This independence between genes can lead to differences in evolutionary changes as noted above causing the placement of the species to differ in each tree. These results strengthen the argument that gene trees do not always accurately infer species evolutionary relationships even when slow evolving orthologs are used (4, 12, 27, 30). Further analyses are needed to reveal the cause for incongruencies observed in the individual trees. For now, our results indicate that single-gene trees may not be precise enough to infer accurate species evolutionary relationships for characterization of *Pasteuria* isolates.

The results of the earliest DNA studies to characterize *Pasteuria* all relied on using bacterial universal primers based on the 16s ribosomal RNA subunit and sequencing the PCR amplified products. These studies all showed *Pasteuria* to be closely related to endospore-producing *Bacilli* (2, 3). The formation of endospores is one of the most complex developmental processes in prokaryotes and involves hundreds of sporulation-specific genes (36). Genomic analysis has shown that many of the genes in the sporulation pathway of these bacteria are conserved, and several recent reports have started to use some of these genes to construct phylogenies. The phylogenetic relationship using *sigE* and *sigF* showed *Pasteuria* to be ancestral to other *Bacilli* and taxonomically in the middle of the *Bacilli,* respectively (28). In a separate study using *spo*0A, *P. penetrans* and *P. ramosa* showed *Pasteuria* to be rooted deeply within the supergenus *Bacilli*. As might be expected, phylogenies based on single genes are likely to produce different results. The results presented here indicate that individual gene trees, although each robustly supported, showed great variation in their placement of species.

We have shown that a concatenation of several genes infers a robust phylogenetic tree, regardless of which analysis method is used. By combining amino acid sequence alignments, the phylogenetic signal is increased and noise is minimized, creating a more accurate phylogeny (4). Maximum likelihood is known to take into account many of the underlying issues concerning amino acid sequence evolution, such as composition bias and rate heterogeneity. It has been reported to be the better choice over maximum parsimony analysis (41) and Bayesian methods (31), but in these analyses, there were no apparent differences in the deep phylogenies inferred. Unfortunately, due to the

computational difficulty for such a large protein dataset, Bayesian inference was only employed on the smallest, least ambiguous dataset. The results from this analysis did, however, mirror those of the maximum likelihood and parsimony analyses. Furthermore, different inference methods were used to reduce the potential of violating the assumptions of a particular method, which may result in an erroneous tree. For example, Bayesian posterior probabilities are thought to overestimate the reliability of a tree, whereas bootstrap values are believed to be an underestimation (38). Our inference of the same tree from a few robust phylogenetic methods increases the credibility of the consensus tree, as demonstrated here.

Not only did the different programs result in the same tree, the same tree was inferred when 40 genes with 33 taxa or 27 genes, with or without indels, and 28 taxa were used. In the smaller analysis, the species that were removed contained the fewest number of genes used in the analysis. Incidentally these all belong in the phylum Crenarchaeota and Euryarchaeota. The taxon selection, in this case, had no influence in branch resolution as otherwise found in the metazoan tree by Rokas and coworkers (31). Likewise, the number of genes, whether 40 or 27, had no effect on the exact branch placement of the species in question. Therefore, both the larger and smaller datasets were adequate in this case, to resolve the tree correctly as opposed to the single-gene analysis (30). Since the number of amino acid in an alignment drastically affects the amount of time for the analysis, running the smaller dataset greatly improves the efficiency of the phylogenetic inference method. Including areas of ambiguity (e.g., indels) also makes the analyses increasingly difficult to compute. Since identical results came from the dataset with and without indels, no extra information was gained from them, and

therefore, they were not necessary. Consequently, there was no loss of information or resolution when removing indels, genes with low species counts and species with low gene counts. There was only a gain of efficiency and computability in the dataset.

Our resolution of the *P. penetrans* phylogenetic position will enable us to exploit comparative genomics to study its relationship to its closest relatives in the *Bacilli* clade and to discover the genetic basis for parasitism of the root-knot nematode. The analysis undertaken so far has been based on genes extracted from a single population of *Pasteuria* (RES147). It is well documented that while some populations of *P. penetrans* attach only to a particular species of RKN (37, 22, 13), others are specific for a particular population within a species of the nematode (37). This intra-specific biological variation and its possible role in horizontal gene transfer may be revealed through genetic comparison to other closely related pathogenetic bacteria identified in this study. It is also interesting to point out that the previous suggestion that *P. penetrans* may be an ancestral member of the *Bacilli*, and not recently evolved, is strongly supported by the results presented herein. Most significantly, the parasitic *P. penetrans* is more closely related to the saprophytic *Bacilli* than to the animal pathogens.

Acknowledgments

This work was supported by the North Carolina Agricultural Research Service and Rothamsted Research, Ltd.

Literature Cited

1. **Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and D. J. Lipman.** 1990. Basic local alignment search tool. J Mol Biol **215**:403-410.
2. **Anderson, J. M., Preston, J. F., Dickson, D. W., Hewlett, T. E., and J.E. Maruniak.** 1999. Phylogenetic analysis of *Pasteuria penetrans* by 16S rRNA gene cloning and sequencing. J Nematol **31**:319-325.
3. **Atibalentja N., Noel G. R., and L.L. Domier.** 2000. Phylogenetic position of the North American isolate of Pasteuria that parasitizes the soybean cyst nematode, *Heterodera glycines*, as inferred from 16S rDNA sequence analysis. Int J Syst Evol Microbiol **50**:605-13.
4. **Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and W. F. Doolittle.** 2000. A kingdom-level phylogeny of Eukaryotes based on combined protein data. Science **290**:972-976.
5. **Chen, Z. X., Dickson, D. W., McSorley, R., Mitchell, D. J., and T. E. Hewlett.** 1996. Suppression of *Meloidogyne arenaria* race 1 by soil applications of endospores of *Pasteuria penetrans*. J of Nematol **28**:159-168.
6. **Cobb, N. A.** 1906. *Fungus Maladies* of the sugar cane, with notes on associated insects and nematodes, 2nd ed. Hawaiian Sugar Planters Assoc. bulletin no. 5, Hawaiian Sugar Planters Association, Honolulu.
7. **Daubin, V., Gouy, M., and G. Perriére.** 2001. Bacterial molecular phylogeny using supertree approach. Genome Informatics **12**:155-164.

8. **Davies, K. G., and M. Redden.** 1997. Diversity and partial characterization of putative virulence determinants in *Pasteuria penetrans*, the hyperparasite of root-knot nematodes. J Appl Microbiol **83**:227-235.
9. **Davies, K. G., Fargette, M., Balla, G., Daudi, A., Duponnois, R., Gowen, S. R., Mateille, T., Phillips, M. S., Sawadogo, A., Trivino, C., Vouyoukalou, E., and D.L. Trudgill.** 2001. Cuticle heterogeneity as exhibited by *Pasteuria* spore attachment is not linked to the phylogeny of parthenogenetic root-knot nematodes (*Meloidogyne* spp.). Parasitology **122**:111-120.
10. **Davies, K. G., Redden, M., and T. K. Pearson.** 1994. Endospore heterogeneity in *Pasteuria penetrans* related to attachment to plant-parasitic nematodes. Lett Appl Microbioly **19**:370-373.
11. **Davies, K. G., Robinson, M. P., and V. Laird.** 1992. Proteins on the surface of spores of *Pasteuria penetrans* and their involvement in attachment to the cuticle of second-stage juveniles of *Meloidogyne incognita.* J Invert Pathol **59**:18-23.
12. **Doyle, J. J.** 1992. Gene trees and species tree: Molecular systematics as one-character taxonomy. Syst Bot **17**:144-163.
13. **Duponnois, R., Fargette, M., Fould, S., Thioulouse, J., and K. G. Davies.** 2000. Diversity of the bacterial hyperparasite *Pasteuria penetrans* in relation to root-knot nematodes (*Meloidogyne* spp.) control on *Acacia holosericea*. Nematology **2**:235-442.
14. **Fox, G. E., Wisotzkey, J. D., and P. Jurtshuk, Jr.** 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Intl J Syst Bacteriol **42**:166-170.

15. **Garcia-Vallvé, S., Romeu, A., and J. Palau.** 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res **10**:1719-1725. http://www.fut.es/~debb/HGT/

16. **Grimshaw, C. E., S. Huang, C. G. Hanstein, M. A. Strauch, D. Burbulys, L. Wang, J. A. Hoch, and J. M. Whiteley.** 1998. Synergistic kinetic interactions between components of the phosphorelay controlling sporulation in *Bacillus subtilis*. Biochemistry **37**:1365-75.

17. **Jones, D. T.**, **Taylor, W. R., and J. M. Thornton.** 1992. The rapid generation of mutation data matrices from protein sequences. Computer Applic Biosci **8**:275-282.

18. **Kobayashi K., Ehrlich S. D., Albertini A., Amati G., Andersen K. K., Arnaud M., Asai K., Ashikaga S., Aymerich S., Bessieres P., Boland F., Brignell S. C., Bron S., Bunai K., Chapuis J., Christiansen L. C., Danchin A., Debarbouille M., Dervyn E., Deuerling E., Devine K., Devine S. K., Dreesen O., Errington J., Fillinger S., Foster S. J., Fujita Y., Galizzi A., Gardan R., Eschevins C., Fukushima T., Haga K., Harwood C. R., Hecker M., Hosoya D., Hullo M. F., Kakeshita H., Karamata D., Kasahara Y., Kawamura F., Koga K., Koski P., Kuwana R., Imamura D., Ishimaru M., Ishikawa S., Ishio I., Le Coq D., Masson A., Mauel C., Meima R., Mellado R. P., Moir A. Moriya S., Nagakawa E., Nanamiya H., Nakai S., Nygaard P., Ogura M., Ohanan T., O'Reilly M., O'Rourke M., Pragai Z., Pooley H. M., Rapoport G., Rawlins J. P., Rivas L. A., Rivolta C., Sadaie A., Sadaie Y., Sarvas M., Sato T., Saxild H. H., Scanlan E., Schumann W., Seegers J. F., Sekiguchi J., Sekowska A., Seror S. J., Simon M., Stragier P., Studer R., Takamatsu H., Tanaka T., Takeuchi M., Thomaides H.**

B., Vagner V., van Dijl J. M., Watabe K., Wipat A., Yamamoto H., Yamamoto M., Yamamoto Y., Yamane K., Yata K., Yoshida K., Yoshikawa H., Zuber U., and N. Ogasawara. 2003. Essential *Bacillus subtilis* genes. Proc Natl Acad Sci (USA) **100**:4678-4683.

19. **Kumar, S., Tamura, K., Jakobsen, I., and M. Nei.** 2001. MEGA2: Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Arizona, USA.

20. **Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and B.G. Spratt.** 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci (USA) **95**:3140-3145.

21. **Mankau, R.** 1975. *Bacillus penetrans* n. comb. causing a virulent disease of plant-parasitic nematodes. J Invertebr Pathol **26**:333-339.

22. **Mendoza de Gives, P., Davies, K. G., Morgan, M., and J. M. Behnke.** 1999. Attachment tests of *Pasteuria penetrans* to the cuticle of plant and animal parasitic nematodes, free living nematodes and *srf* mutants of *Caenorhabditis elegans.* J Helminthol **73**:67-71.

23. **Metchnikoff, E.** 1888. *Pastueria ramose*, un représentant des bactéries à divisions longitudinale. Ann Inst Paster (Paris) **2**:165-170.

24. **Morgenstern, B.** 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15**:211-218.

25. **Nicholas, K. B., and H.B. Nicholas, Jr.** 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Distibuted by the author.

26. **Oostendorp, M., Dickson, D. W., and D. J. Mitchell.**1991.Population development of *Pasteuria penetrans* on *Meloidogyne arenaria*. J Nematol **23**:58-64.

27. **Pamilo, P., and M. Nei.** 1988. Relationships between gene trees and species trees. Mol Biol Evol **5**:568-583.

28. **Preston, J. F., Dickson, D. W., Maruniak, J. E., Brito, J. A., Schmidt, L. M., and R.M. Giblin-Davis.** 2003. *Pasteuria* spp., Sytematics and phylogeny of these bacterial parasites of phytopathogenic nematodes. J Nematol **35:**198-207.

29. **Price, E. W., and I. Carbone.** 2005. SNAP: workbench management tool for evolutionary population genetic analysis. Bioinformatics 21:402-404.

30. **Rokas, A., King, N., Finnerty, J., and S. Carroll.** 2003. Conflicting phylogenetic signals at the base of the metazoan tree. Evol Devel **5**:346-359.

31. **Rokas, A., Williams, B. L., King, N., and S.B. Carroll.** 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425**:798-804.

32. **Ronquist, F., and J. P. Huelsenbeck.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

33. **Sasser, J.N.** 1980. Root-knot nematodes: A global menace to crop production. Plant Dis **64**:36-41.

34. **Sayre, R. M., and M. P. Starr.** 1989. Genus *Pasteuria* Metchnikoff, 1888. S.T. Williams (ed.), Bergey's Manual of Systematic Bacteriology **4**:2601-2615.

35. **Scholl, E. H., Thorne, J. L., McCarter, J. P., and D. McK. Bird.** 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biol **4**:R39.

36. **Stagier, P., and R. Losick.** 1996. Molecular genetics of *Bacillus subtilis*. Ann Rev Genet **30**:297-341.

37. **Stirling, G. R.** 1984. Biological control of *Meloidogyne javanica* with *Bacillus penetrans*. Phytopath **74**:55-60.

38. **Suzuki, Y., Glazko, G. V., and M. Nei.** 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci (USA) **99**:16138-16143.

39. **Thompson, J. D., Higgins, D. G., and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucl Acids Res **22**:4673-4680.

40. **Thorne, G.** 1940. *Duboscqia penetrans* n. sp. (Sporozoa: Microsporidia, Nosematidae), a parasite of the nematode *Pratylenchus pratensis* (de Man) Filipjev. Proc Helminthol Soc Washington **7**:51-53.

41. **Thorne, J.** 2000. Models of protein sequence evolution and their applications. Curr Op Genet Devel **10**:602-605.

42. **Trotter, J. R., and A.H. Bishop.** 2003. Phylogenetic analysis and confirmation of the endospore-forming nature of *Pasteuria penetrans* based on the spo0A gene. FEMS Microbiol Lett **29**:249-256.

43. **Trudgill, D. L., Bala, G., Blok, V. C., Daudi, A., Davies, K. G., Fargette, M., Gowen, S. R., Madulu, J. D., Mateille, T., Mwageni, W., Netscher, C., Phillips, M. S., Abdoussalam, S., Trivino, G. C., and E.Voyoulallou.** 2000. The importance of tropical root-knot nematodes (*Meloidogyne* spp.) and factors affecting the utility of *Pasteuria penetrans* as a biocontrol agent. Nematol **2**:823-845.

44. **Urwin, R., and M. C. J. Maiden.** 2003. Multi-locus sequences typing: a tool for global epidemiology. Trends in Microbiol **11**:479-487.

45. **Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., and B. A. Rapp.** 2002. Database resources of the National Center for Biotechnology Information: 2002 update. Nucl Acids Res **30:**13-16.

46. **Yang, Z.** 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39**:306-314.

47. **Yang, Z.** 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555-556.

48. **Young, J. M.** 2001. Implications of alternative classifications and horizontal gene transfer for bacterial taxonomy. Intl J Syst Evol Microbiol **51**:945-953.

Table 1: Descriptions of genes used in phylogenetic analyses.

***accA**	fatty acid biosynthesis acetyl-CoA carboxylase (α subunit)
adk	purine biosynthesis adenylate kinase
argS	tRNA synthetase arginyl-tRNA synthetase
***dapA**	diaminopimelate biosynthesis dihydrodipicolinate synthase
***dapB**	diaminopimelate biosynthesis dihydrodipicolinate reductase
dnaB	DNA replication-initiation of chromosome replication/membrane attachment
dnaE	DNA replication-DNA polymerase III (α subunit)
~*eno	glycolysis enolase
fmt	tRNA met modification methionyl-tRNA formyltransferase
glmS	aminosugar metabolism L-glutamine-D-fructose-6-phosphate amidotransferase
~*groEL	protein folding class I heat-shock protein (chaperonin)
gyrA	DNA packaging-DNA gyrase (subunit A)
gyrB	DNA packaging-DNA gyrase (subunit B)
infA	translation initiation factor IF-1
***map**	protein modification methionine aminopeptidase
metS	tRNA synthetase methionyl-tRNA synthetase
mraY	peptidoglycan biosynthesis
murB	peptidoglycan biosynthesis UDP-*N*-acetylenolpyruvoylglucosamine reductase
~*murC	peptidoglycan biosynthesis UDP-*N*-acetylmuramate-alanine ligase
murD	peptidoglycan biosynthesis UDP-*N*-acetylmuramoylalanyl-D-glutamate ligase
priA	DNA replication-primosomal replication factor Y
***racE**	peptidoglycan biosynthesis glutamate racemase

rnc	RNA modification-ribonuclease III
rplC	ribosomal protein ribosomal protein L3 (BL3)
*__rplF__	ribosomal protein ribosomal protein L6 (BL8)
~*__rplJ__	ribosomal protein ribosomal protein L10 (BL5)
rplU	ribosomal protein ribosomal protein L21 (BL20)
rpmJ	ribosomal protein ribosomal protein L36 (ribosomal protein B)
rpoB	transcription RNA-polymerase (βsubunit)
rpoC	transcription RNA-polymerase (β' subunit)
rpsB	ribosomal protein ribosomal protein S2
rpsK	ribosomal protein ribosomal protein S11 (BS11)
~*__rpsM__	ribosomal protein ribosomal protein S13
secA	secretion preprotein translocase subunit (ATPase)
sigA	transcription RNA-polymerase major σ factor
~*__tkt__	glycolysis transketolase
topA	DNA packaging-DNA topoisomerase I
~*__trxA__	thioredoxin
tsf	translation elongation factor
tyrS	tRNA synthetase tyrosyl-tRNA synthetase (major)

*Note: These genes were excluded from the concatenation analyses of the 27 genes and 28 taxa.

~Note: These single genes trees have not placed *Pasteuria penetrans* among the low GC-content, gram-positive bacteria.

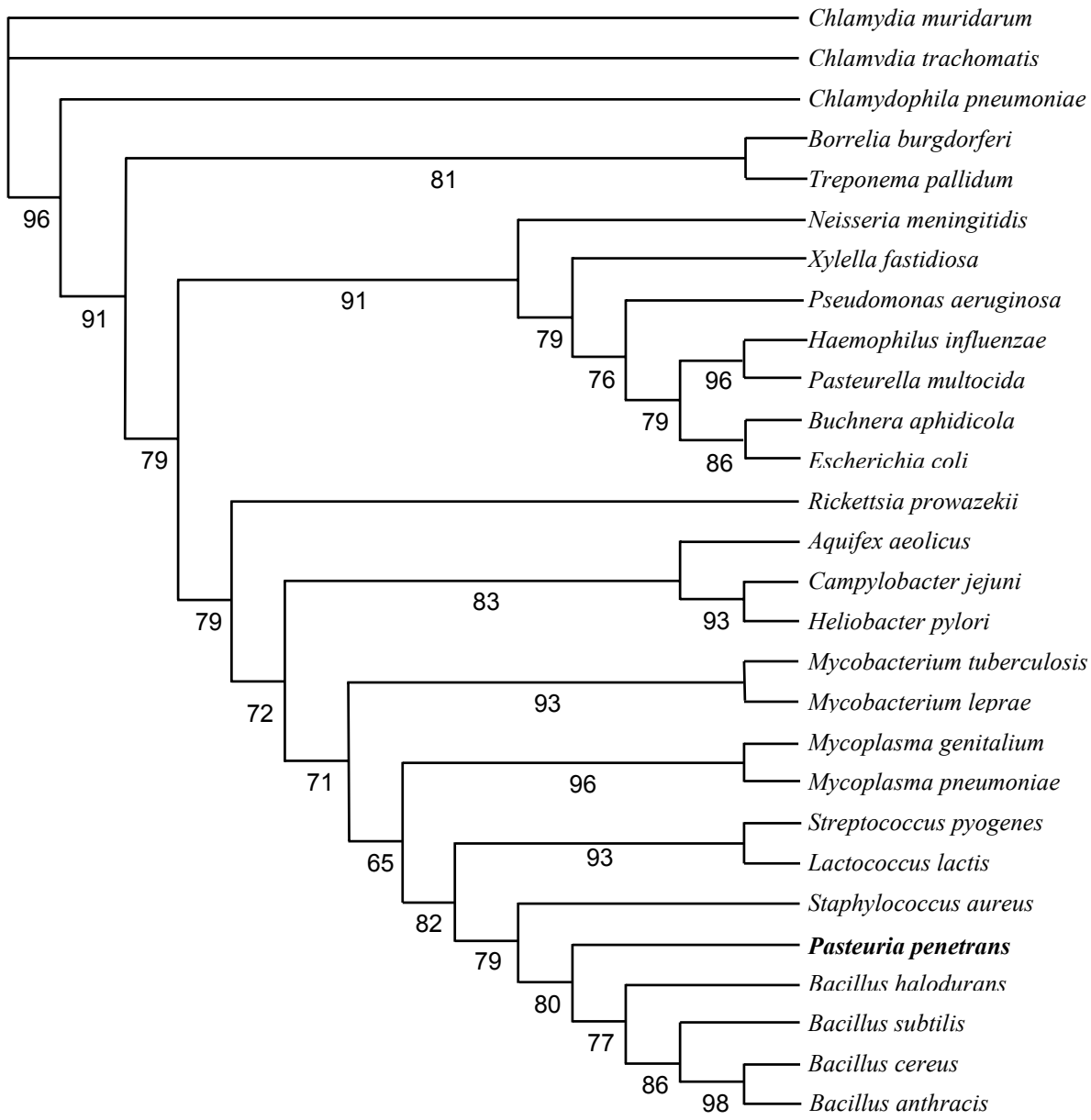

Fig. 1: Bayesian analysis of the 27 independent and concatenated genes with all indels removed for 28 bacterial species. The Bayesian posterior probabilities are shown below the branches; the Markov chain Monte Carlo analysis was run for 10,000 generations.

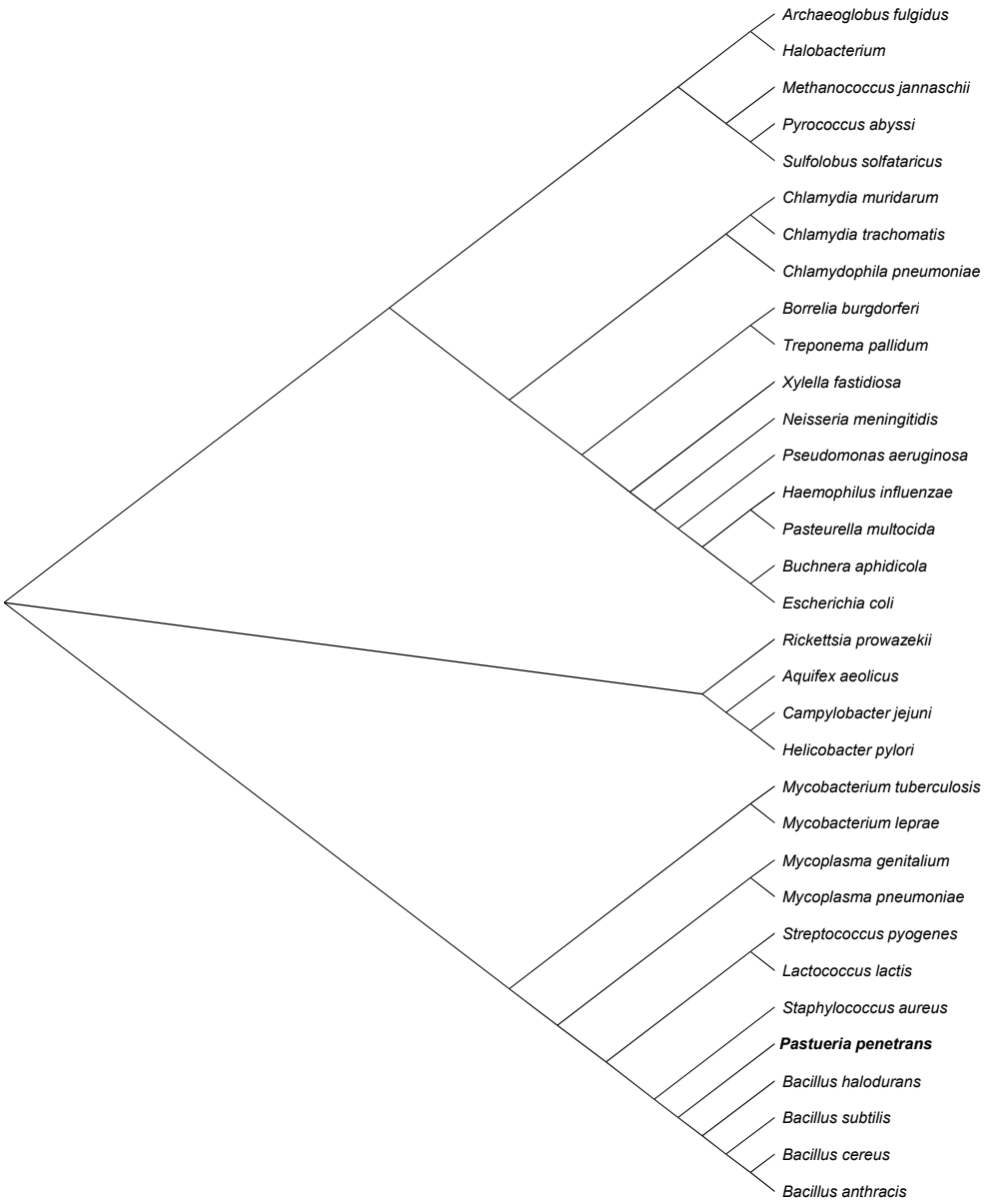

Fig. 2: Maximum likelihood analysis of the 40 concatenated housekeeping genes for 33 bacterial species.

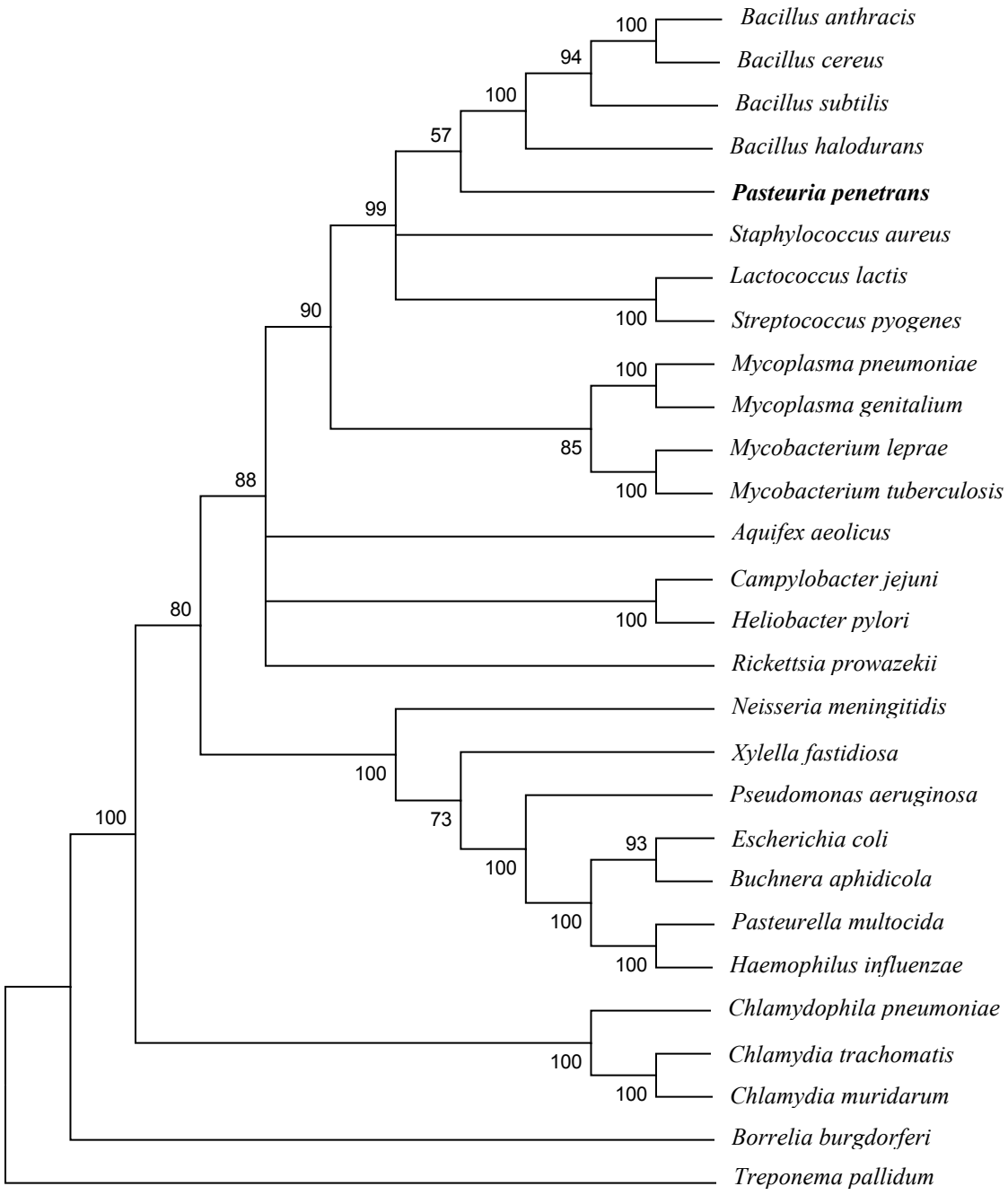

Fig. 3: Maximum parsimony analysis for 28 taxa and a concatenation of 27 full-length genes; only bootstrap values over 50% are shown.

	accA	ack	argS	**dapA**	**dapB**	dnaB	dnaE	**eno**	fmt	glmS	**groEL**	gyrA	gyrB	infA	**map**	metS	mraY	murB	**murC**	murD	priA	**racE**	rnc	rplC	**rplF**	**rplJ**	rplU	rpmJ	rpoB	rpoC	rpsB	rpsK	**rpsM**	secA	sigA	**tkt**	topA	**trxA**	tsf	tyrS	Total for 40 genes
Borrelia burgdorferi	**0**	1	1	**0**	**0**	1	1	**1**	1	0	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**0**	1	**1**	1	1	35
Treponema pallidum	**0**	1	0	**0**	**0**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	36
Chlamydophila pneumoniae	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**0**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**0**	1	**1**	1	1	38
Chlamydia muridarum	**1**	1	0	**0**	**1**	1	1	**1**	1	1	**0**	1	1	1	**1**	1	1	1	**1**	0	1	**0**	1	1	**1**	**0**	1	1	1	1	1	1	**1**	1	1	**0**	1	**1**	1	1	33
Chlamydia trachomatis	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**0**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	39
Aquifex aeolicus	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	0	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	39
Mycoplasma genitalium	**0**	1	1	**0**	**0**	1	1	**1**	1	0	**0**	1	1	1	**1**	1	0	0	**0**	0	1	**0**	1	1	**1**	**0**	1	1	1	1	1	1	**1**	1	1	**0**	1	**1**	1	1	28
Mycoplasma pneumoniae	**0**	1	1	**0**	**0**	1	1	**1**	1	0	**0**	1	1	1	**1**	1	0	0	**0**	0	1	**0**	1	1	**1**	**0**	1	1	1	1	1	1	**1**	1	1	**0**	1	**0**	1	1	27
Bacillus anthracis	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Bacillus cereus	**1**	1	1	**0**	**1**	1	1	**1**	1	1	**0**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**0**	1	**1**	1	1	37
Bacillus halodurans	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Bacillus subtilis	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Pasteuria penetrans	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Staphylococcus aureus	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Lactococcus lactis	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Streptococcus pyogenes	**1**	1	0	**0**	**0**	0	1	**1**	1	1	**0**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	35
Mycobacterium tuberculosis	**0**	1	1	**0**	**1**	1	1	**1**	1	1	**0**	1	1	1	**1**	1	1	0	**1**	1	1	**1**	1	1	**1**	**0**	1	1	1	1	1	1	**1**	1	1	**0**	1	**0**	1	1	33
Mycobacterium leprae	**0**	1	1	**0**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	38
Campylobacter jejuni	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Helicobacter pylori	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Neisseria meningitidis	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Xylella fastidiosa	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**0**	1	1	1	**1**	1	1	1	**1**	1	1	**0**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	38
Pseudomonas aeruginosa	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Buchnera aphidicola	**0**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	0	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	0	**1**	1	1	37
Escherichia coli	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Pasteurella multocida	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Haemophilus influenzae	**1**	1	1	**1**	**1**	1	1	**1**	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**1**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**1**	1	**1**	1	1	40
Rickettsia prowazekii	**0**	1	1	**1**	**1**	1	1	**0**	1	0	**1**	1	1	1	**1**	1	1	1	**1**	1	1	**0**	1	1	**1**	**1**	1	1	1	1	1	1	**1**	1	1	**0**	1	**1**	1	1	35
Total for 28 Taxa	**20**	28	25	**19**	**23**	27	28	**27**	28	24	**21**	28	28	28	**28**	28	26	25	**26**	25	26	**21**	28	28	**28**	**24**	28	28	28	28	28	28	**28**	28	28	**20**	27	**26**	28	28	
Sulfolobus solfataricus	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**0**	**1**	**13**
Pyrococcus abyssi	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**15**
Halobacterium sp. NRC-1	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**1**	**16**
Archaeoglobus fulgidus	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**19**
Methanococcus jannaschii	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**18**
TOTAL for 33 Taxa	**20**	**31**	**29**	**22**	**25**	**27**	**28**	**32**	**28**	**28**	**21**	**30**	**30**	**30**	**33**	**33**	**26**	**25**	**26**	**25**	**27**	**21**	**28**	**30**	**28**	**27**	**28**	**28**	**29**	**32**	**33**	**33**	**33**	**28**	**28**	**23**	**31**	**30**	**30**	**33**	
Alignment size [aa]	178	181	129	63	116	142	197	267	283	123	349	219	151	36	122	415	82	166	160	97	126	163	165	108	42	44	100	37	345	68	197	118	56	239	187	192	195	48	66	64	

Table 2: This is a spreadsheet of Species vs. Genes used in the analyses. The 1 indicates the gene is present, while the 0 indicates that the gene is absent or less than 30% identity. The rows and columns in bold were only used in the analysis of the 40 genes and 33 species.

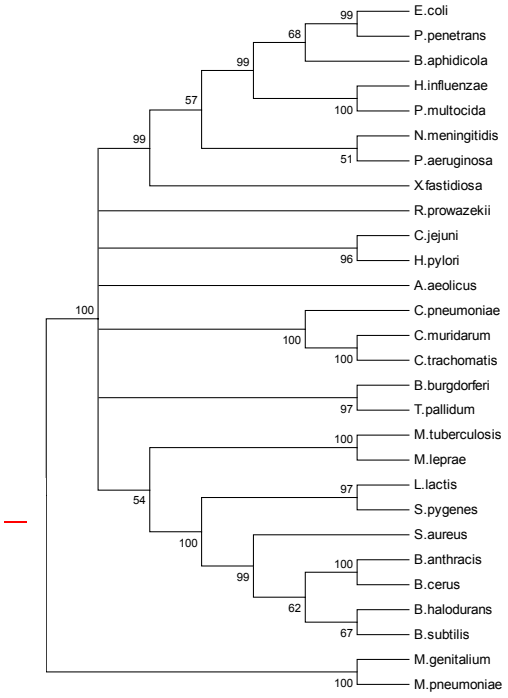

Fig 4. Gene tree for *groEL* inferred using maximum parsimony, positioning *P. penetrans* within the proteobacteria clade with 99% bootstrap support.

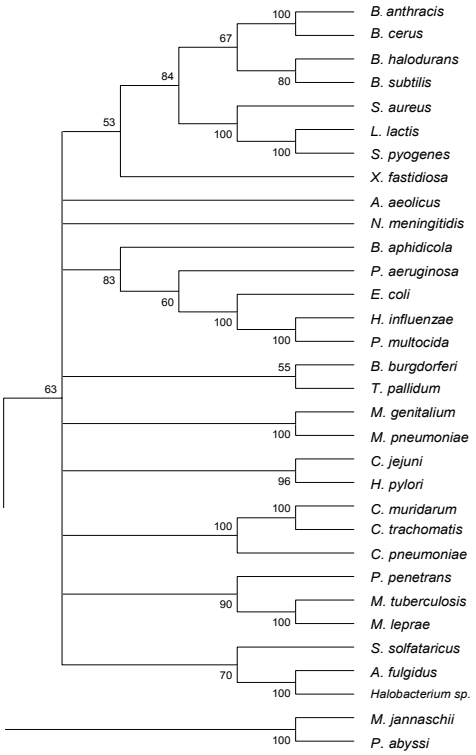

Fig 5. Gene tree for *eno* inferred using maximum parsimony, positioning *P. penetrans* within the high G+C content, gram-postitive clade with 90% bootstrap support.

Gene	# Amino Acids in Alignment	Maximum ln likelihood
infA	36	-810.2
rpmJ	37	-890.1
rplF	42	-954.0
rplJ	44	-1778.7
trxA	48	-2159.7
rpsM	56	-2159.7
dapA	63	-2545.9
tyrS	64	-2584.8
tsf	66	-2266.5
rpoC	68	-1869.1
mraY	82	-2573.5
murD	97	-3905.0
rplU	100	-3560.0
rplC	108	-3843.1
dapB	116	-4849.8
rpsK	118	-3396.3
map	122	-5479.6
glmS	123	-4015.1
priA	126	-4093.3
argS	129	-5692.1
dnaB	142	-2848.0
gyrB	151	-4704.0
murC	160	-5643.3
racE	163	-5122.1
rnc	165	-6251.7
murB	166	-6095.1
accA	178	-4105.2
adk	181	-8024.8
sigA	187	-4839.9
tkt	192	-6785.3
topA	195	-7248.6
dnaE	197	-6409.0
rpsB	197	-6054.5
gyrA	219	-5470.8
secA	239	-5468.1
eno	267	-10165.7
fmt	283	-12315.5
rpoB	345	-8998.4
groEL	349	-7764.8
metS	415	-18388.6
No Indels		
27G/28S	3032	-22783.2
27G/28S	4236	-117243.2
40G/33S	6036	-217221.9

Table 3: The number of amino acids in ascending order for each single gene alignment and the corresponding maximum ln likelihood value obtained from maximum likelihood analysis. All of the highlighted rows were not included in the 27 gene/28 taxa concatenated gene set. The rows highlighted in blue are those genes that also misplaced *Pasteuria* in the single gene tree analyses. The last three rows are for the concatenated datasets where G = genes and S = species.

CHAPTER 3:

Phylogenetic and biochemical across kingdom analysis of collagen-like motifs and their potential virulence role in *Pasteuria penetrans*.

Lauren Charles[1], Ignazio Carbone[2], Keith G. Davies[3], David Bird[1] & Charles H. Opperman[1]*

[1]Center for the Biology of Nematode Parasitism

Department of Plant Pathology

North Carolina State University

Raleigh, NC 27606, USA

[2]Center for Integrated Fungal Research

Department of Plant Pathology

North Carolina State University

Raleigh, NC 27606, USA

[3]Nematode Interactions Unit

Rothamsted Research, Ltd.

Harpenden, Herts. AL5 2JQ, UK

*Correspondence should be addressed to C.H.O. (email: warthog@ncsu.edu; fax: 919.515.9500; phone: 919.515.6699)

ABSTRACT

Recent collagen-like genes have been described as virulence factors in deadly human and insect bacterial pathogens. These diagnostic GXY-motifs have been reported in toxin-forming, parasitic close relatives of *Pasteuria penetrans* (*Bacillus anthracis, B. cereus, B. thuringiensis* and *Streptococcus pyogenes*). Through extensive studies in higher eukaryotes, assumptions were formed about the content, function and structural stability of the triple-helix. With increasing availability of genomes, the true nature of collagen-like repeats have come to light along with their presence in other organisms. While studying *P. penetrans* GXY-repeats for their possible role in virulence, biochemical characteristics and evolution of these motifs in bacteria, vertebrates, invertebrates, and fungi have been revealed.

The GXY-repeat motif in 85 vertebrate, 228 invertebrate, 17 fungal, and 76 bacterial sequences were examined. In the X-position, besides the classic proline, high percentages of alanine, threonine, cysteine, glutamine and glycine were found. Proline, threonine, glycine, alanine, arginine and serine occupied the Y-spot greater than 10% of the time. Each specific group, for evolutionary and functional reasons, utilizes different combinations and numbers of these amino acids.

Maximum parsimony analyses, performed on the collagen-like genes due to the heterogeneity in evolution across kingdoms, inferred trees from 15 *P. penetrans* along with 40 other bacteria, 61 invertebrates, 16 vertebrates, and 17 fungi, individually and together. These results strengthened the similarity of amino acid usage within groups, except for fungi pairing closest to invertebrates. They also infer a pattern of evolution from fungi, invertebrates, vertebrates, to bacteria with *P. penetrans* clustering ancestral to *Bacillus* spp.

INTRODUCTION

Collagen-like genes have recently been identified in parasitic *Bacillus anthracis, B. thuringiensis, B. cereus* and *Streptococcus pyogenes* as possible virulence factors in endospore attachment to their respective mammalian and insect hosts (Sylvestre *et al.*, 2003; Garcia-Patrone and Tandecarz, 1995; Charlton *et al.*, 1999; Lukmoski *et al*., 2000 and 2001). The attachment of this spore is the first stage in *Pasteuria penetrans* infection of its nematode host. Since *P. penetrans* has been found to be ancestral to the *Bacillus* clade (Charles *et al.,* 2005), these GXY-repeat sequences have been further studied in this paper for insight into their evolutionary history and virulence function.

There is potential for the development of *Pasteuria penetrans*, an endospore-forming, gram-positive bacterium, into a biological control agent. It is an obligate parasite of the root-knot nematode *Meloidogyne* spp., which are economically important crop pests (Chen *et al*, 1996; Trudgill *et al*, 2000). Due to its obligate, fastidious nature, it is important to utilize *in silico* approaches to unravel the complexity of this host-parasite interaction, eventually leading to exploitation in the field. The lifecycle of the bacterium has co-evolved with its host and begins when the second-stage juvenile nematodes migrate through the soil.The first stage of the infection process is when the exosporium of the spore-phased bacterium adheres to the nematode cuticle. The exosporium forms wing-like structures containing hair-like filaments providing a solid attachment to the nematode (Fig. 3.1). These filamentous proteins contain triple helices, similar to those found in collagen proteins and, moreover, found in *Bacillus anthracis, B. thuringiensis, B. cereus* and *Streptococcus pyogenes*, close relatives of *P. penetrans*, as a potential virulence factor in endospore attachment (Sylvestre *et al*., 2002

and 2003; Garcia-Patrone and Tandecarz, 1995; Charlton *et al*., 1999; Lukmoski *et al*., 2000 and 2001).

Collagens are filamentous proteins characterized by contiguous GXY-triplet motifs forming triple-helices where G represents glycine and X and Y are variable amino acids. They have many functions including being the major protein of connective tissue and the most abundant protein in animals (Bairati and Garrone, 1985). These collagens were characterized by their helical shape, a right-handed superhelix composed of three left-handed polyproline type II-like chains wound around their central axis. Glycine is present every third amino acid since its small, neutral nature enables the helical protein shape while still maintaining planar peptide bonds. They are stabilized by hydrogen bonds between the backbone groups, the N-H of glycine (donor) and the C=O in the X-position of another chain (acceptor) (Beck and Brodsky, 1998; Xu *et al.,* 2002). The O=H of hydroxyproline can also serve as the donor in addition to water-mediated interactions aiding in the stabilization. A high content of imino acids, proline and hydroxyproline, in the X and Y not only help in hydrogen bonding but create steric repulsion due to the pyrrolidone rings in these residues which force the extended helical form.

Requiring the post-translational hydroxylation of proline to stabilize the triple helical shape, which organisms lacking proline hydoxylases such as bacteria are unable to execute, collagens were thought to be exclusive to the animal kingdom. To this day, there is no evidence for their presence in protozoa, plants, or algae (Beck and Brodsky, 1998; Rasmussen *et al.*, 2003). However, there are recent discoveries of collagen-like genes in other organisms, namely the surface structures or spore components of bacteria, the fungi fimbriae, and other proteins within invertebrate and viral genomes. These findings have lead

to discussion of other means for stabilization of the helical shape and for their evolution (Bann *et al.*, 2000; Xu, *et al.*, 2002; Yang, *et al.*, 1997). For example, the presence of a common ancestor existing before the divergence of these organisms (Celerin *et al.,* 1996) and horizontal gene transfer between organisms or kingdoms (Rasmussen *et al.*, 2003) has been discussed.

When looking at the composition of the X and Y groups, it has been found that the side chains of these residues, always exposed to the solvent, have minor effects on the triple-helix stability and are mainly used for intermolecular interactions (Beck and Brodsky, 1998; and Chan et al, 1997). This allows for a variety of functions the collagen-like proteins perform in organisms, most commonly directed towards interactions utilizing these properties including ligand binding. The Y-position does seem to contribute to stability as shown in melting temperature studies using a variety of GXY triplets, favoring Gly-Pro-Arg and Gly-Pro-Hyp (Yang *et al*., 1997). It is found through calorimetric studies that the major stabilizing factor in these molecules is hydrogen bonding to the glycine backbone structure (Privalov, 1982). Cysteine residues are also thought to contribute to the triple-helix stabilization through covalent inter-chain bonding along with a conservation of polar/apolar nature in the residues (Beck and Brodsky, 1998). For example, it has been noted that if there is a negative amino acid in the X-position, like glutamate or aspartate, there is usually a complimentary positive amino acid present in the Y-position, such as arginine, lysine, that can form stable trimers and vise-versa (Rasmussen *et al.*, 2003). Therefore, the exact amino acids in the X and Y-positions and their side chains seem to show less involvement in the stabilizing of the triple-helix shape, previously thought to require the use of imino acids.

Allowing for substitution in these ancient sites, different evolutionary forces acting on each type of organism provides ground works for phylogenetic analysis.

Glimpses into the use of alternative residues in the GXY motif through biochemical studies and protein models are currently shown in bacteria and invertebrates, but not fungi. The most studied invertebrates are of the Phylum Annelida, including *Nereis sp.*, *Pheretima sieboldi*, i.e. earthworm, and *Riftia pachyptila,* i.e. hydrothermal vent worm. *Nereis sp.* and the earthworm were found to have significant amounts of alanine and serine, with hints of threonine and glutamic acid, in the Y-position. They were resistant to proteolysis by clostridial collagenase, implying that the collagens were extremely stable (Goldstein and Adams, 1970; and Waite *et al.*, 1980). The hydrothermal vent worms, with a very low content of proline and hydroxyproline, use threonine instead in the Y-position creating an incredible amount of stability when glycosylated (Mann *et al*., 1996; and Bann *et al.*, 2000).

Collagen-like repeats are found in only a limited number of microbrial genomes, including the gram-positive bacteria *Bacillus anthracis, B. cereus, B. thuringiensis*, and *Streptococcus pyogenes*. In all bacteria ever studied, proline was still favored in the X-position while threonine dominated the Y location allowing for direct hydrogen bonding to the backbone and a strong collagen-like triple helix structure (Rasmussen *et al.*, 2003; Sylvestre *et al.*, 2003; and Xu *et al*, 2002). The most variations in X and Y-positions were found within these bacterial sequences. For example, *Bacillus anthracis* BclA protein contained aspartic acid and threonine in the X-position (Sylvestre *et al.*, 2003) and the streptococcal Scl1 and Scl2 proteins had significant amounts of arginine, glutamic acid, aspartic acid, and lysine in the X while arginine, aspartic acid and lysine were also found in Y-position (Xu *et al*, 2002). There has also been notation of high amounts of glutamine,

along with threonine, present in the Y-position of gram-positive bacteria when the majority of the X-position is occupied by proline. This same study also noted that if the sequences had greater than fifty-percent of the Y-spot occupied by threonine, then the X was a charged amino acid such as alanine, serine, or proline (Rasmussen *et al.*, 2003).

Since the collagens are very diverse and old molecules, there is a large amount of heterogeneity noted within and throughout the different kingdoms. The beginning and end sections of all the sequences excluding the GXY-repeat are very different with less than 20% identity in most cases. This has lead to two major assumptions: 1.These genes have evolved many different uses, such as for *Pasteuria penetrans* to attach to the nematode verses *Bacillus anthracis* attaching to its mammalian host; and 2. Phylogenetic analysis would be meaningless on such diverse proteins (Rasmussen *et al.*, 2003). Overcoming this last obstacle, the analyses done in this paper excluded those regions of dissimilarity focusing on only the triple-helix motif.

There were 406 sequences in total identified by *in silico* analyses from vertebrates, invertebrates, bacteria, and fungi. They were first dissected to find the percent of amino acids in each position, X and Y. Then, the sequences were narrowed down to a total of 149 based on similarity to *P. penetrans* repeats for phylogenetic analysis. Choosing maximum parsimony over parametric based methods allowed for the heterogeneity seen in the evolution of the sequences while inferring the most reliable evolutionary tree (Kolaczkowski and Thornton, 2004). This type of analysis was done on the *P. penetrans* sequences with the bacteria, invertebrates, vertebrates and fungi individually, along with all of the sequences together.

METHODS

Full gene sequences from bacteria, invertebrates, vertebrates and fungi were collected on the amino acid level to minimize the divergence between species across kingdoms. *Pasteuria penetrans* sequences were gathered by using $(GXY)_4$ as the search pattern in a pBlast query (Altschul, 1990) applying Blossum62 matrix to our in-house database containing 2.5 Mbp of genomic sequence with a 5x coverage (unpublished). The resulting sequences were then evaluated by hand for likelihood of containing the collagen-like motif. Two of these sequences, namely c136.4, containing thirty-four GXY repeats, and c178.1, composed of seventy-nine GXY repeats, were used in NCBI by applying pBlast default searches without a filter and restrictions of a maximum e-value of 1e-20 and minimum score of 100. Searches were done on all sequences available for bacteria, fungi, invertebrates, vertebrates, and plants. Note: there were no matches found for plants or algae. Subsequent blasts were performed on the Wormbase website for *Caenorhabditis briggsae* and *C. elegans* with c178.1 as the query in pBlast with the same restrictions as above.

The collected sequences were then processed for subsequent analyses using manually written Perl scripts (Christiansen and Torkington, 1997) for consistency [See Appendix]. All of the sequences were trimmed so that only the middle section remained, starting at the first $(GXY)_4$ and ending with the last consecutive $(GXY)_4$. They were then imported into Genedoc (Nicholas and Nicholas, 1997) and manually edited by removing amino acids not in the GXY repeat pattern, i.e., globular regions, and that enabled the longest total GXY pattern. The alignment started with the first GXY of every sequence and left gaps only at the end due to differences in their length. GXY repeat motifs that had a 100% identity to one of the other

sequences of its same species were removed. This left the dataset at 406 sequences in total, eighty-five vertebrate, 228 invertebrate, seventeen fungi, and seventy-six bacteria.

Next, the sequences were analyzed for their amino acid usage in the X and Y-position and the number of GXY repeats. By finding the number and percentage of every amino acid in each gene in the X and Y-positions, the average percent of amino acids used in each position were calculated on the level of species, genus and under the classifications of vertebrate, invertebrate, fungi, and bacteria (Table 3.1 and Table 3.2). The average number of GXY repeats per species and group were also calculated. For these analyses, the invertebrate phyla contained seven Arthropoda, three Nematoda, one Annelida, and two Echinodermata sequences. The vertebrate subphylum, Craniata, contained six sequences while the fungal phyla included seven Ascomycota and one Basidiomycota. The bacteria phyla included one Actinobacteria and six Firmicutes sequences.

Lastly, the GXY-repeat motifs were prepared for phylogenetic analyses. Because of the computational load of an evolutionary analysis on 406 sequences, the number of sequences was narrowed down based on similarity to the fifteen *Pasteuria penetrans* GXY-repeats. The *P. penetrans* sequences were broken down into five groups, clustering those with the highest percent identities together. After aligning the five groups individually with the 385 other GXY-repeat sequences, the highest five sequences from each species with identities over twenty percent were collected. This percent identity took into consideration the length of the sequences analogously to the phylogenetic programs. Finally, the sequences with less than twenty-four GXY repeats were eliminated from subsequent analyses to reduce ambiguity in the dataset. We generated the subset of collagen-like genes for phylogenetic analyses by combining the lists of best five matches for each of the five groups of *P.*

penetrans sequences. The criteria resulted in a set containing fifteen *P. penetrans*, thirty-three other bacterial, fifty-four invertebrate, sixteen vertebrate, and fifteen fungal sequences (Table. 3.3).

Maximum Parsimony analyses were performed on the dataset due to the heterogeneous nature of the evolutionary relationships between the sequences chosen (Kolaczkowski and Thornton, 2003). Uniform standard parsimony was done in the same manner on the full dataset (Fig. 3.4) and on *Pasteuria* with bacteria, fungus, invertebrates, and vertebrates, separately (Fig. 3.3). A bootstrap of 1,000 was undertaken concurrently in the phylogenetic analyses performed in MegA2 (Kumar *et al*., 2001) to test the reliability of the trees given. The gaps were treated as missing data since globular regions in the sequences were omitted and only GXY-repeats were included in the analysis. Close-neighbor-interchange was applied with a search level of three and ten replications of random additional trees tested.

Additional analyses were done on the resulting across kingdom phylogenetic tree to ensure its reliability and help unravel the true relationships put forth. First, the tree was split into major clades and number of sequences of each class within each clade was recorded into a table and examined (Table 3.4). A chi-squared test and G-statistic with p-values for both was performed on the resulting table to test the significance of the clades. Added to this table were the percentages of each class found within each clade. To further clarify the relationships in the tree, a matrix of each class and the adjacent neighbors to each of the genes in that class was created (Table 3.5). Again, a chi-squared test and G-statistic with p-values for both was done to ensure the reliability of the table.

RESULTS

The results for the GXY-repeat statistical analysis span twenty *Pasteuria penetrans*, forty-five *Bacillus* spp., eleven other bacterial, 228 invertebrate, eighty-five vertebrate, and seventeen fungal collagen-like sequences. The average number of GXY triplets for the 406 sequences analyzed is thirty-four for *P. penetrans* (range 4-79), 108 for *Bacilli* (range 7-379), 79 for all bacteria (range 4-379), 91 for invertebrates (range 12-476), 289 for vertebrates (range 61-368), and forty-five for fungi (range 10-141).

Proline, as seen in the classical collagen molecule, is the dominating amino acid in the X-position except for the fungal sequences; percent usage ranges from 44% to 24% (Table 3.1). For the vertebrates and invertebrates, alanine and glutamic acid are second with traces of most other amino acids, ranging from 14% to 11% and 10% to 13% respectively. The bacteria overall at 25% and all the *Bacilli* sequences at 27% use alanine as their second choice in the X-spot. *P. penetrans*, on the other hand, uses threonine 18%, alanine 13% and cysteine 11%, as alternatives. Alanine and glycine, both simple amino acids, are preferred by fungal species, 19% and 20% respectively, with proline as a third alternative at 11%.

The Y-position in the GXY triple helix follows a different pattern for the individual groups (Table 3.2). For the invertebrates and vertebrates, proline is the dominating amino acid, at 39% and 35% respectively, as predicted from other studies, with alanine as the second choice, at 10% and 14%. Vertebrates also favor arginine using it 12.3% of the time. The fungal Y-spot contains 24% glycine, 22% proline, 15% serine and 13% alanine. Bacteria, on the other hand, use a very different pattern of amino acids. Threonine is found 65% of the time for *Bacilli*, 49% for all the bacteria, but only 14% of the time for *Pasteuria penetrans* sequences falling in as a forth alternative. Glutamine is the second choice for all

species of bacteria at 26% and specifically for *Bacilli spp.* and *P. penetrans* at 23%. Alanine is favored by *P. penetrans* in the Y-position at 23% and is also used by other bacteria 10% of the time. *P. penetrans,* in addition, utilizes proline in the Y-spot, similar to non-bacterial groups such as the fungi, 22% of the time.

Before performing phylogenetic analyses, the number of sequences was reduced by collecting only those with greatest similarity to the *Pasteuria penetrans* GXY-repeats. The number of possible sequences per class was fifty-four for bacteria, 112 for invertebrates, eighty-five for vertebrates, and seventeen for fungi. Although the amount of high identity matches did not reach these values, the percentage of accepted matches to the collagen-like sequences in each class are as follows: Bacteria (40/54) = 74.1%, Invertebrate (61/112) = 54.5%, Vertebrate (16/85) = 18.8%, and Fungi (17/17) = 100%. Note that even though they showed similarity in this part of the analysis, the sequences with less than twenty-four GXY repeats were eliminated from the analysis to minimize ambiguity in the dataset. This resulting excluded partition included seven bacteria, six invertebrates, five *P. penetrans*, and two fungal sequences resulting in the species numbers found in Table 3.3.

For each of the individual maximum parsimony analyses with *Pasteuria penetrans* and vertebrates, invertebrates, bacteria, and fungi respectively, similar results occur throughout the inferred trees (Fig. 3.3). In all of these with bootstrap support ranging from 92 to 100, the *P. penetrans* sequences cluster together except for the following four GXY-repeats: c178.1, c391.3, c453.1, and c336.2. Following their own pattern of evolution, they pull out with *Streptococcus pyogenes* in the bacterial tree but without any defined attraction to another species within the other individual trees. These sequences have the largest number of GXY-repeats in *P. penetrans*, 79, 62, 54, and 65 respectively, and have the greatest

identity to each other and other species in the GXY-motif alignment. Sequences c178.1, c391.3 and c453.1 always cluster together and use the same unique amino acid pattern as other *P. penetrans* sequences, but not bacteria. Proline, threonine and alanine are in the X-spot with alanine, glutamine and then proline in the Y-position. C336.2, on the other hand, follows a pattern of its own, favoring cysteine in the X-position and cysteine and aspartic acid in the Y-position. This amino acid usage is different than any other sequence analyzed in the project. There is one *P. penetrans* triple helix sequence that always and only lies within the *Bacillus* clade, namely c384.1. Containing 36 GXY-repeats, the triple helix uses proline in the X-position but also serine and arginine, unlike any other gene. Similar to all other bacteria, threonine is the only amino acid occupying the Y-position.

The phylogenetic tree inferred using maximum parsimony on the species present in Table 3.3 showed consistency with the individual tree results with high bootstrap support, i.e., greater than 80 (Fig. 3.4). This tree contained fifteen *Pasteuria penetrans,* thirty-three other bacteria, fifty-five invertebrates, sixteen vertebrates, and fifteen fungi. The *Bacillus spp.* cluster together, along with *Clostridium perfringens, Corynebacterium diphtheriae,* one fungal sequence and c348.1 from *P. penetrans.* The rest of the *P. penetrans* sequences remain ancestral to the bacteria in their own clade except c336.2, which falls adjacent to two *Streptococcus pyogenes* sequences in clade 5B2a. *S. pyogenes* do not cluster with the bacteria but split up into two groups and are found within clade 4 and 5. The invertebrate, vertebrate, and fungal sequences do not split up into perfect individual clusters, but instead seem to share similarities between groups.

When analyzed more carefully, it is clear that only one group, whether bacterial, fungal, vertebrate, or invertebrate, dominates the individual clades in the phylogenetic tree

(Table 3.4). This table, supported by a high G-statistic and chi-squared value both with a p-value of $1.0x10^{-5}$, shows that the results in the phylogenetic tree are significant. Clade 1 is made up of 88% of the bacteria (93% of *Pasteuria penetrans* sequences and 85% of other bacteria), clade 4A and 4B1 are dominated by 63% of the vertebrates, clade 4B2, 5A, 5B1, and 5B2a contain 50% of all of the invertebrates, and 47% of the fungi make up the majority of clade 5B2b. Clade 2 and 3 are both small and seem to be an even mixture of vertebrates, invertebrates, and fungi. Clade 6 and those with higher numbers seem to be outliers of the major clades and contain mostly invertebrates. Following from these results, the oldest collagen genes developed within the fungi, then invertebrates, vertebrates, and finally *P. penetrans* ending with the *Bacilli spp.*

The information from Table 3.5, showing the closest neighbors, identify which group shares the most similarity to the fungi, vertebrates, invertebrates, bacteria, and *Pasteuria penetrans*, respectively. It also addresses the hypotheses for a possible common ancestor or horizontal transfer of genes between groups. Here it is shown, with a high G-statistic and chi-square value both with a p-value of $1.0x10^{-5}$, that the table cannot be explained by a normal distribution and, therefore, the results must be significant. The *P. penetrans*, bacterial, vertebrate, and invertebrate GXY-repeat sequences all share the most homology within their own groups. Fungi, on the other hand, are adjacent to invertebrate sequences more often than their own. The neighboring sequences are not a specific species but a wide range of invertebrate organisms, ranging from flies to nematodes.

DISCUSSION

Collagen genes have been studied in depth in higher eukaryotes, such as humans, leading to assumptions about the content and structural stability of the triple helix GXY repeats. As genomes are sequenced and *in silico* analyses are undertaken, the true nature of collagen-like repeats have come to light along with their presence in other types of organisms, ranging from fungi to bacteria. Taking this approach to study *Pasteuria penetrans* for the possible role in virulence of collagen-like genes as shown in *Bacillus anthracis, B. thuringiensis, B. cereus,* and *Streptococcus pyogenes* (Sylvestre *et al.*, 2003; Garcia-Patrone and Tandecarz, 1995; Charlton *et al.*, 1999; Lukmoski *et al*., 2000 and 2001), more information has been revealed about the true biochemical characteristics and evolution of these motifs in bacteria, vertebrates, invertebrates, and fungi.

The bacterial species in which these GXY-repeat sequences have been identified show some surprising similarities in pathogenicity and lifestyle. There were seven species of Bacteria that contained genes with the GXY-pattern, namely *Pasteuria penetrans, Bacillus anthracis, B. cereus, B. thuringiensis, Streptococcus pyogenes, Clostridium perfringens,* and *Corynebacterium diphtheriae.* All of these are gram-positive bacteria and most are spore-forming, toxin-producing and pathogenic to animals. The exceptions are *B. thuringiensis,* which infects insects and its toxin is also found to effect nematodes (Marroquin *et al*., 2000; Wei *et al*., 2003), and *C. diphtheriae* since it does not produce spores although it is still often encapsulated. The current knowledge of *Pasteuria* also varies slightly from these others since it is believed to infect only nematodes, currently without an identified toxin.

Following sequence collection, the average number of GXY-triplets for the 406 sequences were analyzed and some interesting patterns formed. Conferring with other

studies, there were no collagen-like genes found to be present in plants or algae (Beck and Brodsky, 1998; Rasmussen *et al.*, 2003). There is a large range of possible numbers of GXY-repeats necessary to make a stable collagen-like triple helix shape. For example, the range for *Pasteuria penetrans* is 4 to 79 and for invertebrates it is 12 to 476. From the sequences analyzed here, the highest average of repeats are for vertebrates, invertebrates, then bacteria, fungi, and lastly *P. penetrans*. Furthermore, *P. penetrans* sequences (average 34) resemble in length the fungi (average 45) and *Streptococcus* spp. (average 45) rather than their closest relatives in the *Bacillus* spp. (average 79).

The first set of significant findings from this study revolves around the percent usage of amino acids in the X and Y-position of the GXY repeat sequence of *Pasteuria penetrans*, other bacteria including close relatives of the *Bacilli*, vertebrates, invertebrates, and fungi (Table 3.1 and 3.2) and their biochemical implications. The following results from the analyses done in this paper show overlapping of amino acid usage between species, kingdoms, and previous studies done on each of these groups. Novel usages of amino acids in the X and Y-position have also been revealed, providing further insight into collagen-like repeats. The evidence here supports the hypothesis that the side-chains of amino acids in the GXY repeat are less constricted than previously thought yet are still able to retain their stable helical shape (Beck and Brodsky, 1998). Since these are ancient motifs, each organism has to some extent evolved their own amino acid usage in the variable sites enabling them to generate biochemically distinct, function-specific proteins.

In the vertebrate and invertebrate collagen-like genes analyzed (Table 3.1 and 3.2), there was an overwhelming use of the imino acid, proline, in the X and Y-position. As seen in the classic collagen molecule, most of these amino acids in the Y-spot will be post

translationally hydroxylated for stability (Beck and Brodsky, 1998). Interestingly, alanine was present in the X and Y-position of both groups' GXY motifs greater than 10% of the time. This use of alanine has chiefly been reported in the invertebrate phylum Annelida (e.g., *Nereis* sp. and *Pheretima* sp.) resulting in extremely stable triple helices (Goldstein and Adams, 1970; Waite *et al.*, 1980). Another novel finding was the use of glutamic acid in the X-position of vertebrates, although it has been noted in invertebrates of the phylum Annelida to be extremely stable and resistant to clostridial collagenase (Goldstein and Adams, 1970; Waite *et al.*, 1980). This result was also found in *Streptococcus spp.*, revealing an unknown similarity to bacterial collagen-like genes (Xu *et al.,* 2002). A closer link than previously noted between vertebrate and invertebrate collagen-like sequences is evident through these amino acid usages. Unique to vertebrates, arginine is present in the Y-position shown in melting temperature studies to be as stable as hydroxyproline (Yang *et al*., 1997). The above results are exemplified in the placement of these species within the phylogenetic tree presented in Fig. 3.4.

The seventeen fungal sequences made up of seven Ascomycota and one Basidiomycota are the least studied and, therefore, have the most novel amino acid usage in their GXY triple helix motifs (Table 3.1 and 3.2). Although they utilize proline, the fungal motifs favor the use of simple hydrocarbon side chains of glycine and alanine in the X-position. In the Y-position, glycine is the dominating amino acid and then proline. Proline is already known to hold the shape of the triple helix, yet glycine found in either of the positions is not well discussed. Alanine, with similar chemical composition as glycine, is again used in the Y-spot. However, the most unusual finding for the fungi would be the high percent usage of serine in the Y-position, which contains a hydroxyl group similar to

threonine, an amino acid glycosylated in bacteria for stability. This usage has been noted in *Nereis sp.* and in the earthworm, *Pheretima sieboldi* (Goldstein and Adams, 1970; and Waite *et al.*, 1980), which forms a strong connection between the fungal and invertebrate species as seen in the phylogenetic tree (Fig. 3.4), clade groups (Table 3.4), and adjacent neighbor studies (Table 3.5).

The bacterial part of this study can be broken down into three parts, specifically the twenty-one *Pasteuria penetrans* sequences, the forty-five *Bacillus* genes, and then the overall bacteria containing these sequences along with three Actinobacterial and eight Streptococcal genes. All of the bacterial sequences have a strong favoritism towards proline in the X-position characteristic of any collagen-like repeats (Table 3.1). For the *Bacillus* and overall for bacteria in the X, alanine with a simple hydrocarbon side-chain has the second highest percent usage. This result has been discussed before for bacteria with high incidence of threonine in the Y-position (Rasmussen *et al.*, 2003). It is also the same as the X-position results for invertebrates and vertebrates done in this study, except here at a much higher percent usage of alanine.

P. penetrans, on the other hand, follows its own pattern of amino acid usage after proline in the X-position (Table 3.1). Its second choice is threonine, alanine, and then cysteine. Threonine used in this position has been seen in *Bacillus anthracis BclA* protein (Sylvestre *et al.*, 2003) but cysteine, which contains sulfur, has never been noted in the GXY-repeat X-position. Since cysteine residues can form covalent bonds to one another, this phenomenon may be utilized by *Pasteuria* to form a stronger bond than the classical hydrogen bonding found in collagen.

The *Bacillus* spp. and the overall bacteria follow a pattern of amino acid usage different than that of *P. penetrans* in the Y-position also (Table 3.2). They have a high threonine content corresponding to the proline found in the X-position, which will post-translationally be glycosylated for stability. This pattern of amino acid placement, Gly-Pro-Thr, is characteristic of Firmicutes and almost all other bacterial collagen-like sequences ever studied (Rasmussen *et al.*, 2003; Sylvestre *et al.*, 2003; and Xu *et al*, 2002). *P. penetrans*, although uses threonine in the Y-position, has instead a much higher percentage of alanine, glutamine, and proline equally utilized in the repeat motifs. Alanine and proline are widely used in eukaryotic collagen-like sequences, but glutamine with its amide group is new to this position. Surprisingly, both the *Bacillus* and all bacteria, also utilize glutamine in the Y-spot. This is a novel amino acid to the Y-position and its ability to create stable triple helical structures should be further investigated.

The next step in the analysis looked at the evolutionary relationships of GXY-repeat motifs between *Pasteuria penetrans*, other bacteria, vertebrates, and invertebrates. Using the technique of phylogenetic analysis, the amino acid usage and biochemical make-up along with inheritance inferences leading to comparative functional assignment of these motif patterns were investigated. Collagen-like sequences are extremely old motifs used in a variety of genes and are being examined across kingdoms; therefore, they contain heterogeneity in the rates of evolution for both the X and Y-position. Parametric approaches, such as maximum likelihood and Bayesian statistics, are known to be extremely biased and statistically inaccurate when applying them to sites with different rates of evolution (Kolaczkowski and Thornton, 2004). Because of this inherent characteristic of the dataset, maximum parsimony analysis was chosen as the best alternative for creating a well-

supported and defined phylogenetic tree. [Note: preliminary studies, not discussed here, were performed using Bayesian analysis and resulted in low supported and unresolved trees as expected given the dataset; see Appendix.]

Using maximum parsimony to further understand the relationship between *Pasteuria penetrans* GXY motifs and each group individually and together, the inferred trees provide evidence for and against aspects of the classic vertebrate collagen model. Out of all the sequences collected, based solely on percent identity in the motif alignment, those that match closest to individual *P. penetrans* sequences in descending order are from bacteria, invertebrates, fungi and lastly trailing behind vertebrates. The inferred phylogenetic trees never place *P. penetrans* adjacent to or in a clade group with a vertebrate species (Table 3.4 and 3.5) emphasizing the differences in amino acid usage between the two groups (Table 3.1 and 3.2). Due to *Pasteuria* deviations from the classical triple-helix, new light has been shed on the evolution, structural stability and biochemical function of collagen-like motifs, such as in bacterial pathogenicity.

When examining the individual group trees, whether bacteria, invertebrates, vertebrates, or fungi with *Pasteuria penetrans*, the same results emerged strengthening their interpretation. Figure 3.3 of the fifteen *P. penetrans* sequences and thirty-three other bacteria demonstrates the typical output for all of these inferred trees. The *P. penetrans* GXY-repeat motifs cluster together with the same four outlying sequences falling within the other groups, namely c178.1, c453.1, c391.3, and c336.2. When narrowing down the full dataset for phylogenetic analysis, these four sequences had the highest percent identity to the other groups' sequences, partly explaining why the separation is seen. They all cluster next to *Streptococcus pyogenes* sequences in the bacterial tree (Fig. 3.3) and c336.2 also does in the

complete tree, while the rest of the sequences fall back into the *P. penetrans* clade (Fig. 3.4). Because of this, they are most likely not horizontally gene transferred but just follow a pattern of evolution similar to *S. pyogenes* collagen-like repeats. Therefore, the possibilities that remain for the separation of these four *Pasteuria* sequences are that they evolved slower than the others, they are from an ancient ancestor before the groups differentiated or are a result of convergent evolution.

Two of these sequences, c336.2 and c384.4, not only fall out of the *Pasteuria* clade in the individual group trees but also in the consolidated tree (Fig. 3.4). C336.2 follows a pattern of its own amino acid usage distinct from all other sequences analyzed in the project yet always falls adjacent to a *Streptococcus pyogenes* collagen-like surface protein. It contains cysteine in the X-position and cysteine and aspartic acid in the Y-position, which may explain its unique placement within all inferred trees. The potential need for tight covalent bonding of cysteine residues may be explained through an unknown functional demand on the full gene product. The other *Pasteuria* sequence, c384.4, falls adjacent to pathogenic *Bacillus cereus* in both trees and *B. anthracis* in the bacterial tree. These two genes' patterns of evolution, which closely resemble *S. pyogenes* and *Bacillus* spp. known to utilize this type of motif for virulence, is of great interest. Besides these results, the individual trees of GXY-repeat motifs within and between species of the specific groups show no definite pattern of evolution or clustering.

Placing all of the sequences together and creating a phylogenetic tree across kingdoms, revealed insight into the patterns of evolution of collagen-like repeats for bacteria, invertebrates, vertebrates, and fungi with a high level of confidence (Fig. 3.4). By looking at the inferred tree, the most clear-cut result is the clustering of the *Bacillus* and then the

Pasteuria penetrans sequences at the top of the tree. This confers that *Pasteuria* is ancestral to the *Bacillus* spp. as found in the concatenated phylogenetic analysis by Charles *et al.* 2005. This also implies that the GXY-repeat motif was inherited by a common ancestor of the two genera rather than by developing on their own, through convergent evolution or by horizontal gene transfer. The other relationships are a little harder to distinguish by just looking at the tree without any statistical analyses. By examining Tables 3.4 and 3.5, the relationships between and within fungi, invertebrates, and vertebrates are clarified.

Table 3.4 of Major Groups verse Clade Groups shows that one of the four major groups (bacteria, invertebrates, vertebrates, and fungi) dominate at least one major clade containing over 50% of their sequences. There are, however, two smaller clade groups containing the sequences at the bottom of the tree that seem to be an even mixture of invertebrate, vertebrate, and fungal sequences. These sequences can be considered relatively dissimilar to the rest of the repeats. When the tree is read from the bottom to the top, the collagen-like GXY-repeat sequences follow the divergence from fungi, invertebrates, vertebrates, *P. penetrans*, to bacteria.

The Closest Neighbor Table (Table 3.5) clarifies the true relationship between groups, even though some of the individual sequences may fall into the same clusters within the tree. It is clear for *P. penetrans*, the other bacteria, invertebrates and vertebrates that each group has the greatest similarity to sequences within their own. This is supported by a high number of sequences from each group placed adjacent to themselves. In fact, these results rule out the possibility of horizontal gene transfer (Rasmussen *et al.*, 2003) between these groups since the GXY-repeats have the most similarity within their individual groups, implying that the same pressures and patterns of evolution are in force. There is, however,

something interesting going on between fungi and the invertebrates, as suggested earlier by Celerin *et al.,* 1996. The fungi are more similar to invertebrate collagen-like patterns than to themselves. This is supported by the fact that 93% of the fungal sequences are adjacent to invertebrates while only 53% of the fungi are adjacent to themselves. [Note: the percentages add up to greater than 100 since there can be two or more adjacent sequences per GXY-motif depending on the placement within the tree.] This suggests the presence of a common ancestor existing before the divergence of fungi and invertebrates while leaving room for the idea of horizontal gene transfer and lastly convergent evolution.

This paper introduces a way to utilize amino acid usage and phylogenetic analyses to uncover biochemical adaptations and evolutionary patterns of an ancient gene family, e.g., collagen-like genes. Through these approaches, each organism's individual solutions to creating stability within the triple-helix structure are beginning to be revealed and understood. For example, *Pasteuria penetrans* utilizes cysteine in the X and Y positions potentially creating covalent bonds for added stability. Comparative techniques, used to infer the biological function of the specific genes containing these collagen-like motifs, provide a way to develop hypotheses that can be subsequently tested in the laboratory. In this case, we identified potential virulence factors on the exosporium of *P. penetrans* used to adhere to the cuticle of its nematode host similar to those already identified in its close *Bacilli* relatives.

LITERATURE CITED

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. **215**: 403-410.

Bairati, A. and Garrone, R. (Eds) 1985. Biology of invertebrate and lower vertebrate collagens. Plenum Press, New York.

Bann, J.G., Payton, D.H., and Bachinger, H.P. 2000. Sweet is stable: glycosylation stabilizes collagen. FEBS Letters **473**: 237-240.

Beck, K., and Brodsky, B. 1998. Supercoiled protein motifs: The collagen triple-helix and the α-helical coiled coil. Journal of Structural Biology **122**: 17-29.

Chan, V.C., Ramshaw, J.A.M, Kirkpatrick, A., Beck, K., and Brodsky, B. 1997. Positional preferences of ionizable residues in Gly-X-Y triplets of the collagen triple-helix. Journal of Biological Chemistry **272**: 31441-31446.

Charles, L., Carbone, I., Bird, D., Burke, M., Opperman, C., Davies, K., and Kerry, B. 2005. Phylogenetic Analysis of *Pasteuria penetrans* using multiple genetic loci. Journal of Bacteriology (submitted).

Charlton, S., Moir, A.J.G., Baillie, L., and Moir, A. 1999. Characterization of the exosporium of *Bacillus cereus*. Journal of Applied Microbiology **87**: 241-245.

Chen, Z.X., Dickson, D.W., McSorley, R., Mitchell, D.J., and Hewlett, T.E. 1996. Suppression of *Meloidogyne arenaria* race 1 by soil applications of endospores of *Pasteuria penetrans*. Journal of Nematology **28**: 159-168.

Christiansen, T., Torkington, N., and other contributors. http://www.perl.org Copyright (c) 1997-2003.

Garcia-Patrone, M., and Tandecarz, J.S. 1995. A glycoprotein multimer from *Bacillus thuringiensis* sporangia: Dissociation into subunits and sugar composition. Molecular and Cellular Biochemistry **145:** 29-37.

Goldstein, A. and Adams, E. 1970. Glycylhydroxyprolyl sequences in earthworm cuticle collagen: glycylhydroxyprolylserine. Journal of Biological Chemistry **245(20)**: 5478-5483.

Hoiczyk, E., A. Roggenkamp, M. Reichenbelcher, A. Lupas, and J. Heesemann. 2000. Structure and sequence analysis of Yersinia Yad A and Moraxella UspAs reveal a novel class of adhesions. Embo J. **19**: 5989-5999.

Janulczyk, R. and Rasmussen, M. 2001. Improved pattern for genome-based screening identifies novel cell wall-attached proteins in gram-positive bacteria. Infection and Immunity **69(6)**: 4019-4026.

Kumar, S., Tamura, K., Jakobsen, I., and Nei, M. 2001. MEGA2: Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Arizona, USA.

Kolaczkowski, B. and Thornton, J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**: 980-984.

Lukmoski, S., Nakashima, K., Abdi, I., Cipriano, V. Ireland, R., Reid, S., Adams, G., and Musser, J. 2000. Identification and characterization of Scl gene. Infection and Immunity **68(12)**: 6542-6553.

Lukmoski, S., Nakashima, K., Abdi, I., Cipriano, V., Shelvin, B., Graviss, E., and Musser, J. 2001. Identification and characterization of a second extracellular collagen-like protein made by Group A *Streptococcus*: control of production at the level of translation. Infection and Immunity **69(3):** 1729-1738.

Mann, K., Mechling, D.E., Bächinger, H.P., Eckerskorn, C., Gaill, F., and Timpl, R. 1996. Glycosylated threonine but not 4-hydoxyproline dominates the triple helix stabilizing positions in the sequence of a hydrothermal vent worm cuticle collagen. J. Mol. Biol. **261**: 255-266.

Marroquin, L.D., Elyassnia, D., Griffitts, J.S., Feitelson, J.S., and Aroian, R.V. 2000. *Bacillus thurinriensis* (Bt) toxin susceptibility and isolation of resistance mutants in the nematode *Caenorhabditis elegans*. Genetics **155(4):** 1693-1699.

Nicholas, K.B. and Nicholas, H.B. Jr. 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Distibuted by the author.

Privalov, P.L. 1982. Stability of proteins. Proteins which do not present a single cooperative system. Advances in Protein Chemistry **35**:1-104.

Rasmussen, M., Jacobsson, M. and Björck, L.2003. Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. Journal of Biological Chemistry **278(34)**: 32313-32316.

Sylvestre, P., Couture-Tosi, E. and Mock, M. 2002. A collagen-like surface glycoprotein is a structural component of the *Bacillus anthracis* exosporium. Molecular Microbiology **45(1)**: 169-178.

Sylvestre, P., Couture-Tosi, E. and Mock, M. 2003. Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length. Journal of Bacteriology **185(5)**: 1555-1563.

Tahir, Y., P. Kuusela and M. Skurnik. 2000. Functional mapping of the Yersinia enterocolitica adhesion Yad A. Identification of eight NSVALG…S motifs on the amino-terminal half of the protein involved in collagen binding. Mol.Microbiol. **37**:192-206.

Trudgill, D.L., Bala, G., Blok, V.C., Daudi, A., Davies, K.G., Fargette, M., Gowen, S.R., Madulu, J.D., Mateille, T., Mwageni, W., Netscher, C., Phillips, M.S., Abdoussalam, S., Trivino, G.C. and Voyoulallou, E. 2000. The importance of tropical root-knot nematodes (*Meloidogyne* spp.) and factors affecting the utility of *Pasteuria penetrans* as a biocontrol agent. Nematology **2**: 823-845.

Waite, J.H., Tanzer, M.L., and Merkel, J.R. 1980. *Nereis* cuticle collagen: proteolysis by marine vibrial and clostridial collagenases. Journal of Biological Chemistry **255 (8)**: 3596-3599.

Wei, J.Z., Hale, K., Carta, L., Platzer, E., Wong, C., Fang, S.C, and Aroian, R.V. 2003. *Bacillus thuringiensis* crystal proteins that target nematodes. Proceedings of the National Academy of Sciences of America **100(5):** 2760-2765.

WormBase web site, http://www.wormbase.org, release WS120, date March 1, 2004.

Xu, Y., Keene, D.R., Bujnicki, J.M., Höök, M. and Lukomski, S. 2002. Streptococcal Scl1 and Scl2 proteins form collagen-like triple helices. Journal of Biological Chemistry **277(30)**: 27312-27318.

Yang, W., Chan, V.C., Kirkpatrick, A., Ramshaw, J.A.M., and Brodsky, B. 1997. Gly-Pro-Arg congers stability similar to Gly-Pro-Hyp in the collagen triple-helix of host-guest peptides. Journal of Biological Chemistry **272 (46):** 28837-28840.

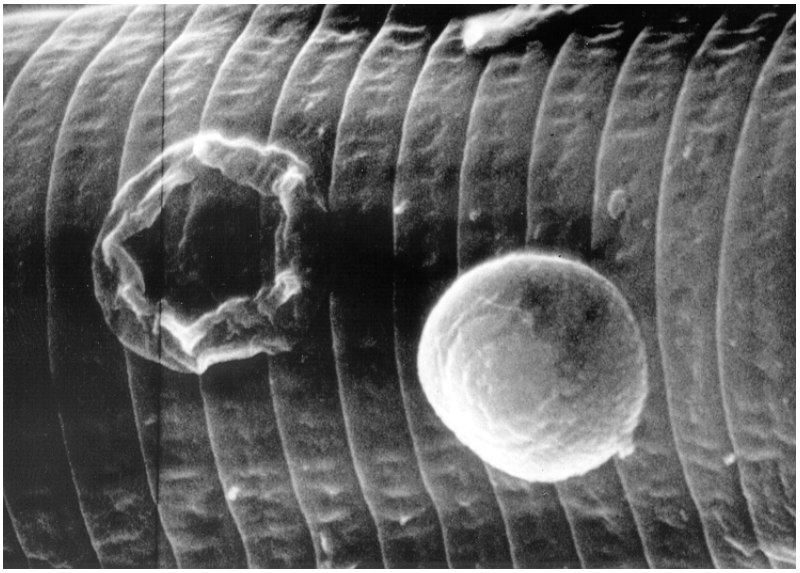

Fig. 3.1: SEM micrograph of an endospore adhering to the cuticle of a second-stage juvenile nematode. The endospore on the left has become detached from the cuticle leaving behind the densely packed hair-like filaments which seemingly form a circular membrane.

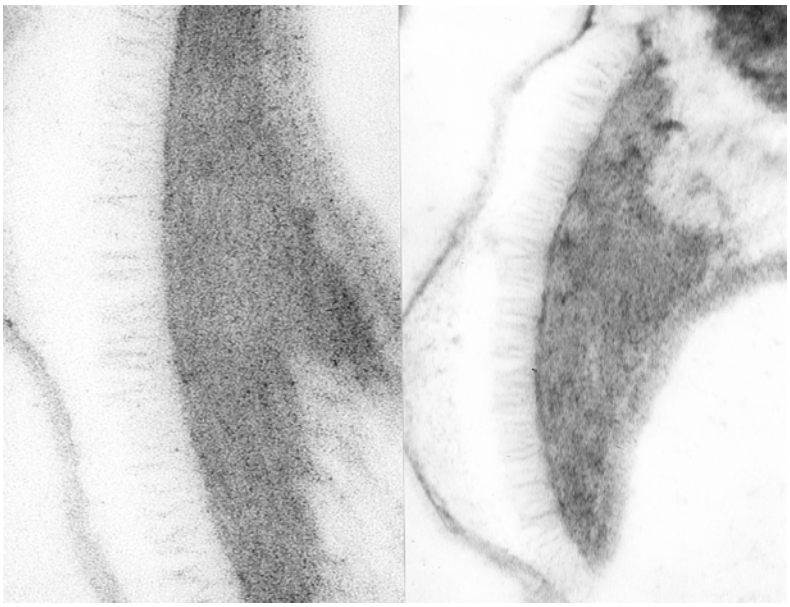

Fig. 3.2: TEM micrographic cross-section of the exosporium. The wing-like structures and the hair-like filaments on the upper and lower surface are shown. The hair-like structures are easily seen on the upper surface but are almost indistinguishable on the bottom surface because they are densely packed together.

*****Fig 3.1 and 3.2: Adapted from 2004 poster "Collagen-like proteins have been identified in the genome of *Pasteuria penetrans*, similar to those in other species of *Bacilli* and are likely to be important in endospore attachment" and used with permission from Keith Davies and Charles Opperman.

		Simple		Acidic	Hydroxyl	Sulfur
	Pro Proline	Ala Alanine	Gly Glycine	Glu Glutamic acid	Thr Threonine	Cys Cysteine
Vertebrate	**33.4**	**14.3**	1.8	**13.2**	1.7	0.0
Inveterbrate	**24.1**	**11.4**	5.8	**9.9**	3.3	0.2
Bacteria	**43.6**	**25.1**	0.7	3.6	3.1	1.3
***Bacillus* spp.**	**38.3**	**26.6**	0.8	2.0	1.9	0.1
P. penetrans	**42.6**	**13.0**	1.2	0.1	**18.0**	**10.6**
Fungus	**11.0**	**18.5**	**19.8**	1.6	3.7	0.1

Table 3.1: Percentage of amino acids in the X-position of the GXY-repeat of all collagen-like sequences analyzed from invertebrates, vertebrates, fungi, bacteria, and *P. penetrans*. Amino acids with percentages over 10% are considered significant, are shown in the table, and are highlighted within their representative groups. The labels on the top part of the graph refer to the type of amino acid below, such as simple, basic and acidic. The abundance of the amino acids in each group is colored accordingly to usage of most to least (red, blue, green and then yellow).

		Simple		Basic	Hydroxyl		Amide
	Pro Proline	Ala Alanine	Gly Glycine	Arg Arginine	Ser Serine	Thr Threonine	Gln Glutamine
Vertebrate	**34.8**	**13.5**	0.9	**12.3**	5.5	4.5	6.5
Inveterbrate	**39.2**	**9.9**	4.9	6.2	4.2	2.5	5.0
Bacteria	6.2	**10.0**	0.9	2.2	1.8	**48.8**	**26.2**
***Bacillus* spp.**	1.4	4.9	0.8	0.5	1.6	**65.0**	**22.7**
P. penetrans	**22.2**	**22.8**	2.9	0.0	1.2	**14.3**	**22.5**
Fungus	**22.2**	**12.5**	**24.4**	1.8	**15.4**	2.6	2.1

Table 3.2: Percentage of amino acids in the Y-position of the GXY-repeat of all collagen-like sequences analyzed from invertebrates, vertebrates, fungi, bacteria, and *P. penetrans*. Amino acids with percentages over 10% are considered significant, are shown in the table, and are highlighted within their representative groups. The labels on the top part of the graph refer to the type of amino acid below, such as simple, basic and acidic. The abundance of the amino acids in each group is colored accordingly to usage of most to least (red, blue, green and then yellow).

Bacteria	48	**Invertebrates**	54			**Vertebrates**	16	**Fungus**	15
Pasteuria penetrans	15	*Anopheles gambiae*	6	*Galleria mellonella*	1	*Bos Taurus*	3	*Aspergillus nidulans*	4
Bacillus anthracis	11	*Paracentrotus lividus*	1	*Alvinella pompejana*	1	*Gallus gallus*	2	*Candida glabrata*	3
Bacillus cereus	10	*Araneus ventricosus*	2	*Argiope trifasciata*	1	*Homo sapien*	5	*Magnaporthe grisea*	3
Bacillus thuringiensis	5	*Hemicentrotus pulcherrimus*	2	*Pig Roundworm*	1	*Mus musculus*	4	*Eremothecium gossypii*	1
Streptococcus pyogenes	5	*Caenorhabditis briggsae*	14	*Bombyx mori*	1	*Rattus norvegicus*	1	*Debaryomyces hansenii*	1
Clostridium perfringens	1	*Caenorhabditis elegans*	13	*Apis mellifera*	2	*Macaca mulatta*	1	*Ustilago maydis*	1
Corynebacterium diphtheriae	1	*Drosophila melanogaster*	8					*Neurospora crassa*	1
		Drosophila yakuba	1					*Yarrowia lipolytica*	1

Table 3.3: Names of the species along with the number of collagen-like sequences used in the phylogenetic tree inferred by maximum parsimony.

	Pasteuria	**Bacteria**	**Invertebrates**	**Vertebrates**	**Fungi**
Clade 1	**14, 93%**	**28, 85%**	0	0	1, 7%
Clade 2,3	0	0	4, 7%	3, 19%	4, 27%
Clade 4A,4B1	0	3, 9%	8, 15%	**10, 63%**	0
Clade 4B2	0	0	**13, 24%**	0	2, 13%
Clade 5A,5B1,5B2a	1, 7%	2, 6%	**14, 26%**	0	0
Clade 5B2b	0	0	7, 13%	0	**7, 47%**
Clade 6+	0	0	8, 15%	3, 19%	1, 7%

Table 3.4: Major groups verse clade groups in maximum parsimony tree (Fig. 3.4). The numbers represent the number of that group in the given clade with the corresponding percentage of the group. The bolded numbers represent which clade the majority of the group falls into. For example, the majority of fungi, 47%, are located in Clade 5B2b of the phylogenetic tree found in Figure 3.4. The chi-squared for this table is 172.9448 and G-statisticis 188.5179 with p-value for both equal to 1.0×10^{-5}.

	Pasteuria	Bacteria	Vertebrates	Invertebrates	Fungi
Pasteuria	**16**	2	0	0	0
Bacteria	3	**37**	0	0	0
Vertebrates	0	0	**10**	8	0
Invertebrates	0	1	7	**47**	14
Fungi	0	2	0	**14**	8

Table 3.5: The adjacent neighbor to each sequence in a group throughout the maximum parsimony tree (Fig. 3.4). The bolded numbers tell which group is closest to each individual group. For example, the bacteria are adjacent to bacteria 37 times, which means that the bacterial GXY-repeats share the most similarity to themselves. For this table, the chi-square is 310.5596 and the G-statisticis 268.9457, both with a p-value of 1.0×10^{-5}.

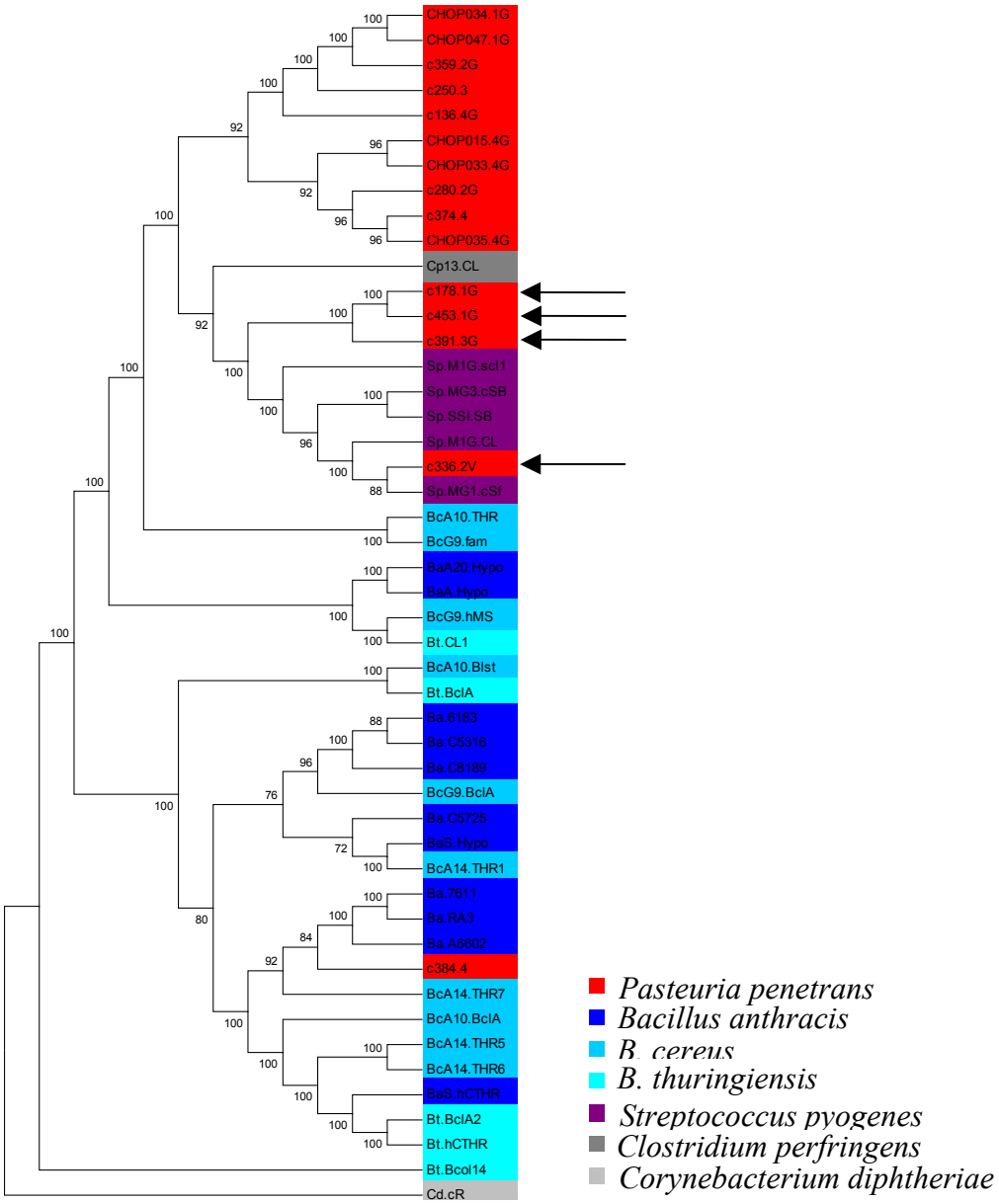

Fig 3.3: Maximum Parsimony analysis of *Pasteuria penetrans* and other Bacteria with bootstrap values on the tree. Arrows are pointing to the unclustered *P. penetrans* sequences, c178.1, c453.1, c391.3 and c336.2, that pull out when running the analysis with either bacterial, vertebrate, invertebrate, or fungal GXY-repeats. The *P. penetrans* sequence c384.4 is always placed closer to the *Bacilli* than any other species.

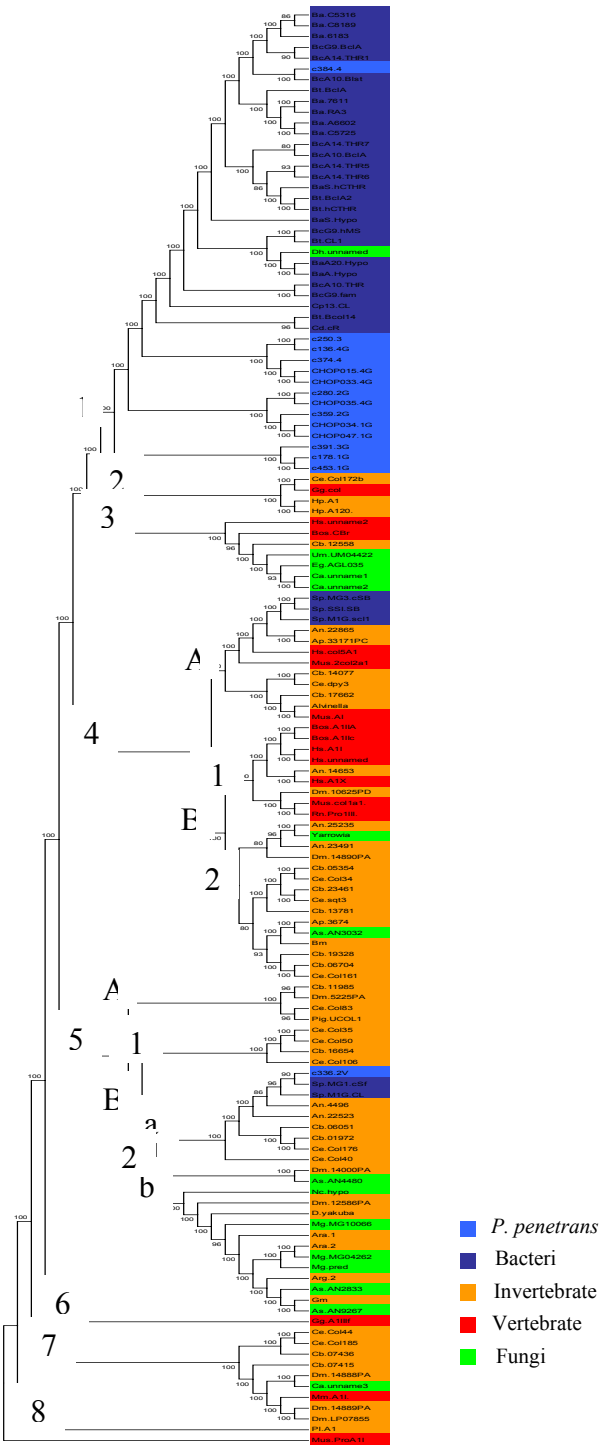

Fig 3.4: Maximum Parsimony tree with bootstrap values and gaps treated as missing data. The numbers and letters seen by the branches are used to label the clades and are referred to in the text.

CHAPTER 4:

CONCLUSIONS

INTRODUCTION

The detrimental application of toxic chemical nematicides into the soil to manage the devastating crop parasite, root-knot nematodes of the *Meloigogyne* spp., desperately needs to be replaced. Discovery of the nematode hyperparasite, *Pasteuria penetrans,* has brought hope of an environmentally-safe method of biocontrol. To be able to exploit the tritrophic relationship between the bacterium, root-knot nematode and plant in the field, the biology and chemistry of the host-parasite interactions must be fully understood. The two major areas that are under investigation for the utilization of *P. penetrans* in biocontrol are the ability to mass produce the fastidious bacteria as well as expanding its limited host range of specific nematode populations. The obligate nature of the bacterium makes it difficult to study *Pasteuria penetrans* alone, *in vitro* and in the field. Although through recent sequencing efforts, the spore DNA of *P. penetrans* can now be studied through genomic and bioinformatics applications to better understand its mechanisms of parasitism.

Phylogenetic Analysis

The first step in the process of genomic analysis is to identify the organism's closest relatives for a deeper understanding of its evolution and for future comparative genomics. Phylogenic analysis, at the protein level, has been done on a total of thirty-three bacterial species, including *Pasteuria penetrans,* using the concatenation of forty housekeeping genes, including two additional subsets and each individual gene (Chapter 2). Through the application of maximum likelihood, maximum parsimony and Bayesian methods on these datasets, *P. penetrans* was found with a high level of confidence to cluster tightly within the

order *Bacilli* of the gram-positive, low G+C content eubacteria, excluding seven of the gene trees. The exact placement of the bacterium as ancestral to *Bacillus* spp. has been resolved.

The results of the concatenated-species-tree-analysis all place *Pasteuria penetrans* among the *Bacilli* clade, between *Bacillus halodurans* and *Saphylococcus aureus*. Surprisingly, the saprophytic extremophiles, *Bacillus halodurans* and *B. subtilis,* are more closely related to *Pasteuria* than to the *Bacillus* mammalian pathogens, including the highly lethal *B. anthracis* and opportunistic *B. cereus*. Unfortunately, the genome of *Bacillus thuringiensis,* an extremely virulent insect pathogen, was not available at the time of this analysis. *B. cereus, B. anthracis,* and *B. thuringiensis* have unusually AT-rich genomes which is most likely responsible for this separation. These important findings facilitate the phylogenetic context of *Pasteuria* to be exploited through further research in comparative genomics and in the development of biocontrol strategies.

Collagen-like Repeat Motifs

The first set of genes to be studied through comparative genomics for *Pasteuria penetrans* was those containing a collagen-like GXY-repeat motif (Chapter 3). Known to be virulence factors in pathenogenic *Bacilli*, these GXY-repeat sequences have been further studied in this paper for insight into their origin and evolutionary history. For example, the parasitic *Bacillus anthracis, B. cereus,* and *B. thuringiensis* all have collagens-like motifs that have been noted as possible virulence factors in endospore attachment (Sylvestre *et al.*, 2003; Charlton *et al.*, 1999; Garcia-Patrone and Tandecarz, 1995). Being the first stage in *P. penetrans* infection of the nematode, a deeper understanding of this interaction is key in expanding its host range.

When utilizing evolutionary tools to perform comparative genomics, it is important that the dataset has enough variation to resolve placement within a tree but not too much as to create large chunks of ambiguity. Since the collagens are very diverse and old molecules, there is a large amount of heterogeneity noted within and throughout the different kingdoms. The non-GXY domains at the amino and carboxy terminus are very distinct with less than 20% identity in most cases. The evolution and incorporation of these triple helix repeats into genes coding for a variety of functions is of great interest yet phylogenetic analysis has been deemed meaningless due to this diversity (Rasmussen *et al.*, 2003).

Overcoming this last obstacle, the comparative analyses done in Chapter 3 excluded those regions of dissimilarity focusing only on the triple-helix motif. The 406 sequences, identified by *in silico* analysis, were from vertebrates, invertebrates, bacteria, and fungi. They were first dissected to find the percent of amino acids in each position, X and Y. Then, the sequences were narrowed down to a total of 149 based on similarity to *P. penetrans* for phylogenetic analysis. Choosing maximum parsimony over parametric based methods allowed for the heterogeneity seen in the evolution of these sequences to produce the most reliable evolutionary tree (Kolaczkowski and Thornton, 2004). This type of analysis was done on the *P. penetrans* sequences with each group individually and all together. There were many interesting results which shed light on the molecular make-up, stability, and evolutionary history of these motifs in *P. penetrans*, along with other bacteria, vertebrates, invertebrates, and fungi.

Collagen genes have been studied in depth in higher eukaryotes, such as humans, which has lead to assumptions about the content and structural stability of the GXY-repeats. Here, it is revealed that different organisms utilize distinct amino acids to achieve the same

type of helical shape and stability. For example, *P. penetrans* prefer the use of alanine, proline, and glutamic acid in the Y spot while other bacteria use mainly threonine; fungi use glycine, and animals and invertebrates use proline. These novel usages of amino acids in both the X and Y-positions provide evidence that the side-chains of amino acids in the GXY repeat are less constricted than previously thought yet are still able to retain their stable helical shape (Beck and Brodsky, 1998). There was also a large range of possible numbers of GXY-repeats to make the stable collagen-like triple helix shape. From the sequences analyzed here, the most average repeats are for vertebrates, then bacteria, invertebrates, fungi, and lastly *P. penetrans.* Surprisingly, *P. penetrans* sequences resemble the fungi in length rather than the *Bacillus* or other bacteria. So, by just looking at the content of the collagen-like sequences information is gained that can be helpful in studying the attachment phase of infection.

When analyzing the different evolutionary trees, the *P. penetrans* GXY-repeat motifs cluster together with the same four sequences falling within the other groups, namely c178.1, c453.1, c391.3, and c336.2. C336.2 follows a pattern of its own, favoring cysteine in the X-position and cysteine and aspartic acid in the Y-position. This amino acid usage is different than any other sequence analyzed in the project, which would explain why it appears different. Interestingly, when narrowing down the full dataset for phylogenetic analysis, these four sequences had the highest percent identity to the other groups' sequences. This reason, in turn, could partly explain why the separation is seen among these GXY-repeats. Another possibility that remains for the differences seen is that these sequences are from an ancient ancestor before the groups differentiated and have evolved slower than the other sequences. They all do cluster next to *Streptococcus pyogenes* sequences in the bacterial tree

(Fig. 3.5) and c336.2 also does in the complete tree, while the rest of the sequences fall back into the *P. penetrans* clade (Fig. 3.6). Because of this, they are most likely not horizontally gene transferred but just follow a pattern of evolution similar to the *S. pyogenes* collagen-like repeats rather than any other species. In addition to these, there is one *Pasteuria* sequence that always falls adjacent to *Bacillus* spp. Important information can be gained from the organisms with similar sequences to these outliers along with further understanding behind their unusual sequence content.

Through these comparative analyses, interesting connections between species other than *P. penetrans* also have been revealed. For example, the fungi are more similar to invertebrate collagen-like patterns than to themselves, as suggested earlier by Celerin *et al.,* 1996. This is supported by the fact that 93% of the fungi sequences are adjacent to invertebrates while 53% of the fungi are adjacent to themselves. In the trees, note that the percentages add up to greater than 100 since there can be two or more adjacent sequences per GXY-motif depending on the placement within the tree. This suggests the presence of a common ancestor existing before the divergence of these organisms and leaves room for the idea of horizontal gene transfer and lastly convergent evolution.

FUTURE DIRECTIONS FOR RESEARCH

By utilizing *Pasteuria penetrans* genomic sequences through comparative gene analyses, almost any part in the organism's life can be studied for a deeper understanding towards its development into a biocontrol agent. This thesis is just a glimpse into the full potential use of DNA and protein sequences to uncover biological phenomenons. There are many areas still left to be analyzed for deeper insight into the processes, chemicals, and proteins used by this nematode hyperparasite that enable it to be a successful pathogen.

Future research is necessary in areas that can improve mass production of the bacteria and identify virulence factors in efforts to expand its limited host range.

Mass Production in vitro

For implementation as a biocontrol agent, *Pasteuria penetrans* must be able to be grown *in vitro* at an exponential rate. This will require the discovery of the nutritional requirements necessary for vegetative growth along with triggers into the sporulation phase for proper distribution and field application. With the ability to utilize genomic sequences from *P. penetrans*, along with other close relatives, these unknown factors can be addressed in a new light.

Culturing *Pasteuria penetrans* without its obligate host, *Meloidogyne* spp., has always been a great obstacle. The DNA sequences, however, can be decoded through comparative genomics to reveal known pathways and necessary components enabling the growth and proliferation of the bacterium. Combining this information with the known biology of the bacteria, a growth medium similar to that devised for the obligate bacterial parasite *Xylella fastidiosa* (Lemos *et al.*, 2003) can be developed for *P. penetrans*. For example, since the bacteria naturally live in the pseudocoelom of the nematode, the physical and chemical makeup of this fluid could also be central to the development of an *in vitro* culture medium. Using genomic sequences and known biology to predict nutritional requirements for steady proliferation is a new but potentially more effective method to obtain results for fastidious, obligate parasites. Once vegetative growth is possible *in vitro*, the transition steps between this stage and sporulation need to be understood for mass distribution of inactive endospores into nematode infested soils.

Sporulation is known to occur when bacteria are deprived of certain nutrients causing them to go into a stationary phase for survival. The genes that are triggered have been well studied in other gram-positive spore forming bacteria. *Clostridium* spp. and *Bacillus* spp., including *B. subtilis,* enter the stationary phase due to nutrient deprivation. It is believed that the actual decision to sporulate depends upon the receptors of an elaborate signal transduction phosphorelay network initiated by two key transcription factors, *SpoOA* and *AbrB* (Phillips and Strauch, 2002). Bacteria can also use quorum-sensing mechanisms to monitor local population densities in attempt to regulate nutrients and ensure their survival. If populations reach their inherent carrying capacity, sporulation is triggered and growth ceases. Here are two examples of major pathways effecting the transition between vegetative and spore stages that can be examined through genetic tools for similar sequences of genes and pathways in *Pasteuria penetrans*.

Virulence Factors

Along with being able to mass produce the endospores for distribution, there are many aspects of the host-parasite interaction between *Pasteuria penetrans* and *Meloidogyne* spp. that still need to be explored. The key factors that influence the attachment to and germination into the host nematode are still not very well understood along with the triggers and chemicals responsible for initiating and causing nematode egg suppression. Examining other virulence factors found in close relatives can also be useful for the discovery of similar gene sequences and pathways utilized by *P. penetrans*. Pathogenicity is often a very complex process with many variables that need to be investigated in as many ways as

possible. Comparative genomics can provide an excellent source of alternative routes to uncovering the secrets to exploit this highly specific yet beneficial interaction.

Pasteuria penetrans spores are only able to infect specific root-knot nematode populations which currently limit its applicability in the field. Due to the heterogeneity in the bacterial spore coat and nematode cuticle, no correlation has thus far been made concerning the attachment or germination of the bacteria to the nematode (Davies *et al*., 1994; Davies and Redden, 1997; Davies *et al*., 2001), besides the *Bacillus* collagen-like genes examined in Chapter 3. However, another property to examine is that isolates of *Pasteuria* have the ability to adapt to host nematodes by propagation on the specific population, increasing in successful attachment and germination each time (Oostendorp *et al.*, 1990). Although this discovery still limits attachment to a specific nematode population, it means that genetic factors are constantly changing while being passed down through generations. A comparison of the genomic and protein sequences of *Pasteuria* isolates before and after successive propagations on a nematode population is one path that might lead to the identification of genes responsible for changes seen in attachment, germination, and pathogenicity.

To find other potential virulence factors for *P. penetrans,* it may be most efficient to look at virulence factors already studied in closely related organisms. When looking at virulence factors in *P. penetrans* close relatives, a variety of pathways are utilized providing a wide range of gene comparisons to be examined. For example, in the *Bacillus* spp. there are known pathogenicity islands along with scattered gene groups and non-specific toxins that enable virulence in these bacteria. Genes such as *plc*R which encodes a pleiotropic regulator of virulence factor expression in *B. thuringiensis* is also present in *B. cereus* and *B. anthracis*, although it is nonfunctional in the latter (Agaisse *et al.,* 1999). This set of

dispersed virulence factors are believed to be inherited through a common ancestor defining organisms possessing these sequences as potential pathogens. Since *Pasteuria penetrans* is ancestral to this class, it is possible that this set of genes may play a role in its parasitism also. In general, *B. thuringiensis* and *B. anthracis* have acquired large extrachromosomal elements or plasmids that encode virulence factors and determinates, such as *Cry* toxins and *pX01*, respectively, while *B. cereus* produces non-specific virulence factors. Ironically, certain *Cry* toxins or crystal proteins produced by *B. thuringiensis* have recently been shown to target free-living bacterial-feeding and parasitic nematodes, such as *Caenorhabditis elegans* and *Nippostrongylus brasiliensis* (Marroquin *et al*., 2000; Wei *et al.*, 2003). The acquisition of these extrachromosomal elements has led to the belief that *B. cereus* is ancestral to the other two pathogens (Agaisse *et al.,* 1999; and Keim *et al.,* 1997) and could provide closer insight into the underlying genetic sequences enabling pathenogenicity in its close relatives. It also infers that *P. penetrans* virulence determinates may be located on plasmids encoding extremely specific virulence factors for each nematode population, such as the Cry toxins. Whatever the case may be, these examples show how the application of comparative genomics can easily lead to the identification of potential virulence factors in the perplexing *Pasteuria penetrans*.

CONCLUSIONS

The unique obligate nature of the soil-borne bacterial pathogen *Pasteuria penetrans* to parasitize root-knot nematodes makes it a prime candidate for biocontrol. Since its spores are metabolically inactive, highly resistant to environmental conditions and pesticide applications, and have a long shelf life, they will provide the perfect substrate for distribution

and application in a field or greenhouse setting. With current laboratory limitations in studying fastidious bacteria, genomics and bioinformatics techniques are proposed to solve the unanswered questions to how to mass produce the bacteria and expand its limited host range.

With the acquisition of increasing numbers of genomic sequences, the lifecycle of *P. penetrans* along with its requirements for survival and pathogenicity are beginning to be unraveled. A deep phylogeny on the protein level has resolved the placement the bacterium as tightly nestled within the *Bacilli* clade and adjacent to *Bacillus halodurans* and *Saphylococcus aureus* (Chapter 2). Further studies examined the collagen-like repeats identified as virulence factors effecting attachment to hosts in specific *Bacillus* spp. and found in pathogenic endospore-forming bacteria along with invertebrates, vertebrates, and fungi (Chapter 3). These studies revealed unique amino acid usage, distinct gene histories, along with genes potentially used in attachment to specific nematode populations.

Future studies have been proposed using the same techniques to identify nutritional requirements and virulence factors necessary for full exploitation as an environmentally-safe pesticide alternative. The ability to identify necessary requirements such as for *in vitro* growth, including the vegetative and sporulation phases, host range expansion, and other potential virulence factors through comparative genomics creates a whole new perspective on issues past found to be futile. With further research on the genomic level utilizing bioinformatics tools, *Pasteuria penetrans* is well on its way to becoming a successful biocontrol agent for the devastating crop parasite, root-knot nematodes.

LITERATURE CITED

Agaisse, H., Gominet, M., Okstad, O.A., Kolsto, A., and Lereclus, D. 1999. PlcR is a pleiotropic regulator of extracellular virulence factor gene expression in *Bacillus thuringiensis*. Molecular Microbiology **32(5):** 1043-1053.

Beck, K., and Brodsky, B. 1998. Supercoiled protein motifs: The collagen triple-helix and the α-helical coiled coil. Journal of Structural Biology **122:** 17-29.

Bishop, A.H. and Ellar, D.J. 1991. Attempts to culture *Pasteuria penetrans in vitro*. Biocontrol Science and Technology **1:** 101-114.

Charlton, S., Moir, A.J.G., Baillie, L., and Moir, A. 1999. Characterization of the exosporium of *Bacillus cereus*. Journal of Applied Microbiology **87:** 241-245.

Davies, K.G., Fargette, M., Balla, G., Daudi, A., Duponnois, R., Gowen, S.R., Mateille, T., Phillips, M.S., Sawadogo, A. Trivino, C., Vouyoukalou, E., and Trudgill, D.L. 2001. Cuticle heterogeneity as exhibited by *Pasteuria* spore attachment is not linked to the phylogeny of parthenogenetic root-knot nematodes (*Meloidogyne* spp.). Parasitology **122 (1):** 111-120.

Davies, K.G. and Redden, M. 1997. Diversity and partial characterization of putative virulence determinants in *Pasteuria penetrans*, the hyperparasite of root-knot nematodes. Journal of Applied Microbiology **83(2):** 227-235.

Davies, K.G., Redden, M., and Pearson, T.K. 1994. Endospore heterogeneity in *Pasteuria penetrans* related to attachment to plant-parasitic nematodes. Letters in Applied Microbiology **19:** 370-373.

Garcia-Patrone, M., and Tandecarz, J.S. 1995. A glycoprotein multimer from *Bacillus thuringiensis* sporangia: Dissociation into subunits and sugar composition. Molecular and Cellular Biochemistry **145:** 29-37.

Keim, P., Kalif, A., Schupp, J., Hill, K., Travis, S., Richmond, K., Adair, D.M., Hugh-Jones, M., Kuske, C.R., and Jackson, P. 1997. Molecular Evolution and Diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers. Journal of Bacteriology **179(3):** 818-824.

Kolaczkowski, B. and Thornton, J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**: 980-984.

Lemos, E.G., Alves, L.M., and Campanharo, J.C. 2003. Genomics-based design of defined growth media for the plant pathogen *Xylella fastidiosa*. FEMS Microbiology Letters **219(1):** 39-45.

Marroquin, L.D., Elyassnia, D., Griffitts, J.S., Feitelson, J.S., and Aroian, R.V. 2000. *Bacillus thurinriensis* (Bt) toxin susceptibility and isolation of resistance mutants in the nematode *Caenorhabditis elegans*. Genetics **155(4):** 1693-1699.

Oostendorp, M., Dickson, D.W., and Mitchell, D.J. 1990. Host range and ecology of isolates of *Pasteuria* spp. from the southeastern United States. Journal of Nematology **22(4):** 525-531.

Phillips, Z.E., and Strauch, M.A. 2002. *Bacillus subtilis* sporulation and stationary phase gene expression. Cell. Mol. Life Science **59(3):** 392-402.

Rasmussen, M., Jacobsson, M. and Björck, L.2003. Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. Journal of Biological Chemistry **278(34):** 32313-32316.

Riese, R.W., Hackett, K.J., Sayre, R.M., and Huettel, R.N. 1998. Factors affecting cultivation of three isolates of *Pasteuria* spp. Journal of Nematology **20:** 657.

Slyvestre, P., Couture-Tosi, E., and Mock, M. 2003. Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in the exosporium filament length. Journal of Bacteriology **185(5):** 1555-1563.

Wei, J.Z., Hale, K., Carta, L., Platzer, E., Wong, C., Fang, S.C, and Aroian, R.V. 2003. *Bacillus thuringiensis* crystal proteins that target nematodes. Proceedings of the National Academy of Sciences (USA) **100(5):** 2760-2765.

Williams, A.B., Stirling, G.R., Hayward, A.C., and Perry, J. 1989. Properties and attempted culture of *Pasteuria penetrans*, a bacterial parasite of root-knot nematodes (*Meloidogyne javanica*). Journal of Applied Bacteriology **67:** 145-156.

APPENDICES

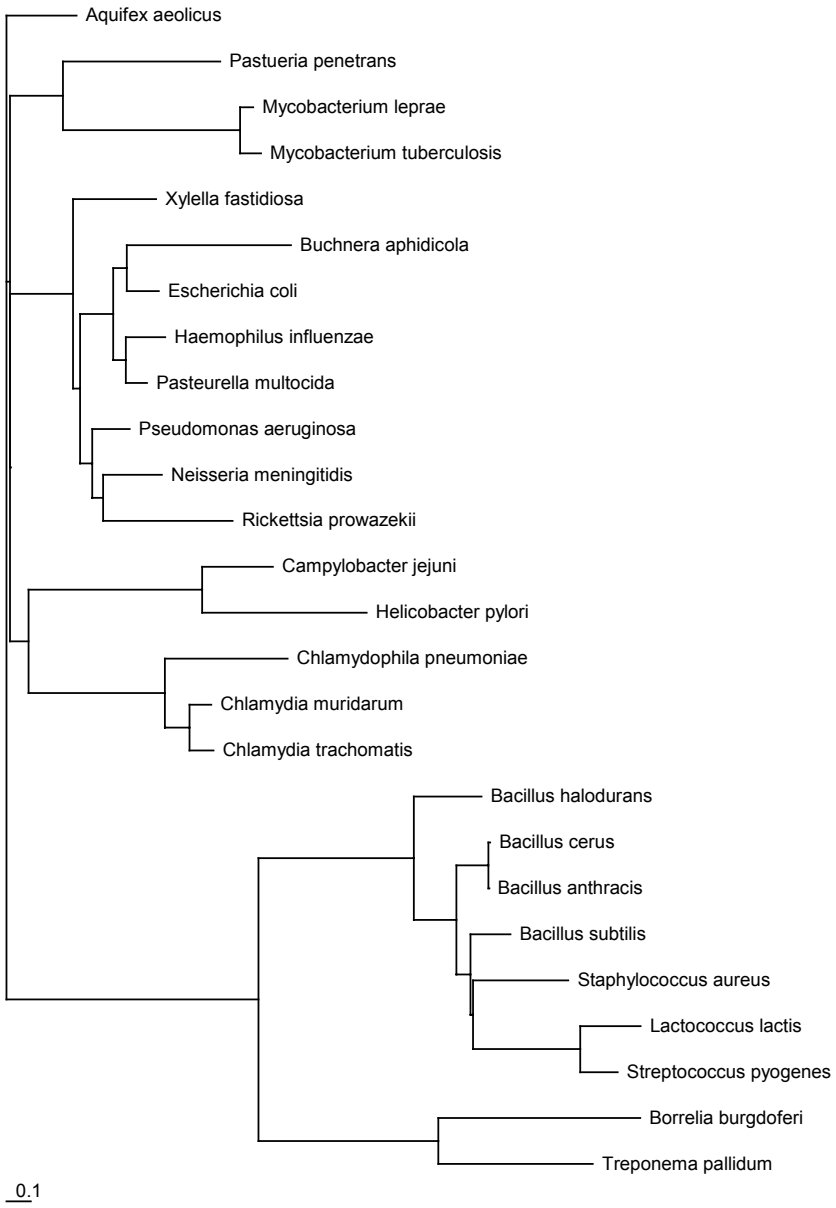

Figure A.2.1: Maximum likelihood single gene tree for *murC* positioning *P. penetrans* with the gram-positive, high G+C content bacteria. Branch lengths are shown.

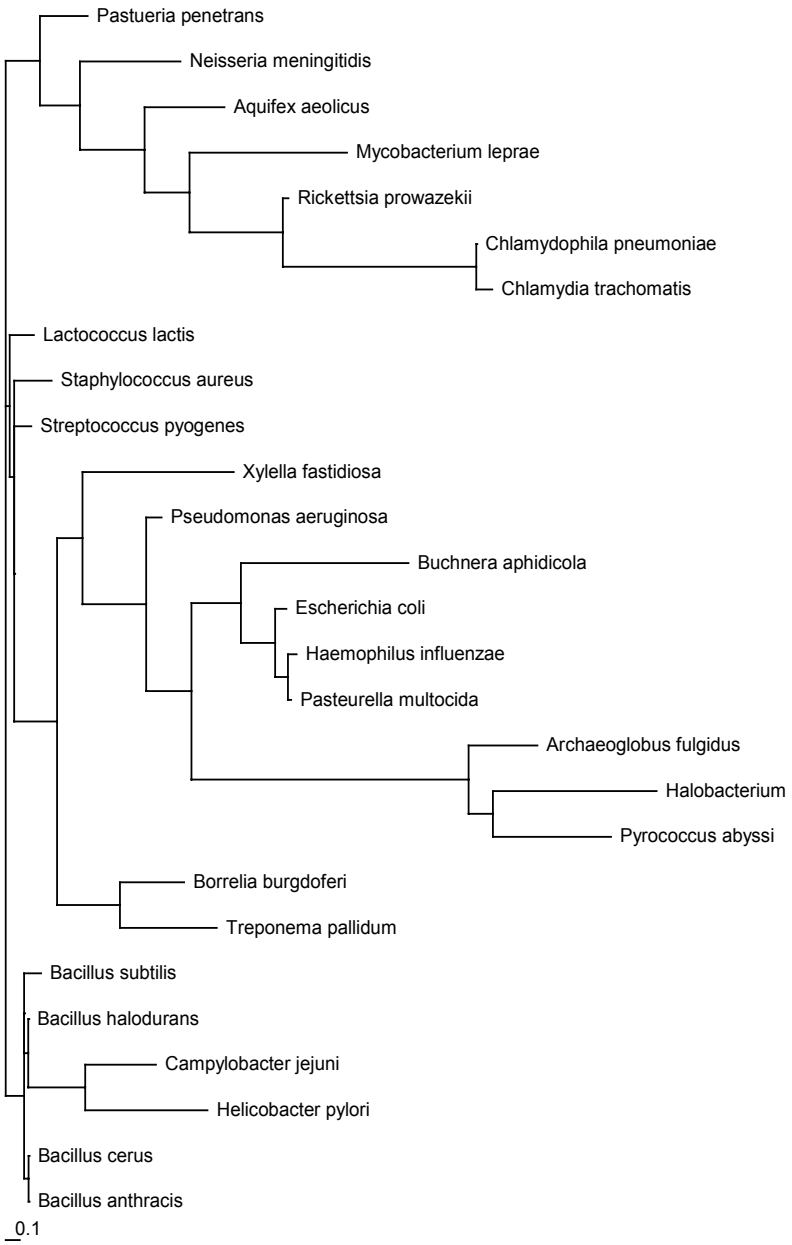

Figure A.2.2: Maximum likelihood single gene tree for *rplJ* positioning *P. penetrans* and other bacteria in unusual clade groupings. Branch lengths are shown.

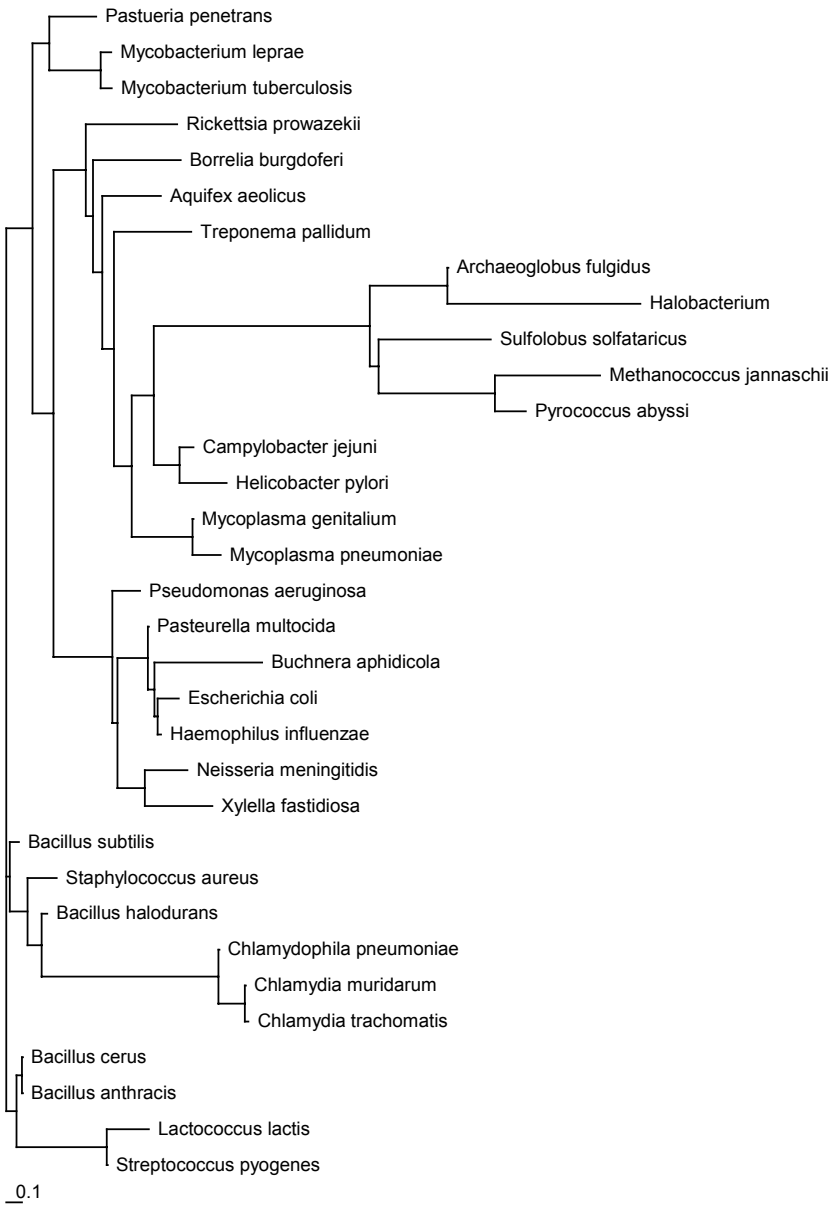

Figure A.2.3: Maximum likelihood single gene tree for *rpsM* positioning *P. penetrans* with the gram-positive, high G+C content bacteria. Branch lengths are shown.

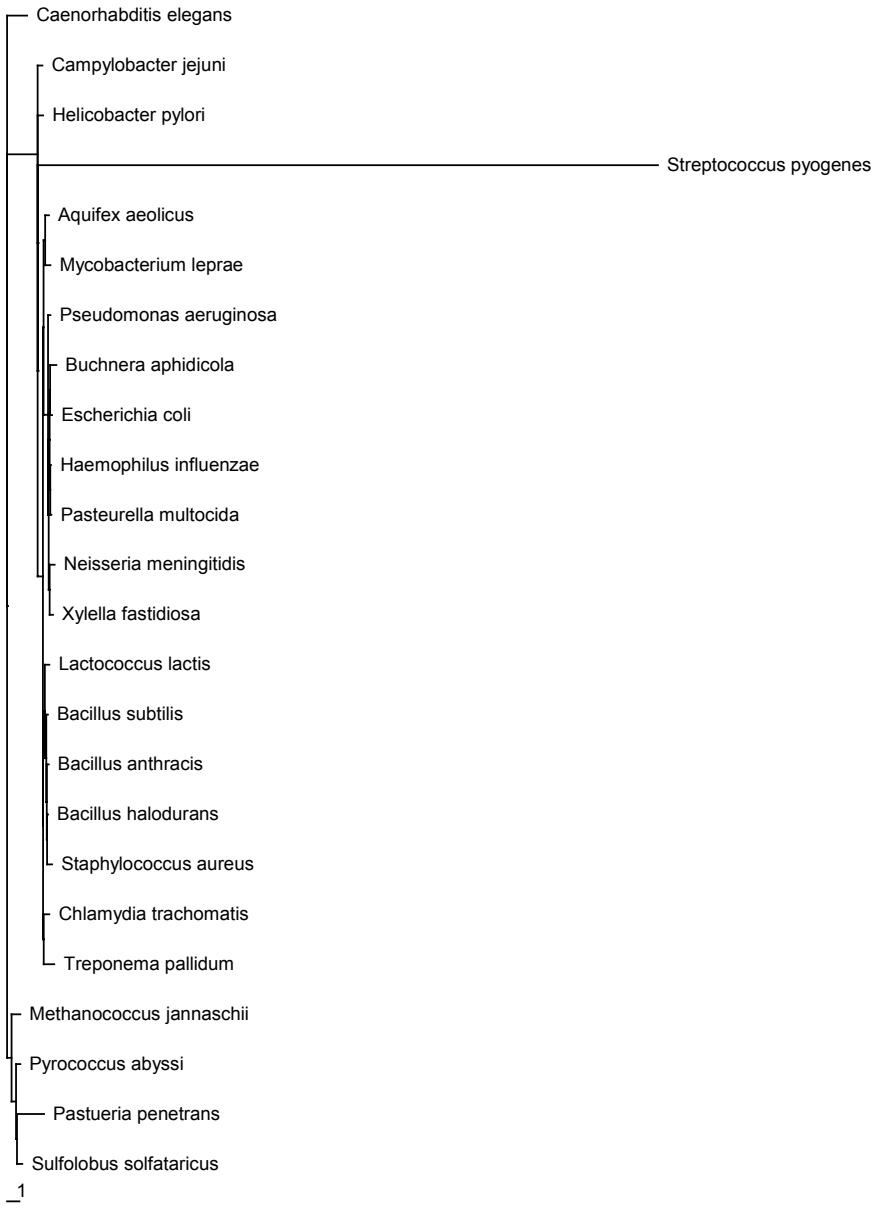

Figure A.2.4: Maximum likelihood single gene tree for *tkt* positioning *P. penetrans* within Crenarchaeota and Euryarchaeota. Branch lengths are shown.

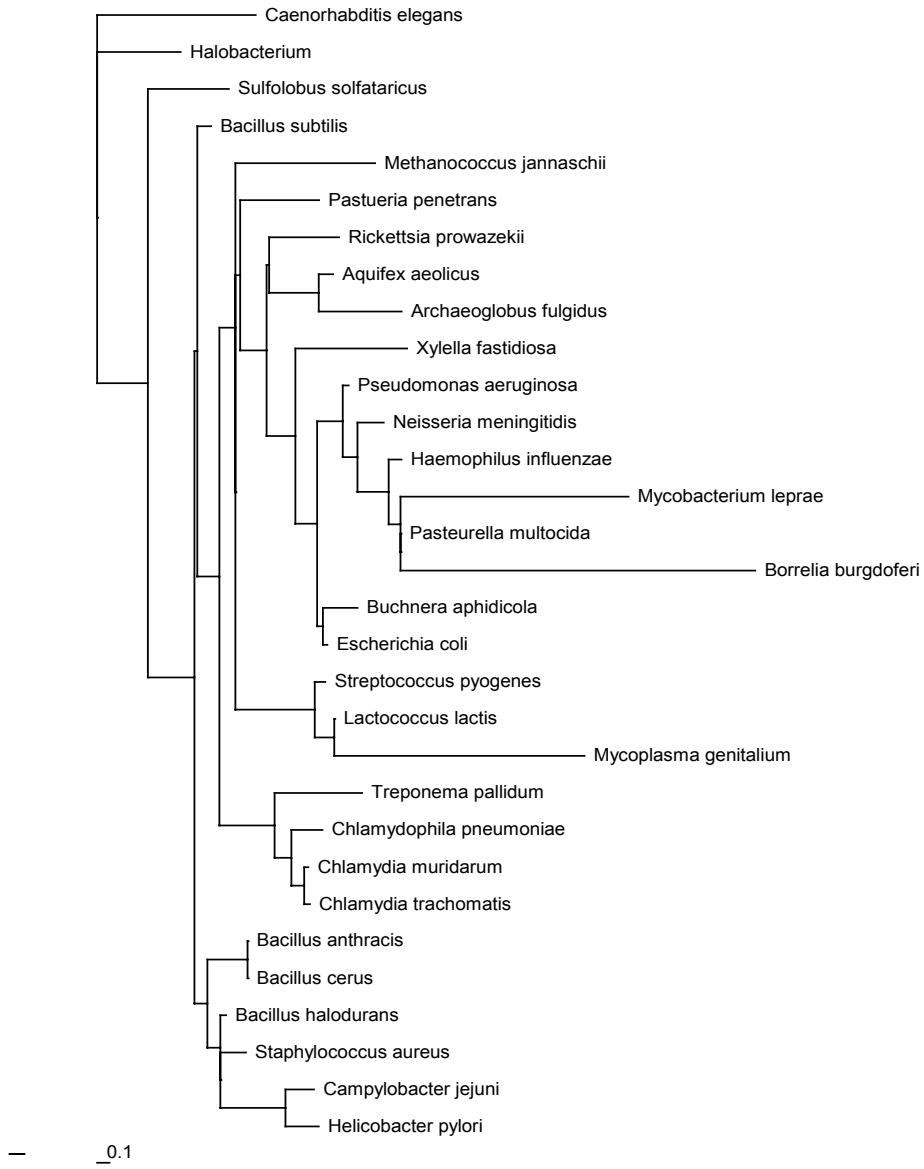

Figure A.2.5: Maximum likelihood single gene tree for *trxA* positioning *P. penetrans* adjacent to the Proteobacteria and Euryarchaeota. Branch lengths are shown.

Table A.3.1: Table of Sequences used in collagen-like sequence analyses.

GROUP	SPECIES	STRAIN	ACCESSION NUMBER	ANNOTATION	GENE	# GXY-REPEATS
Bacteria	*Bacillus anthracis*	str. Ames	gi|30262449|ref|NP_844826.1|	Conserved hypothetical protein	BaA.Hypo	186
		str. A2012	gi|21399133|ref|NP_655118.1|	Hypothetical protein predicted by GeneMark	Ba.A20.Hypo	186
		str. 6602	gi|28475233|emb|CAD56876.1|	BclA protein	Ba.A6602	84
		str. 6183	gi|28475231|emb|CAD56875.1|	BclA protein	Ba.6183	93
		str. 7611	gi|28475235|emb|CAD56877.1|	BclA protein	Ba.7611	81
		str. RA3	gi|28475237|emb|CAD56878.1|	BclA protein	Ba.RA3	87
		str. CIP 5725	gi|28475239|emb|CAD56879.1|	BclA protein	Ba.C5725	75
		str. CIP 53169	gi|28475225|emb|CAD56872.1|	BclA protein	Ba.C5316	198
		str. CIP 8189	gi|28475223|emb|CAD56871.1|	BclA protein	Ba.C8189	219
			gi|49187418|ref|YP_030670.1|	Conserved hypothetical protein, collagen triple helix repeat domain	BaS.hCTHR	141
		str. Sterne	gi|49186559|ref|YP_029811.1|	Conserved hypothetical protein	BaS.Hypo	189
	Bacillus cereus		gi|30020512|ref|NP_832143.1|	Collagen triple helix repeat protein	BcA14.THR1	162
			gi|30020762|ref|NP_832393.1|	Collagen triple helix repeat protein	BcA14.THR5	102
			gi|30021584|ref|NP_833215.1|	Collagen triple helix repeat protein	BcA14.THR6	81
			gi|30022842|ref|NP_834473.1|	Collagen triple helix repeat protein	BcA14.THR7	132
			gi|42780403|ref|NP_977650.1|	Conserved hypothetical protein	BcA10.BIst	111
		str. ATCC 10987	gi|42781416|ref|NP_978663.1|	Collagen triple helix repeat domain protein	BcA10.THR	192
			gi|42782207|ref|NP_979454.1|	BclA protein	BcA10.BclA	198
			gi|47568416|ref|ZP_00239117.1|	BclA protein	BcG9.BclA	117
			gi|47565679|ref|ZP_00236719.1|	PE_PGRS family protein	BcG9.fam	192
		str. G9241	gi|47557938|gb|EAL16263.1|	Hypothetical protein membrane-spanning protein	BcG9.hMS	210
	Bacillus thuringiensis serovar konkukian		gi|23534591|gb|AAN34375.1|	Bcol14-2	Bt.Bcol14	261
		str. 97-27	gi|49477705|ref|YP_036568.1|	BclA protein	Bt.BclA2	168
			gi|49477136|ref|YP_035447.1|	BclA protein	Bt.BclA	120
			gi|49477241|ref|YP_035673.1|	Collagen-like protein	Bt.CL1	420
			gi|49478679|ref|YP_038578.1|	Conserved hypothetical protein, collagen triple helix repeat domain	Bt.hCTHR	216
	Streptococcus pyogenes		gi|15675046|ref|NP_269220.1|	Putative collagen-like protein	Sp.M1G.CL	114
		str. M1 GAS	gi|15675773|ref|NP_269947.1|	Collagen-like surface protein	Sp.M1G.scl1	159
		str. MGAS 10394	gi|50914143|ref|YP_060115.1|	Collagen-like surface protein	Sp.MG1.cSf	159
		str. MGAS 315	gi|21910274|ref|NP_664542.1|	Collagen-like protein SclB	Sp.MG3.cSB	348
		str. SSI-1	gi|28895851|ref|NP_802201.1|	SclB protein	Sp.SSI.SB	222
	Pasteuria penetrans	res147	unpublished data	Collagen-like protein	c136.4G	102
			unpublished data	Collagen-like protein	c178.1G	237
			unpublished data	Collagen-like protein	c250.3	78
			unpublished data	Collagen-like protein	c280.2G	132
			unpublished data	Collagen-like protein	c336.2V	195
			unpublished data	Collagen-like protein	c359.2G	99
			unpublished data	Collagen-like protein	c374.4	105
			unpublished data	Collagen-like protein	c384.4	108
			unpublished data	Collagen-like protein	c391.3G	186
			unpublished data	Collagen-like protein	c453.1G	162
			unpublished data	Collagen-like protein	CHOP015.4G	93
			unpublished data	Collagen-like protein	CHOP033.4G	105
			unpublished data	Collagen-like protein	CHOP034.1G	111
			unpublished data	Collagen-like protein	CHOP035.4G	108
			unpublished data	Collagen-like protein	CHOP047.1G	111
	Clostridium perfringens	str. 13	gi|18309937|ref|NP_561871.1|	Collagen-like protein	Cp13.CL	180
	Corynebacterium diphtheriae		gi|38200670|emb|CAE50366.1|	Collagen-like repeat protein	Cd.cR	123

Table A.3.1 (continued)

GROUP	SPECIES	STRAIN	ACCESSION NUMBER	ANNOTATION	GENE	# GXY-REPEATS
	Anopheles gamiae		gi|31214654|ref|XP_315876.1|	ENSANGP00000004496	An.4496	138
			gi|31195073|ref|XP_306484.1|	ENSANGP00000014653	An.14653	309
			gi|31208179|ref|XP_313056.1|	ENSANGP00000022523	An.22523	120
			gi|31218001|ref|XP_316547.1|	ENSANGP00000022865	An.22865	243
			gi|31240943|ref|XP_320885.1|	ENSANGP00000023491	An.23491	183
			gi|31231764|ref|XP_318589.1|	ENSANGP00000025235	An.25235	159
	Apis mellifera		gi|48098141|ref|XP_393988.1|	Similar to ENSANGP00000003674	Ap.3674	348
			gi|48112109|ref|XP_396317.1|	Similar to CG33171-PC	Ap.33171PC	447
	Araneus ventricosus		gi|47606845|gb|AAT36347.1|	Flagelliform silk protein-1	Ara.1	381
			gi|27228957|gb|AAN85280.1|	Major ampullate gland dragline silk protein-1	Ara.2	339
	Argiope trifasciata		gi|13561980|gb|AAK30593.1|	AF350264_1 flagelliform silk protein	Arg.2	462
	Bombyx mori		gi|542527|pir||S42886	Collagen	Bm	324
	Drosophila melangaster		gi|24647572|ref|NP_650587.1|	CG5225-PA	Dm.5225PA	171
			gi|24658386|ref|NP_729073.1|	CG10625-PD	Dm.10625PD	129
			gi|24644030|ref|NP_649481.1|	CG12586-PA	Dm.12586PA	78
			gi|21356133|ref|NP_650906.1|	CG4000-PA	Dm.14000PA	102
			gi|24647472|ref|NP_650557.1|	CG14888-PA	Dm.14888PA	153
			gi|24647474|ref|NP_650558.1|	CG14889-PA	Dm.14889PA	171
			gi|24647476|ref|NP_650559.1|	CG14890-PA	Dm.14890PA	81
			gi|28317085|gb|AAO39561.1|	LP07855p	Dm.LP07855	237
	Drosophila yakuba		gi|38048369|gb|AAR10087.1|	Similar to Drosophila melanogaster lcs	D.yakuba	90
	Galleria mellonella		gi|9087201|sp|O96614|	SER1_GALME Sericin-1 (Silk gum protein 1)	Gm	84
	Alvinella pompejana		gi|5174770|gb|AAC35289.2|	Fibrillar collagen chain FAp1 alpha	Alvinella	615
	Ceanorhabditis briggsae		BP:CBP06235	CBG01972	Cb.01972	150
			BP:CBP01349	CBG05354	Cb.05354	150
			BP:CBP15518	CBG06051	Cb.06051	150
			BP:CBP15728	CBG06704	Cb.06704	162
			BP:CBP07621	CBG07415	Cb.07415	141
			BP:CBP07626	CBG07436	Cb.07436	120
			BP:CBP09006	CBG11985	Cb.11985	150
			BP:CBP09252	CBG12558	Cb.12558	132
			BP:CBP03357	CBG13781	Cb.13781	150
			BP:CBP03481	CBG14077	Cb.14077	162
			BP:CBP10301	CBG16654	Cb.16654	159
			BP:CBP04218	CBG17762	Cb.17662	156
			BP:CBP11139	CBG19328	Cb.19328	162
			BP:CBP05558	CBG23461	Cb.23461	150
	Ceanorhabditis elegans		WP:CE07013	dpy-3 EGAP7.1	Ce.dpy3	162
			WP:CE05707	sqt-3 F23H12.4	Ce.sqt3	150
			WP:CE07185	col-34 F36A4.10	Ce.Col34	153
			WP:CE08169	col-35 C15A11.1	Ce.Col35	174
			WP:CE13609	col-40 T13B5.4	Ce.Col40	150
			gi|17568109|ref|NP_510274.1|	Collagen structural gene (col-44)	Ce.Col44	141
			gi|17509327|ref|NP_491194.1|	Collagen structural gene (col-50)	Ce.Col50	162
			WP:CE35183	col-83 F33A8.9	Ce.Col83	144
			WP:CE26306	col-106 Y77E11A.15	Ce.Col106	153
			WP:CE08873	col-161 C50B6.4	Ce.Col161	162
			WP:CE34552	col-172 F38B6.5b	Ce.Col172b	90
			WP:CE02380	col-176 ZC373.7	Ce.Col176	150
			WP:CE34569	col-185 H06A10.2	Ce.Col185	141
	Hemicentrotus pulcherrimus		gi|630905|pir||S42731	Collagen alpha 1 chain	Hp.A1	387
			gi|13325096|gb|AAB30065.2|	Fibrillar collagen alpha 120 and 140 chains	Hp.A120	378
	Paracentrotus		gi|280636|pir||A36226	Collagen alpha 1 chain	Pl.A1	477
Invertebrate	Pig roundworm		gi|320995|pir||A44982	Collagen UCOL1	Pig.UCOL1	81

Table A.3.1 (continued)

GROUP	SPECIES	STRAIN	ACCESSION NUMBER	ANNOTATION	GENE	# GXY-REPEATS
Vertebrate	*Bos taurus*		gi|47564048|ref|NP_001001135.1|	Cyanogen bromide	Bos.CBr	279
			gi|28380296|sp||P02459_1	[Segment 1 of 2] Collagen alpha 1(II) chain precursor	Bos.A1IIA	551
			gi|2144804|pir||CGBO6C	Collagen alpha 1(II) chain precursor	Bos.A1IIc	477
	Gallus gallus		gi|63308|emb|CAA23695.1|	Collagen	Gg.col	201
			gi|2119161|pir||I50696	Collagen alpha 1(III) chain	Gg.A1IIIf	309
	Homo sapien		gi|38014150|gb|AAH08760.3|	COL5A1 protein	Hs.col5A1	324
			gi|179594|gb|AAB59373.1|	Alpha-1 type I collagen	Hs.A1I	468
			gi|30095|emb|CAA42933.1|	Collagen subunit (alpha-1 (X)) 3	Hs.A1X	447
			gi|30016|emb|CAA30731.1|	Unnamed protein product	Hs.unnamed	351
			gi|825646|emb|CAA23761.1|	Unnamed protein product	Hs.unname2	483
	Mus musculus		gi|13096810|gb|AAH03198.1|	Col1a1 protein	Mus.col1a1	318
			gi|30353888|gb|AAH52326.1|	Col2a1 protein	Mus.2col2a1	555
			gi|192264|gb|AAA37334.1|	Procollagen type I alpha chain	Mus.AI	396
			gi|192262|gb|AAA37333.1|	Pro-alpha-1 type I collagen	Mus.ProA1I	610
	Rattus norvegicus		gi|57916|emb|CAA49832.1|	Pro1 collagen type III	Rn.Pro1III	369
	Macaca mulatta		gi|13182888|gb|AAK14972.1|	Collagen type I alpha 1	Mm.A1I	185
Fungus	*Aspergillus nidulans*	str. FGSC A4	gi|40744228|gb|EAA63404.1|	Hypothetical protein AN2833.2	As.AN2833	204
			gb|EAA63603.1|	Hypothetical protein AN3032.2	As.AN3032	96
			gb|EAA60823.1|	Hypothetical protein AN4480.2	As.AN4480	174
			gi|40747178|gb|EAA66334.1|	Hypothetical protein AN9267.2	As.AN9267	117
	Candida glabrata		gi|50290767|ref|XP_447816.1|	Unnamed protein product	Ca.unname1	75
			gi|50290763|ref|XP_447814.1|	Unnamed protein product	Ca.unname2	81
			gi|50284839|ref|XP_444847.1|	Unnamed protein product	Ca.unname3	78
	Magnaporthe grisea		gi|38099567|gb|EAA46894.1|	Predicted protein	Mg.pred	198
			gi|38103856|gb|EAA50503.1|	Hypothetical protein MG04262.4	Mg.MG04262	423
			gi|38103166|gb|EAA49902.1|	Hypothetical protein MG10066.4	Mg.MG10066	87
	Eremothecium gossypii		gi|45201061|ref|NP_986631.1|	AGL035Wp	Eg.AGL035	132
	Debaryomyces hansenii		gi|50422935|ref|XP_460045.1|	Unnamed protein product	Dh.unnamed	174
	Ustilago maydis		gi|46100457|gb|EAK85690.1|	Hypothetical protein UM04422.1	Um.UM04422	72
	Neurospora crassa		gi|28917359|gb|EAA27065.1|	Hypothetical protein	Nc.hypo	156
	Yarrowia lipolytica		gi|50554645|ref|XP_504731.1|	Hypothetical protein	Yarrowia	147

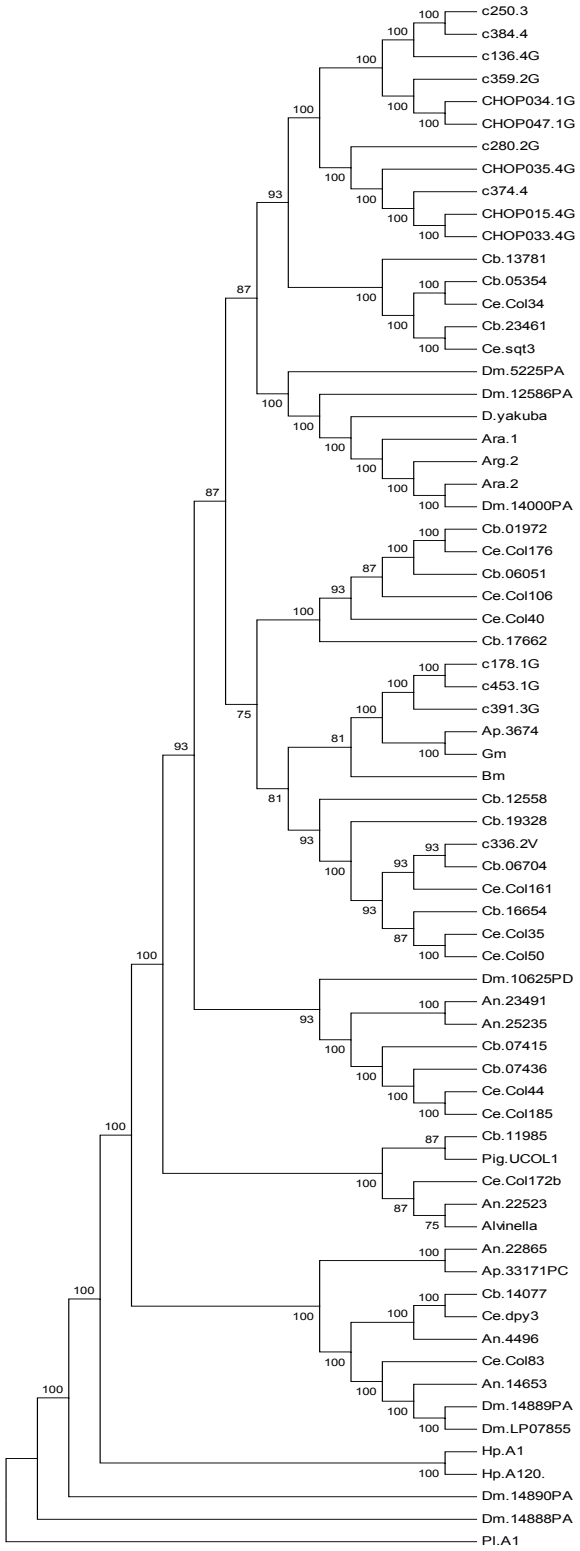

Figure A.3.1: Maximum Parsimony Tree of collagen-like repeat motifs performed on *Pasteuria penetrans* sequences and invertebrates. Bootstrap values over 50 are shown on branches. Refer to Table A.3.1 for full names.

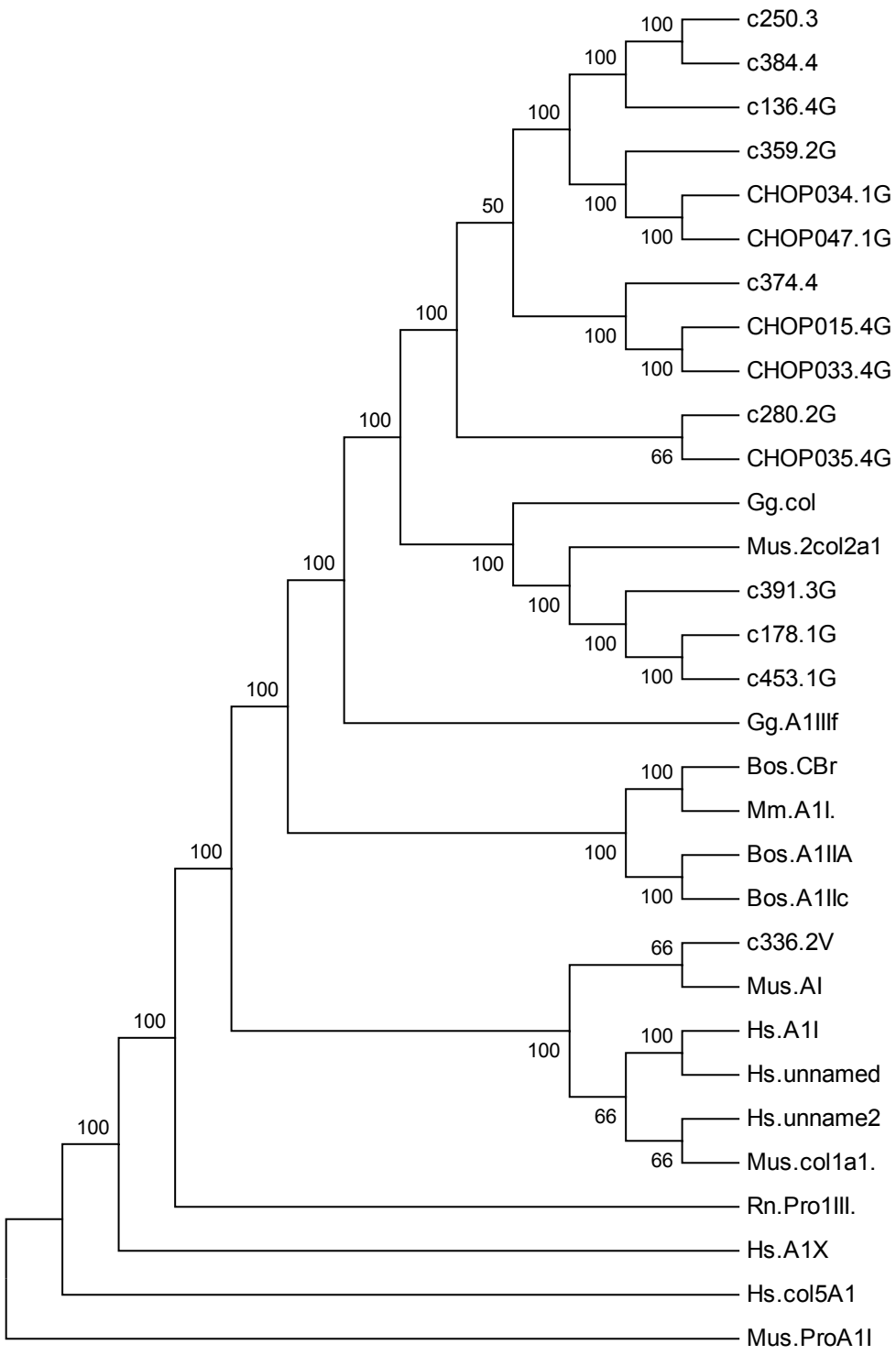

Figure A.3.2: Maximum Parsimony Tree of collagen-like repeat motifs performed on *Pasteuria penetrans* sequences and vertebrates. Bootstrap values over 50 are shown on branches. Refer to Table A.3.1 for full names.

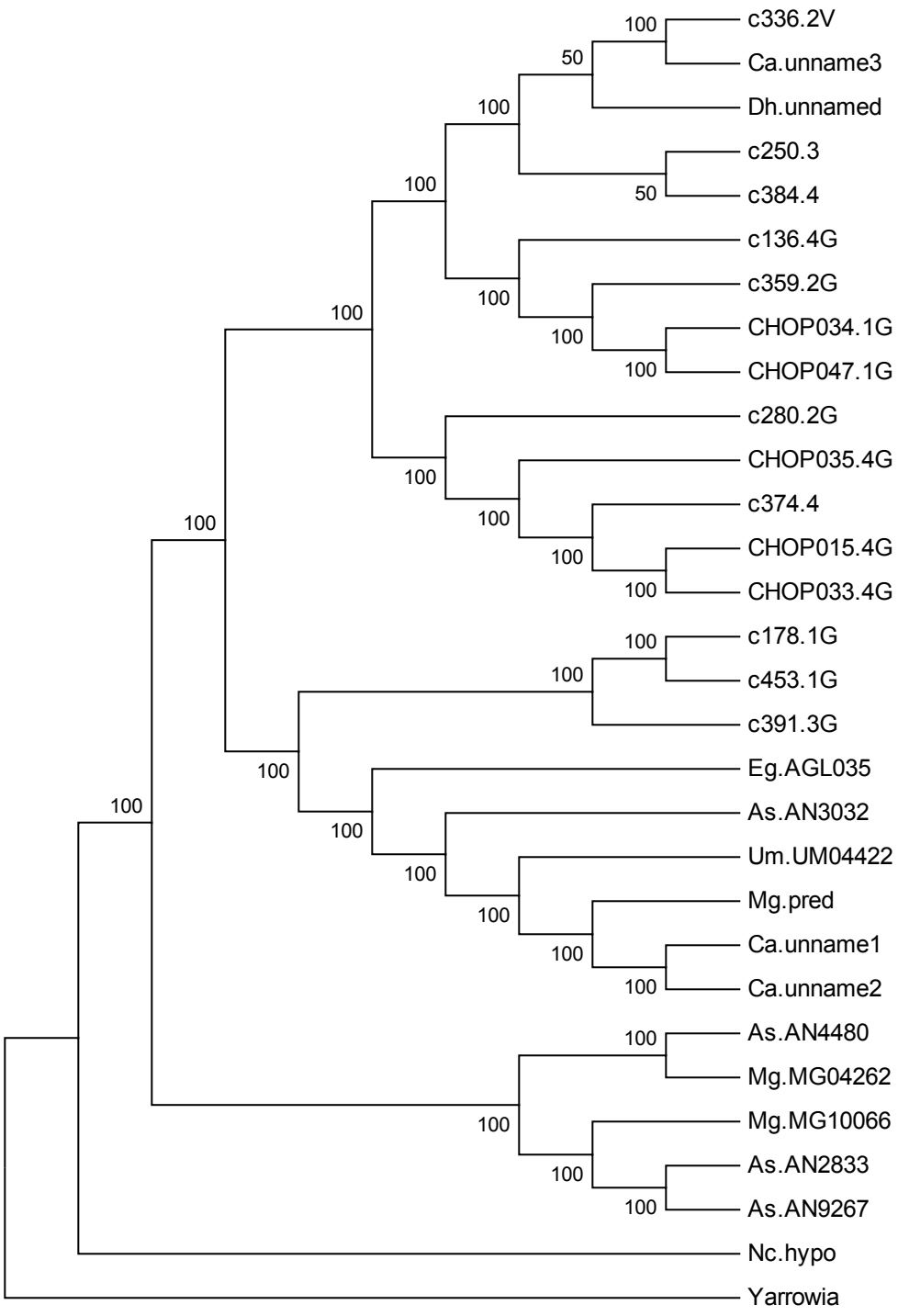

Figure A.3.3: Maximum Parsimony Tree of collagen-like repeat motifs performed on *Pasteuria penetrans* sequences and fungi. Bootstrap values over 50 are shown on branches. Refer to Table A.3.1 for full names.

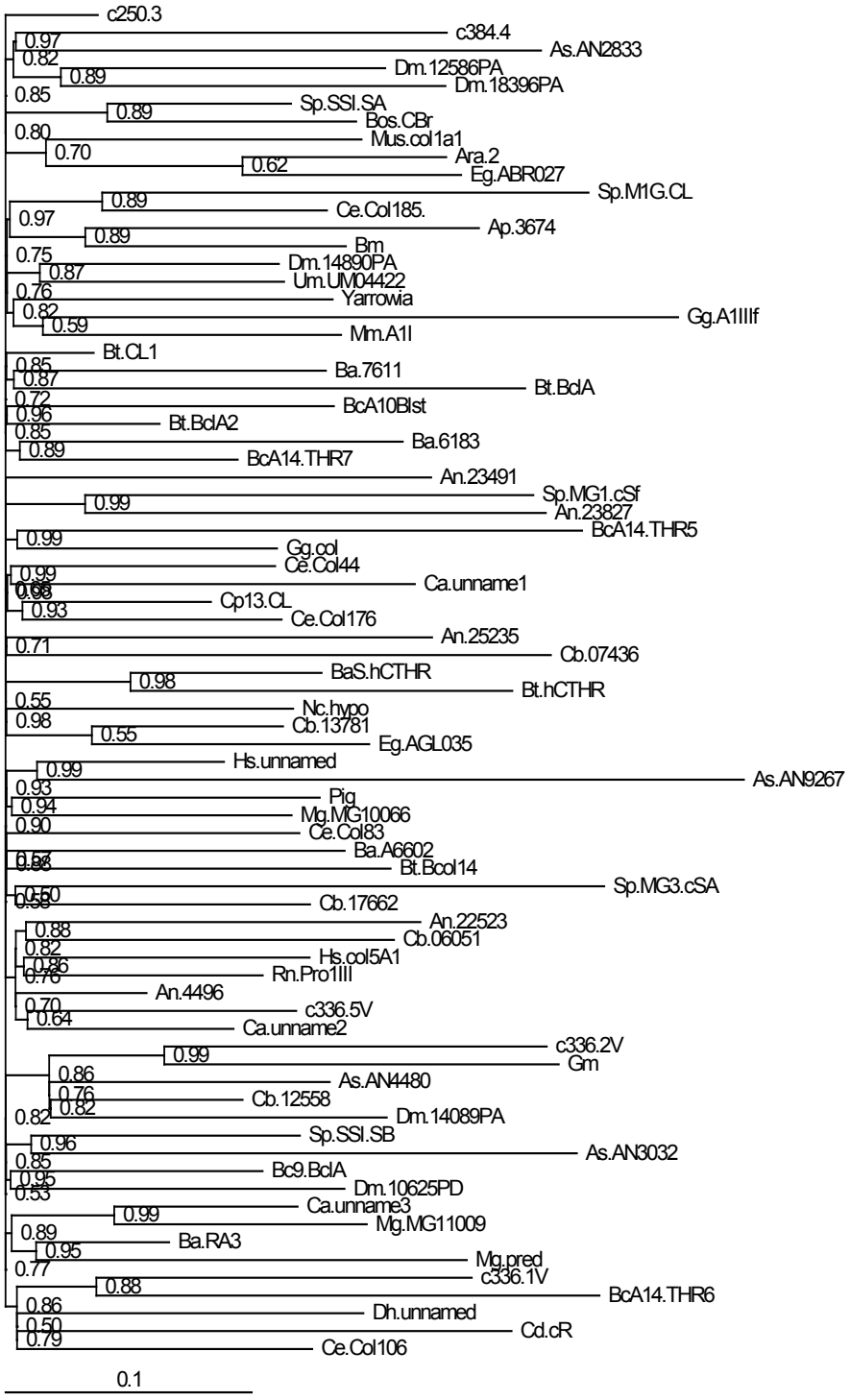

Figure A.3.4: Bayesian Tree of collagen-like repeat motifs performed on *Pasteuria penetrans* Group 1 sequences and the five closest identity sequences from each species. Ran for 100,000 generations, branch lengths are shown wherever possible. Refer to Table A.3.1 for full names.

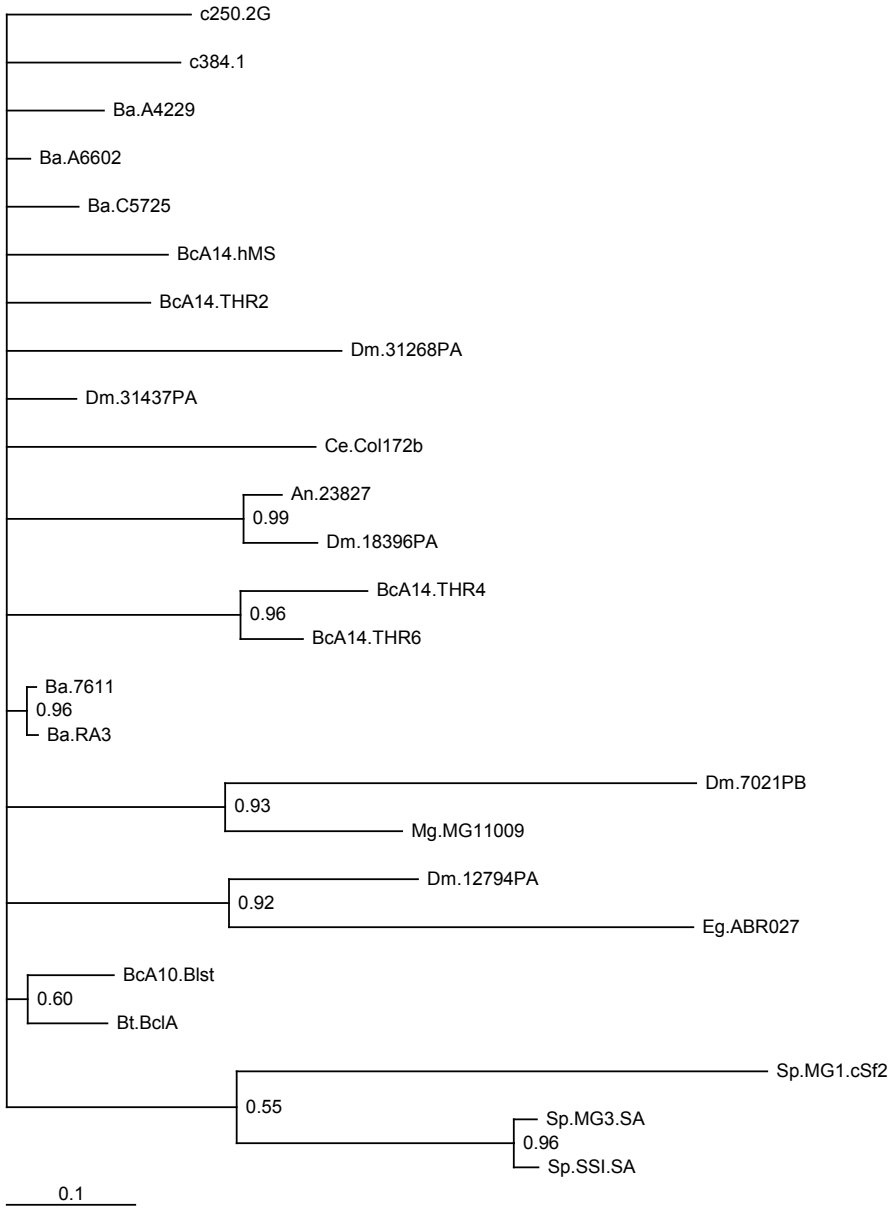

Figure A.3.5: Bayesian analysis of collagen-like repeat motifs performed on *Pasteuria penetrans* Group 2 sequences and the five closest identity sequences from each species. Ran for 100,000 generations, branch lengths are shown wherever possible. Refer to Table A.3.1 for full names.

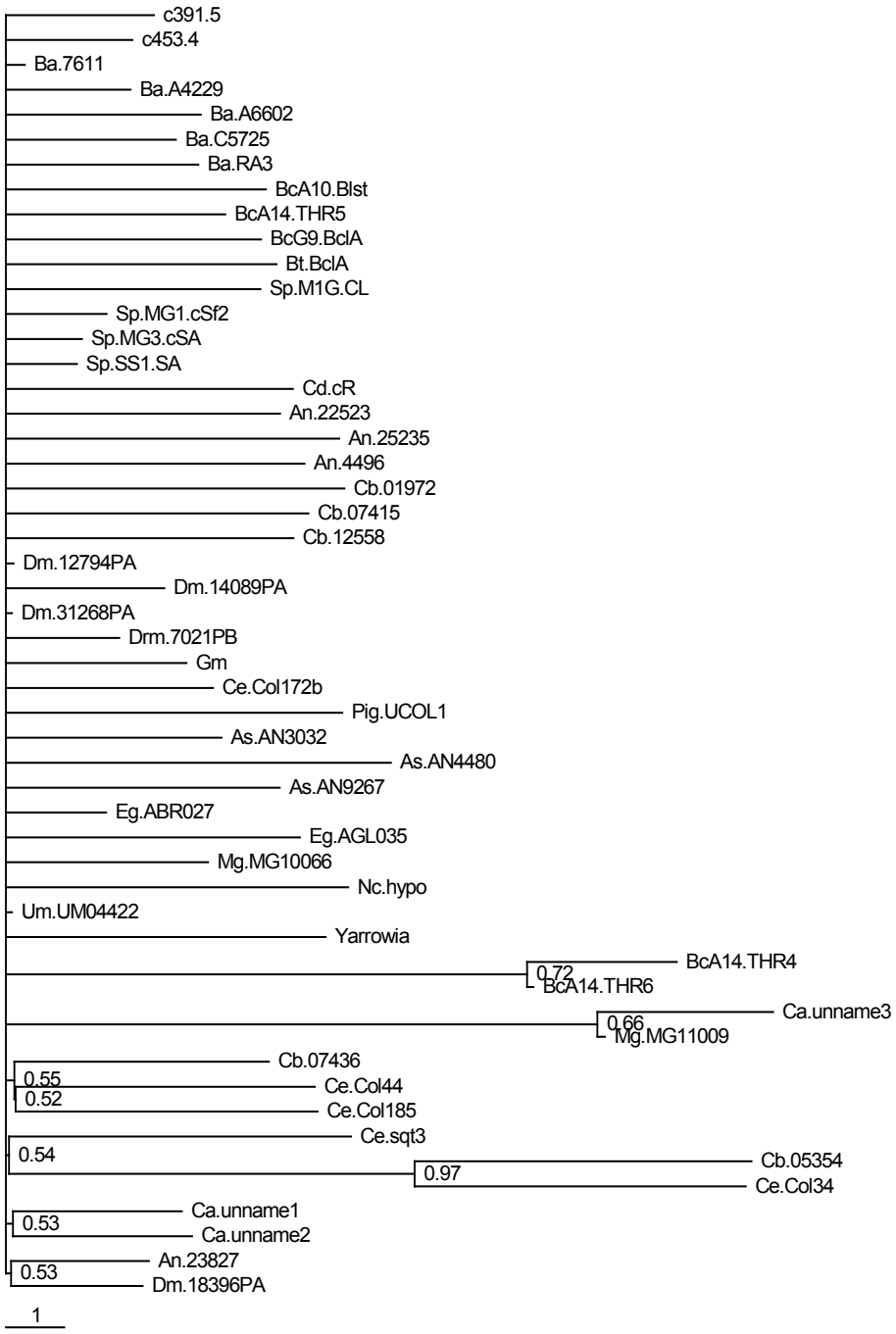

Figure A.3.6: Bayesian analysis of collagen-like repeat motifs performed on *Pasteuria penetrans* Group 3 sequences and the five closest identity sequences from each species. Ran for 100,000 generations, branch lengths are shown wherever possible. Refer to Table A.3.1 for full names.

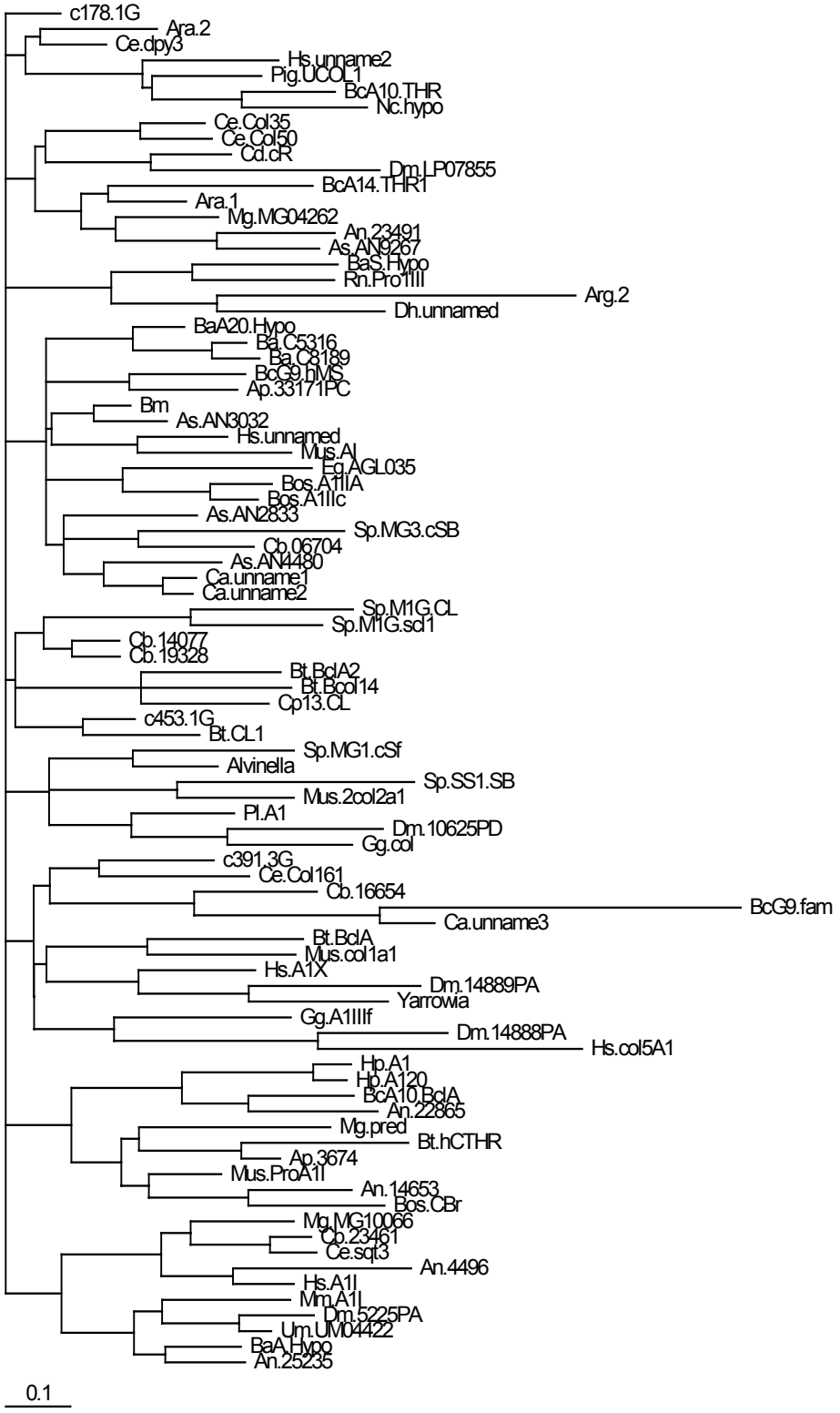

Figure A.3.7: Bayesian analysis of collagen-like repeat motifs performed on *Pasteuria penetrans* Group 4 sequences and the five closest identity sequences from each species. Ran for 100,000 generations, branch lengths are shown wherever possible. Refer to Table A.3.1 for full names.

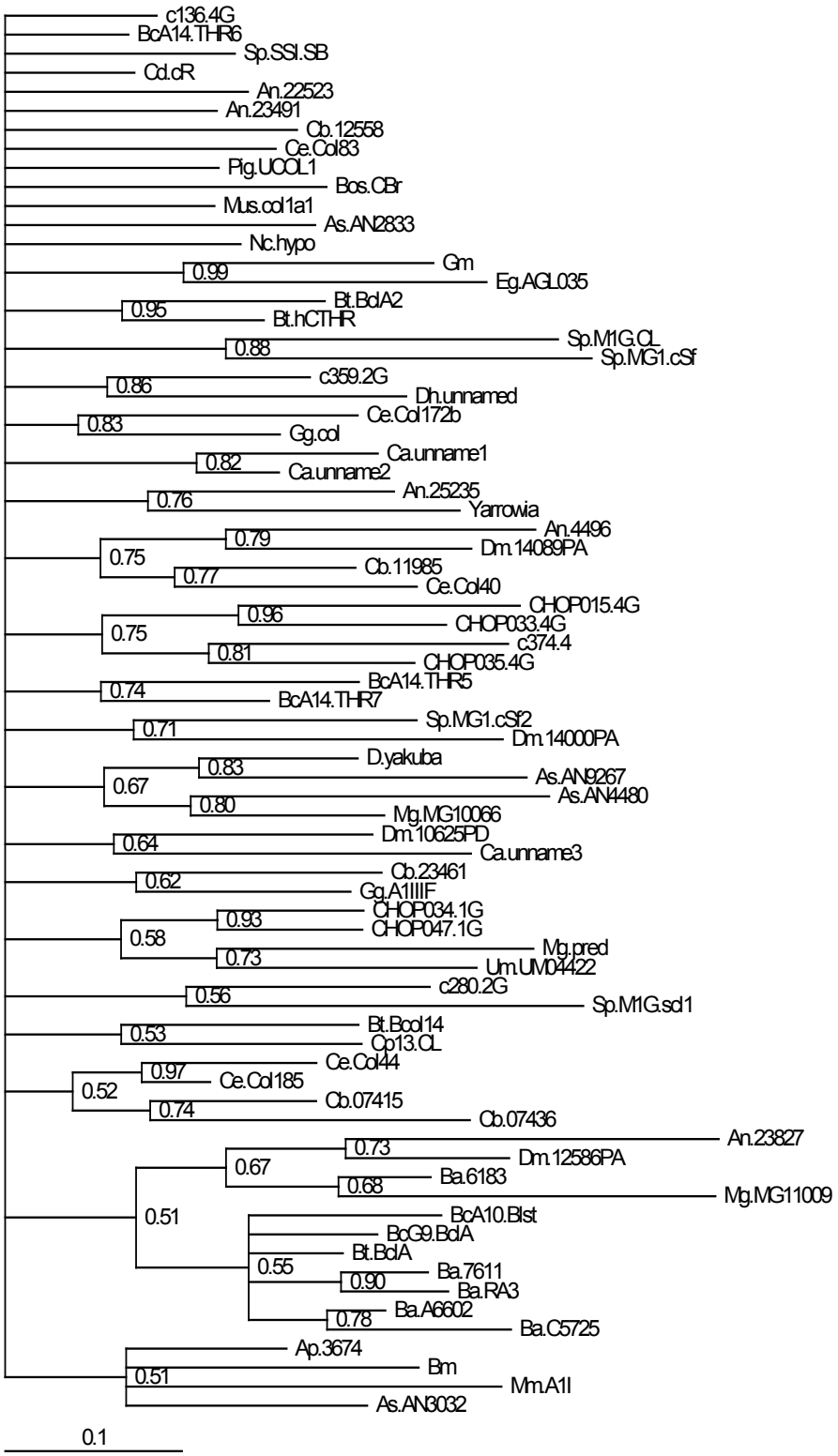

Figure A.3.8: Bayesian analysis of collagen-like repeat motifs performed on *Pasteuria penetrans* Group 5 sequences and the five closest identity sequences from each species. Ran for 100,000 generations, branch lengths are shown wherever possible. Refer to Table A.3.1 for full names.

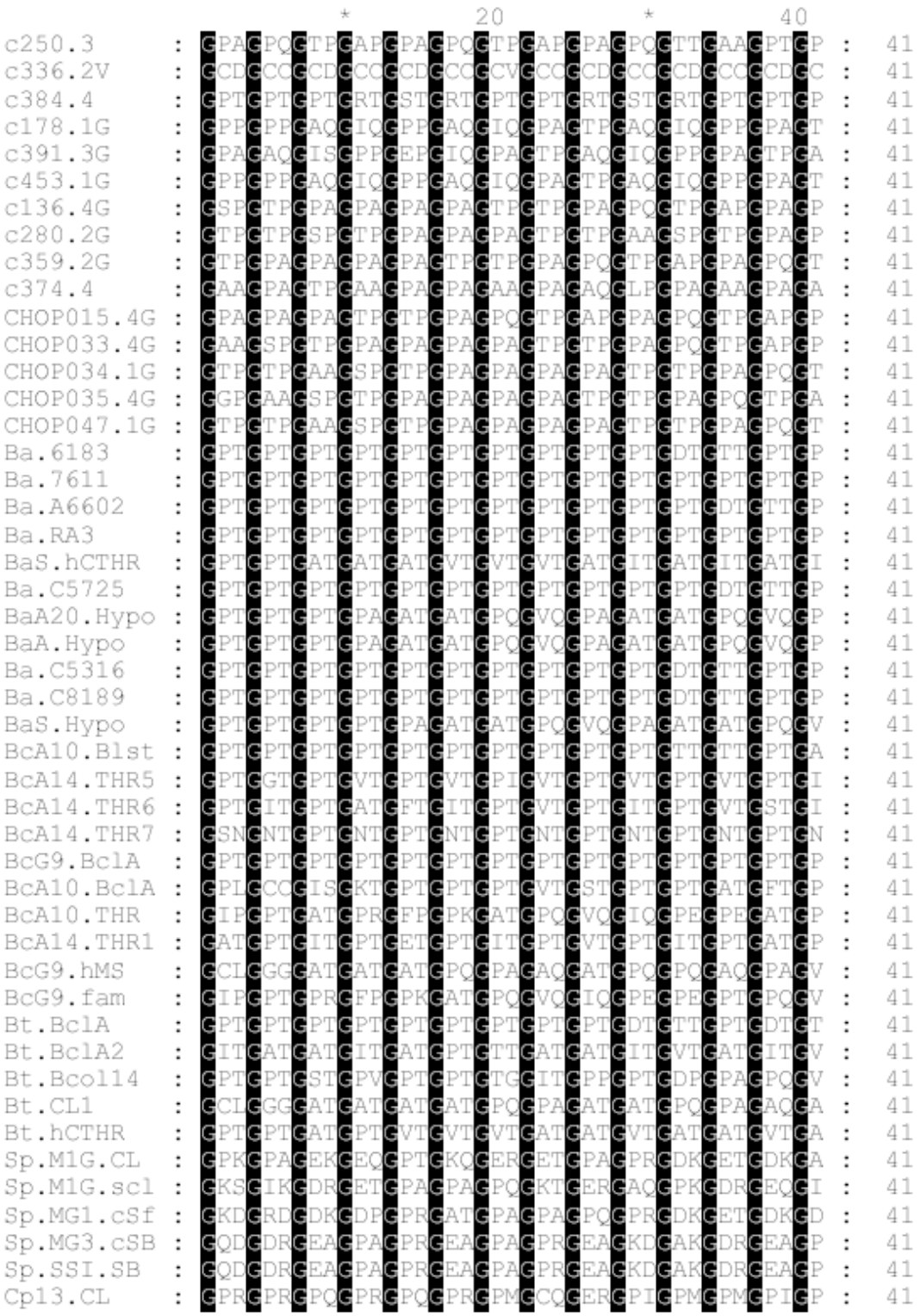

Figure A.3.9: The first forty-one amino acids from the GXY-motif alignment.

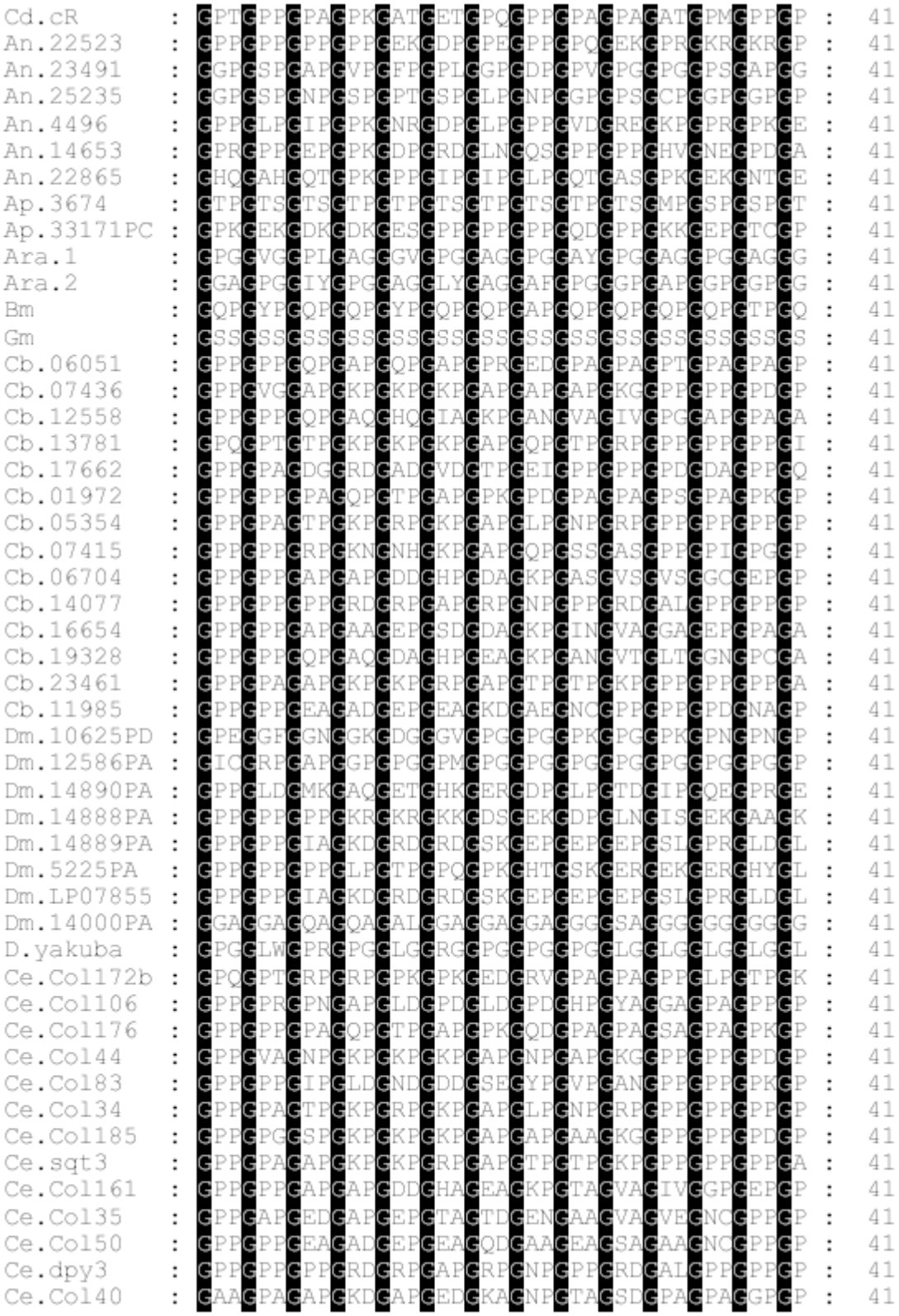

Figure A.3.9 (continued)

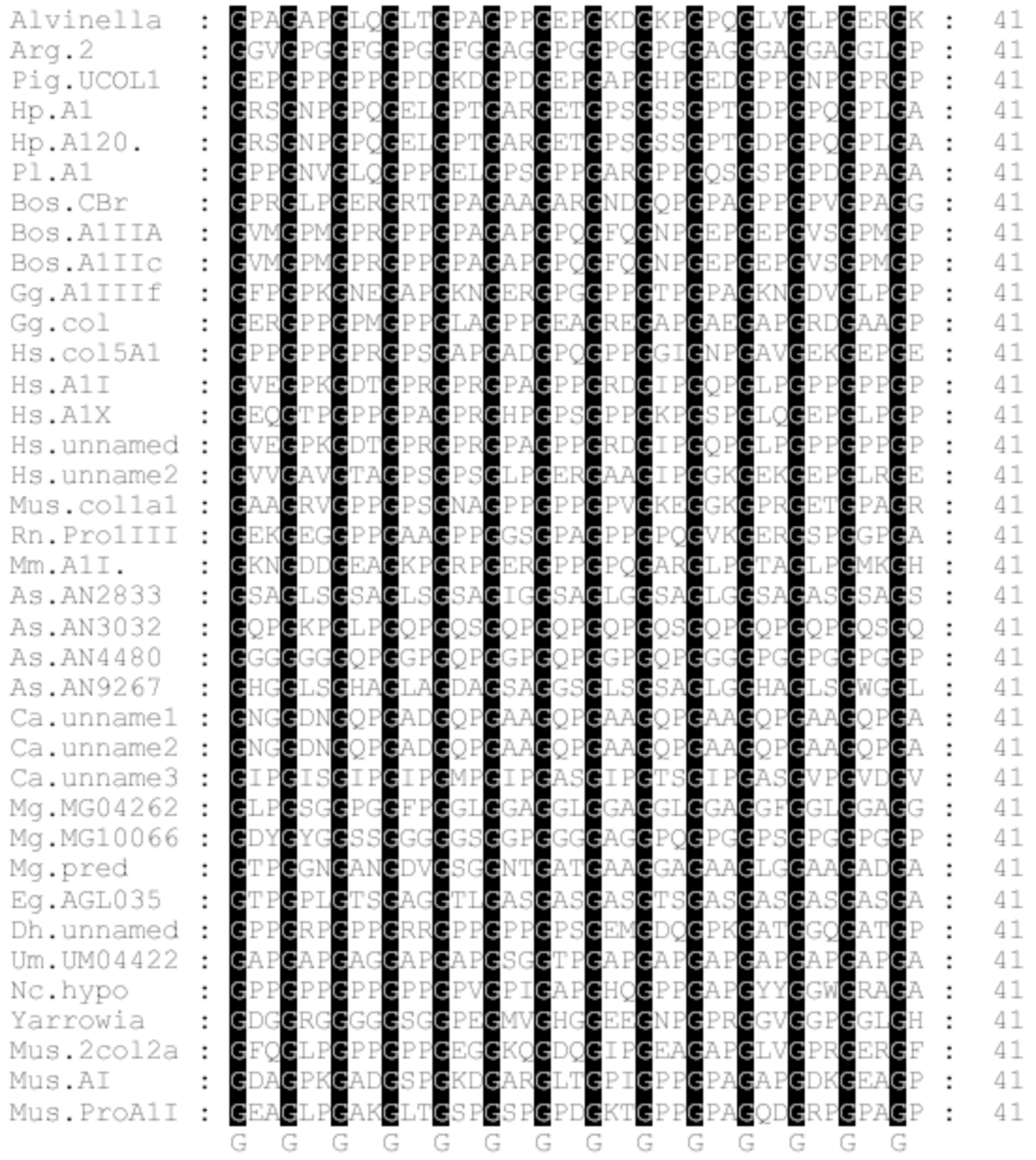

Figure A.3.9 (continued)

Script A.3.1: Pearl script developed to separate the GXY collagen-like repeats from the beginning of the sequences.

```
#!/usr/bin/perl

use strict;
use diagnostics;
use warnings;

##Declare some variables##

my $infile = "";
my $all = "";
my $line = "";
my @sequence = ();
my @fasta = ();
my $midend = "";
my @id = ();
my $num = "";
my $sequencenumber = "";

## Open an outfile ##

open (OUT, ">MidEndSeqs.txt")or die "Can't open: $!\n";

## open each file that ends in .seq, one at a time ##

while ($infile=<*.seq>)
{

        open (IN, "$infile") or die "Can't open: $!\n";

        print "Processing file $infile...\n";

## Go through each line of the input file:##
        while ($line = <IN>)
                {

                chomp $line;
                $all = $all.$line;
                }

## close the infile so you can open a new one next time around...##
        close (IN);
}
while ($infile=<*.seq>)
{
        open (IN, "$infile") or die "Can't open: $!\n";
        while ($line = <IN>)
```

Script A.3.1 (continued)

```
	{
		if($line =~ m/^(>)/)
			{
				push @id, $line;
			}
	}
}
	close (IN);

@sequence = split (/>/, $all);
$num = scalar(@id);
$sequencenumber = 0;

foreach my $sequence (@sequence)
	{


		if($sequence =~ m/((G..){4,}.*)/)
		{
		$midend = $1."\n";
		}

		while($midend =~ s/(.{1,60})//)
		{
		print OUT $1."\n";
		}
		print OUT $id[$sequencenumber];
	++$sequencenumber;
		}

## close the outfile - you're done! ##

close (OUT);
```

Script A.3.2: Pearl script developed to count the amino acids used in each collagen-like sequence.

```
#!/usr/bin/perl

use strict;
use diagnostics;
use warnings;

## Declare Variables##

my $infile = "";
my $all = "";
my $line = "";
my @sequence = ();
my %gene = ();
my @id = ();
my %number_aa_X = ();
my %number_aa_Y = ();
my $aa = "";
my $sequencenumber = "";
my $total_aa = "";

## Open an outfile ##

open (OUT, ">GXYPercent.txt")or die "Can't open: $!\n";

## open each file that ends in .fas, one at a time ##

while ($infile=<*.fas>)
{
	open (IN, "$infile") or die "Can't open: $!\n";
	print "Processing file $infile...\n";

	## Go through each line of the input file:##

	while ($line = <IN>)
		{
			if ($line =~ m/^(>)/)
			{
				$all = $all.$line;
			}
			else
			{
				chomp $line;
				$all = $all.$line;
			}
		}
```

Script A.3.2 (continued)

```
	## Close the infile so you can open a new one next time around##

	close (IN);
}

while ($infile=<*.fas>)
{
	open (IN, "$infile") or die "Can't open: $!\n";
	while ($line = <IN>)
	{
		if($line =~ m/^(>)/)
			{
				push @id, $line;
			}
	}
}
	close (IN);

$sequencenumber = 0;
@sequence = split (/>\S+\s/, $all);
shift @sequence;
print OUT
"NAME\tTOTAL\tA\tC\tD\tE\tF\tG\tH\tI\tK\tL\tM\tN\tP\tQ\tR\tS\tT\tV\tW\tY\t\tA\tC\tD\tE
\tF\tG\tH\tI\tK\tL\tM\tN\tP\tQ\tR\tS\tT\tV\tW\tY\n";
	foreach my $sequence (@sequence)
	{
		$total_aa = 0;
		%number_aa_X = ( A => 0, C => 0, D => 0,E => 0,F => 0,G => 0,H => 0,I
=> 0,K => 0,L => 0,M => 0,N => 0,P => 0,Q => 0,R => 0,S => 0,T => 0,V => 0,W => 0,Y
=> 0);
		%number_aa_Y = ( A => 0, C => 0, D => 0,E => 0,F => 0,G => 0,H => 0,I
=> 0,K => 0,L => 0,M => 0,N => 0,P => 0,Q => 0,R => 0,S => 0,T => 0,V => 0,W => 0,Y
=> 0);
		$aa = $sequence;
		while ($aa=~ s/(.)(.)(.)//)
		{
		$number_aa_X{$2}++;
		$number_aa_Y{$3}++;
		$total_aa = $total_aa+3;
		}
	chomp $id[$sequencenumber];
	print OUT "$id[$sequencenumber]\t$total_aa
\t$number_aa_X{A}\t$number_aa_X{C}\t$number_aa_X{D}\t$number_aa_X{E}\t$numbe
r_aa_X{F}\t$number_aa_X{G}\t$number_aa_X{H}\t$number_aa_X{I}\t$number_aa_X{K
}\t$number_aa_X{L}\t$number_aa_X{M}\t$number_aa_X{N}\t$number_aa_X{P}\t$numb
er_aa_X{Q}\t$number_aa_X{R}\t$number_aa_X{S}\t$number_aa_X{T}\t$number_aa_X{
V}\t$number_aa_X{W}\t$number_aa_X{Y}\t\t$number_aa_Y{A}\t$number_aa_Y{C}\t$n
```

Script A.3.2 (continued)

```
umber_aa_Y{D}\t$number_aa_Y{E}\t$number_aa_Y{F}\t$number_aa_Y{G}\t$number_aa
_Y{H}\t$number_aa_Y{I}\t$number_aa_Y{K}\t$number_aa_Y{L}\t$number_aa_Y{M}\t$
number_aa_Y{N}\t$number_aa_Y{P}\t$number_aa_Y{Q}\t$number_aa_Y{R}\t$number_a
a_Y{S}\t$number_aa_Y{T}\t$number_aa_Y{V}\t$number_aa_Y{W}\t$number_aa_Y{Y}\
n";
        ++$sequencenumber;
        }

#%gene = (NAME => split (/>\S+\s/, $all));
#print OUT $gene{NAME};
        close (OUT);
```