Analysis of the genome of the filarial nematode *Brugia malayi*

David Bernardo Guiliano

A thesis submitted for the degree of Doctor of Philosophy of the University of Edinburgh

Institute of Cell Animal and Population Biology
University of Edinburgh
Ashworth Laboratories
Kings Buildings
West Mains Rd.
Edinburgh EH9 3JT

2002

Acknowledgements

First I would like to thank my supervisor Dr. Mark Blaxter whose help and support both intellectual and financial made this project possible over these past four and a half years. His encouragement has allowed me to push myself farther than I would have ever thought possible.

I would also like to thank the members of the Blaxter lab both past and present (Jen , Aziz, Kate, Daphne, Claire, Marian, Peter, JP, Mark D, Eyualem, Bill and Robin) who continually supported my work with reagents and advise, helped me adjust to Scotland, and made coming to work everyday so much fun.

Thanks to all the people on the third floor in the Allen and Maziels labs who also supplied me with reagents and support particularly Janice, Yvonne, and Bill who where so generous with the *Brugia*

This work would also not have been possible without the collaborations of the Williams and Slatko labs and I owe Steve, Barto, and their labs many beers for all the help and hospitality they have given me.

I would like to thank my friends and parents for the unwavering financial and spiritual support through the dark days of my write up. Special thanks to the dirty ladies of Tackno Jane, Lubby, Cath, Trace, and Cheza for the great nights out that helped me forget my deadlines.

Finally I would like to say thanks to John for his love and support through the whole process and his endless patience during those dark days of separation. John I couldn't have done it without you.

The work presented here was made possible by financial support of the UK Medical Research Counsel.

List of Abbreviations

ADT: Abundant differentially expressed transcript
BLAST: Basic alignment search tool
CLOBB: Clustering on the basis of blast similarity
HSP: High scoring pair
LSU: Large subunit
MCMC: Markov chain Monte Carlo
Mf: Microfilaria
mif: macrophage migration inhibitory factor
mt: mitochondria
MP: Maximum Parsimony
NJ: Neighbor Joining
rp: ribosomal protein
rRNA: ribosomal RNA
SSU: small subunit

Abstract

Filarial nematodes infect over 140 million people worldwide causing widespread morbidity in endemic populations. There is currently no available prophylactic vaccine and drug therapies are not effective against all stages of the parasites. The genomes of two filarial nematodes *Brugia malayi* and *Onchocerca volvulus* are being extensively studied to identify novel drug and vaccine candidates. *B. malayi* and *O. volvulus* have three genomes, the mitochondrial and nuclear genomes which are shared by all animals as well as a third genome contained within a bacterial endosymbiont. The mitochondrial genomes have been fully sequenced and characterized and the genomes of the bacterial endosymbionts are currently being fully sequenced. Full shotgun sequencing of the nuclear genome of *B. malayi* is due to begin within the next year. The filarial genome project laboratories have produced sets of staged cDNA and large insert genomic libraries (in bacterial artificial chromosomes, BACs) for *B. malayi* and *O. volvulus*. Large EST datasets have been generated from both species and a physical map of the nuclear genome of *B. malayi* is currently being assembled. Genomic contigs are being generated through a combination of the end sequencing of BAC clones and hybridization of BAC clone ends and *B. malayi* genes to the genome libraries. The work presented in this thesis builds upon the resources generated by the filarial genome project and focuses the analysis of the gene content of the filarial genomes, the dynamics of gene and genome evolution between nematodes and the conservation of gene order (synteny) and structures between distantly related nematodes.

Automated methods for generating and annotating clusters representative of individual genes from the EST sequences were optimized and implemented. Subsequent analyses of the clustered EST datasets have given insights into the biology of filarial nematodes as well as providing new sets of target molecules for drug or vaccine based intervention strategies. Comparative analysis between the cluster datasets of *B. malayi* and *O. volvulus* and the proteins predicted from the distantly related model nematode *Caenorhabditis elegans* have identified parasite and nematode specific gene families.

Interestingly, when compared to the public databases the majority of the clusters generated from both filarial EST datasets are completely novel.

Two genes that have been discovered in the filarial EST datasets are homologues of the mammalian macropahge migration inhibitory factor (MIF). In vertebrates MIFs play a variety of roles in modulating the activities of immune cells. The filarial MIFs are believed to have potential immunomodulatory, functions. The evolutionary relationship of the filarial MIFs and MIFs from other metazoans, plants and protozoa were analyzed by comparison of the protein sequences and several phylogenetic techniques. The results indicate that MIFs are an ancient gene family that have been duplicated early in the animal lineage. However, while the filarial MIF1s may be evolutionarily distant from the vertebrate MIF1s they are predicted to share similar enzymatic activities and substrate preferences that are distinct from other MIF1 genes. These features are not conserved in the putative MIF orthologues isolated from other protostomes lending some support to the possibility that their conservation is due to their interaction with common receptors in the mammalian immune system.

In collaboration with the Pathogen Sequencing Unit at the Sanger Institute several sections of the *B. malayi* genome have been sequenced and these have demonstrated that there is conservation of long-range and microsynteny between the genomes of *B. malayi* and *C. elegans*. This data is the first demonstration of conserved synteny between two distantly related protostomes and has shown that in nematode genomes intrachromosomal rearrangements are much more common than interchromosomal translocations.

By rationally selecting and screening sets of conserved genes several transcriptional operons from *B. malayi* were isolated. The *B. malayi* operons prove that these unusual genomic features are wide spread in the nematodes. By examining the 5' end of the downstream genes in the operons it has been determined that the usage of alternate spliced leaders in the resolution of operonic messenger RNAs is not a universal feature in nematodes

Chapter 1

Introduction

1.0 Parasitic Diseases and World Health

Parasitic nematodes are acknowledged as being one of the most important groups of human pathogens with over two billion people in developing nations infected with the three most prevalent gastrointestinal nematodes *Ascaris lumbricoides*, *Trichuris trichiura* and hookworm species (*Necator americanus* and *Ancylostoma doudenale*) (Chan, 1997). While these infections are not generally fatal, morbidity associated with high worm burdens has been shown to cause significant socio-economic problems particularly during childhood (Chan, 1997). Table 1.0.1 shows the number of people currently thought to be infected with seven of the most prevalent human parasitic nematodes (Chan, 1997; Molyneux, 1995; Ottesen and Ramachandran, 1995; Richards *et. al.*, 2001; Witt and Ottesen, 2001).

Disease	Causative Organism(s)	Estimated number infections in millions
Human Hookworm	*Necator americanus* and *Ancylostoma doudenale*	1277
Human Roundworm	*Ascaris lumbricoides*	1273
Human Whipworm	*Trichuris trichiura*	902
Human Filariasis	*Wucheria bancrofti*, *Brugia sp.* and *Onchocerca volvulus*	140

Table 1.0.1 Four of the major human nematode pathogens. The diseases common name, the species of nematode(s) and the estimated number of people infected worldwide is given. (Chan, 1997)

Most of these cases occur in developing countries. While the mortality associated with many bacterial and viral diseases is decreasing in these areas the incidence of and level of morbity caused by these nematode infections are increasing as the population grows and shift toward urban environments (Chan, 1997). In addition to directly affecting human health, nematodes that parasitise agriculturally important species can cause failures in crops and livestock that result in economic loss in many areas worldwide.

1.1 Filariasis as a World Health Issue

Tissue dwelling filarial nematodes are estimated to infect at least 140 million people worldwide. Unlike gastrointestinal parasites filarial nematodes can result in long term clinical morbidity and have been cited as a serious impediment to development in many areas of Africa, Asia, the Western Pacific and certain regions in the Americas (Molyneux, 1995; Ottesen and Ramachandran, 1995). Table 1.1.1 lists the characteristics of the two most prevalent filarial diseases.

	Lymphatic filariasis	Cutaneous filariasis
Species	*Wuchereria bancrofti* *Brugia malayi* *Brugia timori*	*Onchocerca volvulus*
Transmitting Vectors	mosquitoes belonging to the genera *Culux*, *Anopheles* and *Aedes*	blackflies belonging to the genus *Simulium*
Mammalian Host(s)	With the exception of *Brugia malayi* which can infect jirds, monkeys and cats the other lymphatic filaria exclusively infect humans and some primates.	*Onchocerca volvulus* can only infect humans and chimpanzees.
Estimated Number of People Infected World Wide	120 million (Ottesen and Ramachandran, 1995)	17.5 million (Molyneux, 1995)
Global Distribution	*Wuchereria bancrofti* is the most widely distributed of the filarial parasites occurring in the tropical zone of Africa, India, Asia and the Indo-Pacific. Small foci exist in South America and the Caribbean. *Brugia malayi* is distributed through Asia and the Indo-Pacific while *Brugia timori* is confined to several small islands of Indonesia	>95% of the 17.5 million people infected with onchocerciasis live in Africa in a zone that extends 15°N and 15°S in the west widening towards the east to reach the southern latitude of Malawi. Nigeria is a hyperendemic area which accounts for over one third of the infected individuals. Small foci exist in Guatemala, Mexico, Venezuela, Colombia and other parts of Central and South America as well as in Yemen and Saudi Arabia in the Middle East.
Pathologies	Mild to severe, chronic infection can lead to permanent disfigurement. These include lymphangitis, funiculitis and hydrocoele, chyluria, tropical pulmonary eosinophilia, lymphoedema to full elephantiasis	Severe, infection can lead to blindness and skin conditions. nodules/granulomas containing adult worms, skin atrophy and hypo-hyperpigmentation, sowda (skin darkens and becomes covered scaly papules), punctuate keratitis, sclerosing keratitis which leads to permanent blindness.
Diagnosis	Geimsa stained blood samples, ELISA based assays for parasite antigens and PCR based assays.	Giemsa stained skin snips, ELISA based assays for parasite antigens and PCR based assays.
Prevention and	DEC (diethylcarbamazine) and	Vector control efforts and yearly

Control	Ivermectin (Mectizan®) can be given periodically (i.e. once per year) to reduce microfilarial burdens and block development of incoming infective L3. However, there are no widely available and safe macrofilaricides. Vector control in urban areas has also helped reduce infection rates	doses of Ivermectin (Mectizan®) to reduce microfilarial burdens and block development of incoming infective L3. However, there are no widely available and safe macrofilaricides.

Table 1.1.1 Summary of the characteristics of the two most prevalent filarial diseases. The, species, transmitting vector, mammalian hosts, estimated number of people infected worldwide, global distribution of the infections, pathologies, diagnosis, prevention and control measures being taken by wold health agencies are listed. All information listed in the tables is taken from Mason et. al 1978 unless otherwise noted (Manson-Bahr and Bell, 1987).

Several global agencies and trusts including the World Health Organization (WHO), World Bank and the Edna McConnell Clark Foundation have sponsored both parasite and vector control programs which have met with success in some regions (Molyneux, 1995; Ottesen, 2000; Ottesen and Ramachandran, 1995; Richards *et. al.*, 2001). However, these agencies recognize that there is a need for an effective prophylactic vaccine and broader spectrum filaricides.

After 30 years of active research a great deal of immunological and epidemiological data has been gathered about these parasites but little progress made towards the development of a vaccine or a drug that is effective against all filarial species or life stages. This is mainly due to the fact that none of the filarial nematodes mentioned above, with the exception of *Brugia malayi*, are amenable to experimentation because viable laboratory-based, non-human lifecycles cannot be maintained. This severely limits the amount of biological materials available to researchers and therefore makes the discovery of suitable drug or vaccine targets difficult.

1.2. The Biology of Filarial Nematodes

All tissue dwelling filarial nematodes spend the majority of their lifecycle in their vertebrate host. However, they must undergo some portion of their development outside of this host in a transmitting vector (an arthropod). Human filarial parasites are classified by the habitat in which the adults reside within the human host. The adults of lymphatic filaria like *B. malayi* and *B. timori* and *W. bancrofti* reside in the lymphatic vessels while the adults of cutaneous filaria (e.g. *O. volvulus*) reside in cutaneous tissues. All human filaria like other nematodes moult four times during their life cycle. The worms have distinctive morphologies between each of these moults and like other nematodes these life periods are designated L1- L5 (or adults).

1.2.1 Filarial Life Cycle:

All filarial nematodes share a common life cycle. Infection of the human host is initiated by the bite of an infected arthropod. Stage 3 (L3) infective larvae are deposited onto the skin while the insect is feeding. The larvae crawl into the bite wound and burrow through the cutaneous tissues. Lymphatic filaria migrate to the lymphatic vessels while *O. volvulus* continues to travel through the cutaneous tissues. During this migration the L3 moult to stage 4 (L4) larvae then to adults. Adult lymphatic filaria reside permanently in the lymphatic vessels while adult *O. volvulus* reside in nodules that form beneath the dermis. After a few months of maturation the adults mate and the females begin producing hundreds of thousands of stage 1(L1) larvae called microfilaria ovoviviparously. The microfilaria of lymphatic filaria circulate through the body in the blood waiting to infect another arthropod through the ingestion of a blood meal, while the microfilaria of *O. volvulus* migrate through the skin and are ingested by feeding blackfly. One major difference between the microfilaria of the lymphatic filaria and *O. volvulus* is that the lymphatic filaria are enclosed by an acellular sheath which is a remnant of the vitelline membrane formed during embryogenesis. After ingestion by an appropriate vector the microfilaria penetrate the stomach walls of the insect (lymphatic filaria shed their sheath at this time) and migrate to the thorax muscles where they moult to stage 2 larvae (L2). After about a week the L2 moult into L3 and migrate to the arthropod's mouthparts where they wait until it takes its next meal. Figure 1.2.1.1 summarizes the lifecycle of filarial nematodes.

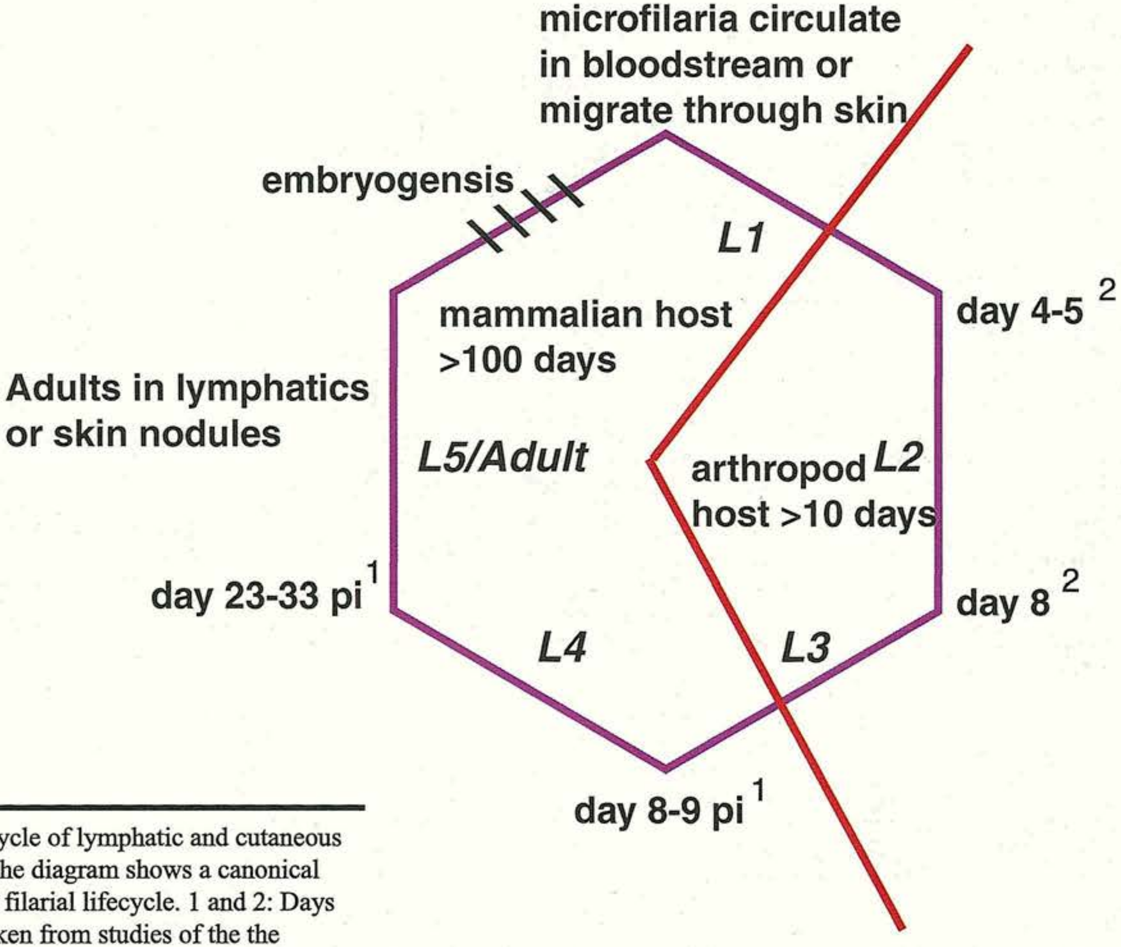


Figure 1.2.1.1 Lifecycle of lymphatic and cutaneous filarial nematodes. The diagram shows a canonical representation of the filarial lifecycle. 1 and 2: Days when molts occur taken from studies of the the lifecycle of *B. pahangi* (**1:** Schacher et. al. 1962a and **2:** Schacher et. al. 1962a).

1.2.2 The evolutionary relationship of filaria and other parasitic nematodes

Nematodes represent an extremely diverse group of animals with species found in almost all environments tested. These species have a variety of trophic habits and recent phylogenetic analysis of the nematode phylum indicates that parasitism has evolved independently in all of the major nematode groups (Blaxter *et. al.*, 1998). Figure 1.2.2.1 adapted from Blaxter *et. al.* 1998 shows the results of the phylogenetic analysis which places the filaria and other spirurids in a group composed exclusively of animal parasitic species (nematode clade III) including the ascarids (roundworms) and oxyurids (pinworms).

The filaria are distantly related to other important human parasitic nematodes such as *T. trichiura* or hookworm species. Filaria are also distantly related to the nematode species that have been used as model organisms such as *Caenorhabdits elegans*, *Caenorhabditis briggsae* and *Pristionchus pacificus.* These model nematodes have been proposed as surrogate test bed for parasitic nematodes. However, the large evolutionary distance between these nematodes and many of the parasitic species as well as the differences in their lifestyles means that some processes will not be conserved. Therefore they may not represent appropriate models for all species even when general biological processes not related to parasitic lifestyles are being examined.

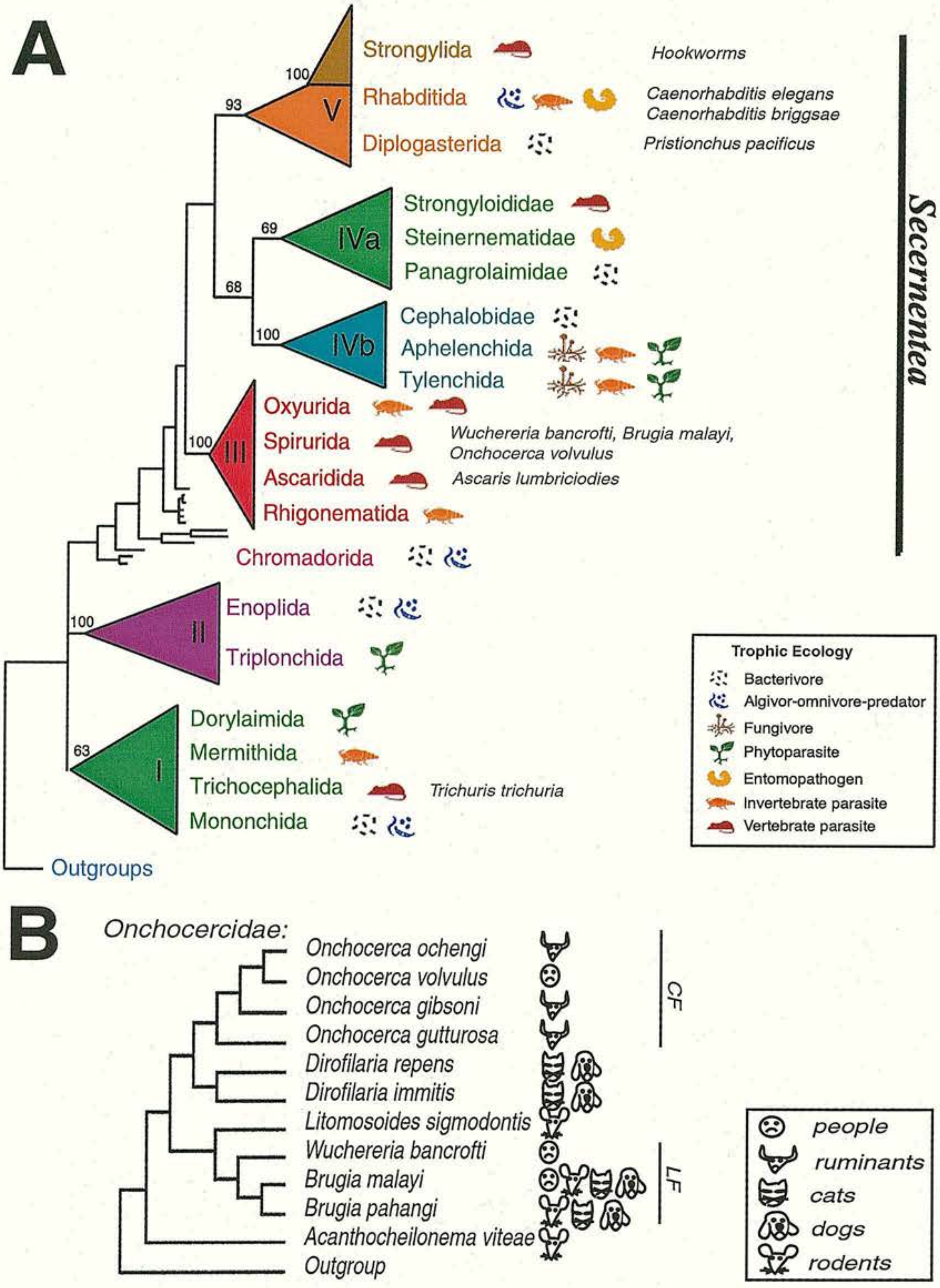


Figure1.2.2.1 The evolutionary relationship of the filaria and other parasitic nematodes **A:** Adapted from Blaxter et. al. 1998 the phylogenetic relationship of important parasitic nematode species based on analyses of the SSU rRNA sequences. The major nematode clades are shown along with the tropic habits of members of the major subgroups. **B:** The phylogenetic relationship of selected filaria belonging to the Onchocercidae based on analyses of the 5S rRNA, COXI, and mitochondrial SSU sequences (Xie et. al. 1994 and Casiraghi et. al. 2000). The mammals infected by the filaria is also shown. **CF:** cutaneous filaria, **LF:** lymphatic filaria. .

Within the filaria, morphological and phylogenetic studies indicate that the human lymphatic and cutaneous species form evolutionarily distinct groups (Anderson and Bain, 1976; Casiraghi *et. al.*, 2001; Xie *et. al.*, 1994). These studies have also confirmed that several species used as models for filarial infection are closely related to the human parasites (*B. pahangi* and *O. ochengi*). However, several of the species that infect rodents are not consistently placed either within the lymphatic/cutaneous groups or at the base of the phylogenetic trees. While the time of divergence between the two groups has not been estimated they are not believed to be closely related.

1.2.3 The genomes and genome projects of filarial nematodes

In December 1994 less than 60 genes had been cloned from filarial nematodes (Blaxter *et. al.*, 1996). Most of these sequences were genes identified as potential or diagnostic or vaccine targets by immuno-screening of filarial cDNA libraries. However, none of these proteins have proved to be effective vaccinogens. In addition all of the currently used anti-filarial compounds have been discovered through random screens. Some are highly toxic to humans and none are effective against all stages of the parasite's development. To identify new vaccine candidates and drug targets as well as stimulate research on basic filarial biology the WHO TDR gave funds to initiate the filarial genome project (FGP). The FGP was organized as a collaborative effort between group of seven laboratories coordinated by Professor Steve Williams (Smith College, USA). The main goal of the project was to utilize genomics techniques to identify new targets and provide resources and reagents to the research community.

Like other animals filarial nematodes have nuclear and mitochondrial genomes. In addition to these genomes most filarial nematodes harbor a third genome belonging to an endosymbiotic bacteria closely related to arthropod pathogens Wolbachia. The WHO, Edna Mc Connell Clarke Foundation, UK Medical Research Council (MRC) and New England Biolabs (NEB) have funded FGP laboratories to launch a filarial gene discovery program and begin the assembly of physical map of the nuclear and endosymbiont genome of *B. malayi*. The physical maps and genomic libraries would in turn be utilized by the research community to study genes of interest, population genetics and serve as a scaffold for genome assembly if the whole genome of *B. malayi* is sequenced.

The previously characterized features of the filarial genomes as well as the goals and methodology utilized by the filarial genome projects are summarized below.

1.2.3.1 Nuclear genome

Nematode nuclear genomes are relatively compact when compared to the genomes of vertebrates or other non-vertebrates. To date, the only the genome of the free-living rhabditid *Caenorhabditis elegans* has been fully sequenced and thus is the only available nematode comparator to the filarial genomes. The *C. elegans* genome is 100 MB in size (Sulston and Horvitz, 1977) and is split into five autosomes and one sex chromosome (consortium, 1998; Herman *et. al.*, 1976). Approximately 20,000 genes have been predicted from the genome sequence (consortium, 1998; Reboul *et. al.*, 2001). An interesting feature of the *C. elegans* genome is the organization of approximately 13% of its genes in transcriptional operons. Unlike bacterial operons, most genes in the nematode operons do not appear to be functionally linked so it is unclear how or why they have become linked. In filaria the calculated genome size appears to be relatively similar to *C. elegans* with estimates ranging from 100 –150 MB (Donelson *et. al.*, 1988; McReynolds *et. al.*, 1986). While there is currently no data on the number of genes contained within the filarial genomes, they are believed to be similar to the number found in *C. elegans*(Blaxter *et. al.*, 2002).

Karyotypes vary between filarial species with *B. malayi* having four and *O. volvulus* three autosomes (Hirai *et. al.*, 1985; Hirai *et. al.*, 1987; Post *et. al.*, 1989; Sakaguchi *et. al.*, 1983). Unlike *C. elegans*, both filaria appear to have dimorphic sex chromosomes with an XX/XY system of sex determination. The Y chromosome of *B. malayi* is currently under investigation (Underwood and Bianco, 1999).

The filarial genomes have a much higher AT content then *C. elegans* ~71% vs 64% (consortium, 1998; Fadiel *et. al.*, 2001; Rothstein *et. al.*, 1988). Unlike the genomes of *D. melanogaster* or *H. sapiens*, a relatively small portion of the *C. elegans* genome is composed of tandem or inverted repeat sequences (~ 17%). This is in contrast to the filaria which have species or genus specific tandem repeat families which make up large portions of their genomes. In *Brugia sp.* the monomorphic 322 bp Hha1 repeat makes up ~10% of the genome and is organized in at least ten tandem arrays (McReynolds *et. al.*, 1986). *In O. volvulus* the O-150 repeat 150 bp is predicted to make up ~ 1% of the genome and is also arranged in four to eight clusters (Meredith *et. al.*, 1991; Perler and Karam, 1986; Shah *et. al.*, 1987). There has been some speculation that these sequences may represent sub-telomeric repeats although this has yet to be demonstrated. While only a few genomic sequences of filarial genes are available they indicate that on average filarial genes are interrupted with introns more frequently than *C. elegans* genes. These introns on average also tend to be longer, averaging 100-300 bp in length (Unnasch and Williams, 2000; Zang *et. al.*, 1999).

1.2.3.2 Mitochondrial genome

The mitochondrial genomes of *O. volvulus* and *B. malayi* have been fully sequenced ((Keddie *et. al.*, 1998) and Daub *et. al.*, unpublished). These genomes share many common characteristics with other the fully sequenced nematode mitochondrial genomes from *Necator americanus*, *Ancylostoma duodenale*, *C. elegans*, *Ascaris suum* and *Trichinella spiralis* (Hu *et. al.*, 2002; Lavrov and Brown, 2001; Okimoto *et. al.*, 1992). When compared to other animal mitochondrial genomes they are relatively compact (~13-16 kb). Unlike the vertebrate mitochondrial genomes the order of the mitochondrial genes appears to be relatively malleable. Interestingly, phylogenies based on the order of mitochondrial genes conflict with the phylogenies based on other sequences such as the SSU rRNA with *A. suum* (nematode clade III) sharing almost exactly the same gene order as the distantly related nematodes *N. americanus*, *A. duodenale* and *C. elegans* (nematode clade V). With the exception of *T. spiralis,* none of the nematode mitochondrial genomes encode a putative ATPase subunit 8 gene. Like the nuclear genome, the filarial mitochondrial genomes show an extreme AT bias (73%) and this is reflected in the codon bias and amino acid composition of the mitochondrial proteins (Keddie *et. al.*, 1998). As additional nematode and animal mitochondrial genomes become available it will become clear whether the high rate of gene rearrangements is a feature unique to the nematodes.

1.2.3.3 Wolbachia endosymbiont genome

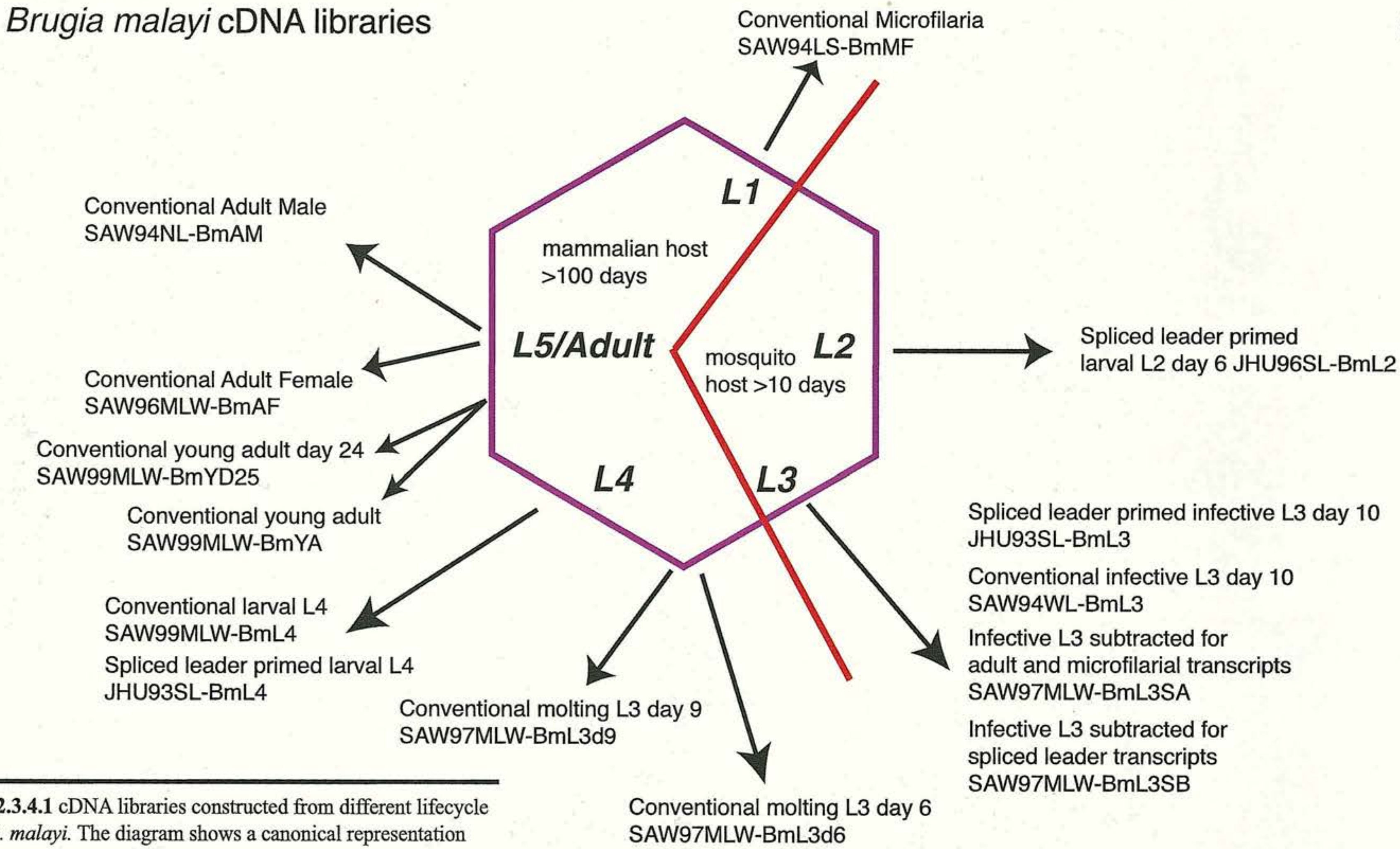
The bacterial endosymbionts of filarial nematodes were initially identified in the mid-1970s during ultra structural studies of various filarial nematodes (Bandi *et. al.*, 2001). These bacteria have been found in all filaria surveyed except *A. vitae* and *Onchocerca flexuosa* (Bandi *et. al.*, 2001). Several recent studies have implicated the release of bacterial toxins after nematode death as a major stimulus for the immune responses that leads to pathologies during filarial infections (Brattig *et. al.*, 2001; Taylor *et. al.*, 2000). Treatment of filaria with antibiotics that kill the bacteria induce sterility and emybrological defects (Bandi *et. al.*, 1999; Brouqui *et. al.*, 2001; Cross *et. al.*, 2001; Hermans *et. al.*, 2001; Hoerauf *et. al.*, 2000a; Hoerauf *et. al.*, 2000b; Langworthy *et. al.*, 2000; Smith and Rajan, 2000; Townson *et. al.*, 2000). However the mechanisms which mediate the effects of the antibiotics are still unknown.

Analysis of genes cloned from the bacteria has shown that they are related to a group of arthropod reproductive parasites called *Wolbachia*. Phylogenetic comparisons of genes cloned from the nematodes and *Wolbachia* show that the evolution of the bacteria mirrors the evolution of their hosts indicating a long-term association between the two organisms (Casiraghi *et. al.*, 2001). A physical map of the *B. malayi Wolbachia* genome has been assembled and is predicted to be between 1-1.2 MB in size (Sun *et. al.*, 2001). The genomes of the *B. malayi* and *O. volvulus* Wolbachia are currently being fully sequenced. Preliminary results indicate that their genes and genomes are very similar to those of other *Wolbachia* however there appears to be little conservation of gene order between the nematode and arthropod *Wolbachia* (Ware *et. al.*, 2002).

Because *Wolbachia* appear to be required for long term survival and reproduction of the nematode, as well as play a role in nematode induced immuno-pathology, they have become attractive drug targets. The *Wolbachia* genome sequences will provide an important resource for target discovery and dissecting the interactions the *Wolbachia* has with its nematode and vertebrate hosts.

1.2.3.4 Filarial gene discovery effort

To facilitate the discovery of new filarial genes cDNA libraries from several lifestages and species have been constructed in the FGP labs (Blaxter *et. al.*, 1996; Lizotte-Waniewski *et. al.*, 2000). Datasets of randomly picked clones (expressed sequence tags, ESTs) have been generated by the collaborating laboratories in the filarial genome network. These sequences have been deposited in GenBanks EST database dbEST and individual clones are available by request from the filarial genome resource centers (Blaxter laboratory Edinburgh, UK and Williams laboratory Northampton, USA). Figures 1.2.3.4.1, 1.2.3.4.2 and tables 1.2.3.4.3 and 1.2.4.1.4 summarize the characteristics of the various filarial genome project cDNA libraries constructed in *B. malayi* and *O. volvulus*.


Brugia malayi cDNA libraries
Conventional Microfilaria
SAW94LS-BmMF
L1
Conventional Adult Male
SAW94NL-BmAM
mammalian host
>100 days
L5/Adult
mosquito
host >10 days
L2
Spliced leader primed
larval L2 day 6 JHU96SL-BmL2
Conventional Adult Female
SAW96MLW-BmAF
Conventional young adult day 24
SAW99MLW-BmYD25
Conventional young adult
SAW99MLW-BmYA
L4
L3
Spliced leader primed infective L3 day 10
JHU93SL-BmL3
Conventional infective L3 day 10
SAW94WL-BmL3
Infective L3 subtracted for
adult and microfilarial transcripts
SAW97MLW-BmL3SA
Infective L3 subtracted for
spliced leader transcripts
SAW97MLW-BmL3SB
Conventional larval L4
SAW99MLW-BmL4
Spliced leader primed larval L4
JHU93SL-BmL4
Conventional molting L3 day 9
SAW97MLW-BmL3d9
Conventional molting L3 day 6
SAW97MLW-BmL3d6

Figures 1.2.3.4.1 cDNA libraries constructed from different lifecycle stages of *B. malayi.* The diagram shows a canonical representation of the *B. malayi* lifecycle and the time points from which cDNA libraries were constructed.

Onchocerca volvulus cDNA libraries

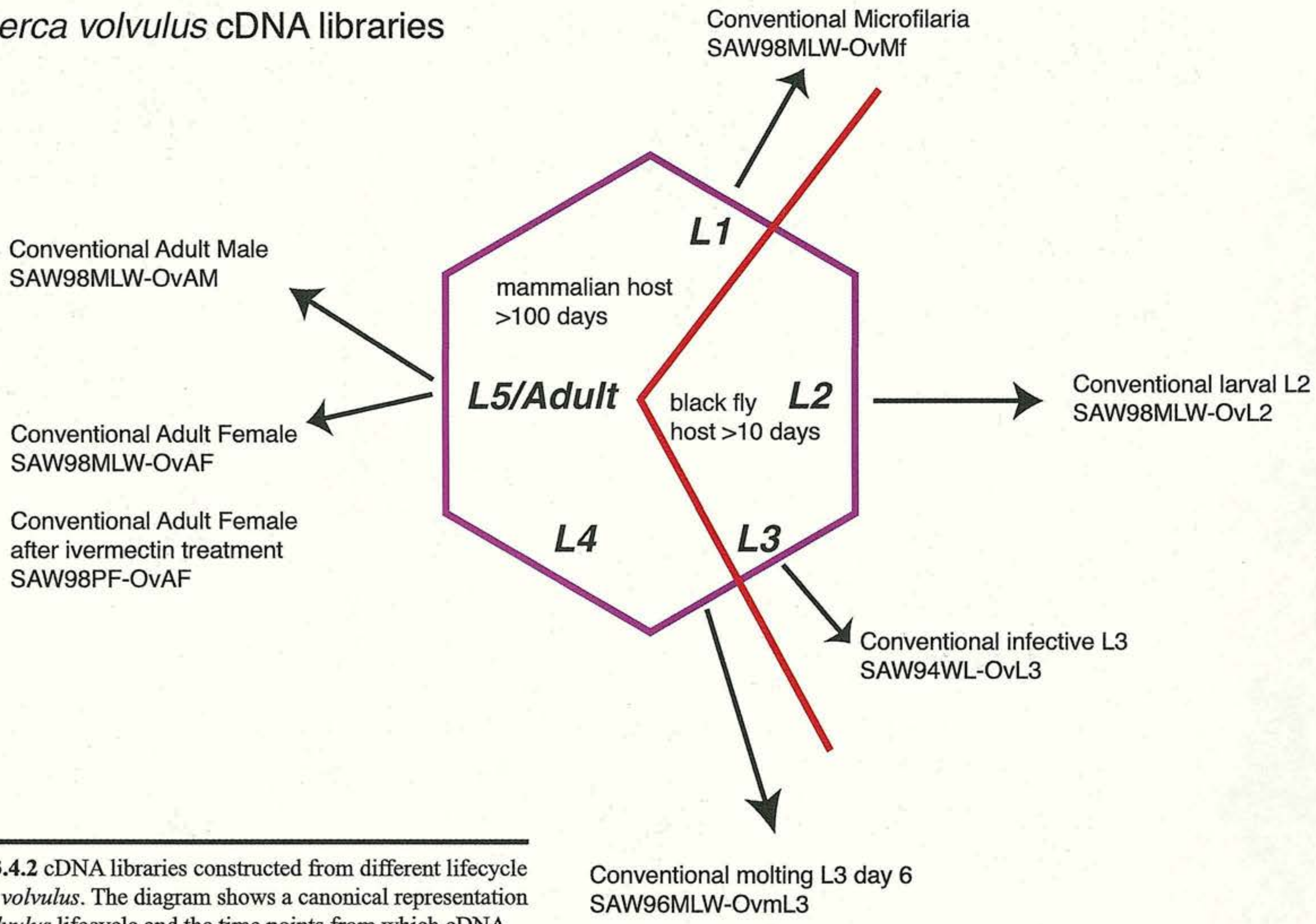


Figure 1.2.3.4.2 cDNA libraries constructed from different lifecycle stages of *O. volvulus*. The diagram shows a canonical representation of the *O. volvulus* lifecycle and the time points from which cDNA libraries were constructed.

Table 1.2.3.4.3 *B. malayi* cDNA libraries

Library	Library ID	
BmMf	SAW94LS-BmMf	cDNA was prepared from microfilariae of isolated from jirds
BmMfZ	SAW94LS-BmMf	18,000 clones isolated from excised SAW94LS-BmMf and gridded at high density. Hybridized with rDNA and other abundant transcripts. Non-hybridizing colonies selected for sequencing
BmL2SL	JHU96SL-BmL2	cDNA was prepared from mosquito derived, second stage larvae (L2) day 6 after infection using PCR with oligo dT and the nematode SL-1 sequence
BmL3	SAW94WL-BmL3	cDNA was prepared from third stage infective larvae isolated from mosquitoes 10 days after infection
BmL3SL	JHU93SL-BmL3	cDNA was prepared from third stage infective larvae isolated from mosquitoes 10 days after infection using PCR with oligo dT and the nematode SL-1 sequence
BmL3SA	SAW97LS-BmL3SA	Microfilaria, adult male and female cDNAs were subtracted from L3 cDNA using the PCR-select cDNA subtraction kit (Clontech)
BmL3SB	SAW97YG-BmL3SB	Nematode spliced-leader L3 cDNA was subtracted from L3 cDNA using the PCR-select cDNA subtraction kit (Clontech)
BmL3Z	SAW94WL-BmL3	18,000 clones isolated from excised SAW94LS-BmL3 and gridded at high density. Hybridized with rDNA and other abundant transcripts. Non-hybridizing colonies selected for sequencing
BmL3D6	SAW96MLW-BmL3d6	cDNA was prepared from third stage larvae isolated from the peritoneal cavity of jirds six days after infection.
BmL3D9	SAW97MLW-BmL3d9	cDNA was prepared from third stage larvae isolated from the peritoneal cavity of jirds nine days after infection.
BmL4	SAW99MLW-BmL4	cDNA was prepared from L4s isolated from the peritoneal cavity of jirds14 days after infection
BmL4SL	JHU93SL-BmL4	cDNA was prepared from fourth stage larvae (L4)14 days after infection using PCR with oligo dT and the nematode SL-1

		sequence.
BmYA	SAW99MLW-BmYA	cDNA was prepared from mixed sex young adult worms isolated from the peritoneal cavity of jirds.
BmYAD25	SAW99MLW-BmYD25	cDNA was prepared from mixed sex young adult worms isolated from the peritoneal cavity of jirds on day 25 after infection.
BmAM	SAW94NL-BmAM	cDNA was prepared from adult males of isolated from jirds
BmAF	SAW96MLW-BmAF	cDNA was prepared from adult females of isolated from jirds.

Table 1.2. 3.4.4 *O. volvulus* cDNA libraries

Library	Library ID	
OvMf	SAW98MLW-OvMf	cDNA was prepared from approximately 200,000 microfilariae isolated from the skin of infected individuals from Kumba, Cameroon
OvL2	SAW98MLW-OvL2	cDNA was prepared from approximately 9,000 L2s isolated from infected black flies from Kumba, Cameroon
OvL3	SAW94WL-OvL3	cDNA was prepared from third stage infectivelarvae of isolated from black flies10 days after infection
OvmL3	SL96MLW-OvmL3	cDNA was prepared from moulting third stage larvae collected after 1,2 and 3 days in culture
OvAM	SAW98MLW-OvAM	cDNA was prepared from adult males isolated from infected individuals from Kumba, Cameroon
OvAF	SAW98MLW-OvAF	cDNA was prepared from adult females isolated from infected individuals from Kumba, Cameroon
OvAFIV	SAW98PF-OvAF	cDNA was prepared from adult females isolated from infected individuals from Kumba, Cameroon after ivermectin treatment

Tables 1.2.3.4.3 and Table 1.2.3.4.4 Characteristics of the FGP *B. malayi* and *O. volvulus* cDNA libraries. The library name, library identifier and a brief description of the materials and methods used to construct the cDNA libraries are listed.

1.2.3.5 Genome mapping and sequencing effort

As well as the gene discovery effort as a prelude to whole genome sequencing, the FGP has generated a set of large insert genomic libraries hosted in BAC vectors (bacterial artificial chromosomes) (Foster *et. al.*, 2001; Guiliano *et. al.*, 1999). The inserts that have been cloned into the BAC vector have been generated by partially digesting genomic DNA with HindIII or SauIIIa and have an average size of 60-80 kb (J.Pope-Chapel, J. Foster, J. Daub and C. Whitton pers. com. 2002, (Foster *et. al.*, 2001; Guiliano *et. al.*, 1999)). Individual clones have been picked and gridded on high density filter arrays. These BACs and filters are being used to assemble a medium resolution physical map. This map is being constructed with three complementary sets of data; BAC end sequences (genome survey sequences (GSSs) generated by J. Daub and C. Whitton, Blaxter laboratory in collaboration with the Pathogen Sequencing Unit (PSU), Sanger Institute), hybridization of BAC end sequences probes to high density filter arrays (generated by J. Daub, Blaxter laboratory) and hybridization of *B. malayi* genes identified in the gene discovery effort (various FGP labs). All three datasets are currently being combined to assemble large clone contigs that will serve as a scaffold in the assembly of the whole genome sequence. Recently, funds have been awarded to The Institute for Genome Research (TIGR) to shotgun sequence the *B. malayi* genome to a 3.5 fold depth of coverage.

1.2.3.6 Other nematode gene discovery efforts

The success of the FGP in rapidly and cost effectively generating large sequence datasets and reagents for the research community has stimulated several additional parasitic nematode gene discovery efforts funded by the National Institutes of Health (NIH) and the Wellcome trust (McCarter *et. al.*, 2000; Parkinson *et. al.*, 2001). These projects are generating EST datasets from a variety of parasitic nematodes that are important in human or agricultural pathogens These include species that infect both plants and animals. Table 1.2.3.6.1 shows the number of nematode ESTs that have been deposited in GenBank.

Species	Common name	Nematode Clade	Number of ESTs
Caenorhabditis elegans	free living	V	191,268
Brugia malayi	human lymphatic filarial nematode	III	22,439
Onchocerca volvulus	human cutaneous filarial nematode	III	14,922
Strongyloides stercoralis	human gut parasite	IV	11,392
Ascaris suum	pig roundworm	III	14,380
Ancylostoma caninum	dog hookworm	V	7,328
Pristionchus pacificus	free living	V	6,932
Meloidogyne incognita	southern root-knot nematode	IV	6,767
Strongyloides ratti	rat gut parasite	IV	6,562
Globodera rostochiensis	cyst nematode	IV	5,934
Meloidogyne javanica	root-knot nematode	IV	5,600
Parastrongyloides trichosuri	possum gut parasite	IV	5,323
Haemonchus contortus	barber's pole worm of sheep	V	4,843
Heterodera glycines	soybean cyst nematode	IV	4,327
Trichinella spiralis	trichina muscle nematode	I	4,247
Meloidogyne arenaria	root-knot nematode	IV	3,334
Caenorhabditis briggsae	free living	V	2,424
Trichuris muris	threadworm of mice	I	2,125
Globodera pallida	Cyst nematode	IV	1,832
Necator americanus	human hookworm	V	961
Nippostrongylus brasiliensis	rat gut parasite	V	734
Toxocara canis	dog gut parasite	III	519
Zeldia punctata	free living	IV	391
Litomosoides sigmodontis	murine filarial nematode	III	198
Wuchereria bancrofti	human lymphatic filarial nematode	III	131
Onchocerca ochengi	bovine cutaneous filarial nematode	III	60

Table 1.2.3.6.1: Nematode ESTs deposited in GenBank. The species, common name, the nematode clade the species have been placed in and number of nematode ESTs deposited in GenBank as of 12/01/2002 are listed

1.2.5 The questions addressed in this thesis

To date the FGP has deposited over 30,000 ESTs and almost 17,000 GSS sequences into GenBank. This represents a tremendous resource to the filarial community. However, examining large number of sequences presents several logistical problems and thus a number of custom built informatics tools are required to properly analyze the dataset.

The work presented in this thesis builds upon the resources generated by the filarial genome project and focuses the analysis of the gene content of the filarial genomes, the dynamics of gene and genome evolution between nematodes and the conservation of linkage groups, gene order (synteny) and structures between distantly related nematodes.

These questions are addressed through the four bodies of work. First, custom informatics tools were designed to analyze the filarial EST datasets and the comparison of filarial genes to sequences in the public databases. An important nematode gene family, the macrophage migration inhibitor factors (*mif*), was selected and their evolution studied using molecular phylogenetic techniques. A large segment of *B. malayi* genomic DNA surrounding the *Bm-mif-1* locus was sequenced and compared to the genome of *C. elegans*. Finally the evolution of nematode operons and poly-cistron resolution was examined through the identification of operon sets which are conserved between *B. malayi* and other distantly related species.

Chapter 2

Materials and methods

2.0 Basic Media and Solution Recipes

For LB, SOC, 2x YT, 1x TE 5x TBE recipes see appendix I.

2.1 BAC Midi preparations

BAC DNA for end sequencing and restriction digests was prepared using the Qiagen Midi Prep Kit (Qiagen) with the following. A BAC colony was inoculated into 200 mL of LB or 2X YT with chloramphenicol (12.5 μg/mL) grown overnight with shaking at 37°C. The bacteria were spun down and the pellet resuspended in 4 mL of P1 buffer. The solution was allowed to sit for 5 min. 4 mL of P2 buffer was then added, the solution inverted 4-6 times and incubated at room temperature for 5 min. 4 mL of chilled P3 was then added, the solution invert 4-6 times and incubated on ice for 15 minutes. The solution was centrifuged in a sterile 50 mL plastic tubes at 13,000 rpm for 30 min. The Qiagen midi column was equilibrated and the supernatant allowed to flow through the column by gravity. The column was washed with 20 mL of QC buffer. The DNA was eluted into 30 mL corex tubes by adding 5 mL of QF buffer pre-warmed to 65° C. BAC DNA was precipitated by adding 3.5 mL of room temperature isopropanol, mixing and centrifuged at 13,000 rpm for 30 minutes at 4°C. The supernatant was carefully removed and the pellet cleaned with 70% ethanol and centrifuged at 13,000 rpm for 15 minutes at 4 °C. The pellet was resuspended in 200-300 μL of 0.1X TE.

2.2 Chemical transformation of *E. coli*

E. coli was made chemically competent using the standard protocol of Maniatis *et. al.* (Maniatis *et. al.*, 1982). An overnight culture of cells was diluted 1:1000 into 100 mL of LB and grown until it reached an OD_{600} of 0.7-0.9. The cells were spun down and resuspended in cold 100 mM $MgCl_2$ and shaken at 4°C for 1 hour. The cells were then pelleted at 4°C and resuspended in cold 50mM $MgCl_2$/ 50mM $CaCl_2$. The cells were then shaken at 4°C for at least 1 hour.

Up to 5 μL of ligation reaction or plasmid DNA was added to 100 μL of chemically competent cells. The cells were then incubated on ice for at least 10 minutes and then heat shocked at 42°C for 30 seconds. The shocked cells were allowed to recover on ice for two minutes and then SOC or LB recovery media was added and the cells incubated at 37°C for at least 30 minutes. The cells were then plated on LB agarose and incubated overnight at 37°C.

2.3 *C. elegans* and *P. pacificus* growth and laboratory maintenance

C. elegans and *P. pacificus* were maintained on MYOB agarose plates seeded with *E. coli* OP50. Plates were kept at 15°C and worms were passaged to fresh plates by picking individual worms to a fresh plate or by taking a piece of the agarose from the old plate and using it to seed a new plate. New plates were seeded every two weeks. *C. elegans* strains and *P. pacificus* were kept in different sealed containers to prevent any cross contamination between species.

MYOB agarose was prepared from a 5.9 g/L dry stock powder (55g Tris-HCl, 24 g Tris-OH, 310 g Bacto-Peptone, 800 mg cholesterol, 200 g NaCl) and 21g /L granulated agar.

E. coli OP50 used to seed the MYOB plates were grown in LB media, pelleted and resuspended in M9 media (3 g KH_2PO_4, 6 g Na_2HPO_4, 5 g NaCl, 1 mL 1M $MgSO_4$ per 1L M9 solution) to a 50X stock solution. ~250 μL of the stock was used to seed the MYOB plates which were then incubated at 37°C overnight to allow bacterial growth.

2.4 Colony boil preparation

Colony boils for PCR were prepared by scraping a colony into 20 μL of sterile ultra pure water and boiling at 95°C for 10 minutes. The boils were then cooled on ice and 1-2 μL of the boil added to a PCR reaction. The boils could be stored at –20°C indefinitely.

2.5 DEPC treatment

Ultra pure MilliQ water and other plasticware were treated with diethylpyrocarbonate (DEPC) to remove any RNAse activity according to Maniatis *et. al* (Maniatis *et. al.*, 1982). Briefly, plasticware was immersed in a 0.1% solution of DEPC and incubated overnight at room temperature. The DEPC solution was then removed and the plastics autoclaved and dried.

2.6 DNAse treatment of RNA

RNA was resuspended in DNAse buffer (50 mM Tris-HCl pH 7.5, 25 mM $MgCl_2$) with 40 U RNAse block (Stratagene), 50 ng BSA (DNAse free, Amersham) and 10U of DNAse (RNAse free, Amersham). The reaction was incubated at 37°C for 30 minutes and then the DNAse was heat killed at 65°C for 10 min.

2.7 DNA cleanup with Microcon-100

To remove primers and unincorporated dNTPs from PCR products the reaction was cleaned using Microcon-100 (Millipore) according to manufacturer's protocols. Briefly the PCR reaction was loaded into the column and brought up to 500 μL with ultra pure MilliQ water. The column was spun at 1,500 rpm for 12 minutes. The filtered liquid beneath the column was discarded and process repeated twice. After three centrifugations the column was reversed into a fresh collection tube and spun for 30 second at 1,500 rpm to retrieve the purified DNA.

2.8 DNA purification from isolated agarose gel fragments

The DNA was run on 1% agarose gel in the presence of ethidium bromide. The DNA was visualized with UV light and cut from the gel with a clean razor blade. The DNA was extracted from the gel fragment using a Ultra-DA extraction column (Millipore). The agarose was loaded into the column and centrifuged at 7,000 rpm for 10 minutes. The centrifugation was repeated until all the agarose was passed through the top of the column. The filtrate with the extracted DNA was then collected and stored at –20°C.

2.9 DNA sequencing

Standard automated cycle sequencing was performed using an ABI 373 or 377 automated sequencer (Applied Biosystems Inc.) using either the PRISM (Applied Biosystems) or Dyenamic ET (Amersham Pharmacia Biotech) cycle sequencing kits according to manufacturer's protocols. Briefly 4 μL of sequencing mix was added to 1 μL of sequencing primer (1.6 pM/μL) and 5 μL of template. The reaction was cycle sequenced in a PCR machine using the following conditions: 25 cycles (96°C for 30 sec, 50°C for 20sec, 60°C for 4 min). After the completion of the cycle sequencing the reaction was purified either by ethanol precipitation or column purification.

For BAC end sequencing the standard reaction conditions were modified according to the protocol published by the Sanger Institute (http://www.sanger.ac.uk/Teams/Team51/PACBACPrep.shtml). Briefly, 12 μL of sequencing mix was added to 1 μL of sequencing primer (30 pM/μL), 40 μL of template (1-2 μg of BAC DNA) and 5 μL of sterile ultra pure water. The reaction was cycle sequenced in a PCR machine using the following conditions: 40 cycles (96°C for 30 sec, 50°C for 20sec, 60°C for 3 min). After the completion of the cycle sequencing the reaction was purified either by ethanol precipitation or column purification.

For ethanol precipitation the standard sequencing reactions were mixed with 1 μL of 3M sodium acetate (pH 4.6) and 50μL of 100% cold ethanol (-20°C). The mixture was frozen at –80°C for at least 1 hour and then centrifuged at 13,000 rpm for 15 minutes. The supernatant was removed, the pellet washed with 500 μL of cold 70% ethanol and recentrifuged at 13,000 rpm for another 15 minutes. The supernatant was removed and the pellets air dried.

Column purifications were preformed with the Performa DTR Gel filtration cartidge (Edge Biosystems). Preprepared hydrated gels were centrifuged at 2,000 rpm for 2 minutes. The sequencing reactions were then loaded onto the column and the column placed in a fresh collection tube. The column was again centrifuged at 2,000 rpm for 2 minutes. The filtrates were air dried either on a heated block or in rotary evaporator.

Dried reaction pellets were submitted to the ICAPB core sequencing facility.

2.10 Hybridization and detection of biotin labeled probes to nylon filters

Each membrane was placed in a hybridization bag or a 22x22 polystyrene tray (Corning) and thoroughly wet with 6X SSC (20X SSC = 3 M NaCl, 0.3 M NaCitrate, pH 7.0), followed by prehybridization with 6X SSC, 5X Denhardt's reagent (50X =10 g ficoll-400, 50 g polyvinylpyrrolidone, 0.5g bovine serum albumin in 500 ml water), 0.5% SDS and 100 μg/ml denatured salmon sperm DNA (0.1 ml of solution per cm^2 membrane) for 1 hour at 55°C. 500 ng – 1 μg of the biotinylated probe was denatured in boiling water for 5 minutes, chilled on ice for 5 minutes, centrifuged briefly and then added to the prehybridization solution.

Hybridization of the probe was allowed to proceed overnight at 55°C, with gentle rocking (Maxi14, Hybaid).

After hybridization, the membrane was removed from the bag or tray, washed twice in 2X SSC, 0.1% SDS at room temperature for 5 minutes each and then washed twice in 0.1X SSC, 0.1% SDS at 60°C for 15 minutes each. The washed membrane was placed in a new hybridization bag or tray for subsequent chemiluminescent detection. The bags were sealed on three sides and on the fourth side a small spout was made to add and remove the detection reagents.

Detection was carried out as described in the Phototope-Star™ Detection manual (NEB Inc). The spouts of the bags were sealed using a dialysis clip between solution removals and additions. Streptavidin, biotinylated alkaline phosphatase and CDP* reagents were sequentially added and removed from the bags or trays, with wash steps in between each addition to remove excess reagent. At each step, the bag or detection tray was rocked for 5 minutes at room temperature with moderate agitation on a shaking rocker. After draining the final detection reagent, the membrane was sealed in the bag and exposed to Hyperfilm MP (Amersham Pharmacia Biotech) X-ray film for 1-2 minutes, before the film was developed in an automated developing processor (CompactX2, X-Ograph).

Following detection with one probe or probe set, the membranes were stripped and washed to remove the probe. The membranes were rinsed in Milli-Q water, incubated in 0.4 N NaOH, 0.1% SDS at 70°C for 30 minutes and then rinsed in 0.2 M Tris-HCI, 0.1X SSC for 30 minutes at 25°C. Membranes were then stored in sealed hybridization bags at -20°C. As many as ten stripping and rehybridizations have been performed without loss of hybridization specificity or efficacy.

2.11 *In vivo* Excision of Phage Clones

EST clones obtained as lambda phage (lambda Uni-Zap, Stratagene) were excised according to manufacturer's protocols summarized below. XL-1 Blue MRF' cells were grown overnight in NYZDT media, spun down (3,000 rpm 5 min) and resuspended in 10 mM $MgSO_4$ to an OD $_{600}$ of 1.0. 200 μL of cells were added to 250 μL of phage stock and 1 μL of ExAssist helper phage (> 1x 10^6 pfu /mL, Stratagene) and incubated at 37°C for 15 minutes. 2 mL of LB was then added and the mixture incubated overnight at 37°C in a shaking incubator. pBluescript phagemid were isolated by heat inactivating the overnight culture at 70°C for 15 min and centrifuging 4000g for 15 min. 100 μL of supernatant was then added to 200 μL of SOLR cells (Stratagene, grown overnight and diluted in 10 mM $MgSO_4$ to an OD_{600} of 1.0). After incubating the mixture for 15 minutes at 37°C the cells were diluted 1:100 and 1:1000 and 10 μL plated on LB-ampicillin. Single colonies were picked and insert size of the plasmid tested by PCR. Plasmid DNA was prepared from positive clones with the QIAprep Spin Miniprep Kit (Qiagen).

2.12 Isolation of genomic DNA

Nematodes were percussively disrupted or homogenized with micro-homogenizers (Biomedix) in worm lysis buffer 2 (110 mM NaCl, 110 mM TrisCl ph 8.5, 55 mM EDTA, 1.1%SDS, 1.1% 2ME, 100 μg /ml proteinase K, 100 μg /mL RNAse). The resulting slurry was incubated at 65°C until the nematode fragments had been completely digested. The genomic DNA was then extracted twice with phenol, once with phenol:choloform and once with chloroform. The aqueous phase was then precipitated with 100% isopropanol (Sigma), pelleted and washed once with 70% ethanol (Sigma). The gDNA pellet was resuspended in water and stored at -20°C until use.

2.13 Messenger RNA Isolation

mRNA was purified from isolated total RNA using the Microfastrack mRNA isolation kit 2.0 (Invitrogen) according to manufacturer's protocols. Briefly total RNA pellets were resuspended in 10 μL elution buffer, 1 mL of lysis buffer and heated to 65°C for 5 min. 63 μL of 5 M NaCl was added, the solution mixed and added to prepared oligo dT cellulose resin. The resin was washed several times with binding buffer and low salt buffer to remove unbound material and non-polyadenylated RNAs. The bound mRNA was eluted with 200 μL elution buffer and the mRNA precipitated 10 μL (2 mg/mL glycogen carrier), 30 μL of 2 M sodium acetate and 600 μL of cold ethanol. The solution was frozen and then centrifuged 13,000 rpm for 15 min and the pellet air dried.

2.14 PCR Protocols (Standard and Long Range)

PCR was performed with either AGS-Gold Taq (Hybaid) or Qiagen Taq (Qiagen,Inc) using the standard PCR reaction conditions (0.2 mM dNTPs, 1.5 mM MgCl, 0.5 pM primer) and cycling conditions (94°C 3 min 1 cycle, 94°C 15 sec. 55°C 20 sec 72°C 3 min 35 cycles, 72°C 10 min 1 cycle).

To amplify products of > 4 kb from complex templates long range PCR was preformed with the Long-Range PCR Kit (Stratagene) using the following conditions suggested by the manufacture using the provided 10X high salt buffer (0.2 mM dNTPs, 600 mM KCl, 2 mM $MgCl_2$, 0.5 pM primer) and cycling conditions (94°C 3 min 1 cycle, 94°C 15 sec. 55°C 20 sec 72°C 5 min 35 cycles, 72°C 10 min 1 cycle).

The PCR reaction was then electrophoresed on a 0.7-1.5% agarose gel in 0.5X TBE with ethidium bromide and the DNA visualized with a UV transilluminator.

2.15 Plasmid Mini Preps

Plasmid DNA was prepared with the QIAprep Spin Miniprep Kit (Qiagen, Inc) according to manufacturer's instructions. A single colony or a frozen bacterial culture were inoculated into 10 mL of LB broth containing the appropriate antibiotic and grown overnight with shaking at 37°C. The cells were pelleted by centrifugation 3,000 rpm for 10 min. After preparation plasmid DNA was stored at –20°C.

2.16 Primer Extension

The primer extension was performed according the protocol described by Leonard and Patient (Leonard and Patient, 1996). Briefly the primer was kinased with ^{32}P-γATP and T4 polynucleotide kinase for 30 min at 37°C. The labeled primer was purified with a TE Midi SELECT-D G-25 column (Eppendorf-5'Prime, Inc), phenol:choroform extracted and ethanol precipitated . The probe was resuspended in DEPC treated water and 5 ~ fM was added to ~10 μg of total RNA in hybridization buffer (400 mM NaCl, 10 mM PIPES pH6.4), denatured at 70°C and then incubated for 3 hours at 55°C. The template and primer were added to the extension buffer (50mM Tris-HCl, pH 8.3, 40 mM DTT, 6 mM $MgCl_2$, 25 μg/mL actinomycin D (Sigma), 0.5 mM dNTPs, 40 U RNAse block (Stratagene)) along with 20 U of AMV-RT (Sigma) and incubated at 42°C for 1 hour. The extension reaction was electrophoresed on a polyacrylamide gel with an M13 sequencing ladder. The gel was dried and visualized with X-ray film (Hyperfilm MP, Amersham Pharmacia Biotech).

2.17 RNA preparation

Total RNA was prepared from all nematodes using RNAstat-60/RNAzol/ TRISOLV (Biogenesis or GibcoBRL) according to manufacturer's protocols. Nematodes were frozen at –80°C in a minimum volume of PBS or DEPC treated water. Four volumes of RNAzol/TRISOLV was added for every volume of tissue and the sample disrupted with either a eppendorf homogeniser (Biomedix) or using a custom built metal percussive disrupter cooled to –80°C. The RNA was extracted from the disrupted samples by adding 100 μL of chloroform (HPLC grade, Sigma) to every 0.5 mL of homogenate. The mixture was shaken vigorously, allowed to sit at room temperature for 3 minutes and then centrifuged at 13,000 rpm (at 4°C) for 10 minutes. The aqueous layer was transferred to a fresh eppendorf and the RNA precipitated with 250 μL of isopropanol (MB grade, Sigma). The RNA was pelleted by centrifuging at 13,000 rpm (4°C) for 10 minutes. The pellet was washed with 1.5 mL 75% ethanol and centrifuged for 5 min at 13,000 rpm (4°C). The ethanol was carefully removed and the pellet air dried. The pellet was then stored at -80°C or resuspended in DEPC treated ultra pure MilliQ water and immediately used.

2.18 Reverse Transcriptase PCR

First strand cDNA was synthesized with the Ultra HF RT PCR Kit (Stratagene) according to manufacturer's protocols. Briefly, 1-6.4 μL of RNA template was added to 1 μL of 10X stratascript RT buffer, 0.6 μL 100 ng/mL of reverse oligo primer (either a specific primer or oligo dT) and 1 μL of 40 mM dNTP mix. The reaction was brought up to a final volume of 9 μL with RNAse free sterile ultrapure water. The reaction was heated to 65°C for 5 min and then cooled to room temperature. 0.5 μL of Stratascript RT (10U) and 0.5 μL (10U) of RNAse block (Stratagene) were then added and the reaction incubated at 42°C for 30 min.

PCR was then performed using 1-2 μL of the first strand cDNA and two specific primers using the conditions described above.

2.19 RACE cDNA synthesis and amplification of RACE fragments

5'-RACE cDNA was synthesized using the GeneRacer Kit(Invitrogen) according to manufacturer's protocols . Briefly 200 ng of nematode polyA+ mRNA was treated with 10U of calf intestinal phosphatase for 1 hr at 50°C. The reaction was extracted with phenol : chloroform and the dephosphorylated RNA precipitated. The mRNA was then decapped with tobacco acid pyrophosphatase for 30 min at 37°C. The reaction was extracted with phenol : chloroform and the decapped mRNA precipitated. The GeneRacer RNA oligo was then ligated to the 5' end of the decapped mRNA using T4 RNA ligase. The reaction was incubated at 37°C for 1 hour. The ligation reaction was extracted with phenol : chloroform and the mRNA precipitated.

First strand cDNA was synthesized with SuperScript II RT (Stratagene) and oligo dT primer or a tagged oligo dT primer. PCR was then performed with specific reverse primers GeneRacer 5' (5'- CGACTGGAGCACGAGGACACTGA -3') primer using 5 μL of RACE cDNA as template. PCR was performed with AGS-Gold Taq (Hybaid) using the standard PCR conditions described above.

2.20 Random Prime Labeling of DNA with Biotin

Purified PCR product was randomly labeled with the NEBlot Phototope Kit (NEB Inc) according to manufacturer's protocols. 500 ng – 1 μg of DNA in 34 μL of water was denatured for 5 min at 95°C. 10 μL of 5X labeling mix (containing biotinylated random octamers), 5 μL of dNTP mix (containing biotin-dATP) and 1 μL of Klenow fragment (3'->5' exo$^-$) were then added to the DNA. The reaction was incubated at 37°C overnight. The reaction was stopped by adding 5 μL of 0.2 M EDTA pH8.0 and precipitated with 5 μL of 4 M LiCl and 150 μL cold ethanol. The labeled DNA was pelleted at 13,000 xg for 30 min and then washed with 70% ethanol. The pellet was resuspended in 20 μL of 1X TE and stored at –20°C until use.

2.21 Restriction Digests

300 ng – 1 μg of DNA was digested with the restriction enzyme according to manufacturer's protocols. 8 μL DNA was added to 1μL of 10X reaction buffer (and 0.5 μL BSA if needed). 1 μL of enzyme was then added and the reaction incubated at

37°C for 3 to 10 hrs until the DNA is fully digested. The reaction was then electrophoresed on a 0.7-1% agarose gel with ethidium bromide and the DNA visualized with a UV transilluminator.

2.22 SAP/EXO Treatment of PCR products

To remove excess primers and dNTPs from PCR reactions before sequencing the products were treated with shrimp alkaline phophatase 1(SAP) and exonuclease I (EXO). To 15μL of PCR product 1μL SAP (1U/ml, Amersham) and 1.5 μL of diluted EXO (0.1 U /μL) was added. The EXO was diluted in dilution buffer provided with the SAP before use. The reaction was incubated on a PCR block using the following conditions 37°C for 30 min and then 80°C for 15 min. The 'cleaned' PCR product was then added directly to the sequencing reaction.

2.23 Southern Blot

Southern Blotting was performed according to Maniatis *et. al.* (Maniatis *et. al.*, 1982). Briefly the digested DNA was electrophoresed on 0.7-1% agarose gel. After electrophoresis the gel was incubated in depurinating solution (0.25 M HCl) for 15 min , washed with water, incubated with denaturing solution (0.5 M NaOH, 1.5 M NaCl) for 20 min, washed with water, washed in neutralizing solution (1 M Tris HCL pH7.5, 1.5 M NaCl) for 20 min and blotted overnight onto a nylon membrane using capillary action in the presence of 10X SSC overnight. The blotted DNA was crosslinked to the membrane using a UV transilluminator. Hybridization and detection with biotin labeled probes was carried out as described above.

2.24 Subcloning PCR products (TOPO-T or pGEM-T)

PCR products amplified with Taq or other polymerases that tailed the DNA with A overhangs were subcloned using the TOPO-T (Invitrogen) or pGEM-T (Promega) cloning kits according to manufacturer's instructions.

1 μL TOPO-T vector was mixed with 1 μL cloning buffer and 4 μL of PCR product (either purified or straight from the PCR reaction). The mixture was incubated at room temperature for 15 minutes. 3 μL of the reaction was then transformed into chemically competent TOP10 cells. The transformed cells were plated on LB-ampicillin-kanamycin and single colonies picked and tested for inserts

by PCR. Plasmid DNA was prepared from positive clones using the QIAprep Spin Miniprep Kit (Qiagen).

1 μL pGEM-T vector was mixed with 1 μL 10X ligase buffer and 8 μL of PCR product (always purified). The mixture was incubated at 4°C overnight. 5 μL of the reaction was then transformed into chemically competent XL-1 blue cells. The transformed cells were plated on LB-ampicillin and single colonies picked and tested for inserts by PCR. Plasmid DNA was prepared from positive clones using the QIAprep Spin Miniprep Kit (Qiagen).

Chapter 3

Design and implementation of an EST clustering algorithm and analysis of the *B. malayi* EST dataset

3.0 Introduction

One of the major goals of the filarial genome project was the development of bioinformatics tools for the analysis and presentation of the filarial genome project EST datasets. These tools include a process for 'clustering' the ESTs into a non-redundant set of gene fragments. Not only would the clustering of the filarial ESTs allow the assessment of the productivity of the gene identification effort but they would provide a smaller more comprehensive dataset of gene fragments that could be used in subsequent analyses or resource development. Initially the publicly available EST clustering algorithms were examined and their suitability for the project assessed. The currently available EST clustering algorithms are described below. Unfortunately the algorithms available at the time did not meet the specifications outlined by the genome project community so the development of novel algorithm was under taken. The development, subsequent refinement of this algorithm and an analysis of the *B. malayi* and *O. volvulus* EST datasets is described.

3.1.0 Publicly available EST clustering methods

Over the past five years various EST clustering algorithms or processes have been described. None of these algorithms present perfect solutions for all EST clustering projects and features such as portability, computational requirements and scalability must be considered when deciding if an algorithm will provide an appropriate solution. The features of several publicly available or published clustering algorithms are outlined below. Table 3.1.0.1 summarizes the relevant features of the described algorithms.

Name of Tool	Description of Algorithm	Are previous cluster numbers retained between builds	Public Availability	Portability	Reference
THC_BUILD	TIGR EST clustering algorithm uses MEGABLAST to construct index tables of sequence overlap. These are used to construct an initial set clusters. CAP3 is then used to refine the initial clusters. EST contigs and consensus sequences generated with CAP3	Yes (a post process event)	No	No	(Liang *et. al.*, 2000; Quackenbush *et. al.*, 2000)
UNIGene	Uses MEGABLAST to determine initial relationship between ESTs.	Yes/No	No	No	(Boguski and Schuler, 1995)
INCA	Uses BLAST searches to build database of overlap relationships with other sequences. Not designed for EST clustering	No	Yes	Yes	(Graul and Sadee, 1997)
SEAL	Set of tools which can be used to construct large sets of automated BLASTs and parsing of results. Not designed for EST clustering	No	Yes	Yes	(Walker and Koonin, 1997)
STACK_PACK	Uses d2_cluster a word based clustering algorithm which uses 6mer overlaps used to construct super clusters. Phrap used to build sequence assemblies. CRAW and CONTIG_PROC used to derive and assess consensus sequence	Yes/No	Yes	Yes	(Miller *et. al.*, 1999)
JESAM	Uses word based clustering algorithm	No	Yes	No	(Parsons and

	with 12 mer overlaps used to build database of super clusters. Smith-Waterman style sequence comparison of super cluster members to verify sequence overlap.				Rodriguez-Tome, 2000)
ICAtools	Fasta like search to measure sequence redundancy in ESTs can extract set of non-redundant sequences from dataset	No	Yes	Yes	(Parsons, 1995)

Table 3.1.0.1. Summary of the relevant features of the seven described EST clustering algorithms. A short description of the process, if cluster numbers are retain between database rebuilds, public availability, portability between systems and references are listed.

3.1.1 BLAST and FASTA based algorithms

THC_BUILD, the TIGR EST clustering process that is used to build the TIGR gene indicies (GIs) uses MEGABLAST (Zhang *et. al.*, 2000) to identify overlapping sequences. These overlaps are stored in a local relational database and those sequences which have >95% identity over 40 nts and have <20 nts non-overlapping sequence are grouped in a cluster. CAP3 (Huang and Madan, 1999) is used to refine the clusters and each assembled contig is assigned its own GI number. These assemblies are then used to generate a consensus sequences. The assigned GI numbers are kept between rebuilds of the database and merger or split events of clusters are logged within the GI database. THC_BUILD is not publicly available and because its functions are intrinsically linked to TIGR's local database it is not portable to other systems.

The NCBIs UNIGene (Boguski and Schuler, 1995) algorithm uses MEGABLAST comparisons of the EST sequences to find sequences that overlap above a preset threshold. Full length cDNAs and 3' EST sequences are used as anchor sequences for the assembly of 5' EST sequences which are associated with the 3' anchor sequences by their clone identities. 5' EST sequences that do not overlap in sequence with the 3' sequences are still placed in the same assembly group. UNIGene cluster numbers are kept from previous database builds. However, events that merge clusters are not tracked, so it is not possible find newly formed clusters using the old UNIGene numbers. The UNIGene process is not publicly available.

Iterative neighborhood cluster analysis (INCA, (Graul and Sadee, 1997)) is a Java based program which performs iterative blast searches beginning with a given starting sequence and proceeding through any other sequence achieving a predefined minimum alignment score. The results of the searches are compiled and a cluster of sequences is defined in which all the constituents are related to at least one other cluster member (within the defined cutoff score). INCA is freely available and easily portable to any system that can run Java and local of versions of BLAST. The length of time taken to assemble the clusters depends on the size of the database being searched by BLAST. While INCA can be used to derive EST clusters it was not initially designed to perform this task and does not have any functions that deal with problems specific to ESTs such as poor sequence data and chimeric clones.

SEALS (System for Easy Analysis of Lots Sequences, (Walker and Koonin, 1997)) is a set of perl based scripts which allow the easy manipulation and analysis of large numbers of sequences. The SEAL component SPLAT is a tool for constructing iterative searches with BLAST and can be use to build clusters of related sequences. However, like INCA it was not initially designed to perform this task and does not have any functions that deal with problems specific to ESTs.

3.1.2 Word based algorithms

STACK_PACK (Miller *et. al.*, 1999) is a suite of tools that will take a set of given sequences and derive a set of super clusters using the d2_cluster algorithm (Burke *et. al.*, 1999). d2_cluster is a word based 'greedy' clustering algorithm which counts word matches (usually 6 mers). Sequences in super clusters are then assembled using Phrap (Phil Green *et. al.* unpublished). Consensus sequences are generated with the tool CONTIGPROC. The Phrap assembles are examined with CRAW which sorts related sequences within the multiple sequence alignment and will generate subclusters if there is heterogeny within the clustered sequences. In the most recent release of STACK_PACK (June 2001) supercluster IDs are kept between rebuilds of the database. Word based algorithms are inherently faster than alignment based algorithms and the entirety of the 3 million human ESTs in dbEST were clustered in 36 hrs on a 126 CPU SGI Origin 2000 (Miller *et. al.*, 1999). STACK_PACK is not available on all Unix platforms which presents portability problems to some systems.

Like STACK_PACK, JESAM is a word based clustering algorithm. JESAM uses a BLAST style dynamic programming algorithm and gap penalty scheme which identifies and stores 12mer alignments (Parsons and Rodriguez-Tome, 2000). The algorithm then uses a database of these 12 mer alignments to construct superclusters. Once the superclusters are defined a secondary assembly program like CONTIGPROC and CRAW, CAP3, or Phrap must be used to derive multiple sequence alignments and consensus sequences. JESAM is freely available and when used in a parallel processing environment can quickly derive superclusters for large numbers of sequences (>120,000 ESTs clustered with 11 CPUs in 5 hrs (Parsons and Rodriguez-Tome, 2000)). However, to run JESAM requires the installation of Java, CORBA, C++ and IDL software and its performance outside of a parallel processing environment has not been published.

3.1.3 Other clustering algorithms

ICAass (Parsons, 1995) a component of the ICAtools package, uses a FASTA like search with an asymmetric scoring scheme that measures redundancy within sequence datasets. Several specialized cluster browsing tools also included with

ICAtools allow the searching or extraction of a non-redundant set of sequences from a starting database. ICAtools is freely available for several Unix platforms and has been ported to MacOS. Unfortunately, ICAtools memory usage scales linearly while the computational time scales quadratically with the database size. Therefore, large datasets (>100,000 ESTs) can take almost a week to analyze on a single machine. The ICAtools package has been superceded by JESAM.

3.2.0 Development of CLOBB

When the filarial genome project began the sequencing and analysis of the *B. malayi* and *O. volvulus* ESTs none of the publicly available clustering algorithms had all the desired specifications needed to produce a robust and updateable clustered dataset. Some like THC_BUILD required computational resources that were unavailable to the genome project. Others like ICAtools did not have the capacity to keep previously assigned cluster numbers consistent between database rebuilds. The following specifications were outlined as being required for the clustering algorithm used to analyze the filarial EST datasets.

- The process must keep previously assigned cluster numbers between database rebuilds.
- The process must have the capacity to deal with library based artifacts particularly chimeric ESTs which are abundant in cDNA libraries constructed with PCR based methodologies.
- The process and its output must be portable to a variety of systems to be fully utilized by the filarial genome community.

To achieve this goal a new EST clustering schema was designed called CLOBB (CLustering On the Basis of Blast Similarity). The BLASTN algorithm was chosen as the tool for generating alignments of EST sequences because of its public availability and its readily parsable output. CLOBBv0.1 was written by D. B. Guiliano (Blaxter Lab, ICAPB). The CLOBBv0.1 code was subsequently revised by J. Parkinson (Blaxter Lab, ICAPB) to the presently used version CLOBBv1.0. Two versions of CLOBB were written in perl and both are described below. All of the analyses with both CLOBBv0.1 and CLOBBv1.0 were performed on either a SGI O_2 workstation running IRIX 6.2 or Pentium 800 workstation running Redhat Linux 7.1.

3.2.1 CLOBBv0.1

Before sequences were clustered with CLOBBv0.1, fragments of ribosomal SSU and LSU RNA and *E. coli* sequences were filtered from the dataset. All ESTs were compared to known SSU and LSU rDNA sequences from *B. malayi* and the whole *E. coli* genome sequence using the BLASTN algorithm (Altschul *et. al.*, 1997). Any sequences showing matches with p-values of ≤ e^{-50} were separated, considered contaminants and removed from further analysis.

CLOBBv0.1 is an iterative process. A query EST is compared to a database of previously clustered sequences with the BLASTN algorithm. If the query sequence is the first sequence to be clustered it is compared to an empty database. The results of the BLASTN search is parsed and any matches with p-values of ≤ e^{-50} are logged. If the sequence does not have any logged matches to sequences to the database the query sequence is given a new cluster number (BMC##### for *B. malayi* sequences or OVC##### for *O. volvulus* sequences). If there are logged matches to sequences in the database they are analyzed. If all matches are from the same previously designated cluster the sequence is assigned to that cluster. If the sequence matches two or more previously designated clusters then the sequence is removed from the analysis and an error logged to allow manual curation of the sequence. After the sequence is assigned a cluster number it is then added to the cluster database. The database is reformatted and next sequence searched against the database. Figure 2.1.1 shows a summary of the CLOBBv0.1 algorithm schema.

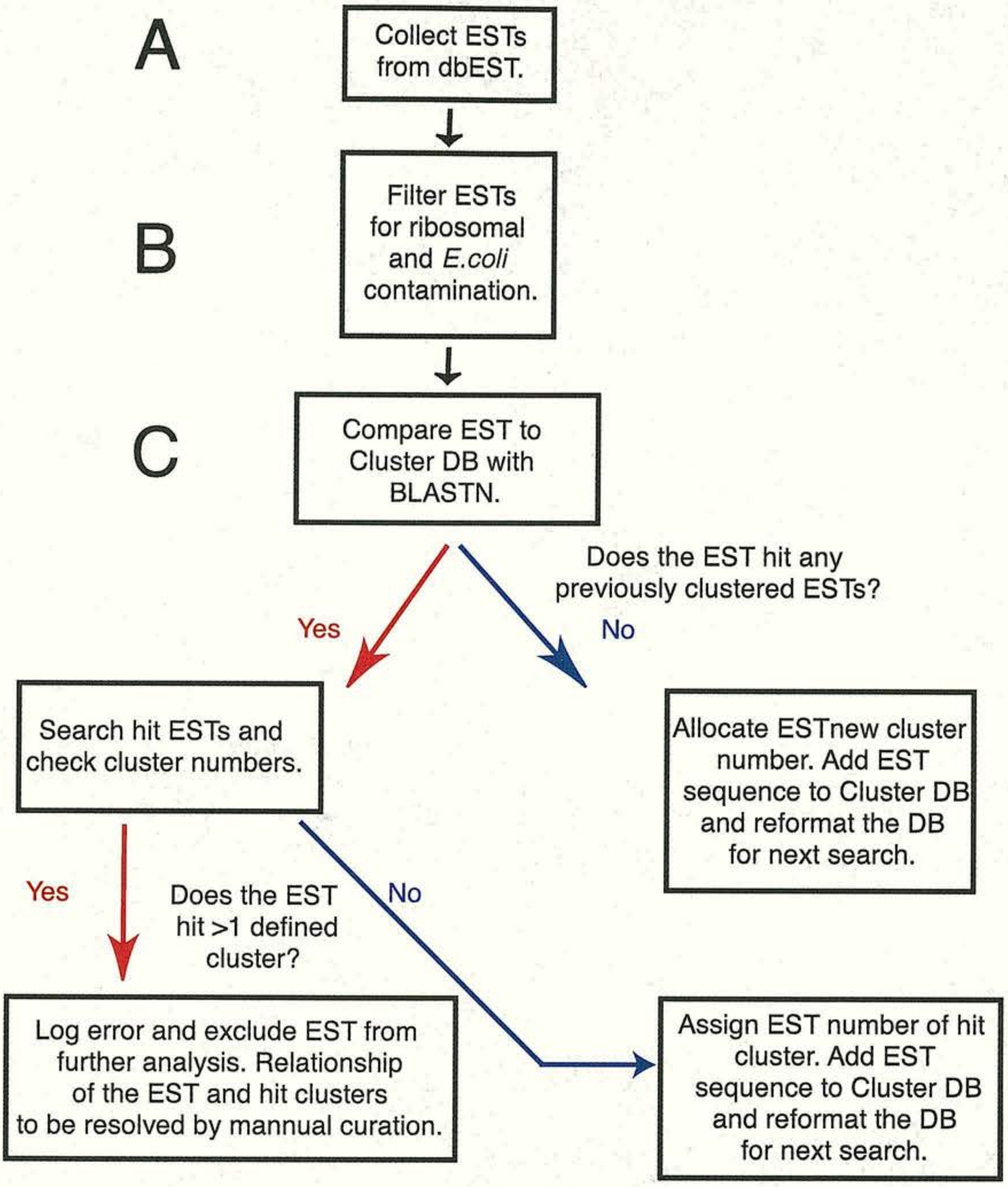

A
Collect ESTs from dbEST.
B
Filter ESTs for ribosomal and *E.coli* contamination.
C
Compare EST to Cluster DB with BLASTN.
Does the EST hit any previously clustered ESTs?
Yes
No
Search hit ESTs and check cluster numbers.
Allocate ESTnew cluster number. Add EST sequence to Cluster DB and reformat the DB for next search.
Yes
Does the EST hit >1 defined cluster?
No
Log error and exclude EST from further analysis. Relationship of the EST and hit clusters to be resolved by mannual curation.
Assign EST number of hit cluster. Add EST sequence to Cluster DB and reformat the DB for next search.

Figure 3.1.1 The CLOBBv0.1 clustering schema. The pre-clustering processes of collecting the ESTs from dbEST **(A)** and filtering the ESTs for rDNA/ *E. coli* contamination **(B)** are also shown. The CLOBBv0.1 process starts at section **(C)** Yes switches are colored red and no switches are colored blue.

2.2.2 CLOBBv1.0

After the initial use of CLOBBv0.1 to cluster the *B. malayi* and *O. volvulus* EST datasets it was decided that while the results of the analysis were sufficiently accurate for a preliminary analysis, there were several intrinsic problems with the algorithm.

First the match criteria (p-value $\leq e^{-50}$) was flawed. The maximum p-value for a match is dependent on the length of the probe sequence and the size of the database being search. Short sequences (<150 bp) cannot generate sufficiently high p-values to meet the match criteria and therefore will never be added to an existing cluster. Hence the number of singleton clusters is artificially inflated with short ESTs. Conversely very long sequences that have sufficient (but not perfect) identity would generate p-values that would pass the match criteria. This becomes problematic when gene families are examined as closely related gene sequences could potentially be placed in the same cluster.

Second CLOBBv0.1 could identify potential problem sequences and isolate them from the rest of the analysis. However, the algorithm did not attempt to resolve the relationships of the problem sequence with the clusters it was matching. This task was performed manually and large datasets sometimes required considerable amounts of curation.

To address these problems Dr. John Parkinson (Blaxter laboratory Edinburgh University) revised several portions of the CLOBBv0.1 algorithm and produced CLOBBv1.0. All ESTs clustered with CLOBBv1.0 were filtered for rDNA *and E. coli* contaminants as described above. Like CLOBBv0.1, CLOBBv1.0 is an iterative process. All of the sequences to be analyzed are placed in a single directory in fasta format. When CLOBB is run a list is made of the files in this starting directory. If the process has been used previously to build clusters with the sequences in the directory the old master file is read to determine the number of the last identified cluster. The first EST is then compared to the current cluster database using BLASTN. The BLAST output is searched for high-scoring segment pairs (HSPs). For all HSPs with an identity of $\geq$ 95% and length >30 nt, the hit sequence is logged as a 'type I' match. Those sequences with no type I matches are assigned a new cluster number. The list of type I matches is then checked to find any problems associated with the matches. The beginning and end positions of the match between the query sequence and the matching sequence are logged. If the overlap of the two sequences does not extend more than 30 bases outside the HSP then the match is designated a 'type II' match. If the positions overlap beyond the HSP match by >30 nt indicating the HSP does not extend over the entire length of the two sequences, the query sequence is checked for the presence of low quality sequence. The quality of the sequence is assessed by the number of Ns in the region of sequence overlap. If the non-HSP overlap includes more than 10% of the query sequence length and the sequence is deemed to be of high quality then the match is designated a 'type III' match. If the non-HSP overlap includes more than 10% of the query sequence length and the sequence is deemed to be of low quality then the match is designated a 'type II' match. The list of type II and type III matches is then analyzed. If the matches are to sequences within a single cluster and all the matches are type II matches then EST is assigned the cluster number of the hit sequences. If the matches are to sequences within a single cluster and there are both type II and type III matches the sequence is assigned a new cluster number. This prevents chimeric ESTs or alternately spliced transcripts from being added to established clusters. These types of matches are catalogued for post-process analysis and manual curation. If two or more type II matches to sequences from different clusters are found then the overlap of the two HSPs is reanalyzed. If they are not overlapping then the query sequence links the clusters and they are merged into a single cluster with the lowest indexed cluster number. If the matches are to overlapping sections of the query sequence then the query sequence is assigned the cluster number of the type II match with the highest blast score. Figure 3.2.1 shows a summary of the CLOBB1.0 algorithm schema. Like CLOBBv0.1once a query sequence has been assigned a cluster number it is added to cluster database which is then reformatted before the next search.

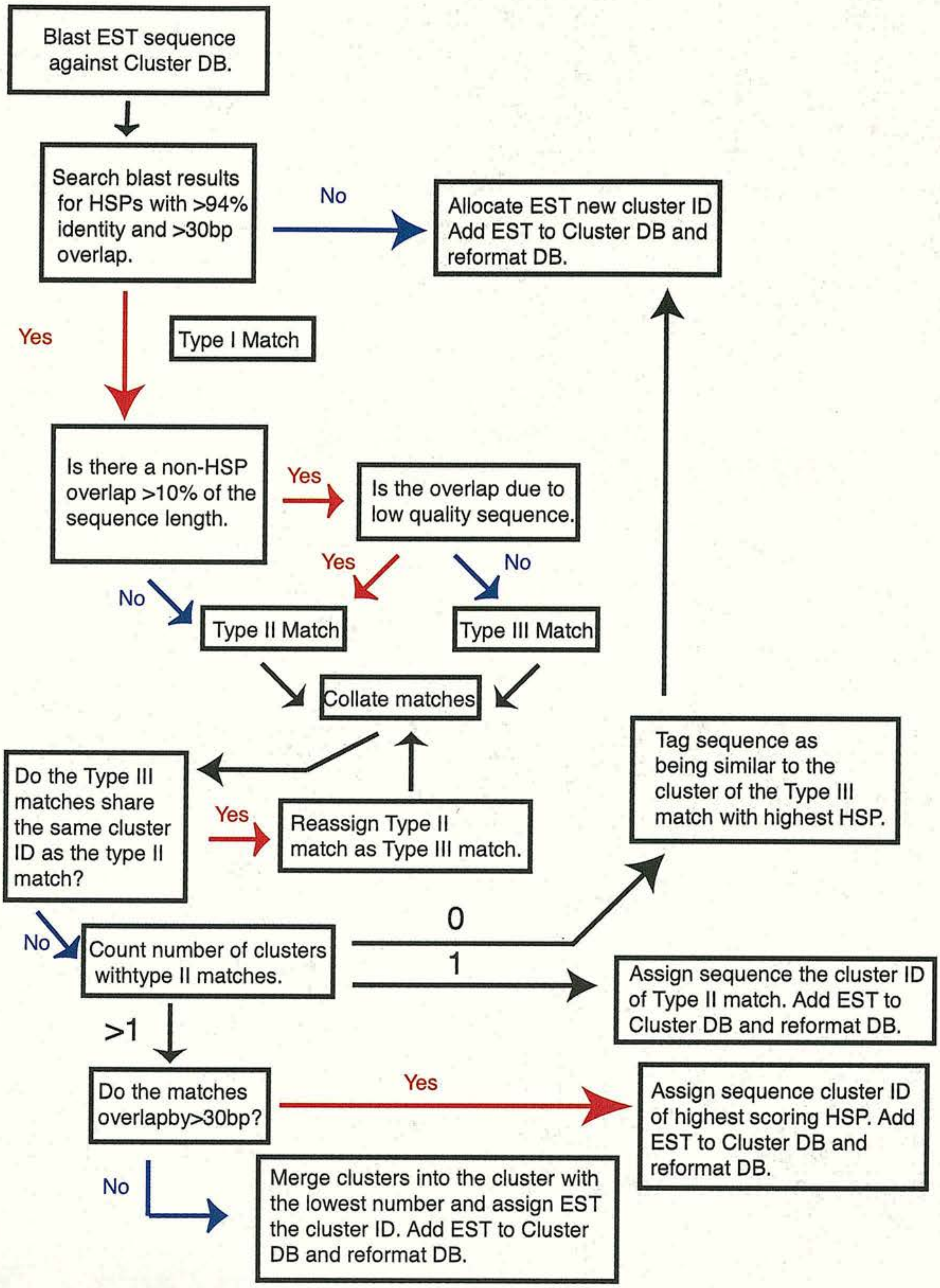
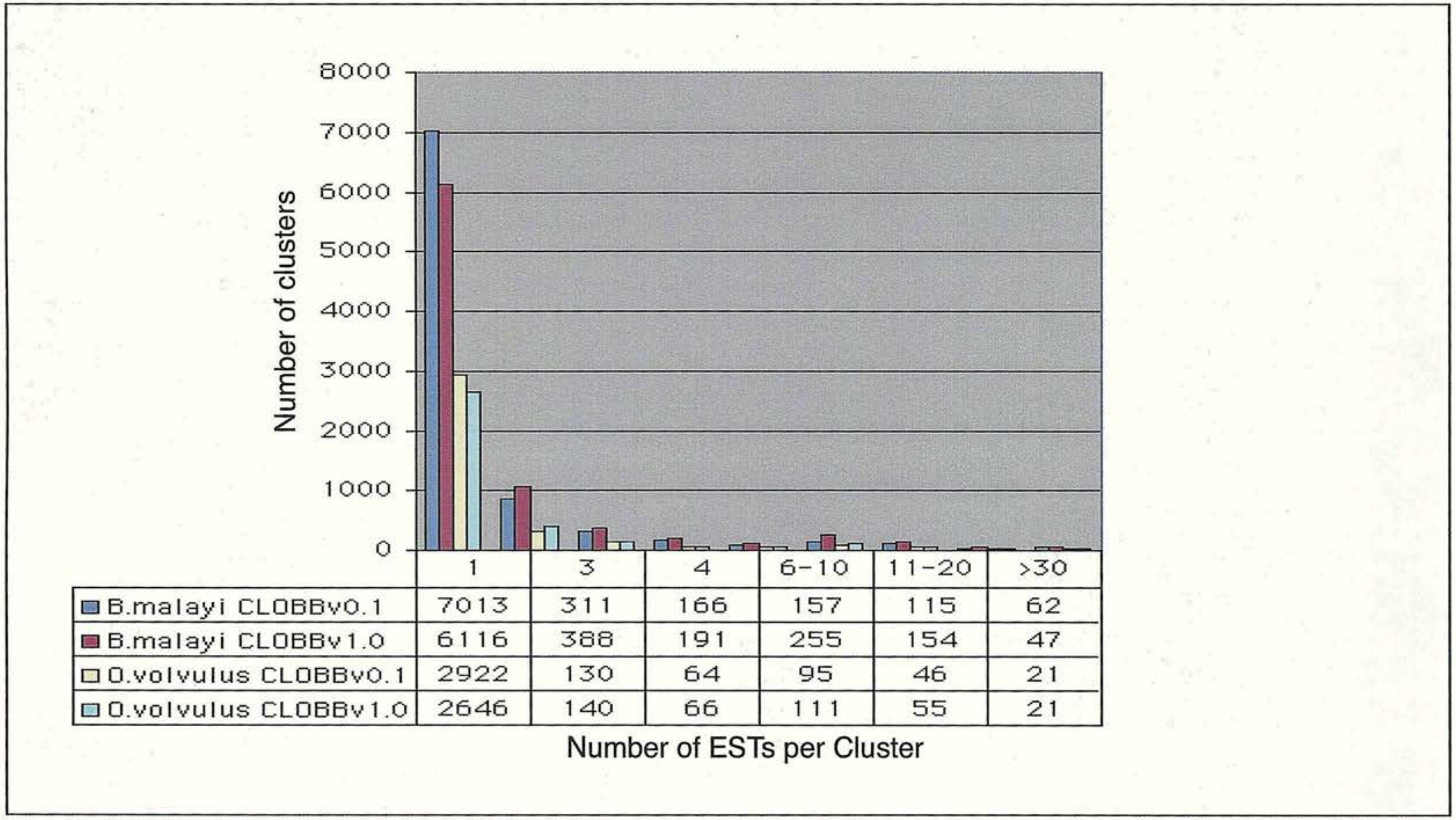

Blast EST sequence against Cluster DB.
Search blast results for HSPs with >94% identity and >30bp overlap.
No
Allocate EST new cluster ID Add EST to Cluster DB and reformat DB.
Yes
Type I Match
Is there a non-HSP overlap >10% of the sequence length.
Yes
Is the overlap due to low quality sequence.
Yes
No
No
Type II Match
Type III Match
Collate matches
Tag sequence as being similar to the cluster of the Type III match with highest HSP.
Do the Type III matches share the same cluster ID as the type II match?
Yes
Reassign Type II match as Type III match.
No
Count number of clusters withtype II matches.
0
1
Assign sequence the cluster ID of Type II match. Add EST to Cluster DB and reformat DB.
>1
Do the matches overlapby>30bp?
Yes
Assign sequence cluster ID of highest scoring HSP. Add EST to Cluster DB and reformat DB.
No
Merge clusters into the cluster with the lowest number and assign EST the cluster ID. Add EST to Cluster DB and reformat DB.

Figure 3.2.1 The CLOBBv1.0 clustering schema. The description of the schema starts at section **C** in the CLOBBv0.1. Yes switches are colored red and no switches are colored blue. Based on figure from Parkinson *et. al.* 2002 submitted.

3.3 Comparison of Clusters generated with CLOBBv0.1 and CLOBBv1.0

To further assess the differences between EST clusters generated with CLOBBv0.1 and CLOBBv1.0 the *B. malayi* and *O. volvulus* EST datasets were clustered with both algorithms. The number of clusters generated and the number of ESTs within each cluster were compared. Figure 3.3.1 shows a comparison of the results of the clustering

Comparison of the EST clusters generated with CLOBBv0.1 and CLOBBv1.0



	1	3	4	6-10	11-20	>30
🟦 B.malayi CLOBBv0.1	7013	311	166	157	115	62
🟥 B.malayi CLOBBv1.0	6116	388	191	255	154	47
□ O.volvulus CLOBBv0.1	2922	130	64	95	46	21
□ O.volvulus CLOBBv1.0	2646	140	66	111	55	21

Figure 3.3.1 Comparison of the CLOBBv0.1 and CLOBBv1.0 clustering of the *B. malayi* and *O. volvulus* EST datasets. The x-axis shows the number of EST sequences in the cluster while the y-axis shows the number of clusters found with those numbers of ESTs. The number of clusters in each group is also shown in the table below the x-axis.

The comparison of the results shows that there were fewer single EST clusters (singletons) in the dataset generated by CLOBBv1.0. In the *B. malayi* dataset twelve percent of the singleton clusters generated in the CLOBBv0.1 analysis were integrated into larger clusters in the CLOBBv1.0 analysis while in the *O. volvulus* dataset nine percent of the singleton clusters were integrated into larger clusters. There are also fewer clusters with more than thirty ESTs in the *B. malayi* dataset generated by CLOBBv1.0. Twenty four percent of these large clusters were split into smaller clusters.

3.4 Renaming CLOBBv1.0 clusters

Analyses of clusters generated with CLOBBv0.1 algorithm have already been published (Blaxter *et. al.*, 1999; Lizotte-Waniewski *et. al.*, 2000; Maizels *et. al.*, 2001; Williams, 1999; Williams *et. al.*, 2000). Keeping the cluster names consistent between database rebuilds is extremely important to the filarial research community so the cluster numbers from the CLOBBv0.1 dataset needed to be assigned to the 'orthologous' cluster built by CLOBBv1.0. To perform this task a perl script called Cluster_name_mover was written. Figure 3.4.1 summarizes the schema of the cluster_name_mover process.

Cluster_name_mover took the clusters generated by CLOBBv0.1 and searched the constituent ESTs against the CLOBBv1.0 cluster database. The names of the clusters assigned to the ESTs in the CLOBBv1.0 database were collected. Then the CLOBBv1.0 cluster number(s) were researched against the CLOBBv1.0 to collect any additional ESTs belonging to the cluster(s). These ESTs were then researched against the CLOBBv0.1 database and any additional cluster numbers collected. If after this first round of searching the list of ESTs collected by Cluster_name_mover were equivalent between the CLOBBv0.1 and CLOBBv1.0 datasets then the script would generate an output of the relationships of the gathered cluster(s). If the number of clusters collected from each dataset was equal to one then the clusters were logged as being equivalent (i.e. they had the same constituents in both analyses) and the CLOBBv0.1 cluster number assigned to the CLOBBv1.0 cluster. If the number of clusters collected from the CLOBBv0.1 dataset was greater than one and only one cluster was found in the CLOBBv1.0 dataset a merge event was logged. The lowest CLOBBv0.1 cluster number was assigned to the CLOBBv1.0 cluster and the remaining CLOBBv0.1 clusters logged as being subsumed into the first CLOBBv0.1 cluster. If the number of clusters collected from the CLOBBv1.0 dataset was greater than one but only one cluster was found in the

CLOBBv0.1 dataset then the a split event was logged. The CLOBBv0.1 cluster number was assigned to the CLOBBv1.0 cluster with lowest ID number and the other CLOBBv1.0 clusters given new numbers to prevent overlap with any numbers already allocated by the process. If the list of ESTs collected from the datasets was not equivalent after the first round of searching, or the relationship between the cluster(s) collected from both datasets did not meet any of the conditions listed above an error was logged and the cluster(s) put on a list called complex_relationships. This list was then searched if the list of ESTs collected were equivalent and the total number of ESTs in the collected clusters were also equal to each other, the process would output a report summarizing the relationships shared by the clusters. The output was analyzed manually and the relationships shared by each of the clusters determined. If the list of ESTs collected were not equivalent or the total number of ESTs in the collected clusters were not equal then the process searched both databases until these conditions were met.

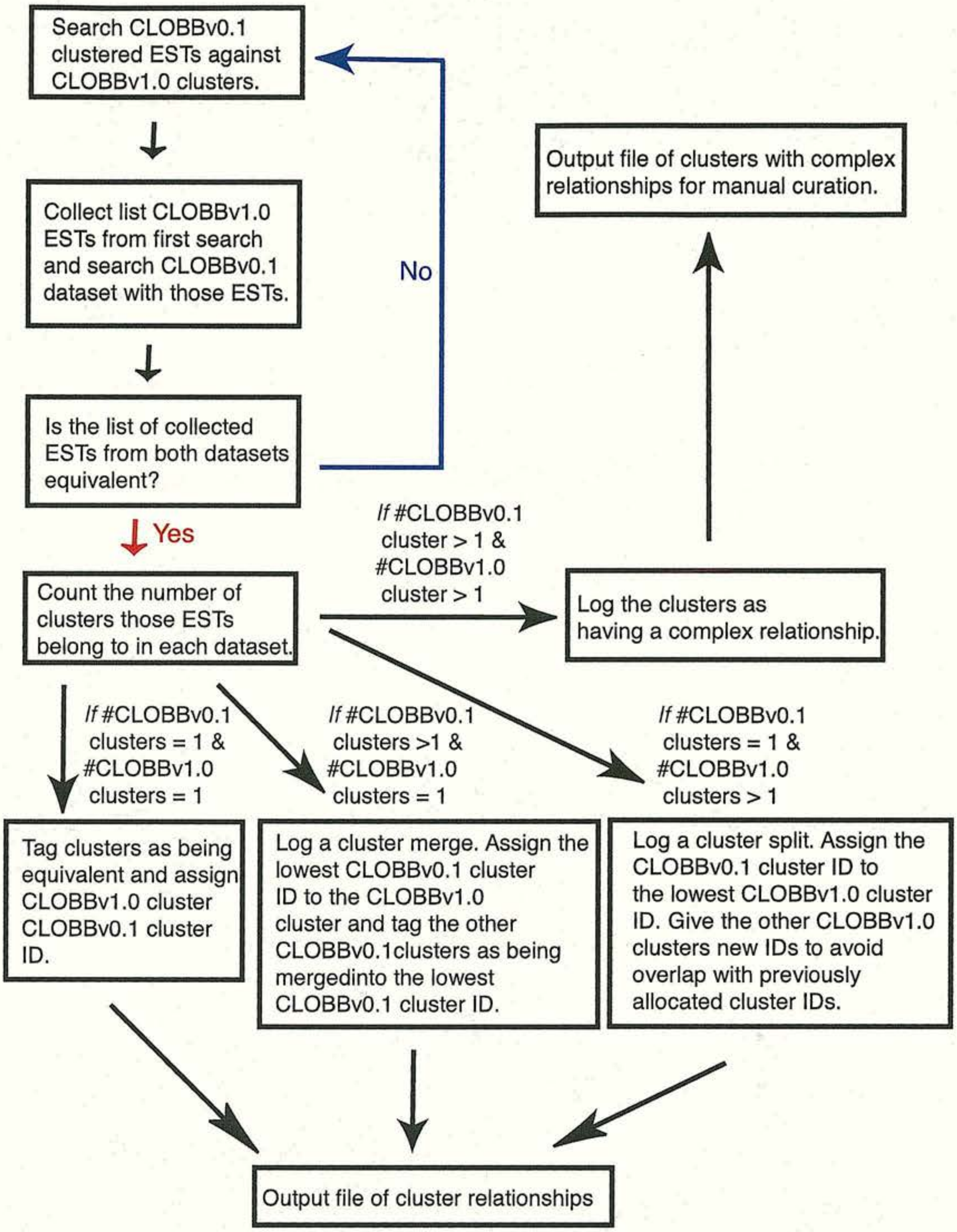

Search CLOBBv0.1 clustered ESTs against CLOBBv1.0 clusters.
Collect list CLOBBv1.0 ESTs from first search and search CLOBBv0.1 dataset with those ESTs.
Is the list of collected ESTs from both datasets equivalent?
No
Yes
Count the number of clusters those ESTs belong to in each dataset.
If #CLOBBv0.1 cluster > 1 & #CLOBBv1.0 cluster > 1
Log the clusters as having a complex relationship.
Output file of clusters with complex relationships for manual curation.
If #CLOBBv0.1 clusters = 1 & #CLOBBv1.0 clusters = 1
If #CLOBBv0.1 clusters >1 & #CLOBBv1.0 clusters = 1
If #CLOBBv0.1 clusters = 1 & #CLOBBv1.0 clusters > 1
Tag clusters as being equivalent and assign CLOBBv1.0 cluster CLOBBv0.1 cluster ID.
Log a cluster merge. Assign the lowest CLOBBv0.1 cluster ID to the CLOBBv1.0 cluster and tag the other CLOBBv0.1clusters as being mergedinto the lowest CLOBBv0.1 cluster ID.
Log a cluster split. Assign the CLOBBv0.1 cluster ID to the lowest CLOBBv1.0 cluster ID. Give the other CLOBBv1.0 clusters new IDs to avoid overlap with previously allocated cluster IDs.
Output file of cluster relationships

Figure 3.4.1 The Cluster_name_mover cluster ID moving schema. Yes switches are colored red and no switches are colored blue. ID: Cluster identifier number.

After all the relationships between the clusters in the CLOBBv0.1 and CLOBBv1.0 datasets were determined the clusters in the CLOBBv1.0 dataset were renamed with the appropriate CLOBBv0.1 cluster number. Table 3.4.2 summarizes the results of this renaming process.

	B. malayi	*O. volvulus*
number of equivalent clusters	6382	2968
number of split clusters	579	187
number of merged clusters	555	233
number of newly allocated cluster numbers	306	118

Table 3.4.2 Results summary of the CLOBBv0.1 and CLOBBv1.0 renaming process for the *B. malayi* and *O. volvulus* datasets. The number of equivalent, split and merged clusters is listed. The number of new cluster numbers allocated to avoid overlap of cluster numbers between the two datasets is also listed.

3.5 Post CLOBB processing of EST sequences and clusters

After the ESTs have been clustered several post-clustering processes are performed. First a set of consensus sequence(s) are predicted for each cluster by aligning the EST sequences within each cluster. These consensus sequences are usually more accurate than individual ESTs as base calls and insertion/deletion events can be assessed with all of the aligned sequences. Overall, these consensus sequences are longer than the individual ESTs and overlaps between ESTs beginning in different sections of the gene often allow the prediction of a consensus sequence that represents almost the entire length of the cDNA. These consensus sequences can then be used in a variety of analyses including extensive database comparisons and the prediction of potential protein sequences. The resulting data can then be parsed into searchable databases that allows the community to access the results of the cluster analysis as well as post clustering analyses

3.5.1 Deriving EST consensus sequences

To derive the consensus sequences for those EST clusters with >1 EST two publicly available assembly algorithms have been tested: Phrap and CAP3 (Phil Green *et. al.* unpublished) (Huang and Madan, 1999). After beta testing both algorithms CAP3 was chosen to perform the contig assembly of the EST cluster because its more stringent assembly parameters gave more reliable consensus sequences with fewer insertion or deletion events. Other groups have also chosen CAP3 to derive consensus sequences from assemblies for this reason (Liang *et. al.*, 2000). Because of this stringency there are instances when CAP3 was unable to produce a single assembly from the ESTs in the cluster. Multiple assemblies are given the cluster number along with a contig number so that each of the contigs can be examined separately (BMC#####_contig1 for example). If CAP3 was not able to derive a contig from the EST cluster then Phrap was used to create an assembly.

3.5.2 Comparison of EST cluster consensus sequences to other publicly available sequences and derivation of potential protein sequences from B. malayi consensus sequences

The *B. malayi* and *O. volvulus* cluster consensus sequences were compared to a set of custom-built databases derived from GenBank. Table 3.5.2.1 lists the databases and what blasts were performed against them with the consensus sequences. BLAST hits with a p-value of $\leq e^{-10}$ were logged as a significant match.

Database	Blast algorithm used to search the database	Description in Venn diagrams	Number of sequences in database	Number of nt or aa in database	Database contents
xrest	blastx	Other Phyla	659,241	195,291,993	All non-nematode protein sequences in genbank
xnon	blastx	Other Nematodes	1,428	354,439	All non-*C. elegans* nemtode protein sequences in genbank
xce	blastx	*C. elegans*	25,645	11,589,474	All predicted *C. elegans* protein sequences in wormpep21
xdm	blastx	*D. melanogaster*	14,348	7,182,582	All predicted *D. melanogaster* protein sequences
xhs	blastx	*H.sapiens*	37,526	15,217,171	All predicted *H. sapiens* protein sequences
txcladeI	tblastx	Clade I	5,085	2,408,302	All Clade I DNA sequences in genbank (mostly ESTs)
txcladeIV	tblastx	Clade IV	29,037	12,878,231	All Clade IV DNA sequences in genbank (mostly ESTs)
txcladeV	tblastx	Clade V	19,776	97,281,089	All Clade V DNA sequences in genbank (many ESTs) including the *C. elegans* genome sequence.
txascarid	tblastx	Ascarid	1625	931,177	All ascarid DNA sequences in genbank (mostly ESTs).
txov	tblastx	*O. volvulus*	15,502	6,558,555	All *O. volvulus* DNA sequences in genbank (mostly ESTs).
txbm	tblastx	*B. malayi*	26,606	12,507,647	All *B. malayi* DNA sequences in genbank (mostly ESTs).

Table 3.5.2.1 Summary of the BLAST searches performed with the *B. malayi* and *O. volvulus* consensus sequences. The table shows the names of the custom databases built for searching with the filarial EST consensus sequences, the blast algorithm used to search the database, the number of sequences in the database, the number of nucleotides or amino acids residues in the database, the description given to the dataset in the Venn diagrams presented in the results and the contents of each database. These databases were constructed on 8/10/01.

The protein sequences predicted from the *B. malayi* cluster consensus sequences were kindly provided by Dr. John Parkinson (Blaxter Lab, ICAPB Edinburgh University). The Decoder algorithm (Fukunishi and Hayashizaki, 2002) was used to predict potential protein sequences from the cluster consensus sequences. Decoder was written by the RIKEN group to make protein predictions from the human and mouse EST datasets. It can utilize chromatograph quality data to help reduce premature truncation of protein translations due to poor quality sequence. Because Decoder only begins protein sequences at a methionine (Met) residue a post process was written that looked upstream of the potential start Met for possible extensions to the Decoder predicted protein (Parkinson *et. al.* 2001 per. com).

3.5.3 Parsing of EST clusters into searchable databases

Custom databases have been constructed in FileMakerPro (v5.0, FileMaker Inc.) to house the EST cluster data and allow the research community access to the clusters and dataset analyses both locally and through the world wide web (Parkinson *et. al.*, 2001). For the analyses described below the EST clusters, consensus sequences and results of the consensus sequence blasts were imported in to a custom built FileMakerPro database. For each EST cluster the length of the longest consensus sequence and predicted protein sequence were calculated and imported into the FileMakerPro database. This dataset was then searched using the query capacity built into FileMakerPro.

3.6 Cluster analysis of *B. malayi* EST datasets and evaluation of gene discovery effort

In total 20,626 *B. malayi* ESTs sequences were analyzed. These sequences originated from fourteen separate cDNA libraries (see figure 1.2.3.4. and table1.2.3.4.3). After filtering for rDNA and *E. coli* sequences 18,740 sequences were grouped into 8,403 clusters. Table 2.6.0.1 shows the results of the cluster analysis for each cDNA library and the total dataset. The redundancies of the datasets were calculated by dividing the number of ESTs in the dataset by the number of clusters generated by CLOBBv1.0.

Library	Total Number of ESTs	Number of rDNA seq.	Percent dataset	Number of *E.coli* seq.	Percent dataset	Number of Clustered ESTs	Number of Clusters	Intra-library Redund.	Total Redund.	Number of Library Specific Clusters	Percent of Library Specific Clusters	Curators Comments
BmMf	1898	289	15	4	<1	1605	1130	1.4	1.7	672	59.5	Very rDNA contaminated
BmMfZ	1402	30	2	0	<1	1372	1019	1.3	1.4	675	66.2	
BmL2	609	0	<1	1	<1	608	241	2.5	2.5	81	33.6	Many chimeras
BmL3	1930	82	4	5	<1	1843	1204	1.5	1.6	700	58.1	
BmL3SL	298	0	<1	0	<1	298	188	1.6	1.6	71	37.8	
BmL3SA	213	4	2	8	4	201	131	1.5	1.6	37	28.2	Some chimeras, Some *E. coli* and rDNA contamination
BmL3SB	334	34	10	2	1	298	162	1.8	2.1	39	24.1	Some chimeras and very rDNA contaminated
BmL3SZ	805	5	1	8	1	792	462	1.7	1.7	243	52.6	
BmL3D6	1491	152	10	9	1	1330	786	1.7	1.9	596	75.8	Very rDNA and *E. coli* contaminated
BmL3D9	240	30	13	14	6	196	147	1.3	1.6	39	26.5	Very rDNA and *E. coli* contaminated
BmL4	664	190	29	5	1	469	277	1.7	2.4	95	34.3	Very rDNA and *E. coli* contaminated
BmL4SL	1061	0	<1	1	<1	1060	463	2.3	2.3	268	57.9	Some chimeras
BmYD25	1072	195	18	22	2	855	477	1.8	2.2	226	47.4	Very rDNA contaminated and some *E. coli* contamination
BmYD29	311	18	5	0	<1	186	146	1.3	2.1	54	37.0	Very rDNA contaminated
BmAM	4660	362	8	6	<1	4292	2642	1.6	1.8	1799	68.1	Some rDNA contaminated
BmAF	3638	291	8	12	<1	3335	2021	1.7	1.8	1201	59.4	Some rDNA contaminated
Total	20626	1682	8	97	<1	18740	8403		2.5			

Table 3.6.0.1 Results summary of the CLOBBv1.0 clustering of the *B. malayi* EST dataset. The number of ESTs sequenced from each library is shown as well as the number of rDNA and *E.coli* sequences removed from the datasets before clustering. The number of clusters containing sequences from each library is listed with the level intra-library sequence redundancy (calculations performed without the filtered rDNA and *E.coli* sequences). The redundancy of the library as calculated by comparison with the complete EST dataset is also shown. The number of clusters containing sequences exclusively from the library and percentage these clusters represent of the library dataset is shown. Curators comments on the general characteristics of the library rDNA /*E.coli* contamination and the incidence of chimeric ESTs is also shown The numbers in red highlight calculations that indicate that the library no longer represents a productive sequencing substrate, either because of high rates of rDNA/ *E. coli* contamination or high levels of redundancy within the dataset.

Besides deriving a non-redundant gene set that can be the subject of other analyses the cluster analysis helps the filarial gene discovery program monitor the productivity of the EST sequencing effort. One of the major problems with extensive EST sequencing from many of the *B. malayi* cDNA libraries is the high level of rDNA clones. In six of the cDNA libraries (BmMf, BmL3SB, BmL3D6, BmL3D9, BmL4 and BmYD25) more than ten percent of the sequenced ESTs have been tagged as rDNA. In total, eight percent of the *B. malayi* EST dataset has been tagged as rDNA. For the BmMf and BmL3cDNA library the subtractive hybridization (BmMfZ and BmL3Z) of rDNA clones and other abundant transcripts increased the rate of new gene discovery. The majority of the cDNA libraries have less than one percent of the clones being tagged as *E. coli* contamination. However, the BmL3D9 cDNA library has six percent of the sequenced clones tagged as being derived from *E. coli*.

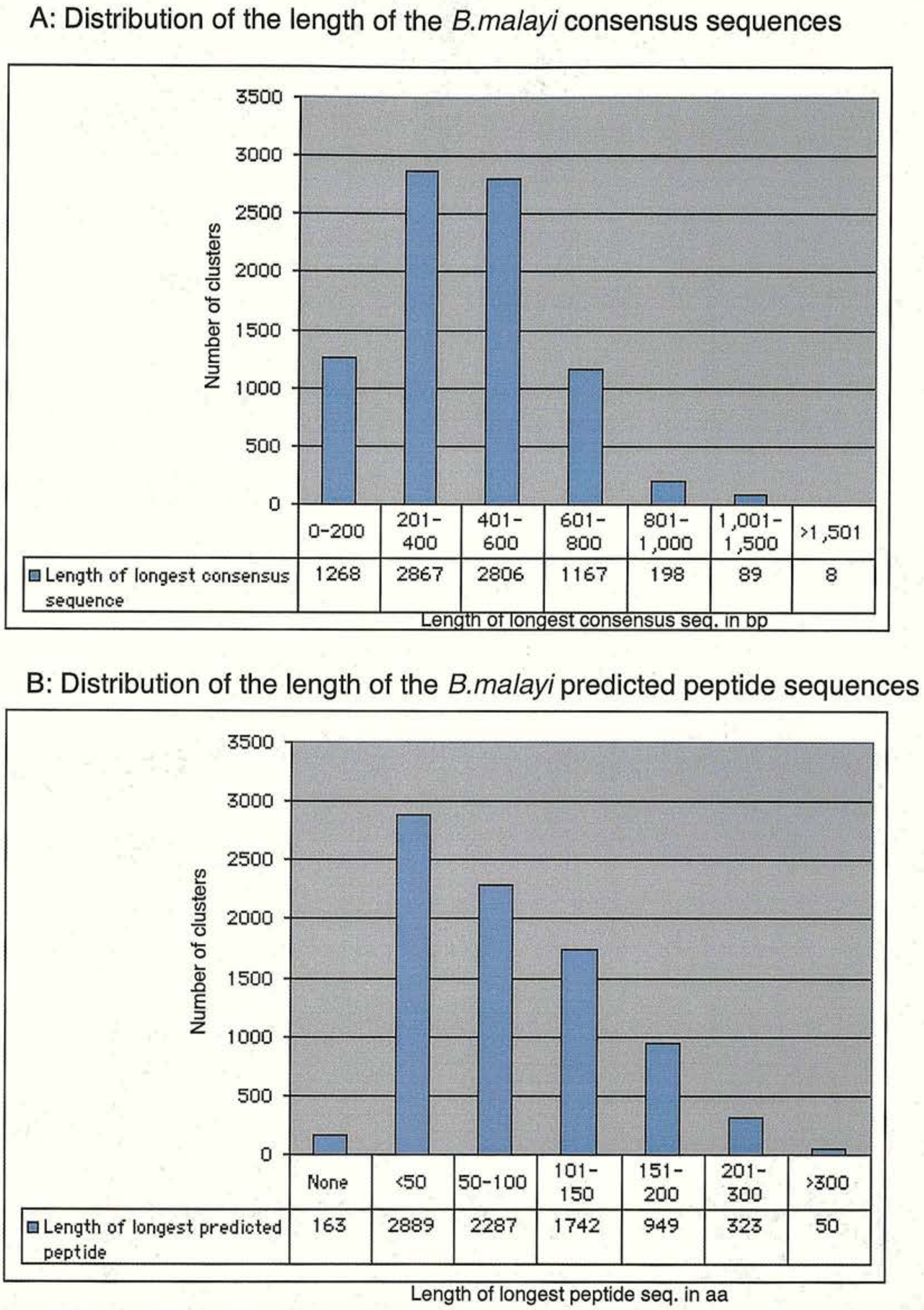
The calculated redundancy of a dataset gives an estimate of the productivity of the gene discovery effort. The overall redundancy of the *B. malayi* EST sequences (2.5) indicates that each cluster contains an average of 2.5 ESTs. Usually sequencing from a library will end after a redundancy of 2 to 2.5 is reached. However, clustering of the *B. malayi* ESTs has shown that 70% of the sequences have been placed in single EST clusters (see table 3.6.0.1). In addition, when the individual library datasets are examined almost half still have redundancies of less than two. This indicates that a small number of highly represented sequences are responsible for high redundancy seen in the total dataset. Remarkably, the redundancy of the BmAM and BmAF cDNA libraries has remained low despite the fact that more than 3000 ESTs have been sequenced from each library.

By examining how many clusters are unique to a library it is possible to get an estimate of how many transcripts in the library might be unique to that stage. Abundant transcripts that have only been observed in a particular library may represent differentially expressed genes. Four of the *B. malayi* libraries (BmMf, BmL3D6, BmAM, BmAF) show higher incidence of unique sequences relative to the other libraries. The majority of these sequences are not clustered with any other EST. It is possible these libraries contain a more diverse set of transcripts because stage specific biology required a larger number of different genes to be expressed at these time points.

Overall, the libraries generated with PCR based methodologies (BmL2SL, BmL3SA, BmL3SB and BmL4SL) show much higher rates of redundancy than the libraries generated by conventional methodologies. This is not unexpected as smaller abundant transcripts are going to amplify more efficiently than larger rarer transcripts. A second problem with the PCR based libraries is that they contain a higher incidence of chimeric clones than the conventional libraries. The clustering of ESTs from these libraries is more problematic because the chimeras can cause the inappropriate merging of gene clusters that do not actually overlap. The stringent HSP overlap rules built into the CLOBBv1.0 algorithm are used to help prevent these mergers. How successful CLOBBv1.0 has been in identifing these chimeric sequences is still under assessment.

3.6.1 Characteristics of B. malayi EST clusters

The *B. malayi* clusters have an average of 2.2 ESTs per cluster (excluding the contribution of the rDNA and *E. coli* contaminants to the dataset redundancy). Figure 3.6.1.1 shows the distribution of the number of ESTs per cluster. The majority of the EST clusters contain only one EST. The average length of the *B. malayi* ESTs is 383 bp in length and the average length of a consensus sequence is 547 bp. The length of the average consensus sequence is significantly longer than the average EST length ($p < 0.0001$ Man-Whitney test, performed in Minitab, Minitab Inc). Figure 3.6.1.1A shows the distribution of the lengths of the *B. malayi* consensus sequences. The majority of the consensus sequences are between 200 and 600 bp in length (67%). The longest predicted consensus sequence is 1,970 bp in length. The average length of the longest protein sequence predicted from each cluster is 88 aa. Figure 2.6.1.1B shows the distribution of the lengths of the protein sequences predicted from the *B. malayi* consensus sequences. The majority of the protein sequences are between 0 and 100 aa in length (62%). The longest predicted protein sequence is 510 aa in length.



A: Distribution of the length of the *B.malayi* consensus sequences

	0-200	201-400	401-600	601-800	801-1,000	1,001-1,500	>1,501
■ Length of longest consensus sequence	1268	2867	2806	1167	198	89	8

B: Distribution of the length of the *B.malayi* predicted peptide sequences

	None	<50	50-100	101-150	151-200	201-300	>300
■ Length of longest predicted peptide	163	2889	2287	1742	949	323	50

Figure 3.6.1.1 Characteristics of the consensus sequences and predicted protein sequences derived from the B. malayi EST clusters. The graphs show the relative distribution of the lengths of the longest consensus sequence assembled from each cluster (**A**) or the longest protein sequences predicted from the consensus sequences (**B**). The x-axis shows the lengths of the sequences while the y-axis shows the number of clusters found with those lengths. The number of clusters belonging to each group is also shown in the table below the x-axis.

The results of the BLAST comparisons of the *B. malayi* consensus sequences to the public databases are summarized in figures 3.6.1.2 and 3.6.1.3. The results of both the whole dataset and several subdivisions of the whole dataset are shown in Venn diagrams. Remarkably, 64% of the EST clusters do not have significant similarities to any sequences in the public databases (see figure 3.6.1.2A). Analysis of the proteins predicted from the whole genome sequences of other animals (*H. sapiens* (Lander *et. al.*, 2001; Venter *et. al.*, 2001), *D. melanogaster* (Adams *et. al.*, 2000) and *C. elegans* (consortium, 1998)) indicate that between 25-40% do not show significant similarities to other sequences in the public databases. Other nematode EST sequencing projects show levels of novel sequences which are consistent with the rates observed in other animal datasets (between 28 and 48% novel sequences Parkinson and Blaxter 2002 pers. com.). This high rate of novel sequences in the *B. malayi* ESTs could be due to several factors that may represent inherent problems with this dataset. Several of these potential problems can be tested for in the dataset. These include the length of the consensus searched against the database, the fidelity of the consensus sequence, or the amount of protein coding potential contained within the consensus sequence.

The length of the consensus sequence is an important factor determining whether BLAST finds a significant match in the database. The calculation of the BLAST p-value is based on the length and composition of the sequence alignment. Short consensus sequences may never achieve high enough scores to be counted as significant. This possibility can be tested by removing those clusters with short consensus sequences (< 300bp) from the analysis.

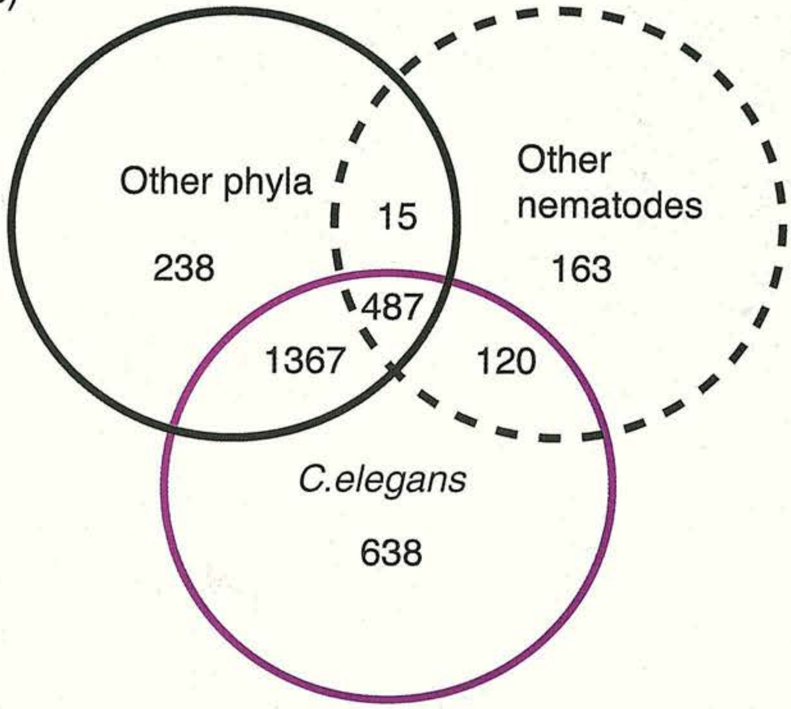
The majority of the *B. malayi* consensus sequences are generated from single EST clusters. The stringent overlap parameters used by CLOBBv1.0 when generating the clusters ensures that low quality sequences will not be clustered with high quality sequences. If there is a high proportion low quality single EST clusters this may artificially inflate the number of novel sequences. By examining clusters with more than one EST these low quality sequences should be excluded from the analysis. However, one caveat to this analysis is that well characterized genes involved in general organism homeostatic functions tend to be abundant transcripts in the EST dataset. Therefore the results of this analysis may be skewed because the sampling of abundant transcripts may enrich for these genes.

Not all transcribed RNAs are translated. Most of these untranslated RNAs are not polyadenylated and therefore should not be present in the filarial genome project cDNA libraries. However, because of the high AT content of the filarial genomes some of these untranslated RNAs may have internal homopolymeric tracks of adenines. During the construction of the cDNA libraries these tracks anneal to the dT oligos used to purify the mRNA away from the total RNA and prime first strand synthesis. In addition, the 5' and 3' untranslated regions (UTRs) of all eukaryotic mRNAs are also non-coding. Nematode 5' UTRs are generally short (consortium, 1998). However, to ensure that UTRs and the contaminating untranslated RNAs are not responsible for the high levels of novel sequences in the *B. malayi* EST dataset the coding potential of each consensus was tested by examining the longest protein predicted by Decoder (Fukunishi and Hayashizaki, 2002). If the consensus sequence had a predicted protein over 50 aa then the sequence was considered to have a high coding potential.

The results of the BLAST searches of the whole consensus sequences dataset against all proteins in GenBank are shown in figure 3.6.1.2A. The results of the BLAST searches of the subdivisions of the consensus sequences dataset are shown in figures 3.6.1.2B-D. The significance of the differences in the percent of novel sequences in each subdivision and the whole dataset was evaluated using Chi^2 analysis (performed in Minitab, Minitab Inc.). In all three subdivisions the percent of the dataset that was novel was significantly lower than the whole dataset. (Chi^2 analysis with p-values < 0.001). The *B. malayi* clusters with >1 EST showed the lowest level of novels (43%).

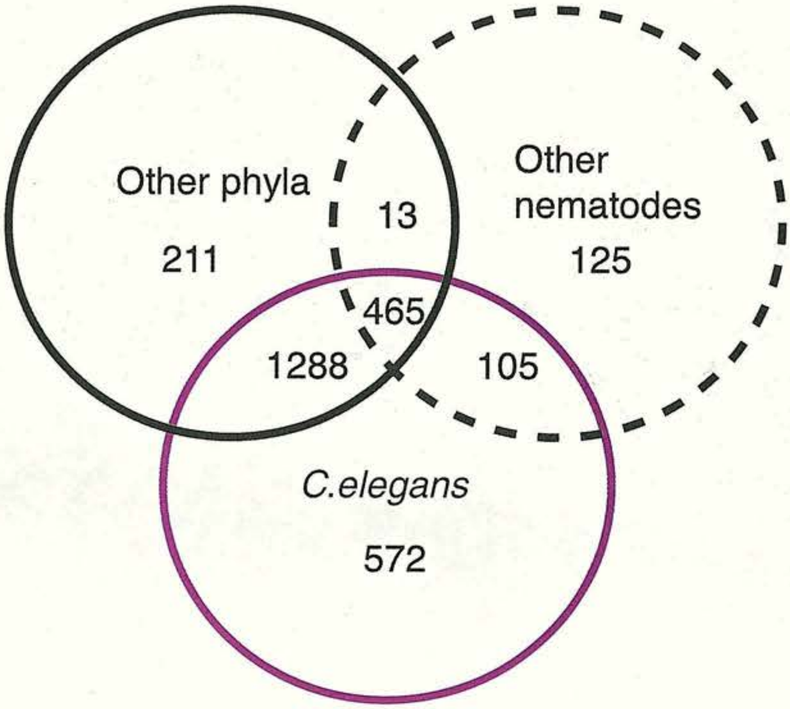
A: All *B. malayi* EST clusters: 8403

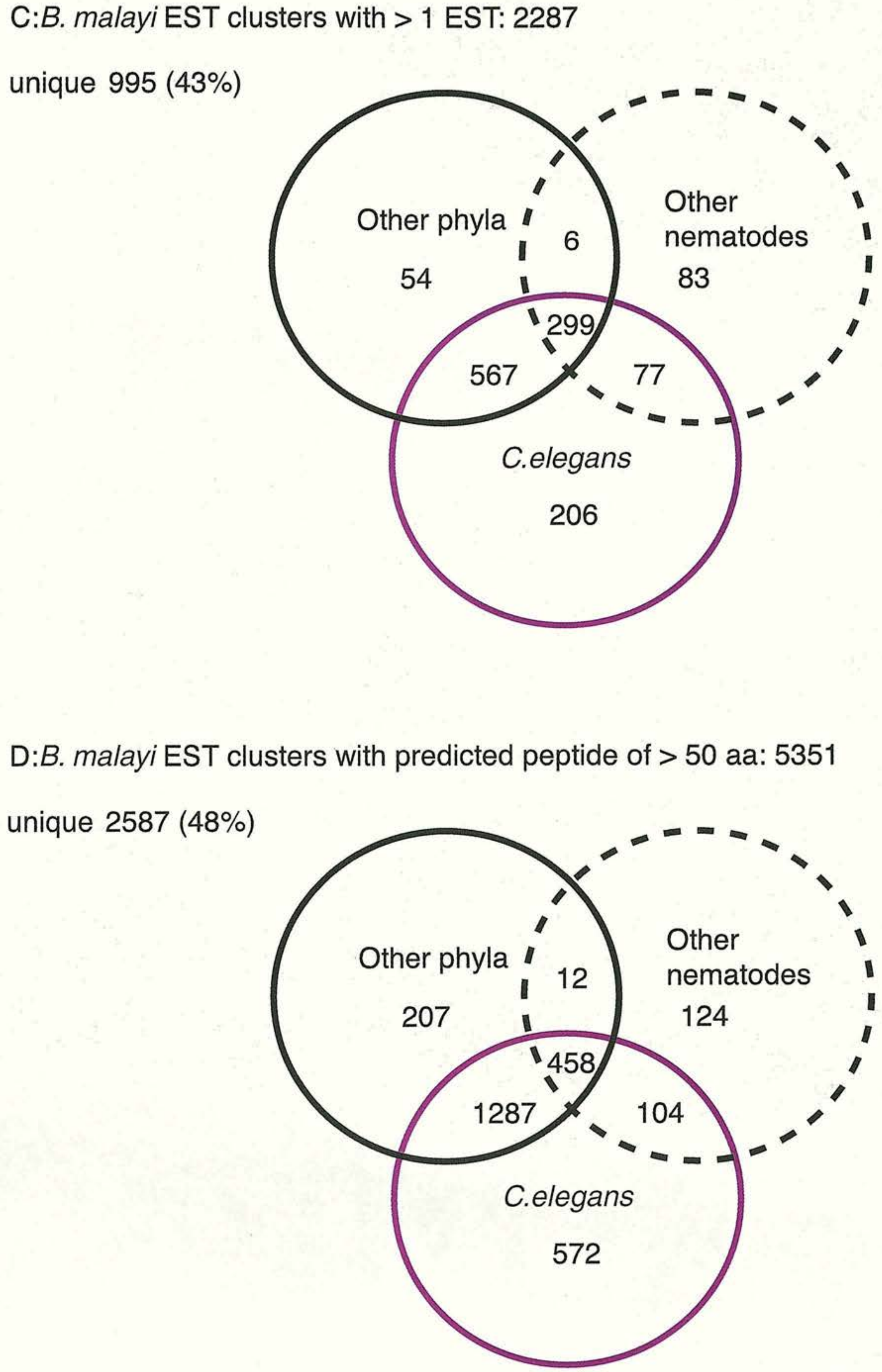
unique 5375 (64%)



B: *B. malayi* EST clusters with consensus sequence > 300 bp: 5814

unique 3035 (52%)



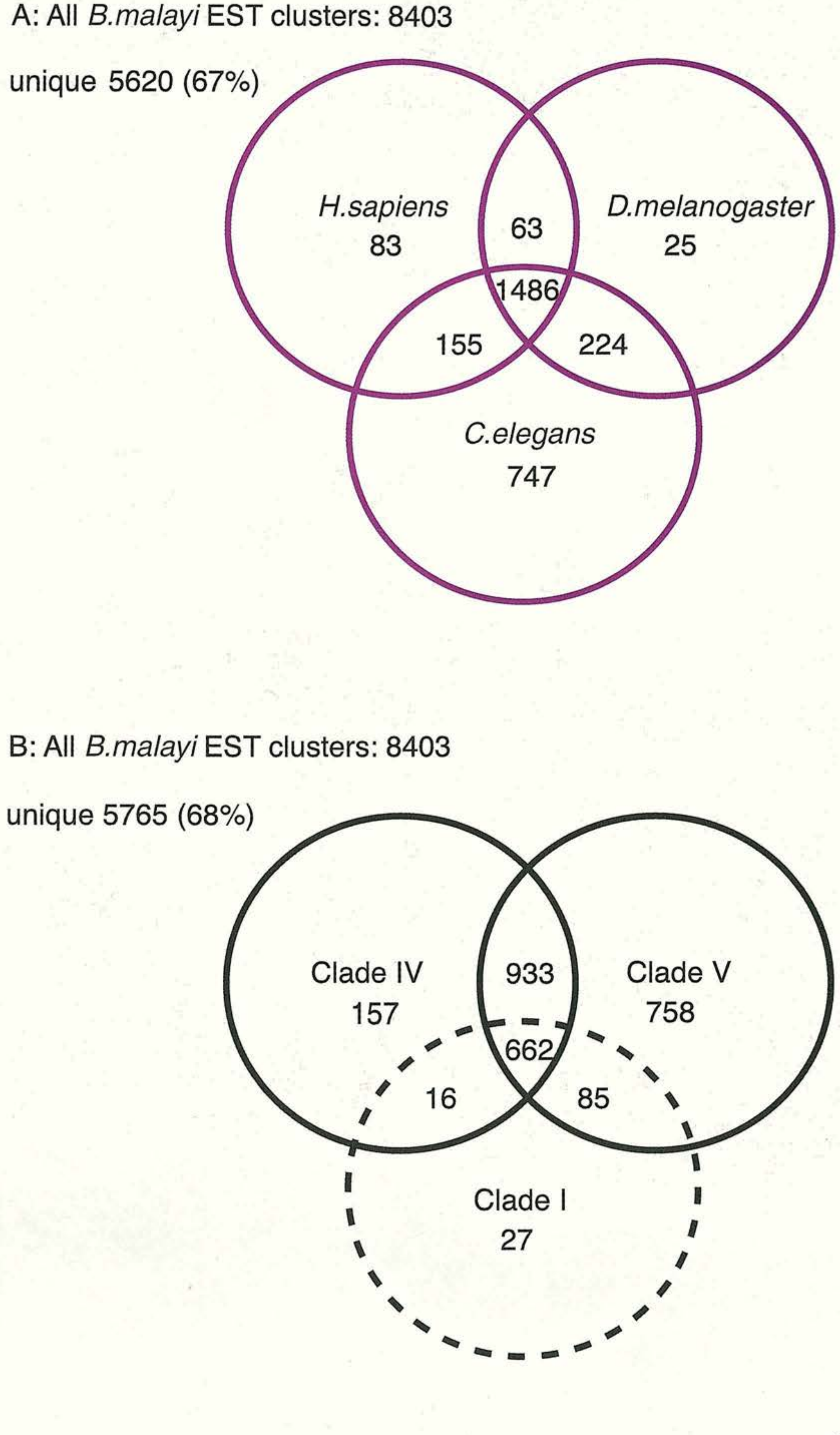


C:*B. malayi* EST clusters with > 1 EST: 2287

unique 995 (43%)

D:*B. malayi* EST clusters with predicted peptide of > 50 aa: 5351

unique 2587 (48%)

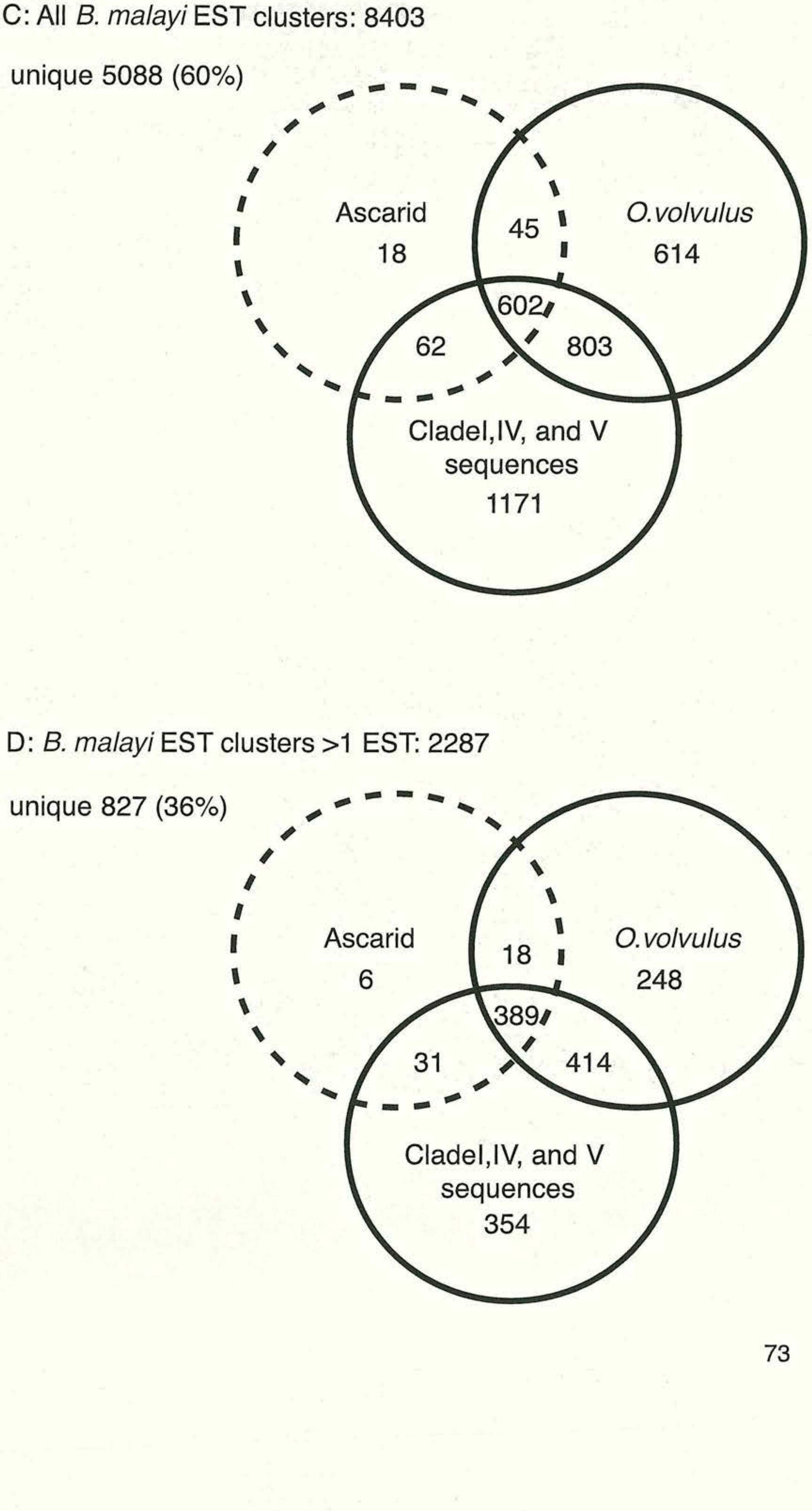


A: All *B.malayi* EST clusters: 8403

unique 5620 (67%)

B: All *B.malayi* EST clusters: 8403

unique 5765 (68%)



C: All *B. malayi* EST clusters: 8403

unique 5088 (60%)

D: *B. malayi* EST clusters >1 EST: 2287

unique 827 (36%)

Figures 3.6.1.2 and 3.6.1.3 Results of the BLAST searches of the *B. malayi* consensus sequences against the public databases. Each Venn diagram shows a three-way comparison of the results of the BLAST searches against three databases. The number of consensus sequences having significant matches to a single database are shown in the center of each circle. The number of consensus sequences showing a significant match (p-value ≤ e^{-10}) to more than one database are listed at the interfaces of each circle. The number of consensus sequences with no significant matches to any of the listed databases is shown to the left of the Venn diagram. Figure 3.6.1.2A shows the results of the BLASTX searches against the xrest, xnem, xce databases. The summary of the whole dataset is shown in figure 3.6.1.2A The results of three subdivsions of the whole dataset , those clusters with consensus sequences with lengths >300 bp Figure 3.6.1.2B, those clusters with >1 EST Figure 3.6.1.2C and those clusters with predicted protein sequences of >50 aa Figure 3.6.1.2D are also shown. Figure 3.6.1.3A shows the results of the BLASTX searches of the consensus sequences against the protein sequences predicted from the *H. sapiens*, *D. melanogaster* and *C. elegans* whole genome sequence. Figure 3.6.1.3.B shows the results of the TBLASTX searches of the consensus sequences against the EST sequences of nematodes from Clades I, IV and V. Figures 3.6.1.3C and 3.6.1.3D show the results of the TBLASTX of the consensus sequences against the EST sequences of Ascarid, *O. volvulus* and Clade I, IV and V nematodes. Purple circles indicate that the dataset is based on proteins predicted from whole genome sequence. Dashed ciricles indicate that the sequence dataset is substantially smaller than the other comparators.

Figure 3.6.1.3A shows the results of the BLASTX searches of the *B. malayi* against the protein sequences predicted from the genomes of *H. sapiens*, *D. melanogaster* and *C. elegans*. Only 33% had significant similarities to proteins in those datasets. The majority of the sequences with significant similarities (17%) had matches to proteins in all three datasets. Nine percent of the consensus sequences had matches to *C. elegans* but not to *H. sapiens* and *D. melanogaster*. These sequences represent a potential group of nematode-specific genes. Two percent of the sequences had significant similarities to sequences in the *C. elegans* and *D. melanogaster* datasets but not the *H. sapiens* dataset. A similar number of sequences have similarities to *C. elegans* and *H. sapiens* but not *D. melanogaster*. Interestingly, a similar small proportion of the dataset (2%) has similarities to *D. melanogaster* and/or *H. sapiens* but no similarity to *C. elegans* proteins. These may represent a group of genes that have been lost from *C. elegans* are still present in other animal genomes.

Figures 3.6.1.3B-D show the results of the TBLASTX comparisons of the *B. malayi* consensus sequences against the EST datasets from other nematodes. Figure 3.6.1.3B shows the results of searches against ESTs from the major nematode clades with species that have EST sequences deposited in genbank (clades I, IV, V). Like the searches against the protein sequences, the majority of the sequences (68%) do not have significant similarity to any of the clade I, IV and V nematode EST sequences. The majority of the similarities are found to ESTs from clade IV and V nematodes, which is not surprising because they represent the most heavily sampled groups. Figures 3.6.1.3C and D show the results of the TBLASTX searches against the clade III ascarid, *O. volvulus* and the clade I, IV and V nematode EST datasets. Figures 3.6.1.3C shows the results of the total dataset while figure 3.6.1.3D shows a subdivision of the *B. malayi* consensus sequences assembled from clusters with >1 EST. Like the searches against other datasets the ESTs clusters with >1 EST are much more likely to have similarities to sequences in the dataset (64% vs 30%). In both datasets approximately 45% of the clusters had similarities to *O. volvulus* ESTs. This indicates that unlike results of comparisons to other nematode datasets both abundant and rare transcripts are as likely to have similarities to EST sequences in the *O. volvulus* dataset.

3.6.2 Analysis of abundant and differentially expressed transcripts of B. malayi

One of the main goals of the gene discovery effort is to isolate the next generation of vaccine and drug candidates as well as provide insights into filarial nematode biology that would not be elucidated with directed research programs. The level of mRNA in a cell often correlates with the level of production of protein. Thus the EST datasets provide a platform which allows the researcher to identify the proteins the nematode produces either throughout the life cycle or in discrete developmental stages.

3.6.2.1 Abundant transcripts of B. malayi

To isolate the most abundant transcripts the *B. malayi* cluster dataset was searched and those clusters with > 40 ESTs examined. Table 3.6.2.1.1 lists the fifteen clusters identified in this search. Figure 3.6.2.1.2 shows the relative expression pattern of the twelve non-mitochondrial EST clusters.

Cluster	Number of ESTs	Percent of Total EST Dataset	Similarities	Gene Names	References
BMC00185	174	0.93	RNA binding protein (polyadenylation complex subunit)	*rbp-1*	(Gregory *et. al.*, 1997; Martin *et. al.*, 1996)
BMC01618	134	0.72	Novel	*aaf-1*	
BMC01601	131	0.70	Novel	*aad-1*	
BMC00060	100	0.53	ribosomal protein	*rpl-36*	
BMC04376	83	0.44	cytidine deaminase	*cdd-1*	(Anant *et. al.*, 1997; Gregory *et. al.*, 1997)
BMC00188	80	0.43	ribosomal protein	*rps-12*	
BMC00213	76	0.41	abundant larval transcript 2	*alt-2*	(Gregory *et. al.*, 2000; Gregory *et. al.*, 1997)
BMC00211	67	0.36	thioredoxin peroxidase	*tpx-2*	(Ghosh *et. al.*, 1998)
BMC00071	52	0.28	mitochondrial gene	*cox-2*	(Keddie *et. al.*, 1998)
BMC00351	51	0.27	vespid venom-associated-allergen-like (activation secreted protein)	*vah-1*	(Murray *et. al.*, 2001)
BMC01688	47	0.25	similar to cuticular collagen	*col*	
BMC00352	46	0.25	mitochondrial gene	*cox-3*	(Keddie *et. al.*, 1998)
BMC00802	43	0.23	mitochondrial gene	*ssu-mt*	(Keddie *et. al.*, 1998)
BMC00030	42	0.22	similar tropomyosin proteins	*tin-1*	
BMC00849	41	0.22	ribosomal protein	*rps-7*	

Table 3.6.2.1.1 Abundantly expressed transcripts discovered in the *B. malayi* EST sequencing project. The table lists the fourteen identified *B. malayi* EST clusters with >40 ESTs. The cluster ID, number of ESTs in the cluster, the percent of the dataset the cluster comprises, similarities to proteins in the public databases, gene names if assigned and any relevant references are shown.

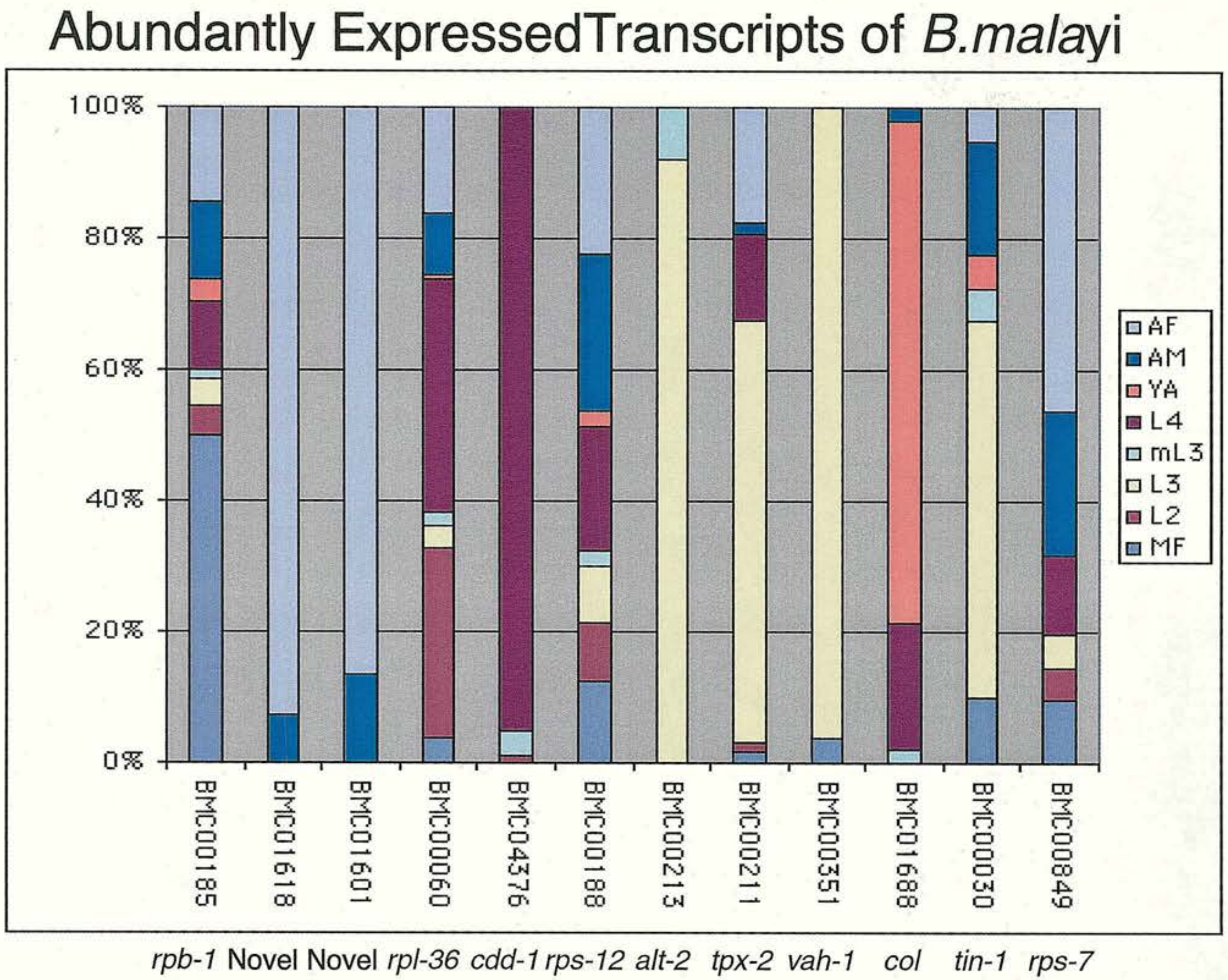


Figure 3.6.2.1.2 The relative expression pattern of the abundantly expressed non-mitochondrial *B. malayi* EST clusters presented in table 3.6.2.1.1. For each of the clusters the number of ESTs originating from different lifecycle stages is presented. The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. MF: microfilaria, L2: day six vector derived stage 2 larvae, L3: day twelve vector derived infective stage 3 larvae, mL3: molting stage 3 larvae, L4: day fourteen post infection stage four larvae, YA: young adult, AM: adult male, AF: adult female.

Six of these abundant clusters have homologies to mitochondrial (mt) or ribosomal protein (rp) genes. These highly expressed transcripts appear to be major components of most EST datasets examined to date. The comparison of the ESTs in the *cox-2*, *cox-3* and *ssu-mt* clusters with the *B. malayi* mt-genome indicates the sequences originate from mt- mRNAs not mt-genomic contamination (Keddie *et. al.*, 1998). The remaining clusters can be divided into three categories: proteins of unknown function (novels), conserved metabolic enzymes and structural proteins and potential mediators of host-parasite interactions.

Two abundant EST clusters BMC01618 and BMC01601 do not have any similarities to proteins in GenBank. Both clusters may not be coding as no large open reading frames can be identified in the consensus sequences. Both transcripts are hyper-abundant in the adult female EST dataset.

Four of the hyper abundant EST clusters BMC00185 (*rbp-1*), *BMC04376 (cdd-1*), BMC00030 (*tin-1*), BMC01688 (*col*) show similarities to proteins in the database that indicate they may serve roles as metabolic enzymes or structural proteins. *Bm-rbp-1* encodes a small RNA binding protein that has similarities to a sunbunit of the polyadenylation complex. The ESTs suggest that *Bm-rbp-1* is more abundantly expressed during the microfilarial stage of the parasites lifecyle. An ortholog of this gene has been cloned from *Brugia pahangi* where it was isolated L4 stage as abundant SL-1 *trans*-spliced transcript (Anant *et. al.*, 1997; Gregory *et. al.*, 1997). *Bm-ccd-1* is predominately expressed at the L4 stage of the parasite. It shows similarities to cytidine deaminases, enzymes involved in RNA metabolism and editing. The *B. pahangi* ortholog of *cdd-1* has been shown to be enzymatically active however no evidence for RNA editing properties was detected (Anant *et. al.*, 1997). EST abundances indicate *Bm-tin-1*(BMC00030, tropomyosin) is upregulated in the L3 stage of the parasite while BMC01688 (cuticular collagen) is upregulated during the young adult stages of the parasite. It is unclear why *tin-1* would be upregulated at the L3 stage of development. Young adults are rapidly elongating and therefore must synthesize large amount of new cuticle components. BMC01688 may encode a collagen gene that is incorporated into this lengthening cuticle.

Filarial nematodes interact with the host immune system and specifically down regulate host responses to filarial antigens (Allen and Loke, 2001). It is believed that proteins secreted by the parasite function as immunomodulators which induce this down regulation and skew the host immune response in a manner which perpetuates parasite survival (Maizels *et. al.*, 2001). The ESTs have identified a number of candidate immunomodulatory molecules. Several of these molecules are highly expressed these include *alt-2* (BMC00213), *tpx-2* (BMC00211) and *vah-1* (BMC00351) genes. *Bm-alt-2* is a hyper-abundant transcript which is restricted to the L3 stage of the lifecycle. *Bm-alt-2* and a second member of the *alt* gene family *Bm-alt-1*, have been shown to secreted by filarial nematodes after they enter the mammalian host (Gregory *et. al.*, 2000). Their functions remain elusive. However, a recently discovered *alt* homologue in the *C. elegans* genome sequence may give provide some clues (Gregory *et. al.* pers. com.). The *alts* are promising vaccine candidates and their efficacy is currently being tested in rodent and bovine infection models (Gregory *et. al.*, 2000). *Bm-tpx-2* is also abundantly expressed in the L3 stage of the parasite. Thioredoxin peroxidases (or peroxidoxins) are oxyradical detoxifying enzymes and *tpx-2* is believed to play a role in protecting the nematode from host immune responses (Ghosh *et. al.*, 1998). *Bm-vah-1* is a homologue of a family of proteins identified in hookworms that are released by invading L3 larvae (ancylostoma secreted proteins (Bin *et. al.*, 1999; Daub *et. al.*, 2000; Hawdon *et. al.*, 1996; Hawdon *et. al.*, 1999; Moyle *et. al.*, 1994)). The proteins have been shown to interact with surface receptors on immune cells and modulate their functions. While *Bm-vah-1* has been shown to be expressed at all stages of the parasite's lifecycle it appears to be upregulated in the L3 larvae (Murray *et. al.*, 2001).

Interestingly, with the exception of the ribosomal protein encoding clusters all of the other hyper-abundant ESTs appear to be unevenly distribution through the different lifecycle stages. Why would enzymes involved in detoxification processes or RNA metabolism such as *tpx-2* or *cdd-1* need to be synthesized in large amount at specific stages of the parasites development? This may indicate they play a role in a specific biology unique to that stage.

3.6.2.2 Abundant differentially expressed transcripts of B. malayi

To further explore what the *B. malayi* EST datasets can tell us about biological processes which may be specific to particular stages of the parasites development. The dataset was searched for clusters which are abundantly expressed ($\geq$ 6 ESTs) but are composed of ESTs from a single stage in the parasites lifecycle.

Patch 4

With the L3 and adult datasets additional searches were performed to isolate transcripts that were upregulated the particular portions of the dataset (i.e. infective vs moulting, young adult vs. mature adult, or male vs female). Tables 3.6.2.2.1A-J list the clusters identified as abundant differentially expressed transcripts. The tables are organized by the lifecycle stage from which the clusters originate. There were no clusters identified from the L2 and L4 dataset that fit the criteria listed above. For those searches which yielded ESTs from more than one library the number of ESTs from each library was determined. The relative expression pattern of these clusters as infered by the number of ESTs from each library is presented in figures 3.6.2.2.2-4.

A: Abundant differentially expressed transcripts found in the *B. malayi* microfilarial EST dataset

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent Total Mf Dataset	Similarities	*Gene Names*	References
BMC12282	33	0.18	1.49	microfilarial serine proteinase inhibitor serpin	*spn-2*	(Zang *et. al.*, 1999)
BMC00312	11	0.06	0.19	Novel protein with PSCL		
BMC00546	8	0.04	0.50	similar to *W.bancrofti* repeat sequence, protein has PSCL	*wrr-2*	(Siridewa *et. al.*, 1996; Siridewa *et. al.*, 1994)
BMC11791	6	0.03	0.06	cathepsin L like cysteine proteinase	*cpl-2*	

B: Abundant differentially expressed transcripts found in the *B. malayi* infective L3 datasets

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent L3 ESTs Dataset	Similarities	*Gene Names*	References
BMC04934	44	0.23	2.05	cathepsin L like cysteine proteinase	*cpl-1*	
BMC00123	33	0.18	1.54	abundant larval transcript 1	*alt-1*	(Gregory *et. al.*, 2000)
BMC04886	24	0.13	1.12	similar to aldo-keto reductase		
BMC00136	16	0.09	0.32	small abundant glycine tyrosine rich protein	*gya-1*	(Gregory *et. al.*, 1997)
BMC04832	14	0.07	0.65	serine proteinase inhibitor (serpin)	*spn-1*	(Yenbutr and Scott, 1995)
BMC11994	13	0.07	0.61	similar to malate dehydrogenase	*mld-1*	
BMC00133	12	0.06	0.56	Novel protein with PSCL		
BMC04956	7	0.04	0.33	Novel protein with PSCL		
BMC00178	6	0.03	0.28	cystatin type cysteine proteinase inhibitor	*cpi-1*	(Gregory *et. al.*, 1997)
BMC00271	6	0.03	0.28	abundant larval transcript 8	*alt-8*	

C: Abundant differentially expressed transcripts found in the *B. malayi* moulting L3 datasets

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent moulting L3 Dataset	Similarities	*Gene Names*	References
BMC12497	6	0.03	0.39	similar to cuticular collagen col-34	*col*	

D: Abundant differentially expressed transcripts found in the *B. malayi* L3 datasets

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent total L3 Dataset	Similarities	*Gene Names*	References
BMC00213	76	0.41	1.53	abundant larval transcript 2	*alt-2*	(Gregory *et. al.*, 2000)
BMC05110	23	0.12	0.46	similar to pyruvate dehydrogenase	*pyd-1*	
BMC00075	10	0.05	0.20	similar to *C. elegans* protein F53A9.10		
BMC04888	8	0.04	0.16	similar to cuticular collagen	*col*	
BMC12127	8	0.04	0.16	similar to calcium binding protein	*cab*	
BMC11971	7	0.04	0.14	similar to cuticular collagen	*col*	
BMC00068	6	0.03	0.12	Similar to novel unpredicted *C. elegans* gene on cosmid F54C9		

E: Abundant differentially expressed transcripts found in the *B. malayi* young adult dataset

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent total Young Adult Dataset	Similarities	*Gene Names*	References
BMC08961	6	0.03	0.58	similar to cuticular collagen	*col*	

F: Abundant differentially expressed transcripts found in the *B. malayi* mature adult male dataset

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent Adult Male Dataset	Similarities	*Gene Names*	References
BMC01685	32	0.17	0.75	major sperm protein 2	*msp-2*	
BMC03373	23	0.12	0.54	Novel Ser and Arg rich protein	*aam-15*	(Michalski and Weil, 1999).
BMC11914	12	0.06	0.28	Novel		
BMC03411	11	0.06	0.26	Novel Ser and Arg rich protein	*aam-10*	
BMC03274	9	0.05	0.21	Novel protein with PSCL		
BMC04232	9	0.05	0.21	Novel protein with PSCL	*aam-5*	
BMC03272	8	0.04	0.19	Novel protein with PSCL	*aam-12*	
BMC03552	7	0.04	0.16	Novel protein with PSCL	*aam-8*	(Michalski and Weil, 1999).
BMC06017	7	0.04	0.16	Novel		
BMC03393	6	0.03	0.14	similar to transcriptional repressor		
BMC03788	6	0.03	0.14	Novel protein with PSCL		
BMC11932	6	0.03	0.14	Novel		

G: Abundant differentially expressed transcripts found in the *B. malayi* mature adult female dataset

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent Adult Female Dataset	Similarities	*Gene Names*	References
BMC01764	21	0.11	0.63	Novel protein with PSCL		
BMC01695	16	0.09	0.47	microfilarial sheath protein	*shp-1*	(Selkirk *et. al.*, 1991)
BMC01967	8	0.04	0.24	microfilarial sheath protein		
BMC01750	7	0.04	0.21	similar to *C.elegans* protein Y41D4B.12		

H: Abundant differentially expressed transcripts found in the *B. malayi* mature adult datasets

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent Mature Adult Dataset	Similarities	*Gene Names*	References
BMC01618	179	0.96	2.35	Novel	*aaf-1*	
BMC01601	67	0.36	0.88	Novel	*aad-1*	
BMC12409	16	0.09	0.21	similar to major sperm protein 2	*msp*	
BMC01765	14	0.07	0.18	calreticulin-like antigen	*sxp-1*	(Chandrashekar *et. al.*, 1994)
BMC02125	17	0.09	0.22	similar to *C.elegan* protein F55C5.1, msp like	*ssp-1*	
BMC02058	10	0.05	0.13	similar to *C.elegan* protein F49F1.1	*aam-7*	
BMC01596	8	0.04	0.10	similar to protein-tyrosine phosphatase	*typ*	
BMC02040	8	0.04	0.10	similar to ubiquitin	*ubq*	
BMC02801	7	0.04	0.09	similar to calmodulin	*cal*	
BMC01441	6	0.03	0.08	Novel		
BMC02543	6	0.03	0.08	similar to beta-mannosidase	*bmd*	

I: Abundant differentially expressed transcripts found in the *B. malayi* young and mature adult datasets

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent Total Adult Dataset	Similarities	*Gene Names*	References
BMC01892	10	0.05	0.12	similar to BCL7B PROTEIN	*bcl-7*	
BMC06467	8	0.04	0.09	similar to cytochrome P450	*cyt*	
BMC02367	7	0.04	0.08	similar to nucleosome assembly protein	*nsa-1*	
BMC12436	7	0.04	0.08	cuticlular glutathione peroxidase	*gpx-1*	(Cookson *et. al.*, 1992)
BMC02423	6	0.03	0.07	Novel		
BMC02427	6	0.03	0.07	similar to calcium binding protein	*cab*	

Tables 3.6.2.2.1A-J Abundant differentially expressed transcripts discovered in the *B. malayi* EST dataset. The clusters are placed in separate tables based on which stage(s) the constituent ESTs are derived from. **A:** microfilaria, **B:** infective L3, **C:** moulting L3, **D:** all L3, **E:** young adult, **F:** adult male, **G:** adult female, **H:** mature adult (adult male and female) and **I:** young and mature adult. The cluster ID, number of ESTs in the cluster, the percent of the dataset the total and specific stage(s), similarities to proteins in the public databases, gene names if assigned and any relevant references are shown. PSCL: N-terminal secretion signal predicted by PSORTII (Nakai and Horton, 1999).

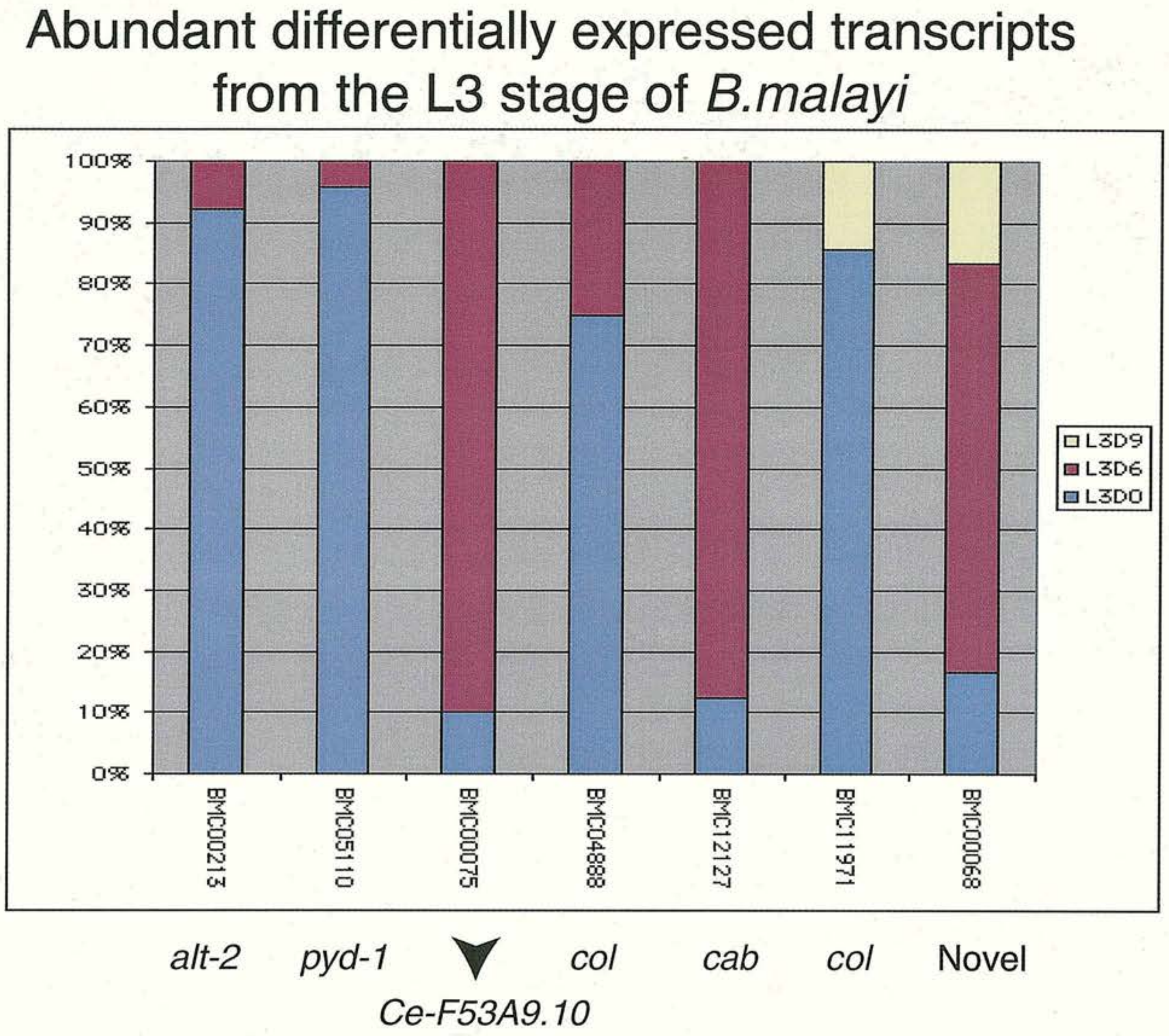


Figure 3.6.2.2 Relative expression of the differentially expressed abundant L3 transcripts. For each of the clusters the number of ESTs originating from different lifecycle stages is presented. The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. L3D0: vector derived infective L3, L3D6: molting L3 six days post infection, L3D9: molting L3 nine days post infection.

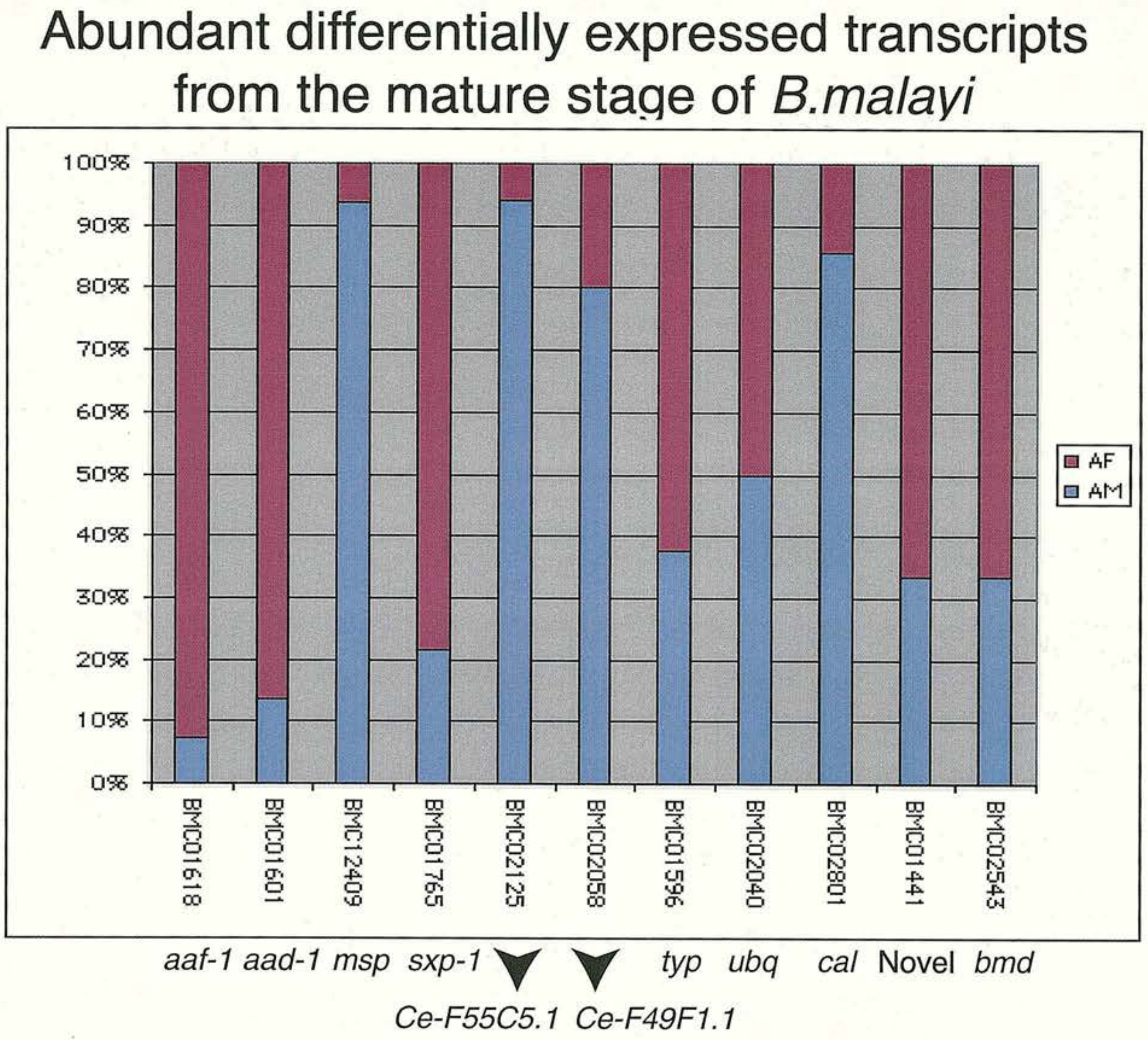
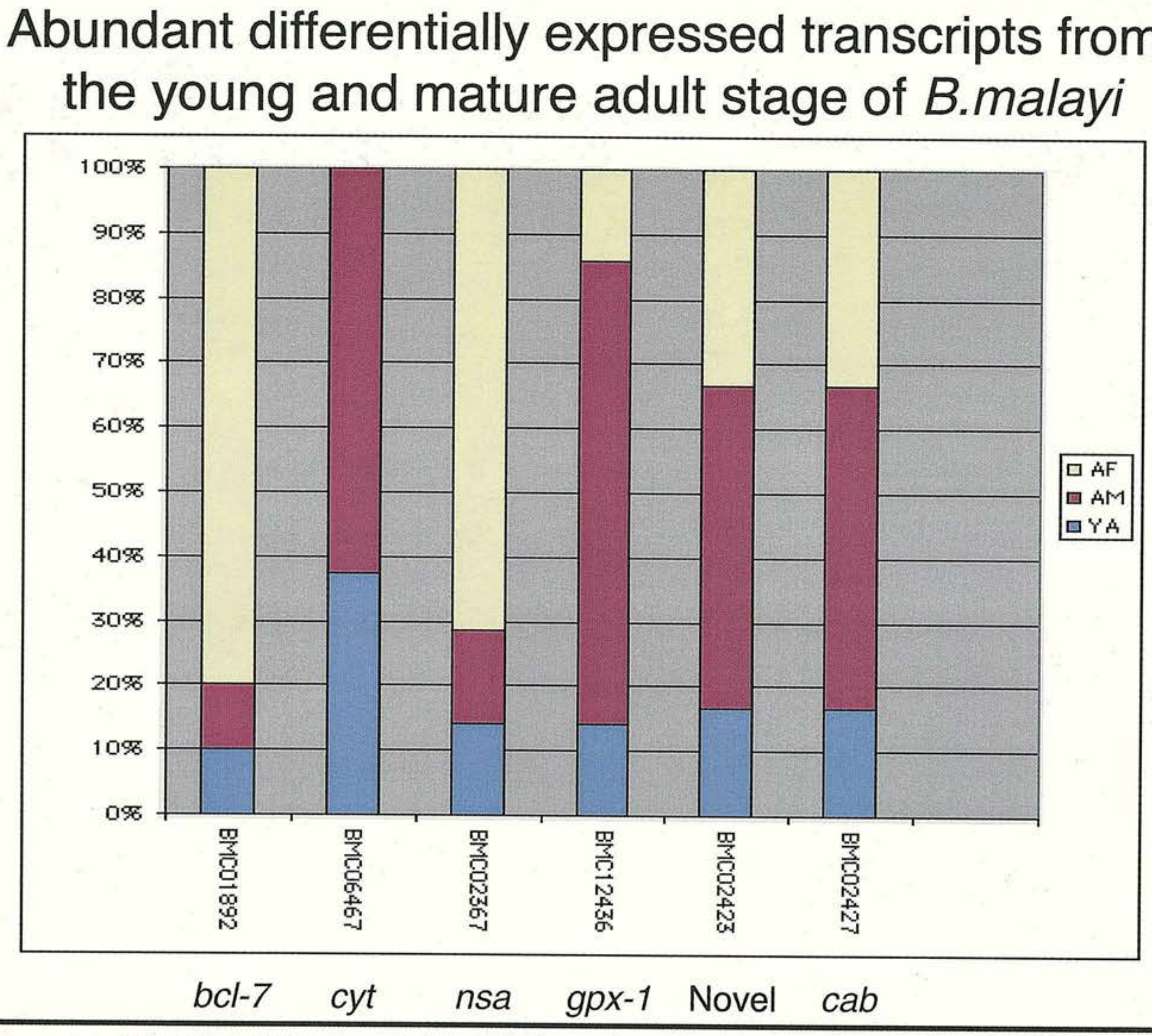


Figure 3.6.2.2.3 Relative expression of the differentially expressed abundant mature adult transcripts. For each of the clusters the number of ESTs originating from different lifecycle stages is presented. The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. AM: adult male, AF: adult female.



3.6.2.2.4 Relative expression of the differentially expressed abundant young and mature adult transcripts. For each of the clusters the number of ESTs originating from different lifecycle stages is presented. The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. YA: young adult, AM: adult male, AF: adult female.

Like the abundantly expressed transcripts the abundant differentially expressed transcripts (ADTs) can be can be divided into three categories based on their similarities to sequences in the public databases, conserved metabolic enzymes and structural proteins, potential mediators of host-parasite interactions and proteins of unknown function (novels).

Almost one half of the isolated ADT clusters have similarities to conserved metabolic enzymes or structural proteins. Their expression patterns and similarities suggest potential functions for several of these genes.

Several of the clusters differentially expressed during periods when the nematode is synthesizing or elongating its cuticle have similarities to proteins known to be involved in cuticle biology. Four clusters (BMC12497, BMC04888, BMC11971 and BMC08961) are similar to *C. elegans* cuticular collagens. The majority of the ESTs from these clusters have been isolated from the moulting L3 stages or young adult stages of the parasite's development. Biochemical studies have shown that cysteine proteinases are important for the moulting of filarial stage 3 larvae (Lustigman *et. al.*, 1996; Richer *et. al.*, 1993). BMC04934 (*cpl-1*) encodes a cathepsin L like cysteine proteinase which is differentially expressed at the L3 stage of the parasite's development. Therefore it possible that *Bm-cpl-1* may have role in moulting.

Two of the clusters differentially expressed in the adult female stage of the parasite are similar to components that are incorporated into the eggshell of developing microfilaria. One of these sheath components, *Bm-shp-1* (BMC01695), has been previously characterized (Selkirk *et. al.*, 1991).

Two clusters (BMC01685, BMC12409) which are differentially expressed in the adult male stage of the parasite are similar to the major sperm protein (*msp*) gene family. MSPs are components of a motility system unique to the amoebiod sperm of nematodes. BMC01685 (*msp-2*) orthologue from *O. volvulus* has been previously characterized while BMC12409 has not (Scott *et. al.*, 1989). Interestingly, BMC12409 contains a single EST from the adult female library (see figure 3.6.2.2.3). It is possible that this EST arose from the accidental inclusion of an adult male in the materials used to construct the adult female library. Alternatively mRNA from sperm carried in the female nematodes may have contained *msp-2* transcripts.

A number of the ADTs that have been characterized are believed to play a role in mediating host parasite interactions. Three of the ADTs have similarities to protease inhibitors (BMC12282, BMC04832, BMC00178). Two of these inhibitors are similar to the serpin (serine proteinase inhibitor) family. These serpins have been shown to be expressed at different developmental stages (Yenbutr and Scott, 1995; Zang *et. al.*, 1999). *Bm-spn-2* (BMC12282) is expressed in the microfilarial stage while *Bm-spn-1* (BMC04832) is expressed in the infective L3 and shortly after the nematode enters the mammalian host. Both proteins are secreted by the parasites and are believed to interact with serum components or proteases derived from immune cells (Yenbutr and Scott, 1995; Zang *et. al.*, 1999). BMC00178 (*cpi-1*) shows similarity to the cystatin (cysteine protease inhibitor family). *Bm-cpi-1* and a second cystatin *Bm-cpi-2* (BMC01649) are upregulated in the infective L3 larvae (Gregory *et. al.*, 1997). Both proteins are secreted by the larvae and are believed to have potential roles in inhibiting host enzymes or proteases involved in the L3 moulting process (Gregory *et. al.* pers. com., Manoury *et. al.*, 2001). Unlike, *Bm-cpi-2* which is expressed and secreted throughout the period the nematode spends in the mammalian host, *Bm-cpi-1* expression is restricted to the L3 stage of the parasite's development.

The *alts*, whose expression and secretion from the L3 stage of the parasite were mentioned previously, make up a second group of ADTs that may function as potential mediators of host parasite interactions. *Bm-alt-1* and *2* (BMC00123 and BMC00213) have been described as potential vaccine candidates (Gregory *et. al.*, 2000). However, additional alt -like genes can be found in the EST dataset indicating *alts* may represent a large gene family in filaria. One of these alts *Bm-alt-8* is abundantly expressed in the infective L3 stage of parasite. Whether this expression is restricted to the L3 stage of development like *Bm-alts-1* and *2* is still to be determined.

Examination of the relative expression of those ADTs that were isolated from searches that encompassed more than one stage has shown that there are fold differences in the transcript abundance between stages. For instance BMC000213 (*alt-2*), BMC00075 and BMC12127 show an almost 10 fold difference in the number of ESTs found in the infective L3 vs moulting L3 datasets. RT-PCR data supports these findings and have shown that *alt-2* expression is higher in infective vs moulting

L3 (Gregory *et. al.*, 2000). Similarly in the adult dataset the *Bm-gpx-1* cluster (secreted glutathione peroxidase , BMC12436) has at least 60% of its ESTs originating from the adult male library (see figure 3.6.2.2.4). Published studies indicate that *Bm-gpx*-1 is much more abundant on the surface of adult males and the difference in transcript abundance between the stages is consistent with this observation (Cookson *et. al.*, 1992).

Remarkably, more than one third of the clusters isolated in the search for ADTs have no similarities to proteins in the public databases. Almost half of the protein predicted from these clusters are predicted to have potential N-terminal secretion signals. Some stages have higher numbers of these novel and potentially secreted proteins (microfilarial, infective L3, adult male). The adult male dataset is particularly rich in these sequences with 80% of the ADTs being novel. Half these sequences are predicted to be secreted. A subsequent survey has shown that at least two of these proteins (BMC03373 and BMC03552) are specific to the male stage of development (Michalski and Weil, 1999).

Chapter 4

Analysis of the *O. volvulus* EST dataset and comparison of the filarial EST datasets

4.0 Analysis of the O. volvulus EST dataset with CLOBBv1.0 and comparison of the filarial EST datasets

This chapter builds on the work presented in chapter 3 by presenting the results of the clustering of the *O. volvulus* EST dataset and comparing the abundant and abundant differentially expressed transcripts found in the two datasets.

4.1.0 Cluster analysis of O. volvulus EST datasets and evaluation of gene discovery effort

In total 8,876 *O. volvulus* ESTs sequences were analyzed. These sequences originated from seven different cDNA libraries (see figure 1.2.3.4.2 and table 1.2.3.4.4). After filtering for rDNA and *E. coli* sequences 7,909 sequences were grouped into 3,504 clusters. Table 4.1.0.1 shows the results of the cluster analysis for each cDNA library and the total dataset.

Library	Total Number of ESTs	Number of rDNA seq.	Percent dataset	Number of *E.coli* seq.	Percent dataset	Number of Clustered ESTs	Number of Clusters	Intra-library Redund	Total Redund	Number of Library Specific Clusters	Percent of Library Specific Clusters	Curators Comments
OvMf	184	26	14	4	2	154	127	1.2	1.4	64	50.4	Some *E. coli* contamination and very rDNA contaminated
OvL2	120	12	10	3	3	105	60	1.8	2.0	27	45.0	Some *E. coli* contaminated
OvL3	2747	151	5	75	3	2521	1175	2.1	2.3	836	71.1	Some *E. coli* contamination
OvmL3	3413	279	8	24	1	3110	1598	1.9	2.1	1173	73.4	
OvAM	88	7	8	4	5	77	71	1.1	1.2	44	62.0	Very *E. coli* contaminated
OvAF	2195	344	16	36	2	1815	1117	1.6	2.0	818	73.2	Very rDNA contaminated
OvAFIV	129	3	2	1	1	125	94	1.3	1.4	41	43.6	
Total	8876	822	9	147	2	7909	3504		2.5			

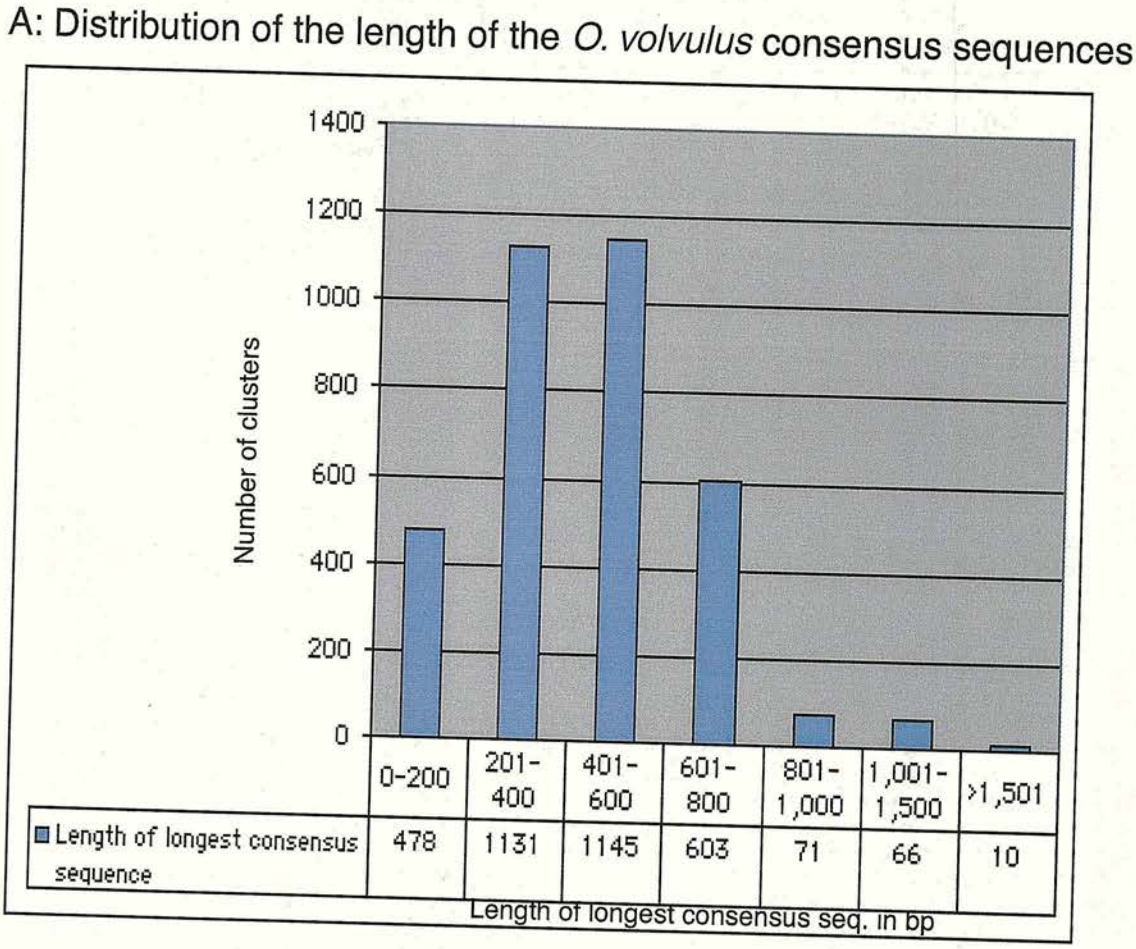
Table 4.1.0.1 The table summarizes the results of CLOBBv1.0 clustering of the *O. volvulus* EST dataset. The number of ESTs sequenced from each library is shown as well as the number of rDNA and *E.coli* sequences removed from the datasets before clustering. The number of clusters containing sequences from each library is listed with the level intra-library sequence redundancy (calculations performed without the filtered rDNA and *E.coli* sequences). The redundancy of the library as calculated by comparison with the complete EST dataset is also shown. The number of clusters containing sequences exclusively from the library and percentage these clusters represent of the library dataset is shown. Curators comments on the general characteristics of the library rDNA /*E.coli* contamination and the incidence of chimeric ESTs is also shown The numbers in red highlight calculations that indicate that the library no longer represents a productive sequencing substrate, either because of high rates of rDNA/ *E. coli* contamination or high levels of redundancy within the dataset.

CLOBBv1.0 grouped the 7,909 *O. volvulus* ESTs into 3,504 clusters. Like the *B. malayi* clusters the majority of the *O. volvulus* clusters (75%) contain one EST sequence. Like the *B. malayi* dataset several of the cDNA libraries have high rate of rDNA contamination (OvMf, OvL2 and OvAF). Nine percent of the EST sequences are tagged as rDNA contamination. In general the *O. volvulus* libraries have higher incidences of *E. coli* contamination than the *B. malayi* libraries. Two percent of the total dataset has been tagged as *E. coli* contamination.

The overall redundancy of the cDNA EST dataset is 2.5 which is much higher than the *B malayi* dataset when the relative numbers of ESTs analyzed is compared (18,740 vs 7,909). All of the libraries that have been heavily sampled (>2,000 ESTs) have redundancies ≥ 2 indicating the relative level of transcript diversity in these libraries is not as high as the *B. malayi* datasets.

4.1.1 Characteristics of O. volvulus EST clusters

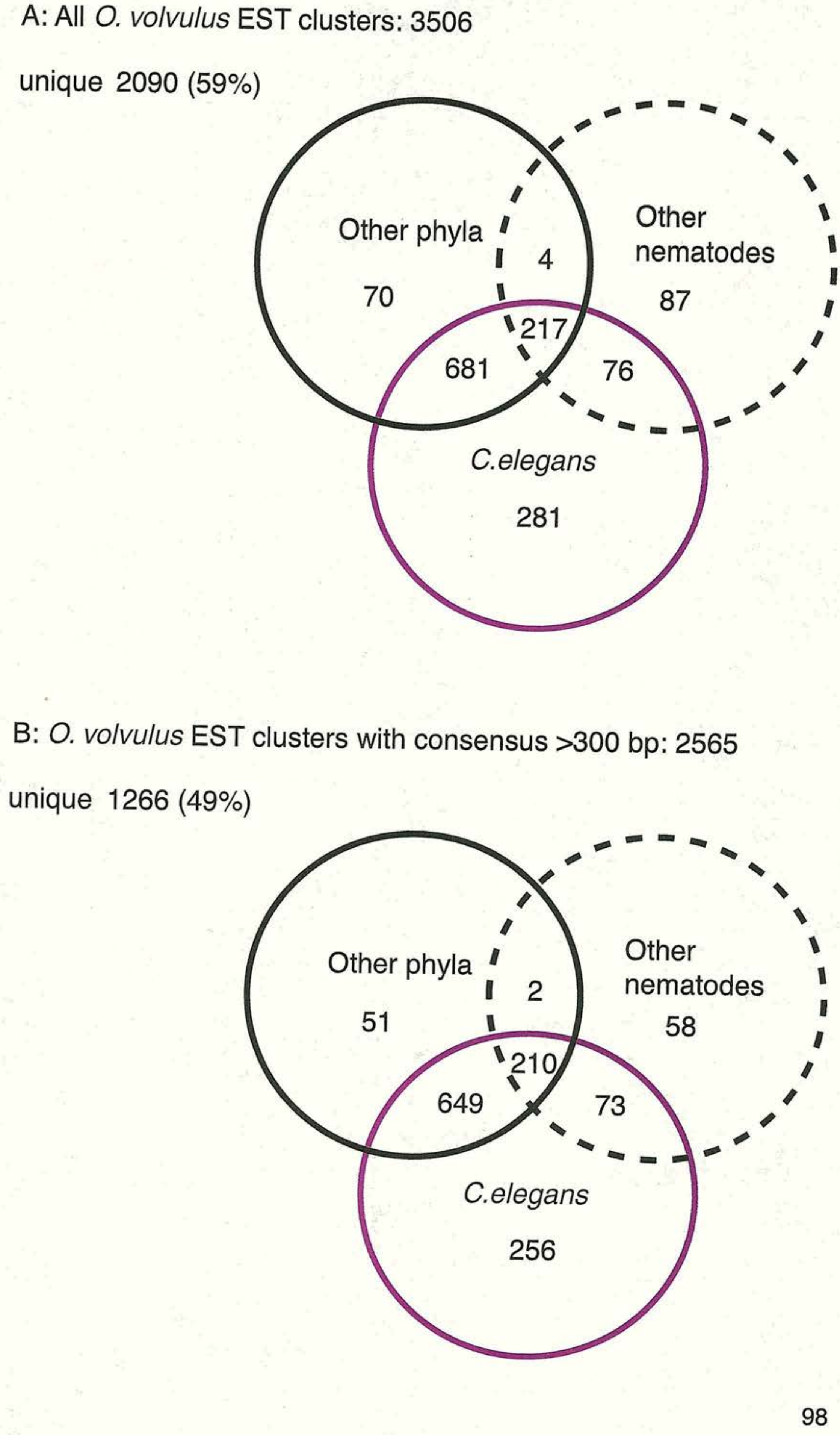
The *O. volvulus* clusters have an average of 2.2 EST per cluster (excluding the contribution of the rDNA and *E. coli* contaminants to the dataset redundancy). Figure 3.3.1shows the distribution of the number of ESTs per cluster. The average length of an*O. volvulus* EST is 417 bp in length and the average length of a consensus sequence is 578 bp. The length of the consensus sequences are significantly longer than the average EST length (p < 0.0001 Man-Whitney test, performed in Minitab, Minitab Inc). Figure 4.1.2.1A shows the distribution of the lengths of the *O. volvulus* consensus sequences. Like the *B. malayi* dataset majority of the consensus sequences are between 200 and 600 bp in length (64%). The longest predicted consensus sequence is 2,156bp in length.



A: Distribution of the length of the *O. volvulus* consensus sequences

Figure 4.1.1.1 Characteristics of the consensus sequences derived from the *O. volvulus* EST clusters. The graphs show the relative distribution of the lengths of the longest consensus sequence assembled from each cluster (A). The x-axis shows the lengths of the sequences while the y-axis shows the number of clusters found with those lengths. The number of clusters belonging to each group is also shown in the table below the x-axis.

The results of BLAST comparisons of the *O. volvulus* consensus sequences to the public databases are summarized in figures 4.1.1.2 and 4.1.1.3. The results of both the whole dataset and several subdivisions of the whole dataset are shown in Venn diagrams. Remarkably, the results of the *O. volvulus* searches were extremely similar to the results of the *B. malayi* searches. More than 60% of the sequences do not have homologues in the public databases (see figure 4.1.1.2A). Like *B. malayi,* if the dataset is subdivided and those only sequences with >300bp consensus sequences or clusters with >1 EST are examined the level of novels drops significantly (49% and 33% novels respectively see figures 4.1.1.2B and C). The *O. volvulus* dataset also has similar proportions of sequences which have matches to the proteins predicted from the three fully sequenced animal genomes (see figure 4.1.1.3A). Thirty-two percent of the consensus sequences were tagged as having significant similarities when compared to EST datasets of Clades I, IV and V nematodes (see figure 4.1.1.3B). When compared to the *B. malayi* EST dataset 40% of the sequences were found to have significant matches (see figure 4.1.1.3C). Interestingly when compared to the ascarid, *B. malayi and* other nematode EST datasets only 49% of the sequences did not have significant matches which is a lower proportion than the *B. malayi* dataset. Like the *B. malayi* dataset if the clusters with >1 EST are examined there is a dramatic increases the proportion of novel sequences (49% vs 22%, see figure 4.1.1.3D).



A: All *O. volvulus* EST clusters: 3506

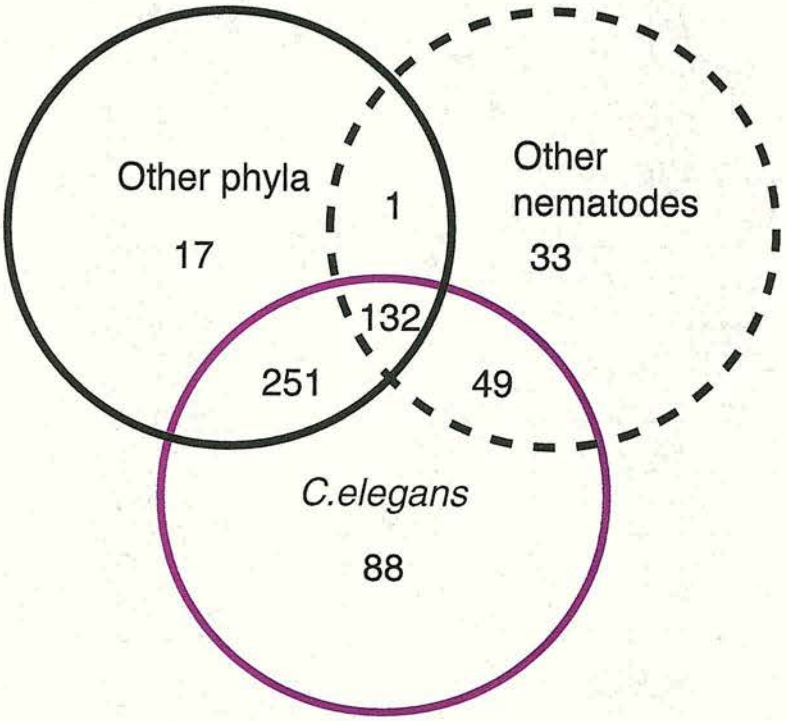
unique 2090 (59%)

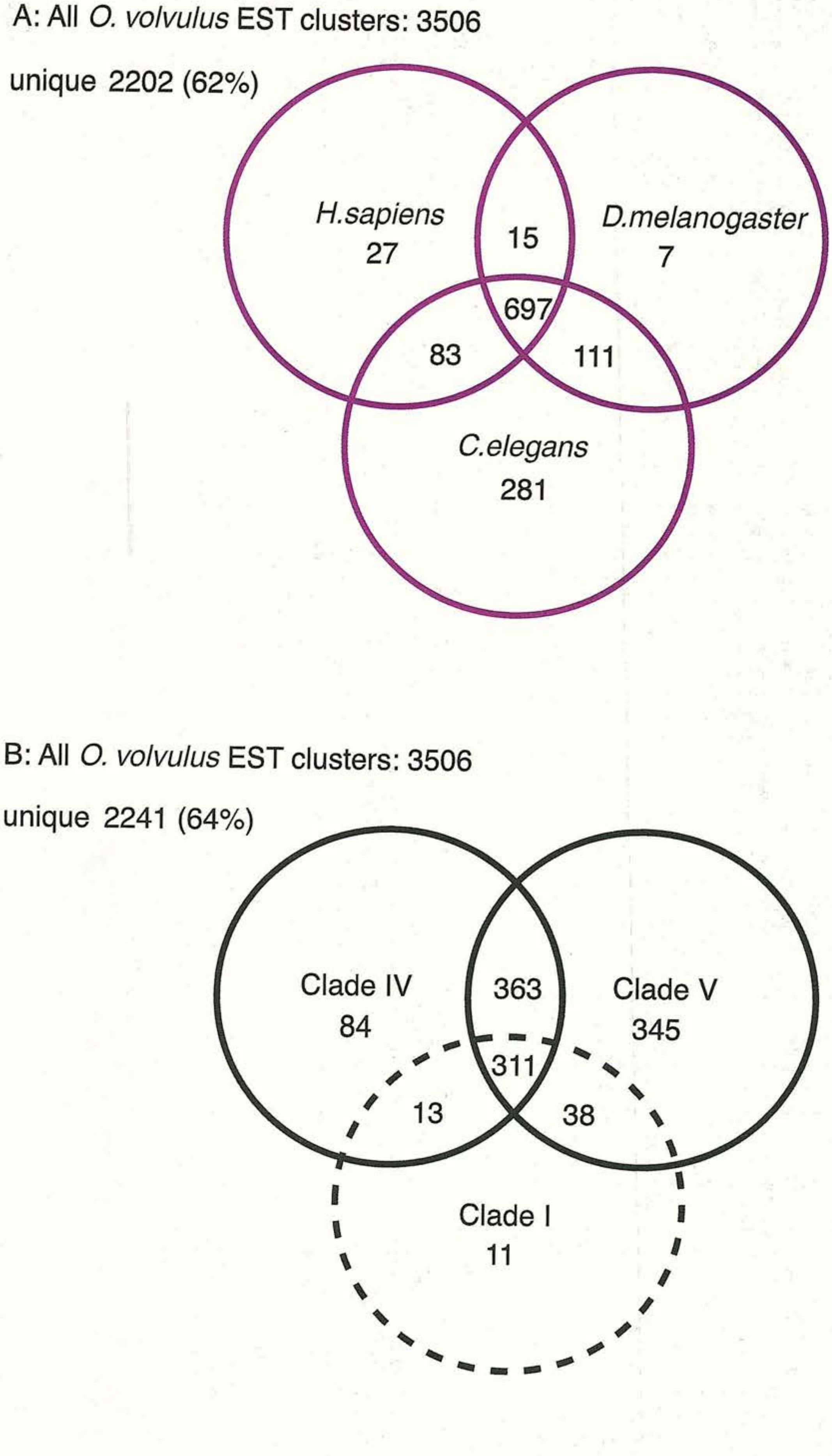
B: *O. volvulus* EST clusters with consensus >300 bp: 2565

unique 1266 (49%)

C:*O. volvulus* EST clusters with > 1 EST: 858

unique 287 (33%)



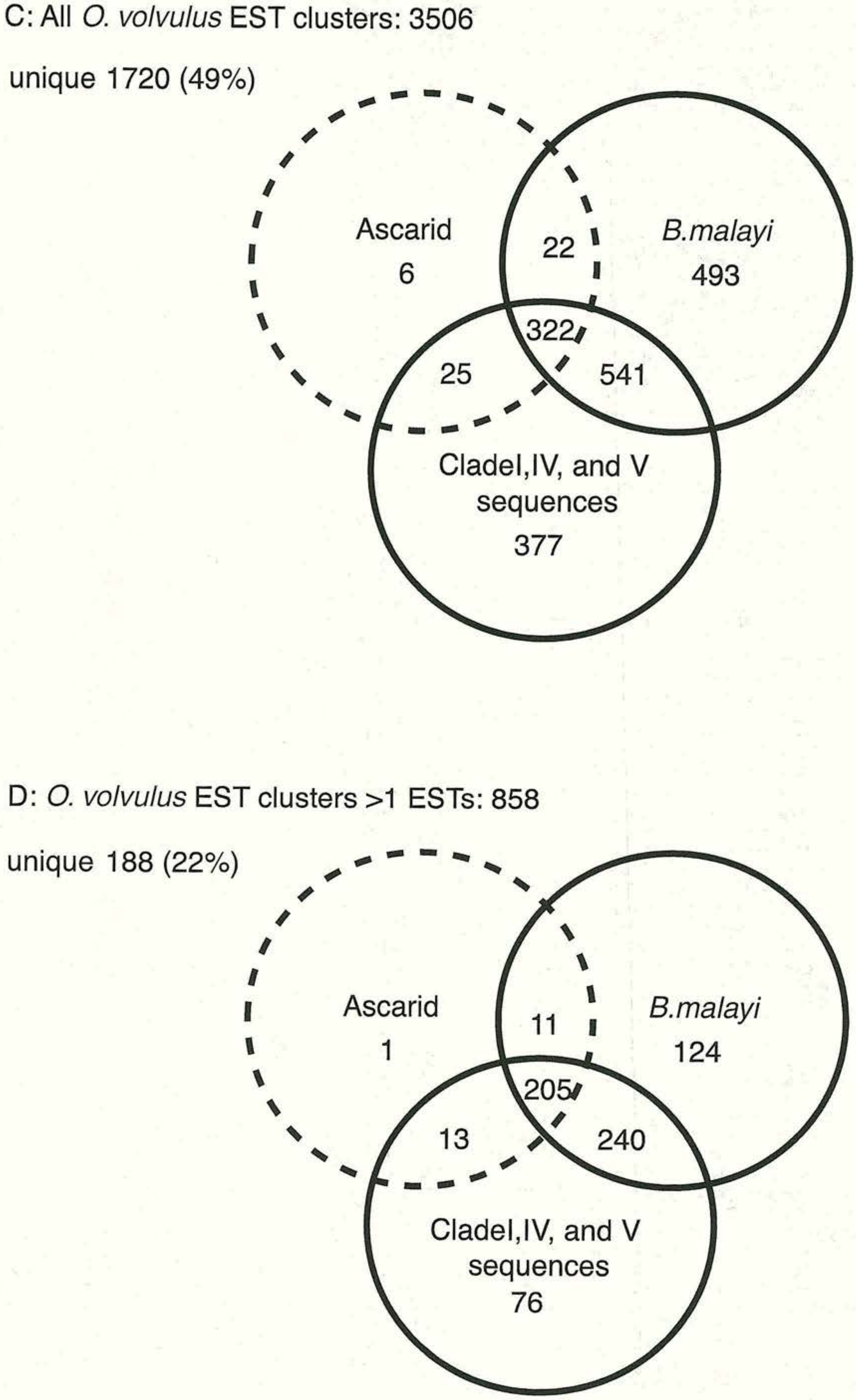


A: All *O. volvulus* EST clusters: 3506

unique 2202 (62%)

B: All *O. volvulus* EST clusters: 3506

unique 2241 (64%)



C: All *O. volvulus* EST clusters: 3506

unique 1720 (49%)

D: *O. volvulus* EST clusters >1 ESTs: 858

unique 188 (22%)

Figures 4.1.1.2 and **4.1.1.3**. Results of the BLAST searches of the *O. volvulus* consensus sequences against the public databases. Each Venn diagram shows a three-way comparison of the results of the BLAST searches against three databases. The number of consensus sequences having significant matches to a single database are shown in the center of each circle. The number of consensus sequences showing a significant match (p-value $\leq e^{-10}$) to more than one database are listed at the interfaces of each circle. The number of consensus sequences with no significant matches to any of the listed databases is shown to the left of the Venn diagram. Figure 4.1.1.2A shows the results of the BLASTX searches against the xrest, xnem, xce databases. The summary of the whole dataset is shown in figure 4.1.1.2A. The results of three subdivsions of the whole dataset , those clusters with consensus sequences with lengths >300 bp. 4.1.1.2B and those clusters with >1 EST 4.1.1.2C are also shown. Figure 4.1.1.3A shows the results of the BLASTX searches of the consensus sequences against the protein sequences predicted from the *H. sapiens, D. melanogaster* and *C. elegans* whole genome sequence. Figure 4.1.1.3.B shows the results of the TBLASTX searches of the consensus sequences against the EST sequences of nematodes from Clades I, IV and V. Figures 4.1.1.3C and 24.1.1.3D show the results of the TBLASTX of the consensus sequences against the EST sequences of Ascarid, *B. malayi* and Clade I, IV and V nematodes. Purple circles indicate that the dataset is based on proteins predicted from whole genome sequence. Dashed ciricles indicate that the sequence dataset is substantially smaller than the other comparators.

4.1.2 Analysis of abundant and differentially expressed transcripts of O. volvulus

To examine the similarities and differences between the abundant and differentially expressed genes in *B. malayi* and *O. volvulus* datasets the searches performed on the *B. malayi* clusters were also performed on the *O. volvulus* clusters.

4.1.2.1 Abundant transcripts of O.volvulus

To isolate the most abundant transcripts the *O. volvulus* cluster dataset was searched and those clusters with > 40 ESTs examined. Table 4.1.2.1.1 lists the thirteen clusters identified in this search. Figure 4.1.2.1.2 shows the relative expression pattern of the ten non-mitochondrial EST clusters.

Cluster	Number of ESTs	Percent of Total EST Dataset	Similarities	Gene Name	Reference
OVC00048	297	3.76	abundant larval transcript 1/2	*alt-1/2*	(Joseph *et. al.*, 1998)
OVC00032	168	2.12	Novel peptide with PSCL		
OVC00039	143	1.81	Ribosomal protein	*rpl-12*	
OVC00018	99	1.25	thioredoxin peroxidase	*tpx-2*	(Lu *et. al.*, 1998)
OVC00128	93	1.18	RNA binding protein (polyadenylation complex subunit)	*rbp-1*	
OVC00036	85	1.07	similar to cuticular collagen	*col*	
OVC00060	72	0.91	similar to small heat shock protein 25	*hsp-25*	
OVC00025	99	1.25	small abundant glycine and tyrosine rich protein	*gya-1*	
OVC00142	67	0.85	*O. volvulus*cystatin	*cpi-2*	(Lustigman *et. al.*, 1992)
OVC00021	64	0.81	Mitochondrial gene	*cox-2*	(Keddie *et. al.*, 1998)
OVC00121	43	0.54	Mitochondrial gene	*ndh-1*	(Keddie *et. al.*, 1998)
OVC00265	42	0.53	Novel		
OVC00460	41	0.52	Mitochondrial gene	*atp-6*	(Keddie *et. al.*, 1998)

Table 4.1.2.1.1. Abundantly expressed transcripts discovered in the *O. volvulus* EST sequencing project. The table lists the thirteen identified *O. volvulus* EST clusters with >40 ESTs. The cluster ID, number of ESTs in the cluster, the percent of the dataset the cluster comprises, similarities to proteins in the public databases, gene names if assigned and any relevant references are shown. PSCL: N-terminal secretion signal predicted by PSORTII (Nakai and Horton, 1999).

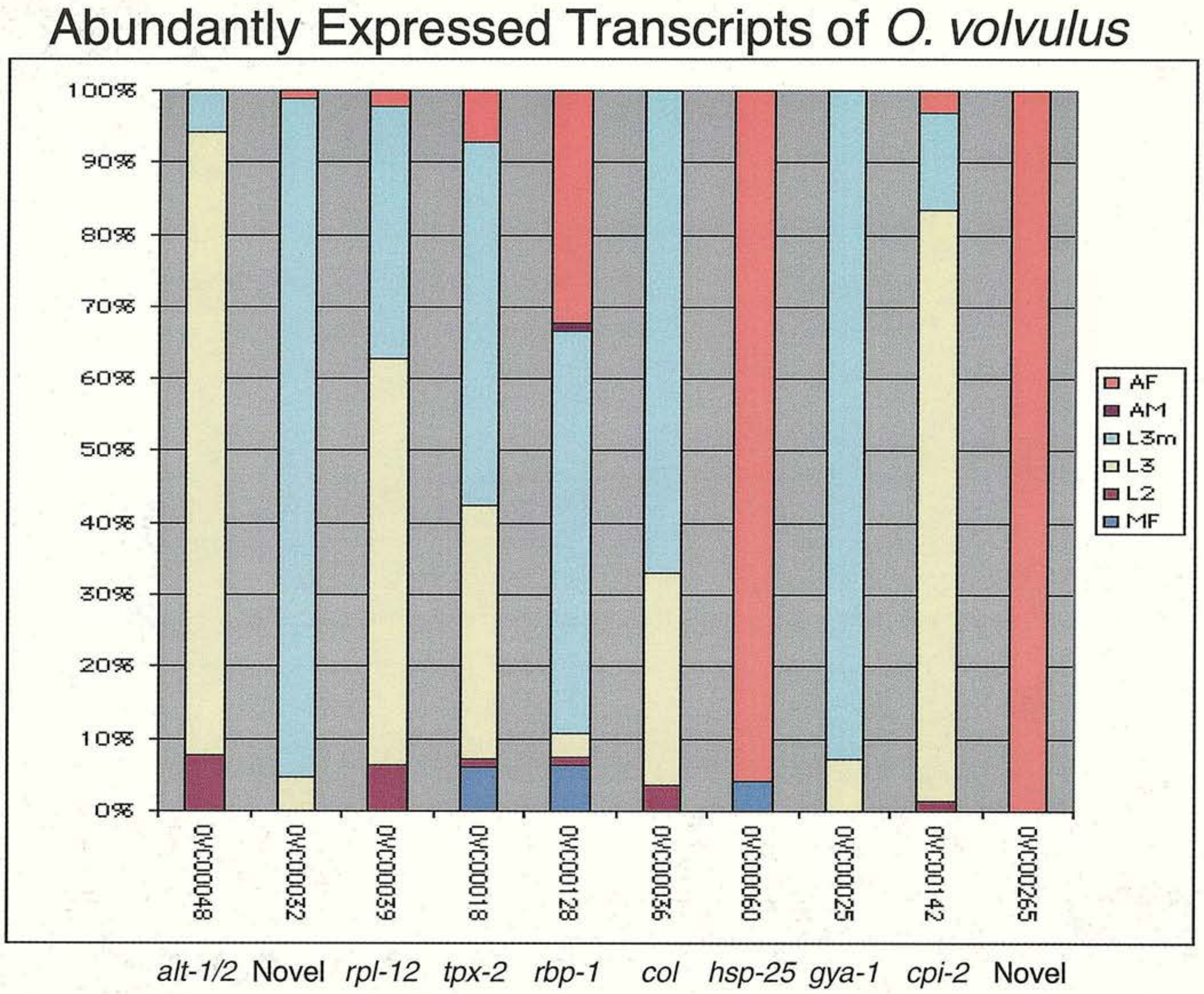


Figure 4.1.2.1.2. The relative expression pattern of the abundantly expressed non-mitochondrial *O. volvulus* EST clusters presented in table 4.6.2.1.1. For each of the clusters the number of ESTs originating from different lifecycle stages is presented . The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. MF: microfilaria, L2: vector derived stage 2 larvae, L3: day twelve vector derived infective stage 3 larvae, L3m: molting stage 3 larvae days 1,2, and 3, AM: adult male, AF: adult female.

Like the *B. malayi* abundant transcripts, four of the *O. volvulus* abundant transcripts encode ribosomal proteins or mitochondrial genes. The rest can be assigned to the three functional catagories described above. *B. malayi* orthologs of several of these abundant transcripts have already been described. These include *Ov-tpx-2*, *rbp-1* and *cpi-2*. Presumably, their functions are very similar in both organisms. The most abundant transcript in the *O. volvulus* dataset is similar to the *B. malayi alt* genes (Joseph *et. al.*, 1998). Phylogenetic analysis has not been able to assign clear orthology between this gene and any of the *B. malayi alts* (Gregory *et. al.* 2002 pers. com.). However there are several ESTs abundant in the *O. volvulus* that were not present in the list of abundant transcripts from *B. malayi*. These include *hsp-25* homologue, a homologue of the *Bm-gya-1* (small abundant glycine and tyrosine rich potein) which was found to be differentially expressed in L3 (Gregory *et. al.*, 1997), a cuticular collagen and several novel proteins. One of the novel proteins, OVC00032, may be secreted. Like the *B. malayi* dataset all of the non-rp and mt hyper-abundant clusters appear to have uneven expression through the lifecycle. Most are derived from the L3 or molting L3 datasets (see figure 2.7.2.1.2). This may be because these are the two most heavily sampled time points.

4.1.2.2 Abundant differentially expressed transcripts of O. volvulus

The *O. volvulus* dataset was searched for clusters that are abundantly expressed ($\geq$ 6 ESTs) but are composed of ESTs from a single stage in the lifecycle. With the L3 dataset an additional search was preformed to isolate transcripts that were upregulated in infective and molting larvae. Tables 4.1.2.2.A-D list the clusters identified as abundant differentially expressed transcripts. The tables are organized by the lifecycle stage from which the clusters originate. There were no clusters identified from the Mf, L2 and AM datasets that fit the criteria listed above. Presumably this is because such a small number of EST sequences from those libraries were clustered. For the L3 search the relative expression pattern of these clusters as inferred by the number of ESTs from each library is presented in figure 4.1.2.2.2.

A: Abundant differentially expressed transcripts in the *O. volvulus* infective L3 datasets

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent L3DO ESTs Dataset	Similarities	Gene Name	Reference
OVC03893	18	0.23	0.71	vespid venom-allergen-like (activation associated secreted protein)	*vah-1*	(Tawe *et. al.*, 2000)
OVC00092	17	0.21	0.67	Small Novel Gly rich peptide with PSCL		
OVC00324	15	0.19	0.60	Novel		
OVC04021	8	0.10	0.14	Novel peptide with PSCL		
OVC00704	7	0.09	0.28	LIM domain containing protein OvL3-1	*lim-1*	(Oberlander *et. al.*, 1995)
OVC00129	6	0.08	0.24	Novel peptide with PSCL		
OVC00409	6	0.08	0.24	similar to calponin		
OVC00472	6	0.08	0.24	Novel peptide with PSCL		

B: Abundant differentially expressed transcripts in the *O. volvulus* molting L3 datasets

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent L3D1-3 ESTs Dataset	Similarities	Gene Name	Reference
OVC03901	31	0.39	1.00	similar to cuticulin	*cut-1*	
OVC00156	18	0.23	0.58	Novel peptide with PSCL		
OVC00441	11	0.14	0.35	similar to cuticulin	*cut-2*	
OVC00093	10	0.12	0.32	similar to cuticular collagen	*col*	
OVC00130	9	0.11	0.29	Novel peptide with PSCL		
OVC00510	8	0.10	0.26	similar to osteonectin	*ost-1*	
OVC00115	6	0.08	0.19	similar to Y102A11A.5 novel immunogenic protein		(Lizotte-Waniewski *et. al.*, 2000)
OVC00181	6	0.08	0.19	Novel		
OVC00322	6	0.08	0.19	similar to cystathionine gamma-lyase		
OVC00711	6	0.08	0.19	Novel Gln rich peptide with PSCL		
OVC00753	6	0.08	0.19	similar to cuticle collagen	*col*	

C: Abundant differentially expressed transcripts in the *O. volvulus* L3 datasets

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent total L3 ESTs Dataset	Similarities	Gene Name	Reference
OVC00025	99	1.25	1.76	small abundant glycine and tyrosine rich protein	*gya-1*	
OVC03892	28	0.35	0.50	vespid venom-allergen-like (activation associated secreted protein)	*asp-1*	(Tawe *et. al.*, 2000)
OVC00109	25	0.32	0.44	similar to abundant larval transcript	*alt*	
OVC00762	24	0.30	0.43	similar to cuticle collagen	*col*	
OVC01409	22	0.28	0.39	similar to *O. volvulus* abundant larval transcript 1	*alt*	
OVC03876	19	0.24	0.34	Ov-16 antigen	*peb-1*	(Erttmann and Gallin, 1996)
OVC00657	18	0.23	0.32	cathepsin L like cysteine proteinase	*cpl-1*	
OVC03971	14	0.18	0.25	extracellular superoxide dismutase	*sod-1*	(Henkle *et. al.*, 1991)
OVC04023	14	0.18	0.25	similar to troponin	*tin*	
OVC00238	12	0.15	0.21	similar to cuticular collagen	*col*	
OVC00295	11	0.14	0.20	Novel		
OVC00340	11	0.14	0.20	similar to histone H3	*his-3*	
OVC00413	11	0.14	0.20	similar to *C. elegans* protein R07E5.13		
OVC04050	11	0.14	0.20	similar to thioredoxin	*thi-1*	
OVC00005	10	0.12	0.18	similar to MIP family protein	*mip-1*	
OVC00164	9	0.11	0.16	similar to glutathione reductase	*gtr-1*	
OVC00667	9	0.11	0.16	similar to LIM domain protein	*lim*	
OVC00634	8	0.10	0.14	similar to nematode myoglobin	*glb*	
OVC00364	7	0.09	0.12	similar to inx-9 GTP-binding protein	*inx-9*	
OVC00434	7	0.09	0.12	similar to transketolase	*tkt*	
OVC00465	7	0.09	0.12	similar to translationally controlled tumor protein	*tph-1*	
OVC00758	7	0.09	0.12	similar to RAS-related protein	*rab-1b*	
OVC00707	6	0.08	0.11	Novel peptide with PSCL		
OVC00712	6	0.08	0.11	Novel		
OVC00811	6	0.08	0.11	similar to *C. elegans* protein T27C4.1		
OVC00951	6	0.08	0.11	similar to monooxygenase	*mox*	
OVC01239	6	0.08	0.11	Novel Pro,Tyr, Phe, Gly rich peptide with PSCL		
OVC03916	6	0.08	0.11	proteosome regulatory subunit S12	*rpn-8*	
OVC03990	6	0.08	0.11	similar to col-34 cuticular	*col*	

		collagen		

D: Abundant differentially expressed transcripts in the *O. volvulus* mature adult female dataset

Cluster	Number of ESTs	Percent of Total EST Dataset	Percent Adult Female ESTs Dataset	Similarities	Gene Name	Reference
OVC00265	42	0.53	2.16	Novel		
OVC00113	15	0.19	0.77	Novel		
OVC00041	14	0.18	0.72	microfilarial sheath protein	*shp-1*	
OVC03912	11	0.14	0.57	similar to protein Y5F2A.1 transthyretin-like protein	*tsy-1*	
OVC04013	11	0.14	0.57	similar to small heat shock protein 25	*hsp*	
OVC01323	10	0.12	0.52	Novel peptide with PSCL		
OVC01335	10	0.12	0.52	Novel peptide with PSCL		
OVC01355	9	0.11	0.46	Novel		
OVC02395	8	0.10	0.41	Novel peptide with PSCL		
OVC00150	7	0.09	0.36	Novel peptide with PSCL		
OVC01404	6	0.08	0.31	Novel peptide with PSCL		
OVC01561	6	0.08	0.31	Novel peptide with PSCL		

Tables 4.1.2.2.1A-D Abundant differentially expressed transcripts discovered in the *O. volvulus* EST dataset. The clusters are placed in separate tables based on which stage(s) the constituent ESTs are derived from. A: infective L3, B: molting L3, C: all L3, D: adult female. The cluster ID, number of ESTs in the cluster, the percent of the dataset the total and specific stage(s), similarities to proteins in the public databases, gene names if assigned and any relevant references are shown. PSCL: N-terminal secretion signal predicted by PSORTII (Nakai and Horton, 1999).

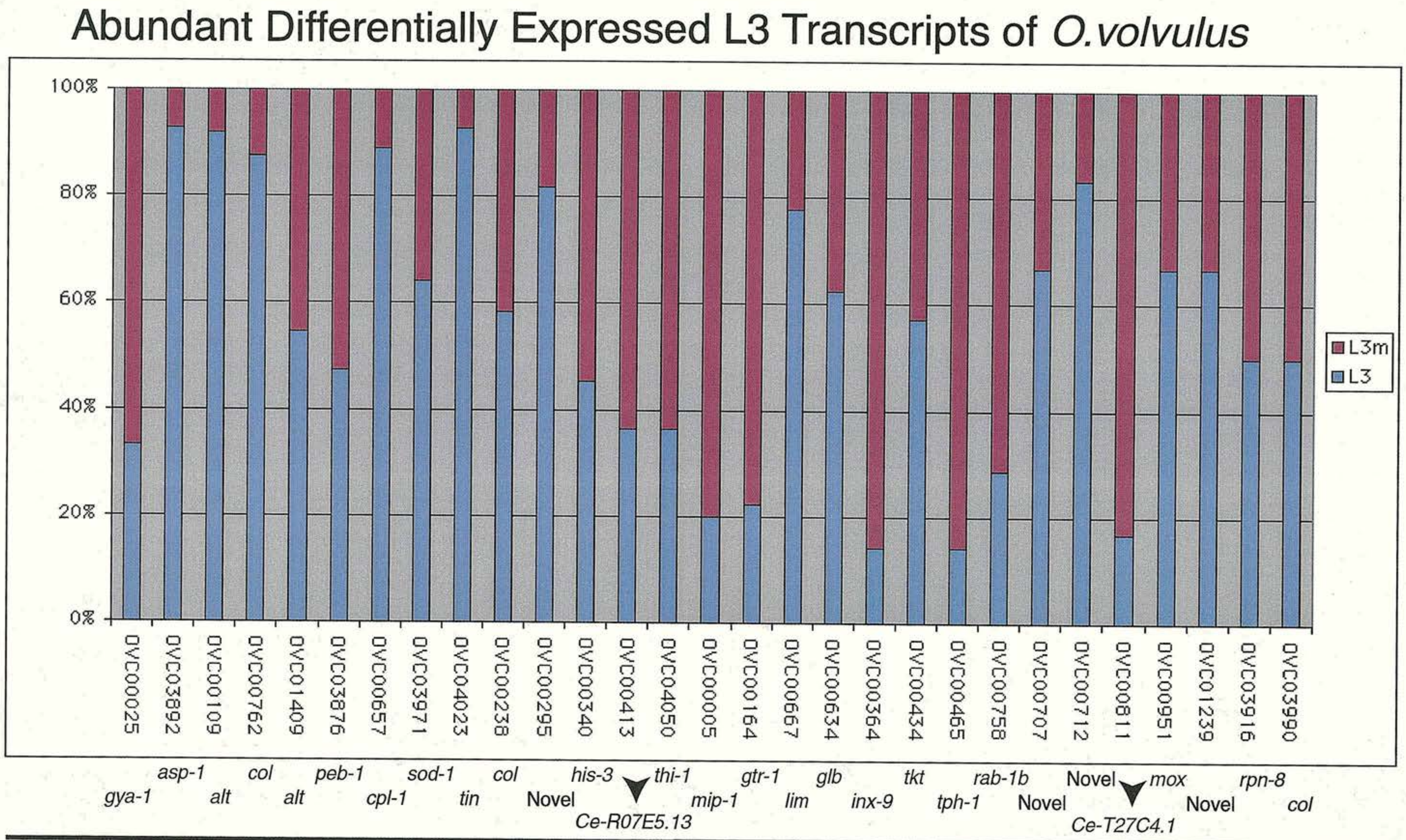


Figure 4.1.2.2.2 Relative expression of the differentially expressed abundant L3 transcripts. For each of the clusters the number of ESTs originating from different lifecycle stages is presented. The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. L3: vector derived infective L3, L3m: molting L3.

The ADTs isolated from the *O. volvulus* and *B. malayi* infective and molting L3 dataset share many common features. Both contain members of gene families that are believed to be important mediators of host parasite interaction in *B. malayi* (*vahs* and *alts)* as well as homologues of genes whose functions are currently unknown *(gyas)*. There are also large numbers of highly expressed novel genes. In the infective L3 many of the novel ADTs have potential secretion signals. The *O. volvulus* L3 are also expressing large number of proteins predicted to be involved in cuticle biology. Five different collagens (OVC00093, OVC00753, OVC00762, OVC00238, OVC03990), two cuticulins (*cuts*, OVC03901, OVC00441) and one osteonectin (*ost-1*, OVC00510) genes were present in the dataset. Interestingly, *B. malayi* homologues of the *cuts* and the *ost-1* genes have not been isolated from the molting L3 dataset. Assuming the relative biology of molting is similar between these species this indicates that while both sets of libraries were derived from molting worms the differences in the timing of the isolation of the nematodes has produced two very different datasets. One major difference between the datasets is that almost four times the number of ADTs that have been isolated from the *O. volvulus* molting L3 and the infective/molting L3 datasets. Part of this difference is due to the fact that there are almost twice as many molting L3 ESTs in have been sequenced in *O. volvulus* then *B. malayi*. In *O. volvulus* many of the ADTs isolated from the infective/molting L3 dataset have similarities to proteins which would not be expected to be differentially expressed (*his-3, thi-1*, *tph-1*, *rpn-8*, *sod-1*). The high level of ADTs these stages is probably due to the fact that other stages from the *O. volvulus* lifecycle have not been sampled heavily (see tables 3.6.0.1 and 4.1.0.1).

Like the adult females from *B. malayi* one of the ADTs from the *O. volvulus* dataset is similar to the microfilarial sheath protein gene family. However, one major difference between the microfilaria of *O. volvulus* and *B. malayi* is that the former do not retain their egg shells and thus are not sheathed when they leave the adult female (Selkirk *et. al.*, 1991). It is unclear whether this *shp* homologue will be retained and perhaps incorporated in the microfilarial cuticle or if it is simply shed with the eggshell. Interestingly, unlike the set of proteins isolated from *B. malayi* most of the ADTs found in the adult female ESTs are novel sequences many of which are predicted to have potential secretion signals.

4.2 Comparative analysis of *B. malayi* and *O. volvulus* EST cluster datasets.

B. malayi and *O. volvulus* are believed to be relatively closely related nematodes. However, there are many differences between their lifecycles and survival strategies. The analysis of the EST has revealed that only 40-45% of the identified genes are common to both datasets. The EST datasets from *B. malayi* and *O. volvulus* individually contain a wealth of information about the expression of genes in the nematodes at the sampled lifecycle stages. However, by comparing the two datasets a new layer of information can be added. This comparative 'transcriptomics' will identify genes that are common to both species as well as highlight those which are differentially expressed between the species. This strategy has been successfully used to identify genes involved in differences in pathogenicity between closely related species or strains of bacteria and is currently being applied to the study of closely related parasitic protozoa.

4.2.1 Comparative analysis of the abundant transcripts

The abundant transcripts (>40 ESTs) isolated from both datasets were compared. Table 2.8.1.1 lists the fourteen identified non-rp or mt clusters. Figure 2.8.1.2 shows the relative expression pattern of the EST clusters in both species.

B. malayi Cluster	Number of ESTs	Percent Total Dataset	*O. volvulus* Cluster	Number of ESTs	Percent Total Dataset	similarities	Gene Name
NO	-		OVC00048	297	3.76	abundant larval transcript1/ 2	*alt-1/2*
BMC06846	3	0.01	OVC00032	168	2.12	Novel peptide with SECL	
BMC00211	67	0.36	OVC00018	99	1.25	thioredoxin peroxidase	*tpx-2*
BMC00185	174	0.93	OVC00128	93	1.18	RNA binding protein (polyadenylation complex subunit)	*rbp-1*
BMC12467	7	0.03	OVC00036	85	1.07	similar to cuticular collagen	*col*
BMC00498	20	0.1	OVC00060	72	0.91	similar to small heat shock protein 25	*hsp-25*
BMC00136	16	0.09	OVC00025	99	1.25	small abundant glycine and tyrosine rich protein	*gya-1*
BMC01649	20	0.1	OVC00142	67	0.85	*O. volvulus* cystatin	*cpi-2*
None	-	-	OVC00265	42	0.53	Novel	
BMC04376	83	0.44	None	-	-	cytidine deaminase	*cdd-1*
BMC00213	76	0.41	NO	-	-	abundant larval transcript 2	*alt-2*
BMC00351	51	0.27	OVC03893	18	0.23	vespid venom-allergen-like (activation associated secreted protein)	*vah-1*
BMC01688	47	0.25	OVC00363	8	0.1	similar to cuticular collagen	*col*
BMC00030	42	0.22	OVC00071	40	0.5	similar to tropomyosin	*tin-1*

Table 4.2.1.1 Comparison of the abundantly expressed transcripts discovered in the *B. malayi* and *O. volvulus* EST sequencing projects. The table lists the fourteen identified non-rp and mt clusters. For each cluster the closest homologue from the reciprocal dataset was identified.The cluster ID, number of ESTs in the cluster, the percent of the dataset the cluster comprises, similarities to proteins in the public databases, gene names if assigned and any relevant references are shown. The three genes which comprise equivalent proportions of both datasets are shown in blue text. NO: Unable to infer orthology in gene family; PSCL: N-terminal secretion signal predicted by PSORTII (Nakai and Horton, 1999).

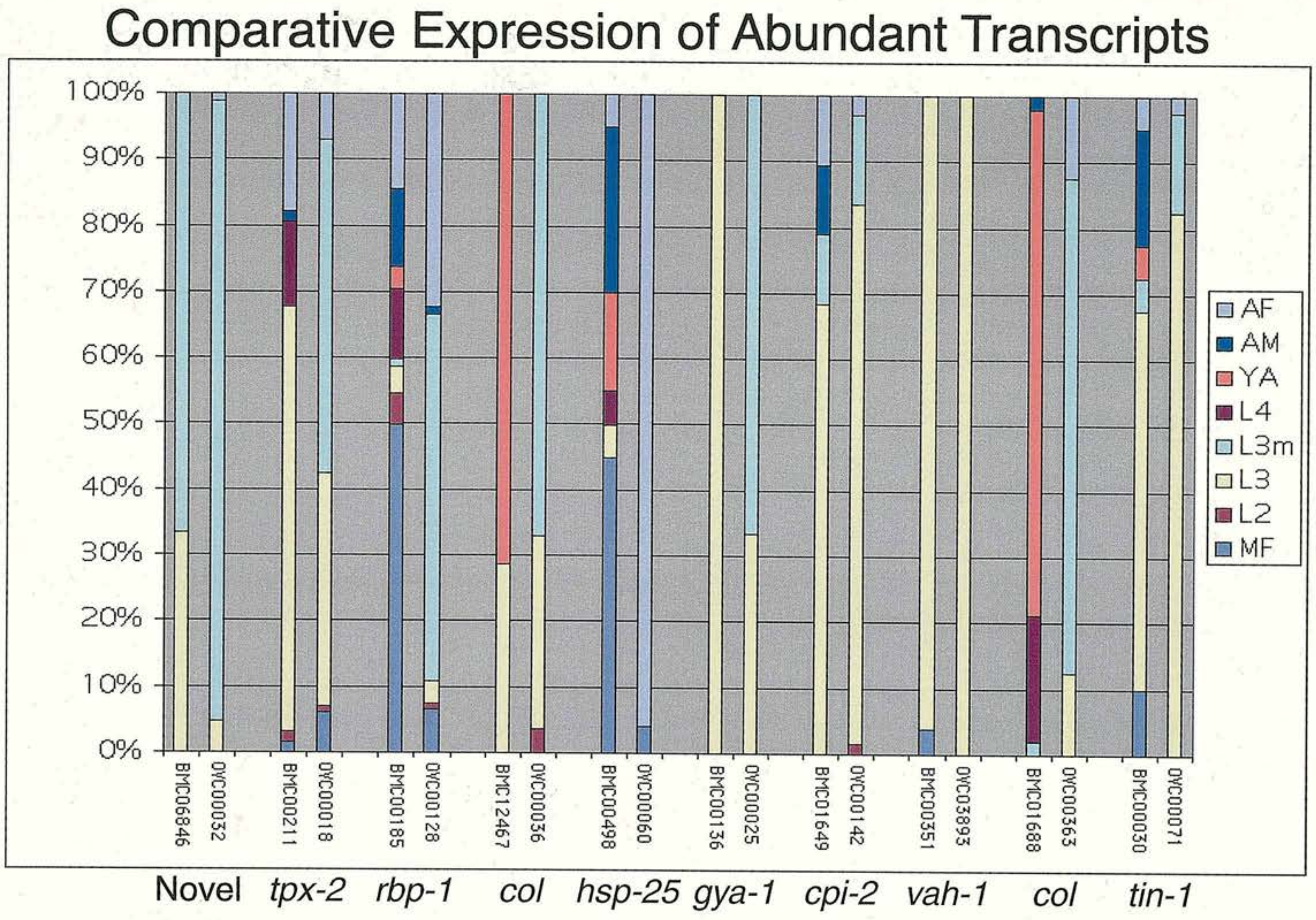


Figure 4.2.1.2 Comparison of the relative expression pattern of the abundantly expressed non-rp and mt *B. malayi* and *O. volvulus* EST clusters. For each of the clusters the number of ESTs originating from different lifecycle stages is presented . The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. MF: microfilaria, L2: vector derived stage 2 larvae, L3: day twelve vector derived infective stage 3 larvae, mL3: molting stage 3 larvae, L4: stage 4 larvae, YA: young adult, AM: adult male, AF: adult female.

Of the fourteen abundantly expressed genes examined only three were found to be upregulated in both datasets (*tpx-2*, *rbp-1* and *tin-1*). All of the other genes were found to be hyper-abundant in only one species. Homologues could not be identified for two of the genes in both dataset (BMC04376, cdd-1 and OVC00265, novel). The expression of *cdd-1* in *Brugia* is believed to be restricted to the L4 stage of development (Anant *et. al.*, 1997). There currently are no *O. volvulus* ESTs originating from this stage so if the expression of *Ov-cdd-1* is similarly restricted it would not be present in the current dataset. When the relative expression patterns of the abundantly expressed transcripts are examined only those transcripts which are hyper-abundant in the L3 stage of the parasites development have similar expression patterns.

4.2.2 Comparative analysis of abundant differentially expressed transcripts in the L3 and adult female stages

The abundant differentially expressed transcripts (≥ 6 ESTs) isolated from both datasets were compared. Because only the L3 (combined infective and molting) and adult female stages have been heavily sampled in *O. volvulus* these were the only groups compared. Tables 4.2.2.1A and B lists clusters compared in this analysis. Figures 4.2.2.2 and 4.2.2.3 shows the relative expression pattern of the EST clusters in both species. Because of the large numbers of clusters isolated from the L3 datasets only those clusters with ≥ 10 ESTs were included in tables 4.2.2.1A and figure 4.2.2.2.

A: Comparison of abundant differentially expressed L3 transcripts from *B. malayi* and *O. volvulus*

B. malayi Cluster	Number of ESTs	Percent Total L3 Dataset	*O. volvulus* Cluster	Number of ESTs	Percent L3 Total Dataset	Similarities	Gene Name
BMC00351	51	1.03	OVC03893	18	0.32	vespid venom-allergen-like (activation associated secreted protein)	*vah-1*
None	-	-	OVC00092	17	0.30	Small Gly rich peptide with SECL	
None	-	-	OVC00324	15	0.27	Novel	
None	-	-	OVC03901	31	0.55	similar to cuticulin	*cut-1*
None	-	-	OVC00156	18	0.32	Novel peptide with SECL	
None	-	-	OVC00441	11	0.20	similar to cuticulin	*cut-2*
BMC01962	29	0.58	OVC00093	10	0.18	similar to cuticular collagen	*col*
BMC00136	16	0.32	OVC00025	99	1.76	small abundant glycine and tyrosine rich protein	*gya-1*
None	-	-	OVC03892	28	0.50	vespid venom-allergen-like (activation associated secreted protein)	*asp-1*
NO	-	-	OVC00109	25	0.44	similar to abundant larval transcript	*alt*
BMC02934	25	0.50	OVC00762	24	0.43	similar to cuticle collagen	*col*
NO	-	-	OVC01409	22	0.39	similar abundant larval transcript	*alt*
BMC00143	5	0.10	OVC03876	19	0.34	Ov-16 antigen	*peb-1*
BMC04934	44	0.89	OVC00657	18	0.32	cathepsin L like cysteine proteinase	*cpl-1*
BMC00677	9	0.18	OVC03971	14	0.25	extracellular superoxide dismutase	*sod-1*
BMC00135	22	0.44	OVC04023	14	0.25	similar to troponin	
BMC00160	6	0.12	OVC00238	12	0.21	similar to cuticular collagen	
None	-	-	OVC00295	11	0.20	Novel	
BMC01609	8	0.16	OVC00340	11	0.20	similar to histone H3	*his-3*
BMC03794	2	0.04	OVC00413	11	0.20	similar to protein R07E5.13	
BMC00153	30	0.61	OVC04050	11	0.20	similar to thioredoxin	*thi-1*
BMC02613	2	0.04	OVC00005	10	0.18	similar to MIP family protein	*mip-1*
BMC00123	33	0.67	NO	-	-	abundant larval transcript 1	*alt-1*
BMC04886	24	0.48	None	-	-	similar to aldo-keto reductase	
BMC04832	14	0.28	OVC00784	3	0.05	serine proteinase inhibitor (serpin)	*spn-1*
BMC11994	13	0.26	OVC00366	3	0.05	similar to malate dehydrogenase	*mld-1*
BMC00133	12	0.24	None	-	-	Novel peptide with SECL	
BMC00213	76	1.53	NO	-	-	abundant larval transcript 2	*alt-2*
BMC05110	23	0.41	None	-	-	similar to pyruvate dehydrogenase	

BMC00075	10	0.20	OVC00332	17	0.30	similar to protein F53A9.10	

B: Comparison of abundant differentially expressed adult female transcripts from *B. malayi* and *O. volvulus*

B. malayi Cluster	Number of ESTs	Percent Adult Female Dataset	*O. volvulus* Cluster	Number of ESTs	Percent Adult Female Dataset	Similarities	Gene Name
None	-	-	OVC00265	42	2.16	Novel	
BMC01965	16	0.48	OVC00113	15	0.77	Novel	*slt-1*
BMC01695	16	0.48	OVC00041	14	0.72	microfilarial sheath protein	*shp-1*
BMC03479	33	0.99	OVC03912	11	0.57	similar to protein Y5F2A.1 transthyretin-like protein	*tsy-1*
BMC00498	20	0.60	OVC04013	11	0.57	similar to small heat shock protein 25	
None	-	-	OVC01323	10	0.52	Novel peptide with SECL	
None	-	-	OVC01335	10	0.52	Novel peptide with SECL	
None	-	-	OVC01355	9	0.46	Novel	
None	-	-	OVC02395	8	0.41	Novel peptide with SECL	
None	-	-	OVC00150	7	0.36	Novel peptide with SECL	
BMC00408	12	0.36	OVC01404	6	0.31	Novel peptide with SECL	
BMC02980	3	0.09	OVC01561	6	0.31	Novel peptide with SECL	
BMC01764	21	0.63	None	-	-	Novel peptide with SECL	
BMC01967	8	0.24	None	-	-	microfilarial sheath protein	
BMC01750	7	0.21	None	-	-	similar to protein Y41D4B.12	

Table 4.2.2.1A and B Comparison of the abundant differentially expressed L3 and adult female transcripts. The clusters are placed in separate tables based on which stage(s) the constituent ESTs are derived from. A:L3 and B: adult female. The cluster ID, number of ESTs in the cluster, the percent of the dataset the total and specific stage(s), similarities to proteins in the public databases, gene names if assigned and any relevant references are shown. Genes whose relative expression patterns are similar between both species are shown in red text. NO: Unable to infer orthology in gene family; PSCL: N-terminal secretion signal predicted by PSORTII (Nakai and Horton, 1999).

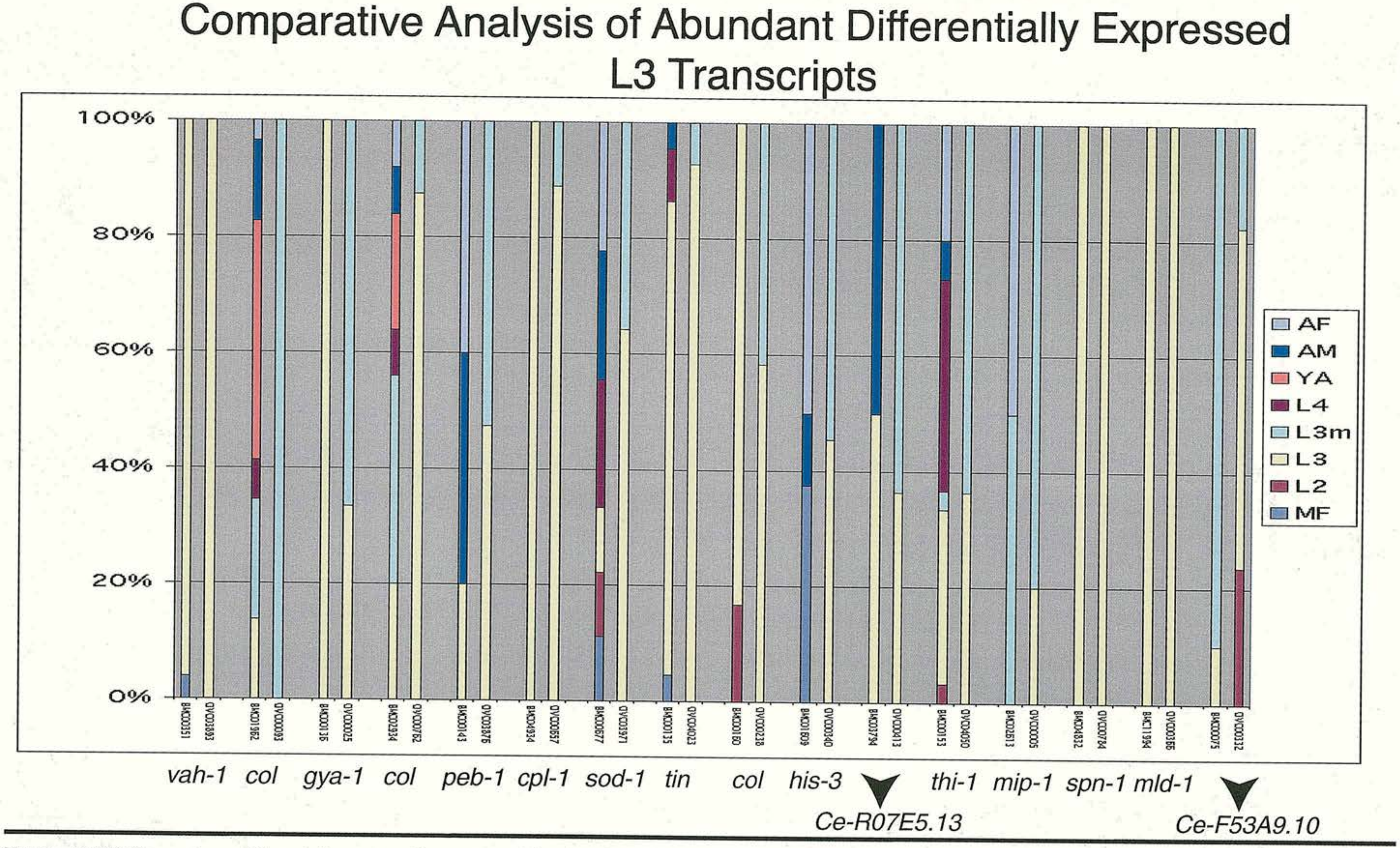


Figure 4.2.2.2 Comparison of the relative expression pattern of the abundant differentially L3 transcripts. For each of the clusters the number of ESTs originating from different lifecycle stages is presented. The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. MF: microfilaria, L2: vector derived stage 2 larvae, L3: day twelve vector derived infective stage 3 larvae, mL3: molting stage 3 larvae, L4: stage 4 larvae, YA: young adult, AM: adult male, AF: adult female.

Comparative Analysis of Abundant Differentially Expressed Adult Female Transcripts

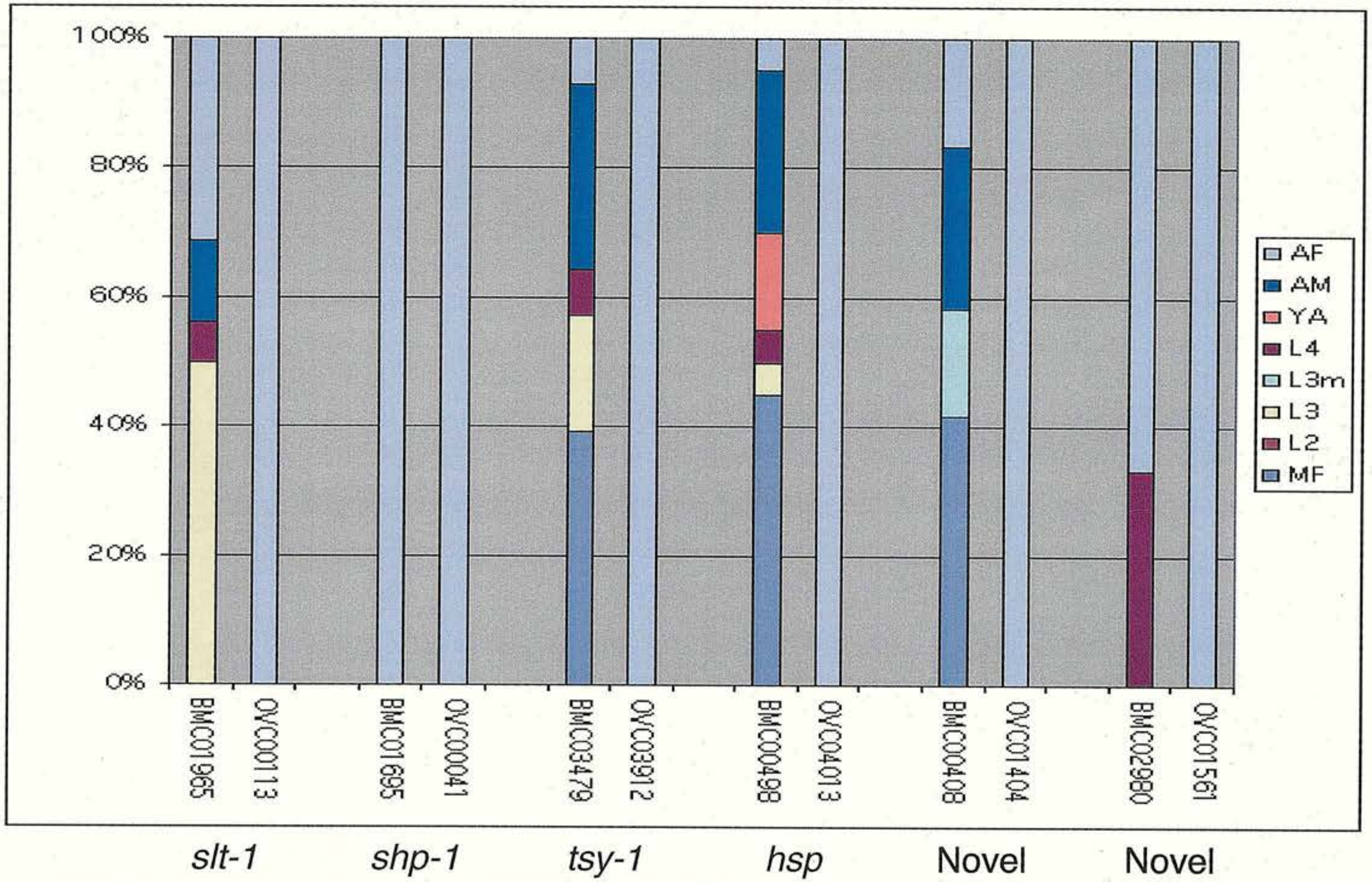


Figure 4.2.2.3 Comparison of the relative expression pattern of the abundant differentially adult female transcripts. For each of the clusters the number of ESTs originating from different lifecycle stages is presented . The relative proportion of the ESTs per stage is plotted as a percentage of the cluster total. MF: microfilaria, L2: vector derived stage 2 larvae, L3: day twelve vector derived infective stage 3 larvae, L3m: molting stage 3 larvae, L4: stage 4 larvae, YA: young adult, AM: adult male, AF: adult female.

Analysis of the relative expression of the ADTs of both species revealed many surprising differences between the two datasets. Homologues in both organisms could not be identified for over one third of the genes in the L3 dataset and over half of the genes in the adult female dataset. In both datasets many of these species specific clusters were novel sequences (L3 50% and adult female 80%). In both datasets only about 30% of the genes made up similar percentages of the total datasets. When the relative expression patterns of the genes were compared most of the clusters which were consistent between species originated from the L3 dataset. Most of these have been mentioned previously (*vah-1*, *gya-1*, *spn-1* and *cpl-1*). However a few other genes also showed similar expression patterns in the two species. These included troponin, collagen, malate dehydrogenase homologues and a homologue of the anonymous *C. elegans* protein F53A9.10. Despite the fact that it would be unexpected to find the expression of an enzyme like malate dehydrogenase restricted to one stage the fact that the ESTs from two species show similar expression profiling lends additional strength to the expression patterns observed in the single species. Additionally, the observation that a novel gene (BMC02980/OVC01561; see figure 4.2.2.3) is upregulated in the adult female stage of both species implies it may serve a function linked to a biological process conserved between both species.

4.3 Chapters 3 and 4 General Discussion

No clustering process will ever provide a complete solution to the problem of generating non-redundant gene sets from complex and often problematic datasets like ESTs. Very strict algorithms will produce robust clusters. However, because the quality of EST sequences is extremely variable, some sequences which should be grouped with other sequences may remain separated. This artificially inflates the number of genes predicted to be contained within the dataset and increases the number of clusters that have to be handled in post-clustering events. Less strict algorithms will produce larger clusters. However, biologically relevant sequence features such as closely related gene families and alternatively spliced transcripts may be lost in the clustering process. Chimeric sequences and other library based artifacts may also cause problems for these algorithms. The CLOBB algorithm is relatively strict and examples of the inappropriate exclusion of ESTs from clusters have been observed. We are currently assessing the extent of this problem and implementing an additional CLOBB function that will tag clusters of related sequences that may need to be merged. The output of the CLOBB process is easily curated so when these problems are identified they can be rectified. However, when compared to the other publicly available EST clustering algorithms CLOBB provides a portable and easily adapted solution that will cope with most sequence datasets.

The clustering process has proved extremely useful in assessing the productivity of the filarial gene discovery effort. Problems with sequencing substrates such as high levels of rDNA/*E. coli* contamination or highly redundant libraries have been observed and measures taken rectify those difficulties. Subtraction protocols that eliminate these contaminates or abundant sequences from the datasets have proved an efficient way of boosting the rate of gene discovery. The current clustering analysis indicates the redundancy in the EST dataset is due to the presence of a small number of highly abundant transcripts. A new set of subtracted or normalized cDNA libraries will need to be constructed if the EST sequencing is to continue to be efficient.

The general characteristics of the genes discovered in the EST datasets has provided some surprises. Over 70% of the ESTs sequenced have been placed in single EST clusters. This high level of singletons could partially be due to the strict overlap rules written into the CLOBBv1.0 algorithm. However, this high rate of singleton clusters implies that the majority of the transcripts found in the filarial cDNAs libraries are derived from a diverse set of genes with relatively low levels of expression.

Interestingly, more than 60% of the clusters from both *B. malayi* and *O. volvulus* do not have significant matches to any other sequence in the public databases. Examination of subdivisions of the total datasets suggests that sequence artifacts (small sequences, low fidelity sequence and sequences with low coding potential) may be partially responsible for the high rate of novels. However, the rates of novels in these subdivisions is still unexpectedly high (30-50%). Comparisons reveal that only 30% of the clusters have homologues in other animal genomes. Between 8-9% of these sequences appear to be nematode specific gene families whose functions could be tested in the model nematode *C. elegans*. Between 40-45% of the clusters had homologues in both species of filaria. Interestingly, when the expression patterns of the abundant or differentially expressed clusters were compared between the two species the majority showed very different profiles. Despite the fact the both organisms are closely related and must share many common biological process it is probable that the differences in their parasitic lifestyles means that orthologous genes may play very different functions in the two organisms. Many of the identified ADTs are novel genes that have only been identified in a single species. These genes represent some of the most interesting sequences in the dataset because they may be species specific and fulfill some biological process unique to one of the filaria.

Assuming filarial nematodes possess a similar complement of genes as the model nematode *C. elegans* the ESTs clusters indicate that the discovery effort has identified almost 30% of the genes present in the *B. malayi* genome. The ESTs have revealed hundreds of new genes. The similarities these genes possess to other sequences in the public databases along with the available expression data has allowed putative functions to be assigned to some of these genes. For instance the molting dataset contains large numbers of protein families known to be structural components of the *C. elegans* cuticle (collagens and cuticulins etc.) as well as enzymes or enzyme inhibitors that could be involved in the cuticle remodeling process (*cpl-1* and the *cpis*). Other datasets such as the adult male or female datasets contain large numbers of genes that are similar to anonymous genes predicted from the *C. elegans* genome sequence. Recently published *C. elegans* microarray data along with their expression profiles in the filarial datasets supports them having potential roles in adult reproductive functions ((Michalski and Weil, 1999), Kamal *et. al.* 2002 *in press*). These proteins provide a wealth of potential new targets for nematicides that will either interrupt reproduction or kill adult worms. The forthcoming availability of RNAi data for all *C. elegans* genes (Fraser *et. al.*, 2000; Maeda *et. al.*, 2001) provides the filarial community with an easy *in silico* screening process which can be used to create a short list of targetable candidates.

One of the main goals of the gene discovery effort is to provide the next generation of vaccine targets. The ESTs have provided an excellent source of new targets and many studies are currently underway which include proteins that have been isolated from these datasets. The analyses that have been performed in these studies have identified large numbers of transcripts that can serve as additional candidates. Isolating prophylactic vaccinogens from the L3 stage(s) is particularly important and a number of studies have been undertaken to isolate candidates from this stage. The EST datasets have revealed that in both species there are sets of proteins which are highly expressed and in some cases specific to this stage. While mRNA abundance is not always correlated with protein levels additional studies of several abundant transcripts have supported the pattern seen in the EST datasets (Gregory *et. al.*, 2000; Lustigman *et. al.*, 1992). The stage specificity observed with some L3 ADTs is often conserved between both filaria. This contrasts with patterns observed in ADTs from other stages indicating these genes may share functions in L3 biology that are conserved between the species. Examination of additional species of filaria has reinforced some of these observations (Allen *et. al.*, 2000; Frank and Grieve, 1991; Pogonka *et. al.*, 1999). Many of these proteins are present in the secretions of both species (Frank and Grieve, 1991; Pogonka *et. al.*, 1999). The potential roles these proteins play in mediating host-parasite interactions at this critical stage of the parasites development are still under investigation. However preliminary vaccination experiments with two of these proteins has yielded promising results (Gregory *et. al.*, 2000; Murray *et. al.*, 2001). However, the L3 EST dataset still has a wealth of potential candidates which have not yet been examined and both *in silico* and laboratory based screens are currently underway to isolate additional proteins secreted from this stage.

4.4 Chapters 3 and 4 Conclusions

The filarial EST datasets offer a tremendous resource to the research community. As more nematode EST datasets and genomes become available their value will increase as comparative analyses will allow nematode, parasite and filarial specific gene families to be defined. The ESTs (both the sequences and clones) will serve as a core reagent for the elucidation of the biology of these gene families. Functional genomics studies are being planned or are currently underway which are using the ESTs as a base resource. *B. malayi* microarrays are currently being constructed with oligos designed from the EST consensus sequences. The expression profiling performed with these arrays can offer much more accurate estimates of relative gene expression of transcripts. It will be interesting to compare the results of these analyses to the *in silico* analyses that have been presented here. The recent award of money to shotgun sequence the whole *B. malayi* genome means that large contigs of genomic DNA will soon be available. Again the ESTs and clusters will be extremely useful in the identification of genes and the annotation of these contigs. The ESTs have also provided an extremely effective catalyst to the filarial research community and many recent publications are based on genes or gene families that have been identified initially by the ESTs. They also have provided an important resource for new filarial vaccine candidates and several genes discovered from the L3 ESTs are currently under evaluation. The analyses presented here only scratch the surface of the information available in the cluster datasets. However, the public availability of these resources will allow the members of the community the opportunity to perform custom searches that address the questions specific to their fields of research.

Chapter 5

Phylogenetic analysis of the nematode MIF gene family

5.0 Discovery of MIF Gene Family

Macrophage Migration Inhibitory Factor (MIF) was first described during the mid-1960s as a product of activated lymphocytes that inhibited random migration of cultured monocytes (David, 1966) (Bloom and Bennett, 1966). MIF was the first cytokine ever characterized *in vitro* and its activities on immune cells were the subject of many subsequent studies. However, because of difficulties in isolating large quantities of MIF, its protein and nucleic acid sequence were not determined until 25 years later (Weiser *et. al.*, 1989). MIFs have remained enigmatic molecules, with unusual properties. Recently the influx of whole genome data and large EST sequencing projects have led to the discovery of MIF-like molecules in a variety of nonvertebrate metazoans, protozoa and plants. Most research on the MIF family has focused on vertebrate homologues and their functions as immune modulators or growth factors. Recently, MIFs from several parasitic nematode species have been discovered and proposed to have potential roles in immune modulation of the host. Very little is known about how the MIF gene family has evolved in nematodes and comparisons of the nematode MIFs with MIFs from other species may give any clues as to their functions. A survey was performed to collect all MIF sequences in the publicly available databases. The MIF sequences were aligned and compared. The alignment was also used to perform a phylogenetic analysis that has given some clues as to how this large gene family has evolved in nematodes and other eukaryotes.

5.0.1 Biochemical and Structural Properties of MIFs

Most cytokines act by binding receptors on the surface of target cells after secretion by effector cells. This binding initiates a signaling cascade inside the target cell via secondary mediators, which evokes specific responses. MIFs are unusual cytokines they have no defined surface receptors and have several distinct enzymatic activities that are required for some of their biological function.

All MIFs tested can tautomerize several small compounds. MIFs and a family of proteins related to MIF, DDTs (D-dopachrome tautomerase) can catalyses the conversion of the non-physiological substrate D-dopachrome to 5,6-dihydroxyindole

(Sugimoto *et. al.*, 1999). *Hs*-MIF and *Hs*-DDT share common tautomerase activities but *Hs*-DDT has an additional decarboxylase activity. A naturally occurring substrate, phenylpyruvate, has been identified but the reaction kinetics are not believed to be physiologically relevant (Rosengren *et. al.*, 1997; Stamps *et. al.*, 2000). The kinetics and substrate specificities of individual MIFs differ a great deal and this may be important for defining genuine substrates (Pennock *et. al.*, 1998a) (Tan *et. al.*, 2001). Mutational studies of human and murine MIFs have established that tautomerase activity is dependent on the presence of a proline at the amino terminus of the molecule (Pro2; the N-terminal Met is removed from the molecule). Pro2 is believed to serve as the catalytic base in the reaction and other tautomerases such as *E.coli* 4-oxalocrotonate tautomerase (4-OT) and 5-(carboxymethyl)-2-hydroxymuconate isomerase (CHMI) also rely on an N-terminal proline residue for their activity (Bendrat *et. al.*, 1997; Lubetsky *et. al.*, 1999; Taylor *et. al.*, 1999). Other residues (*Rn*–MIF-1: Lys32, Ile64, Tyr95 and Asn97) which line the substrate binding pocket may help determine substrate preference. However, none of these residues are required for catalysis (Lubetsky *et. al.*, 1999). The C-terminal portion of *Hs*-MIF is required for catalysis (Bendrat *et. al.*, 1997). However its function is unclear as the tertiary structure of *Hs*-MIF is undisturbed in the C-terminal deletion mutants CΔ104 and CΔ109 (Mischke *et. al.*, 1998).

A second enzymatic activity, thiol-oxioreductase, is present in some mammalian MIFs (Kleemann *et. al.*, 1998a; Kleemann *et. al.*, 1998b). This activity depends on the CXXC motif in the fourth β-sheet of the molecule (*Hs*-MIF residues 57 –60). The CXXC is conserved in several known thiol-oxioreductases such as thioredoxin, protein disulphide isomerases and glutaredoxins (Kleemann *et. al.*, 2000b) where it functions as the active site. In *Hs*-MIF both residues are important for catalysis. However, Cys57Ser mutants still show reduced but significant oxioreductase activities (Kleemann *et. al.*, 2000b). *Hs*-MIFs is active on small molecules like dihydroxyethyldisulphide (HED) or proteins such as insulin. However, unlike other oxioreductases it has a strong preference for larger reduction cofactors such as glutathione (GSH) or dihydrolipoamide (DHL) (Kleemann *et. al.*, 1998b). The residues Ala58 and Leu59 have been implicated in determining the substrate specificity of *Hs*-MIF-1 and Ala58Gly;Leu59Pro or

Ala58Gly;Leu59His could not reduce large substrates such as insulin but were still functional in biological assays which required oxioreductase activity (Kleemann *et. al.*, 1999). Neither the N or C-terminus are required for oxioreductase activity and β–sheets 1 and 6 deletion mutants show normal activity (Kleemann *et. al.*, 1998b; Kleemann *et. al.*, 2000b). However the removal of β-sheet 1 dramatically reduces the oxioreductase activity of, *Hs*-MIF when using HED as a substrate indicating it may be required to stabilize some interaction with this small molecule (Kleemann *et. al.*, 2000b).

The elucidation of the crystal structure of several MIFs (Sugimoto *et. al.*, 1996; Sugimoto *et. al.*, 1999; Sun *et. al.*, 1996; Suzuki *et. al.*, 1996; Tan *et. al.*, 2001) has provided additional clues to how these molecules might function (see figure 5.0.1.1). Despite the fact that some of these proteins share less than 14% sequence identity MIFs retain a conserved protein architecture (see figure 5.0.1.2). A MIF monomer is composed of two alpha helices (α1 and α2) and six beta sheets (β1-6). The core of the molecule contains two β/α/β motifs which combine to make a four-stranded β-sheet and two antiparallel helices. β3 is flanked on one side by the central core of the molecule and joins the two β/α/β motifs. The C-terminal structure of the MIFs differs between species. However in all proteins β6 interacts with β5 on an adjacent subunit. Like *Ec*-CHMI, MIFs and DDTs form homotrimers with a symmetric barrel-like structure (see figure 5.0.1.2). *Ec*-4-OT has a similar tertiary structure, but is made up of six β/α/β subunits

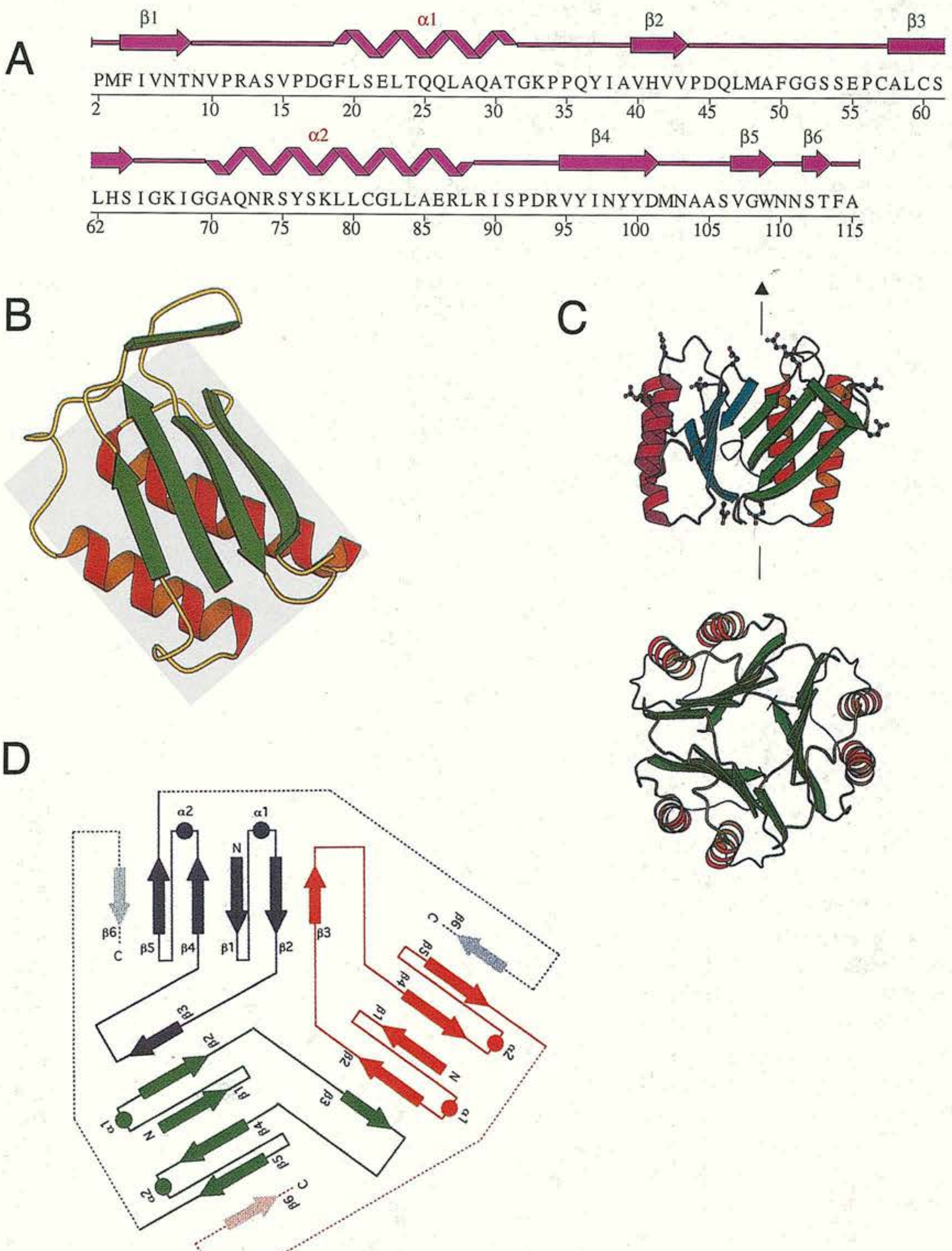
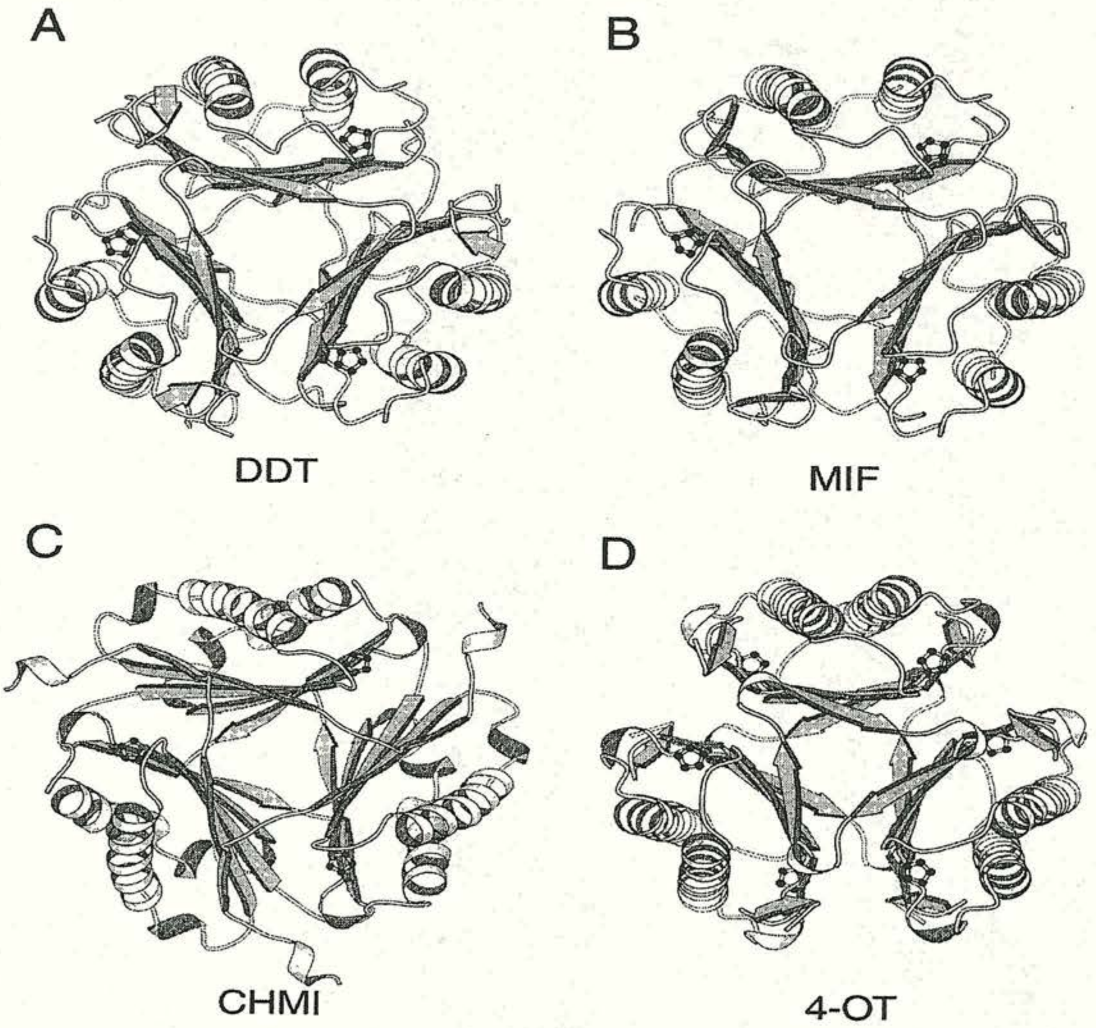

A
β1
α1
β2
β3
PMF I VNTNVPRASVPDGFLSELTQQLAQATGKPPQY I AVHVVPDQLMAFGGSSEPCALCS
2 10 15 20 25 30 35 40 45 50 55 60
α2
β4
β5
β6
LHS I GKI GGAQNRSYSKLLCGLLAERLRI SPDRVY INYYDMNAASVGWNNSTFA
62 70 75 80 85 90 95 100 105 110 115
B
C
D
α2
α1
N
β6
C
β5
β4
β1
β2
β3

Figure 5.0.1.1 Graphical representations of human and rat MIF-1 secondary and tertiary structures. **A:** The relationship between the secondary structure elements of *Hs*-MIF1 to the primary protein sequence. This cartoon was adapted from the pdb summary of *Hs*-MIF-1 taken from the CATH database (http://www.biochem.ucl.ac.uk/bsm/pdbsum/1mif/main.html and Orengo *et. al.* 1997). **B:** The structure of the *Rn*-MIF-1 monomer. α-helicies in red, β-sheets in green, unordered in yellow. **C:** The structure of the *Rn*-MIF-1 trimer seen from the side and from above. **D:** A flattened schematic of the *Rn*-MIF-1 trimer showing the points of contact between the β-sheets of adjacent monomers (β2 with β3 and β6 with β5). The graphics B, C, and D are taken from Suzuki *et. al.* 1996.



E

1 CHMI_ESHCO	-			
2 DDT1_HOMSA	13.6	-		
3 MIF1_HOMSA	14.8	32.2	-	
4 MIF1_RATNO	14.8	31.3	90.4	-
	1	2	3	4

Figure 5.0.1.2 Comparison of the tertiary structures of human MIF and DDT and *E.coli* CHMI and 4-OT. **A:** DDT1_HOMSA, **B:** MIF1_HOMSA, **C:** CHMI_ESHCO, and **D:** 4-OT_ESHCO. The cartoon shows each structure facing down the open "barrel" of the trimer (or hexamer in the case of 4-OT). The three fold symmetryof the molecules and the overall structural similarities despite the absence of high levels of sequence conservation is readily apparent. This figure has been taken from Sugimoto *et. al.* 1999. **E:** Pairwise similarity matrix showing the percent identity of the protein sequences of the three crystallized vertebrate MIFs and *E. coli* CHMI. The protein sequences have been aligned based on their secondary and tertiary structures (see figure 5.2.1)

Other growth factors such as interleukin 1B (IL-1B), fibroblast growth factor (FGF) and tumor necrosis factor (TNF) also form barrel structures with three fold symmetry. The barrel of vertebrate MIFs and DDTs forms an open, solvent accessible channel, which has been compared to a two-sided funnel. This channel is believed to be important for binding of substrates. However, the recently published structure of a MIF isolated from the parasitic nematode *Trichinella spiralis* shows that its channel is blocked in the center by the protrusion of several residues, suggesting that movement of substrates through its channel may not be vital for activity (Tan *et. al.*, 2001). In human and rat MIF the catalytic Pro2 residue lies in a region surrounded by two hydrophobic pockets. The first pocket is composed of Met3, Phe50, Ile65, Tyr96 and Val107, and the second of Pro34, Tyr37, Trp109 and Phe114 (Orita *et. al.*, 2001). In *Hs*-DDT these pockets are primarily composed of hydrophilic residues and it is believed that these differences can explain the affinity of MIFs for some substrates (Sugimoto *et. al.*, 1999). Two polar residues (Lys22 and Ser63) which are adjacent to the pockets have been conserved between the human MIF and DDT but their functions have not been determined. One of the cysteine residues essential for thiol-oxioreductase activity (Cys57) lies on the opposite side of the MIF trimer relative to Pro2. The other (Cys60) protrudes into the center of the cavity, presumably giving it access to substrates within it. No substrates or inhibitors of thiol-oxioreductase activity have been co-crystalized with MIF so it is still unclear what residues outside of the CXXC motif are important in determining substrate specificity.

Besides its enzymatic properties MIF has been shown to bind a variety of small molecules and proteins, including glycolipids such as gangliosides (Liu *et. al.*, 1982), haematin (Pennock *et. al.*, 1998b) and long chain fatty acids such as oleic acid (Bendrat *et. al.*, 1997). Haematin and oleic acid both inhibit tautomerase activity of *Hs*-MIF. However, it has not been established if these compounds are generally effective against MIFs and some nematode MIFs do not appear be as sensitive to haematin as their vertebrate homologues (Pennock *et. al.*, 1998a).

Human MIF has been shown to associate with several proteins under physiological conditions. *Hs*-MIF co-purifies with sarcolectin, a serum component with immunomodulatory and growth factor properties (Zeng *et. al.*, 1993). Sarcolectin has capacity to bind sugars like N-acetylneuraminic acid and it is unclear if *Hs*-MIF is binding directly to sarcolectin or sugars that may have co-purified with it.

MIF has also been shown to form covalent associations with peroxiredoxin (PAG) (Jung *et. al.*, 2001). While this interaction was initially found in a yeast two-hybrid screen both proteins could be coprecipitated *in vivo*. This binding is dependent on redox status and reducing agents such as 2-ME and DTT inhibited association. Association between MIF and PAG is also dependent on the active site cysteine of PAG (Cys173) and PAG Cys173Ser mutants do not coprecipitate. It is believed that this association is mediated via MIFs Cys60. The association inhibits the thiol-oxioreductase activities of MIF and PAG which suggests that the complex may have a regulatory function (Jung *et. al.*, 2001). It is not known whether MIF associates with CXXC domains of proteins other than PAG.

Two-hybrid screens have also revealed that human MIF directly interacts with Jab1, a component of the COP9/signalosome complex (Kleemann *et. al.*, 2000a). *In vivo* binding of MIF to Jab1 inhibits subsequent phosphorolaytion of c-Jun and the activation of AP-1 transcription complex. While it is unknown if the tautomerase activity is required for Jab1 binding and repression, the *Hs*-MIF Cys60Ser mutant was unable to repress Jab1 but still associated with it, implicating the thiol-oxioreductase activity in this function. One of the most interesting features of this interaction is the fact that MIF acts directly on Jab1 after crossing the cell membrane. The mechanisms which allow its translocation and subsequent release in the cytosol remain undefined.

5.0.2 The functions of MIFs in vertebrates

MIF and DDT are ubiquitously expressed in almost all vertebrate tissue types tested. Their functions in most of these tissues are unknown. MIFs have been shown to possess potent autocrine and paracrine properties. The most extensive studies of MIF function are those which have centered on MIF activities in the modulation of immune responses and immune cells in humans or rodents. However, MIF expression and secretion has been linked to a variety of processes outside of the immune system including growth, differentiation, neuronal function, apoptosis, oxidative stress and metabolism (reviewed in (Fingerle-Rowson and Bucala, 2001; Nishihira, 2000)). MIFs are potent signaling molecules that act on a wide variety of cell types. Despite the broad range of cells it can act upon the effects of MIF are context dependent. Slight differences in the environment of the target cell can elicit very different responses not all of which appear to be beneficial (Bozza *et. al.*, 1999; Roger *et. al.*, 2001). In at least one context MIF appears to be linked to the pathological aspects of inflammation (Froidevaux *et. al.*, 2001). Depletion of MIF either by treatment with anti-MIF antibodies, RNAi, or generation of MIF knock out mice has shown that MIF potentiates the inflammatory response, antagonizes glucocorticoid (GC) suppression of inflammation and increases bacterial toxin induced lethality in mouse models (Bozza *et. al.*, 1999; Froidevaux *et. al.*, 2001; Mitchell *et. al.*, 1999). The activities of MIF do not always appear to help in the resolution of infections and removal of MIF actually results in faster clearance of bacterial infections and reductions in inflammation induced pathology (Bozza *et. al.*, 1999). However, the context of MIF's activities appears to be pathogen dependent and killing of some intracellular pathogens like *Leishmania major* appear to be stimulated by MIF so it is unclear when MIF function is required for successful pathogen control (Bozza *et. al.*, 1999).

One of the most intriguing aspects of MIFs functions is that many of its activities are not dependent on its enzymatic activities (Bendrat *et. al.*, 1997; Hermanowski-Vosatka *et. al.*, 1999; Hudson *et. al.*, 1999; Kleemann *et. al.*, 2000b). Unfortunately, most of these observations are incomplete and no comprehensive studies have been undertaken that test the biological effects of systematically removing MIFs tautomerases and oxioreductase activities. Table 5.0.2.1 summarizes the different activities ascribed to vertebrate MIFs that have been examined in conjunction with structure function analyses.

Cell Type		Context	Activity	Species	Structure function relationship	References
Many cells	Chemokine Action	many	chemoattractant	murine and human	Pro2Gly abolishes chemotactic activity in many cell types. Pro2Phe MIF had chemotactic activity, stopped random migration of macrophages and antagonized the chemokine activity of MCP1.	(Hermanowski-Vosatka *et. al.*, 1999; Shimizu *et. al.*, 1999)
macrophages	Immune Response	inflammation	Blocks GC suppression of inflammatory cytokine production	murine	(Pro2Ser, Pro2Ala and N-terminal deletions NΔ2-6MIF) still blocked GC suppression of TNFα production in LPS stimulated monocytes. However, Cys60Ser showed reduced blocking of blocked GC suppression.	(Bendrat *et. al.*, 1997; Calandra *et. al.*, 1995; Kleemann *et. al.*, 1999; Kleemann *et. al.*, 2000b).
		pathogen killing	Macrophage killing of *Leishmania major* is stimulated by MIF via TNFα and NO^-	murine	MIF-/- macrophages or Cys57Ser mutants are no longer able to stimulate *L. major* killing	(Juttner *et. al.*, 1998; Kleemann *et. al.*, 1998a; Satoskar *et. al.*, 2001)
Macrophages and Natural killer cells	Immune Response	respiratory burst	MIF induces NO- and SOD production	murine and human	Pro1Gly mutants had reduced SOD in primed cells.	(Bernhagen *et. al.*, 1994; Swope *et. al.*, 1998)
Synovial fibroblasts	Immune Response / Pathology	inflammation	MIF treatment induces upregulation inflammatory cyctokines in cells collected from patients with rheumatoid arthritis	human	Pro2Ala mutant does not induce the upregulation	(Onodera *et. al.*, 2000).
Fibroblasts	Cell Cycle Control		MIF binds Jab1 and prevents Jab1 dependent upregulation AP-1 activity.	murine	Cys60Ser mutants still bind Jab1 but do not repress AP-1 activity	(Kleemann *et. al.*, 2000a)

Table 5.0.2.1 Summary of some the published activities of vertebrate MIFs that have included structure function analyses. The table lists the cell types the activity is found in, the context of the activity, the species the activity was found in, available structure function data associated with the activity and the pertinent references. GC: glucocorticoid SOD: super oxide dismutase,.

5.0.3 MIFs in other organisms

MIF-like genes have recently been cloned from several non-vertebrate metazoans. These MIFs have only been functionally characterized from nematodes and one arthropod where they may be involved in host-parasite interactions.

MIF from the hard-bodied tick *Amblyomma americanum* had tautomerase activity, inhibited random migration of human monocytes and was localized in the midgut and salivary glands (Jaworski *et. al.*, 2001).

Several MIFs have been characterized from free-living and parasitic nematodes. Nematode MIFs were first identified in the human parasite *Brugia malayi* (Blaxter *et. al.*, 1996; Pastrana *et. al.*, 1998). *Bm*-MIF-1 is present in all stages of developement and has been localized to the lateral chords and uterine wall of adult females and the developing larvae *in utero*. *Bm*-MIF-1 is secreted into the mammalian host by both larval and adult parasites, has tautomerase activity and is chemotactic for human monocytes. *Bm*-MIF-1 has been shown to induce expression of Ym-1 (an eosinophil chemotactic factor) expression in alternately activated macrophages and is able to recruit eosinophils *in vivo*. Eosinophil recruitment is an important feature of nematode infection and *Bm*-MIF-1 provides a potential link between parasite products and the recruitment of these cells (Falcone *et. al.*, 2001). Mutant *Bm*-MIF-1 (Pro2Gly) lacking tautomerase activity was unable to perform either of these activities (Falcone *et. al.*, 2001). *B. malayi* also expresses a second MIF homologue, *mif-2* which has similar enzymatic activities, but its role in interactions with the host immune system has not been assessed. A MIF with tautomerase activity has been isolated from the secretions of the muscle and intestinal parasite *Trichinella spiralis.* Like *Bm*-MIF-1 is active in chemotactic assays with human monocytes (Tan *et. al.*, 2001). The crystal structures of *Ts*-MIF-1 and *Bm*-MIF-2 reveal tertiary structures extremely similar to vertebrate MIFs, indicating that stringent sequence conservation is not required for the maintenance of structural and enzymatic function (Zang X.X. and Maizels R. submitted 2002 and (Tan *et. al.*, 2001)).

Four MIFs are found in the free-living nematode *Caenorhabditis elegans* (Marson *et. al.*, 2001). RNAi of the MIF genes did not show any visible phenotypes, but all four are expressed across the major stages of post embryonic development.

Ce-mif-1 is upregulated in adult worms while *Ce-mif*-2 and *mif-3* are upregulated in the dauer larvae or during environmental stress. *Ce-mif-2* and *mif-3* are expressed in body wall and vulva muscle. *Ce-mif-2* was also expressed in the hypodermis while *Ce-mif-3* was expressed in the pharynx and in embryos after the 8 cell stage. *Ce*-MIF-3 was associated with embryonic nuclei indicating a possible role in the control of early development. *Ce*-MIF-4 is unusual because a catalytic proline has been replaced with a glutamine. Because of this substitution it is not predicted to have tautomerase activity.

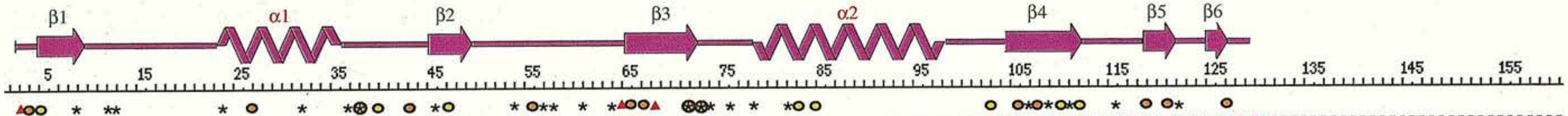
To help better understand how the nematode MIF and other MIFs sequences are related, a survey to isolate MIF genes available in the public databases was undertaken. The isolated MIF sequences were aligned and analyzed for conserved structural features. The alignment was used for phylogenetic analyses which have given some clues as to how the MIF gene families have evolved in nematodes and other organisms.

5.1 Isolation of MIF gene family members from the public databases

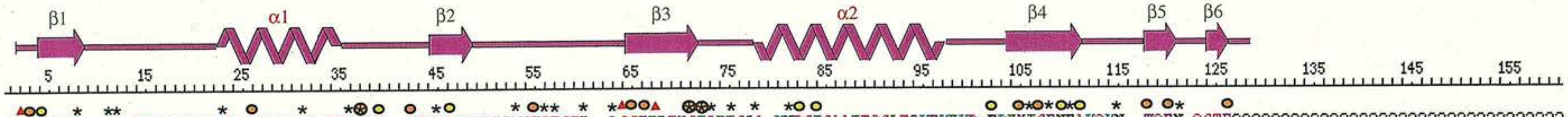
To isolate *mif* genes the non-redundant nucleotide, protein and EST databases (GenBank, 10/01/2002) were searched iteratively with TBLASTN, BLASTN and PSI-BLAST using vertebrate, nematode and plant MIFs as probe sequences (Altschul *et. al.*, 1990; Altschul *et. al.*, 1997). In the case of ESTs, where there were multiple sequences from one gene, assemblies were made using AssemblyLIGN (Oxford Molecular). Seventy-one distinct sequences from forty-two species were identified (see appendix II table 5.1.1).

5.2 Multiple sequence alignment of the MIF protein and cDNA sequences

A multiple sequence alignment was constructed using all seventy-one protein sequences using CLUSTAL X (Thompson *et. al.*, 1997) and optimized by hand (figure 5.2.0.1). The coding sections of the cDNAs used to predict the amino acid sequences of the MIF genes were aligned using the protein multiple sequence alignment as a guide. All gaps used in the protein alignment were reflected in the cDNA alignment (see figure 5.2.0.2 appendix III).



```
CMI1_ESCEC MPHFIVECSDNIREEADLPGL-FAKVNPTLAATGIFPLAGIRSRVHWVDTWQMADGQHDY--ASVHMTLK-IGAGRSLESRQQAGEMLFELIKTHFAALMESRLLALSFEIEELHPTLNFKQNNVHALFK??????????????????????????????
MIF1_MUSMU MPMFIVNT--NVPRASVPEGF-LSELTQQLAQATGKPAQ--YIAVHVVPDQLMTFSGTND--PCALCSLHSIGKIGGA-QNRNYSKLLCGLLSDRLHISP-DRVYINYYDMNAAN--VGWN-GSTFA?????????????????????????????????
MIF1_RATRA MPMFIVNT--NVPRASVPEGF-LSELTQQLAQATGKPAQ--YIAVHVVPDQLMTFSGTSD--PCALCSLHSIGKIGGA-QNRNYSKLLCGLLSDRLHISP-DRVYINYYDMNAAN--VGWN-GSTFA?????????????????????????????????
MIF1_HOMSA MPMFIVNT--NVPRASVPDGF-LSELTQQLAQATGKPPQ--YIAVHVVPDQLMAFGGSSE--PCALCSLHSIGKIGGA-QNRSYSKLLCGLLAERLRISP-DRVYINYYDMNAAN--VGWN-NSTFA?????????????????????????????????
MIF1_MERUN MPMFIVNT--NVPRSSVPEGL-LSELTQQLAQATGKPAQ--YIAVHVVPDQLMTFSGSSD--PCALCSLHSIGKIGGA-QNRTYSKLLCGLLADRLRISP-DRIYINYYDMNAAN--VGWN-GSTFA?????????????????????????????????
MIF1_SUSSC MPMFVVNT--NVPRASVPDGF-LSELTQQLVQAMGKPAQ--YIAVHVVPDQLMAFGGSSE--PCALCSLHSIGKIGGA-QNRSYSKLLCGLLAERLRISP-DRIYINYYDMNAAN--VGWN-G?????????????????????????????????????
MIF1_BOSTA MPMFVVNT--NVPRASVPDGL-LSELTQQLAQATGKPAQ--YIAVHVVPDQLMTFGGSSE--PCALCSLHSIGKIGGA-QNRSYSKLLCGLLTERLRISP-DRIYINYYDMNANF--VGWN-GSTFA?????????????????????????????????
MIF1_GALGA MPMFTIHT--NVCKDAVPDSL-LGELTQQLAKATGKPAQ--YIAVHIVPDQMMSFGGSTD--PCALCSLYSIGKIGGQ-QNKTYTKLLCDMIAKHLHVSA-DRVYINYFDINAAN--VGWN-GSTFA?????????????????????????????????
MIF1_XENLA MPVFTIRT--NVCRDSVPDTL-LSDLTKQLAKATGKPAE--YIAIHIVPDQIMSFGDSTD--PCAVCSLCSIGKIGGP-QNKSYTKLLCDILTKQLNIPA-NRVYINYYDLNAAN--VGWN-GSTFA?????????????????????????????????
MIF1_DANRE MPMFVVNT--NVAKDSVPAEL-LSEATQELAKAMGKPQQ--YIAVQVVPDQMMMFGGKGD--PCALCSLTSIGKISGA-QNKQYSKLLMGLLNKHLGVSA-DRIYINFVDMDPAN--VAWN-NSTFG?????????????????????????????????
MIF1_PSEAM MPMFVVNT--NVAKGDVPAAL-LSEATEELAKEMGKPAQ--YIAVHINPDQMMMFGGKGD--PCALCSLHSIGKISGA-QNKKYSKLLCGLLNKHLGISP-DRIYINFFDMDAAN--VAWN-NSTFA?????????????????????????????????
MIF1_BRUMA MPYFTIDT--NIPQNSISSAF-LKKASNVVAKALGKPES--YVSIHVNGGQAMVFGGSED--PCAVCVLKSIGCVGPK-VNNSHAEKLYKLLADELKIPK-NRCYIEFVDIEASS--MAFN-GSTLG?????????????????????????????????
MIF1_BRUPA MPYFTIDT--NIPQNSISSAF-LKKASNVVAKALGKPES--YVSIHVNGGQAMVFGGSED--PCAVCVLKSIGCVGPK-VNNSHAEKLYKLLADELKIPK-NRCYIEFVDIEASS--MAFN-GSTFG?????????????????????????????????
MIF1_WUCBA MPYFTIDT--NKPQDSISSAF-LKKAPNVVPKALGKPES--YVSIHVNGGQPMVFGGSED--PCPVCVLKSIGCVGPK-VNNSHAEKLYKLLADELKIPK-NRCYIESVDIEASS--MAFN-GSTFG?????????????????????????????????
MIF1_ONCVO MPAFTINT--NIPQSNVSDAF-LKKASSTVAKALGKPES--YVAIHVNGGQAMVFGGSTD--PCAVCVLKSIGCVGPN-VNNSHSEKLFKLLADELKIPK-NRCYIEFVNIDAST--MAFN-GSTFG?????????????????????????????????
MIF1_ASCSU MPVLTINT--NVPSDKVPQDF-LKKTSALVAKSLSKPES--YVAVRVNPDQQMTFGGSAD--PCAVCTLESIGAVGGS-RNNAHAEKLYNHLNETLGIPK-NRMYISFVDIDPTT--MAYN-GSTFA?????????????????????????????????
MIF2_CIOIN MPEITIQT--NVSSDKIASDL-QEIVVELVSQHLNKPKA--NICVTVLTDLWMSFGESEE--PCACCTVTSIVDFNAE-TCEKLAALLMPPLQKALGVSG-TRFYLQFHEITAGI--MGFQ-GTTVKVVRERKQSS????????????????????????
MIF1_CIOIN MPHLFVKT--NVAKDKLPKSILQDLTKLVSSTIPNKPEK--YVCVTVVPDVWMSFGGTEE--PCAAAVLTSISDFNAE-TCTTYAEAMLGEIYKLLGVAQ-DRMYLEFHEATRET--MGYN-GTTFHQLAAKK???????????????????????????
MIF2_TRISP MPIFTIIT--N--KKTAPKDF-HRLLTDLLAELLKKPKE--LVVVDLLLDQKMEFGGADD--PCLIGVVRAVGRISAE-ENAQYAERLSEFLHQQLGILP-QRMYIRYLNMDGFY--VGWS-GCLRA?????????????????????????????????
MIF1_HAECO MPVFSFHT--NVAASKVTPDL-LKQISALVARILHKKKS--YVCVHVVPDQHMIFAGTDE--PCGVGVLKSIGGVGGS-KNNEHAKALFALIKDHLGISG-NRMYVEFIDIGAAD--IAFN-SKTFA?????????????????????????????????
MIF1_AMBAM MPTLTINT--NIPASKIPNDF-LKTTANVVADSLGKPLS--YVVVHINADQLLSFGGTDD--PCAIANLYSIGCLSPK-ENKKHSAVLFEHIEKTLGIKE-NRMYINYFDMPASD--VGYN-GKTFAG????????????????????????????????
DDT1_HOMSA MPFLELDT--NLPANRVPAGL-EKRLCAAAASILGKPAD--RVNVTVRPGLAMALSGSTE--PCAQLSISSIGVVGTAEDNRSHSAHFFEFLTKELALGQ-DRILIRFFPLESWQ--IGKI-GTVMTFL???????????????????????????????
DDT1_RATNO MPFVELET--NLPASRIPAGL-ENRLCAATATILDKPED--RVSVTIRPGMTLLMNKSTE--PCAHLLISSIGVVGTAEQNRSHSSSFFKFLTEELSLDQ-DRIIIRFFPLEPWQ--IGKK-GTVMTFL???????????????????????????????
DDT1_MUSMU MPFVELET--NLPASRIPAGL-ENRLCAATATILDKPED--RVSVTIRPGMTLLMNKSTE--PCAHLLVSSIGVVGTAEQNRTHSASFFKFLTEELSLDQ-DRIVIRFFPLEAWQ--IGKK-GTVMTFL???????????????????????????????
DDT1_ORYLA MPFVELQT--NLPGSSFNEDF-LKKLCSCVASTLSKPEE--RMNVVVKPGLPMLMAGSCS--PCVILSVAAISVTDSADKNKQHSAKIFEFLTKELSMTE-DRILIKFDELQPHQ--VGKK-GTVMSFL???????????????????????????????
MIF1_CAEEL MPVFSINV--NVKVPAEKQNEILKELSTVLGKLLNKPEQ--YMCIHFHEDQGILYAGTTE--PAGFAVLKSIGGVGSAKQNNAISAVVFPIIEKHLGIPG-NRLYIEFVNLGAAD--IAYN-GQTFA?????????????????????????????????
MIF1_TRITR MPIFTFST--NVPSENISVDF-LKSTSKLIAGMLGKPES--YVAVHINGGQKITFGGTDA--PAGFGQLLSLGGVGGE-KNRSHSAKLFKHLTDGLGIPG-NRMYINFVDMRGSD--VGYN-GSTF??????????????????????????????????
MIF1_TRISP MPIFT?NT--NIKATDVPSDF-LSSTSALVGNILSKPGS--YVAVHINTDQQLSFGGSTN--PAAFGTLMSIGGIEPS-RNRDHSAKLFDHLNKKLGIPK-NRMYIHFVNLNGDD--VGWN-GTTF??????????????????????????????????
MIF1_TRIPS MPIFTFNT--NIKATDVPSDF-LSSTSALLADILSKPES--YVAVHLNTDQQLTFGGNTS--PAAFGSLMSIGGIEAS-RNRDHSTKLFDHINKKLGIPK-NRMYIHFVNLRGND--VGWN-GTTF??????????????????????????????????
MIF2_CAEEL MPMVRVAT--NLPNEKVPVDF-EIRLTDLLARSMGKPRE--RIAVEIAAGARLVHGATHD--PVTVISIKSIGAVSAE-DNIRNTAAITEFCGKELGLPK-DKVVITFHDLPPAT--VGFN-GTTVAEANKK????????????????????????????
MIF2_STRSE MPYVRLFS--NLPETSFTDAF-CTQFTDLLAEKLHKDKS--RIVMLVQPHTMMSSGGVPN-QPSIWIEINNVGQLSPR-QTQELSRDLTHFVMEQTTVPR-ESVSILYFDMSPDM--VARG-GITIAESIAGLK??????????????????????????
MIF2_HETGL MPFINLWT--NLPELKLDNEF-RRTFLATVANSMQKPVE--STALIVQSGPNCCQIGSDPAEPAMILQIKSIGCVSAD-ENVIHCKNISEFVQSRLGIPP-DRVMIHFQSLEKHE--VGKG-GTTVEKMCQ?????????????????????????????
MIF3_CAEEL MPVIKVQT--NV--KKVSDGF-EVRLAIHMAKVMKRPES--QIFVSLDMNSRMTRGQLTD--PLAVLDVTSSTVLTPI-LTEEYTVALCEFFSQELALDS-DAVLINYRSLSPEL--IGFN-GHILTENRPFISTDRARFIIGVLGIAFLAFLLQFLKYI
MIF4_CAEEL MQVVRIQT--NIRSADIPEKF-EQDVIYNLSVVMELPAD--KFVIIVEPAVRMRIGFENKEIPVAIVNFQTTRPSSRI-ENDSYAKKLTSVLNEQLKLDP-AHIFISFDFKDAKS--FATQ-GKTIASLYE?????????????????????????????
MIF2_BRUMA MPLITLAS--NVPASRFPSDF-NVQFTELMAKMLGKPTS--RILLLVMPNAQLSHGTTEN--PSCFTVVKSIGSFSAD-KNIEYSSSISEFMKKTLDIDP-AHCIIHFLNLDPEN--VGCK-GTTMKVLMKK????????????????????????????
MIF2_ONCVO MPLITLAS--NVLASGFPTDF-SVQFTKLMAELLGKPIS--RITLLVTPSAQLSRGATQD--PTCLIVIKSIGSFSAD-KNIKYSGSISEFIKKTLNIDP-AYCIIHFLDLNPED--IGCN-GTTMKELMKK????????????????????????????
```



```
MIF1_GIRLA MPCAIVTT--NADFTKDQADAFCLDMGQVLAKETGKPVS--YCMAGVRKAD?MSFGTSTD--LCCFVDFYCIGVISQA-KNPSISAAITGCLTQHFKVKP-ERVYISFNEAKGHN--WGFN-GSTF???????????????????????????????????
LS1_EMETE  MPLCQIVC--NTQVESGAEEAFLAAVESGLSKILGKPTQ--YITVTLTRGSVRH-SGSCD--PAASVSVHSIGGISSR-TNNMICAEVAALCQQHLKVPV-DRVFFHFADVSAAN--IGI?-GSRVFG?????????????????????????????????
LS1_PLAFA  MPCCEVIT--NVNLPDDNVQSTLSQIENAISDVMGKPLG--YIMSNYDYQKNLRFGGSNE--AYCFVRITSIGGINRS-NNSAL-ADQITKLLVSNLNVK-SRRIYVEFRDCSAQ--NFAF-SGSLFG?????????????????????????????????
LS1_PLAYO  MPCCELIT--NISIPDDKAQNALSEIEDAISNVLGKPVA--YIMSNYDYQKNLRFSGSNE--GYCFVRLTSIGGINRS-NNSSL-ADKITKILSNHLGVK-PRRVYIEFRDCSAQ--NFAF-SGSLFG?????????????????????????????????
LS1_PLABU  MPCCELIT--NISIPDDKAQNTLSEIEDAISNILGKPVA--YIMSNYDYQKNLRFSGSNE--GYCFVRLTSIGGINRS-NNSLL-ADKITKILSNHLSVK-PRRVYIEFRDCSAQ--NFAF-SGSLFG?????????????????????????????????
LS1_PHYSO  MPNVQVTS--NVPSSGVDKAKAMAAISKGVATALGKSEQ--VVMVHLNLDTPMLFQASDA--PCAMIQLKSIGKVDAQ-HNPTT-ASILTETVSQELNVP-KDRIFMNIDDVQRS--N-WA-KGGVLIPEPKQ????????????????????????????
LS1_PINTA  MPSLSIST--NVSLDGFNTSEILSETSKNVAKIIGKPEA--YVMVQLKGSVAISFGGTEE--PAAYGELVSIGGLGSD-TNKKLSSAIANVLETKLGVSK-SRFYIKFYDVKRSD--FGWN-GTTF???????????????????????????????????
LS1_CYRJA  MPSLSIST--NVPLDGVNTSGILSQASKSVAQIIGKPEA--YVMVQLKGSVAISFGGTED--PAAYGELVSIGGLSAD-TNKKLSAAISSILESTLSVPK-SRFYIKFYDVKGSN--LGYN-GSTF???????????????????????????????????
LS1_TRIAE  MPCLNVST--NVNLEGVDTSAVLADASSTVATIIGKPEN--YVMVVLKGSVPMAFGGTQE--PAAYGELVSIGGLNPD-VNKKLSAGIASILESKLSIPK-SRFYLKFHDSKRSD--FGWN-GSTF???????????????????????????????????
LS2_TRIAE  MPCLNVST--NVNLEGVDTSAVLADASSTVATIIGKPEA--YVMVVLKGSVPMAFGGTQE--PAAYGELVSIGGLNAD-VNKKLSAGIASILESKLSIPK-SRFYLKFHDSKRSD--FGWN-GSTF???????????????????????????????????
LS1_ZEAMA  MPCLNVST--NVNLEGVDTSAILAEASKSVANIIGKPEA--YVMVVLKGSVPMAFGGTQE--PAAYGELVSIGGLNPD-VNKKLSAGIASILESKLSVPK-SRFYLKFYDSKRSD--FGWN-GSTF???????????????????????????????????
LS1_HORVU  MPCLNVST--NVNLEGVDTSAVLADASSTVATIIGKPEG--YVMVVLKGSVPMAFGGTQE--PAAYGELVSIGGLNPD-VNKKLSAGISSILESKLSISK-SRFYLKFHDSKRSD--FGWN-GTTF???????????????????????????????????
LS1_ORYSA  MPCLNVST--NVNLDGVDTSAVLADASKTVATIIGKPEA--YVMVVLKGSVPMAFGGTQE--PAAYGELVSIGGLNPD-VNKKLSAGIASILESKLSIPK-GRFYLKFYDSKRSD--FGWN-GTTF???????????????????????????????????
LS1_SORBI  MPCLNVST--NVNLEGVDTSVILAEASKSVANIIGKPEA--YVMVVLKGSVPMAFGGTQE--PAAYGELVSIGGLNPD-VNKKLSAGIASILESKLSVPK-SRFYLKFYDSKRSD--FGWN-GSTF???????????????????????????????????
LS1_ARATH  MPCLNLST--NVNLDGVDTSSILSEASSTVAKIIGKPEN--YVMIVLKGSVPMSFGGTED--PAAYGELVSIGGLNAD-VNKKLSAAVSAILDTKLSVPK-SRFFLKFYETKGSF--FGWN-GATL???????????????????????????????????
LS2_ARATH  MPCLYITT--NVNFDGVNTDPFYSEVTKAVASIVGRPQN--LVMVVLKGSVEIVFGGNKE--AAAYAEIVSMGGITKQ-VKRELIATVGSILHTHFSIHP-TRFIFKVFDINSLP--LPSK-L??????????????????????????????????????
LS3_ARATH  MPTLNLFT--NIPVDAVTCSDILKDATKAVAKIIGKPES--YVMILLNSGVPIAFAGTEE--PAAYGELISIGGLGPG-VNGKLSETISEILQIKLSIDS-SRFYIKFYDSPRPF--FGYN-GSTF???????????????????????????????????
LS1_GOSAR  MPCLNLST--NVNLDGVDTSAILSEATSSVAKLIGKPEA--YVMIVLKGSVPMSFGGTEQ--PAAYGELVSIGGLNPD-MNKKLSAAIAAILETKLQVPK-SRFFLKFYDTKGSN--FGWN-GSTF???????????????????????????????????
LS2_GOSAR  MPCLNLST--NVNLDGVDTSAILSEATSSGAKLIGKPGA--YVMIGLKGSIPMSFGGTEQ--PACYGELVSIGGLNPE-YEQETECLQLLEFLKPSYKCL-SHGSSSNSMTPRVP--TLDG-TDPPSEPGMF?????????????????????????????
LS1_GLYCL  MPTLNLFT--NVPVDTVVASDILRDATKAVAKIIGKPES--YVMILLNGGVPIAFAGTEE--PAAYGELISIGGLGPS-VNGKLSSTIAEILQTKLYIDS-SRFYIKFYDVQRSF--FGFN-GSTF???????????????????????????????????
LS1_LOTJA  MPCLNLST--NVNLDGVDTSSILSEATSTVATLIGKPEA--YVMIVLKGSVPVSFGGTEQ--PAAYGELVSIGGLNPD-VNKKLSAAIASILETKLSVPK-SRFFLKFYDTKGSN--FGWN-GST????????????????????????????????????
LS1_HELAN  MPCLTLST--NVNLDGLDSSSILSDATSAVAKIIGKPEA--YVMIVLKGSIPISFGGNEQ--PAAYGELVSIGGLNAD-VNKKLSAAVAEILETKLSVPK-SRFYLKFYDTKGSF--LGWN-GSTF???????????????????????????????????
LS1_LYCES  MPCLNIST--NVNLEGVDTSSVLSEATSTVAKLIGKPEA--YVMIVLKGSVPMAFGGTEQ--PAAYGELVSIGGLNAD-VNKKLSAAIADILETKLSIPK-SRFFLKFYDTKGSF--FGWN-GSTF???????????????????????????????????
LS1_GLYMA  MPCLNLST--NVSLEGVDTSSILAEATSSVANIIGKPEA--YVMIVLKGSVPIAFGGNEQ--PAAYGELVSIGGLNPS-VNKELSAAIASILETKLSVPK-SRFFLKFYDTKGSN--FGWN-GSTF???????????????????????????????????
LS2_GLYMA  MPCLNLST--NVNLDGVDTSSILSEATSTVASIIGKPEA--YVMIVLKGSVPISHGGSEQ--PAAYGELVSIGGLSPD-VNKKLSAGIASILENKLSVPK-SRFYLKFYDTKGSN--FGWN-GSTF???????????????????????????????????
LS3_GLYMA  MPCLNLNT--NVSLDGVDTSSILAEATSSVANIIGKPAA--YVMIVLKGSVPIAFGGNEQ--PAAYGELVSIGGLNPS-VNKELSAAIASILETKLSVPK-SRFFLKFYDTKGSN--FGWN-GSTF???????????????????????????????????
LS1_MESCR  MPCLNIST--NVSLDGVDTSAILSEATSSVANIIGKPEA--YVMVVLKGSVPMAFGGTEQ--PAAYGELVSIGGLNPD-TNKKLSAAIAAILESKLSVPK-SRFFLKFYDTKGSF--FGWN-GSTF???????????????????????????????????
LS2_MESCR  MPCLNIST--NVSLDGVDTSAILSEATSSVANIIGKPEA--YVMVVLKGSVPMAFGGTEQ--PAAYGELVSIGGLNPD-TNKKLSAAIAAILESTSPVPK-SRFFLKFYDTKGPS--LRSN-GSTSQTKPVV?????????????????????????????
LS1_MEDTR  MPCLNLST--NVNLEGVDTCSILSEATSTVATLIGKPES--YVMIVLKGSVPISFGGTEQ--EAAYGELVSIGGLNPD-VNKKLSAAIAAILETKLSVPK-TRFFLKFYDTKGSN--FGWN-GTTF???????????????????????????????????
LS1_SOLTU  MPCLNIST--NVNLEGVDTSSVLSEATSTVAKLIGKPEA--YVMIVLKGSVPMAFGGTEQ--PAAYGELVSIGGLNAD-VNKKLSAAIADILETKLSIPK-SRFFLKFYDAKGSF--FGWN-GSTL???????????????????????????????????
LS1_ROBPS  MPCLTIST--NVSLDGVDTSTILSEATSTVAKLIGKPEA--YVMIVLKGSVPMAFGGTEQ--PAAYGELVSIGGLNPN-VNKKLSAAIAEILETKLSVPK-SRYFLKFYDTKGSD--FGWN-GSTF???????????????????????????????????
MIF2_MELJA ??????DH--FPPKIFHIPPF-QGGLQPPLATSEKIKFN--FSKVLVNAGNVGCFGGSTD--PFIYAELQSIGGFTDPN-KVTGEMTKLFTEHFGVPGSRV-YMKLTG---PDANK--FAHN-GKTFA?????????????????????????????????
MIF1_ANCCA ?????????????????????????????????????????????????????????????????????ESIGALSAD-DNIRHTQKITQFCQDTLKLPK-DVIITYF?DLQPIH--VGFN-GTTVAAATM??????????????????????????????
MIF2_ANCCA ?????????????QNKVTPDL-LKQISELVARILHKPES--YVAVHVVPDQKMTFAGTDA--PCGIGILKSIGGVGGS-QNNSHAKALFALIKDHLGIEG-NRMYIEFVDIGASD--IAHN-GRTFA??????????????????????????????????
LS1_TOXGO  MPKCMIFC--PVAATRRNRTPS?RTPKKPSQSILGKPLS--YVMVGYSQTGQMRFGGTAT--RVRSFALLPLEASPVP-RTAKSPLLSPLHANAPGRPQE-PHLHDIHKQRAPLK--LGHG-GPEL???????????????????????????????????
```

- * Residues conserved in the majority of MIFs
- ▲ Residues Required For Enzymatic Activity
- ● Group Specific Synapomorphies
- ● VariableResidues Lining Substrate Binding Pocket
- ⊛ Conserved Residues Lining Substrate Binding Pocket

Figure 5.2.1 The MIF protein multiple sequence alignment. The amino acid residues in the alignment were colored using the default settings in Seqpup v0.6 (D.G. Gilbert ,Biology Dept., Indiana University). Small uncharge residues A, G, P, S, T: magenta; positively charged H, K, R: light blue; polar neutral or negatively charged D, E, N, Q: black; hydrophobic residues I, L, M,V: green; hydrophobic aromatic residues F, W, Y: blue; cysteine C: red The sequences are named following table 5.1.1.

Comparison of the of the sequences in the alignment indicate that the MIF gene family is extremely polymorphic with many of the sequences showing very low overall sequence similarity to each other (range of 10-99%). However, examination of the sequence alignment has revealed that there are at least eighteen residues that are conserved in most of the isolated MIF sequences. Mapping of these conserved residues back to a canonical MIF secondary structure indicates the majority of these conserved residues are found in four sections of the molecule; β-sheet 1, the regions adjacent to β-sheet-2, β-sheet 3, and the regions adjacent to β-sheet 4 (see figure 5.2.0.2). The residues critical for MIFs enzymatic activities are found near the conserved residues in β-sheets 1 and 3 while several residues lining the substrate binding pocket are found in β-sheets 2 and 4. Eleven group specific synapomorphies were also identified. Like the conserved residues they map to areas adjacent to residues important for catalysis or substrate binding. The group specific synapomorphies identified are summarized in table 5.2.0.3 along with substitutions observed in the active site residues and residues lining the substrate binding pockets. Interestingly, despite the diversity in sequence, examination of the position of gap/insertions in the MIF alignment revealed that in all but one instance they occurred in loop regions outside of the six β-sheets and two α-helices. The exception was found in the *Plasmodium sp.* and *P. sojae* sequences where a gap, relative to the other MIF sequences, is predicted in the middle of α-helix two.

Feature	#		Group or species	Comments on possible structure function relationships
Active site residues	**2**	Pro -> Gln	MIF4_CAEEL	Single base pair transversion CCA -> CAA likely to abolish tautomerase activity
	64	Cys -> Ala and other residues	All plant sequences many nematode sequences	Reduced oxioreductase activity
	67	Cys -> Gly and other residues	All plant, all DDT, many nematode sequences	No oxioreductase activity
Group specific synapomorphies	**4**	Aromatic	vertebrate and most ecdysozoa MIF1s	Near the Pro2 may interact with substrates
		Hydrophobic	DDTs, nematode MIF2 and plant MIFs	
	39	Cys	Apicomplexa	On the surface of MIF facing solvent
		Positvely charged	Vertebrate MIFs, DDTs and nematode MIF2s	
		Small uncharged	Most other sequences	
	46	Positively charged	Ecdysozoan MIF1s and vertebrate MIF1s	Faces the center of the trimer
		hydrophobic	Most plant sequences	
	82	Positively charged	Vertebrate MIFs, DDTs and plant MIFs	
		Polar or negatively charged	many nematode MIF1s	
	84	aromatic	Vertebrate MIFs	
		Positively charged	DDTs and Ecdysozoan MIF1s	
		Positively charged or aromatic	nematode MIF2s	
		hydrophobic	Plant and protozoan MIFs	
	102	Polar or negatively charged	Animal MIFs	Faces residue 55 may interact
		Small uncharged	Plant LSs	
	109	Aromatic	Vertebrate MIFs, DDTs, *Plasmodium sp* and many plant MIFs	Faces interior of the trimer
	111	Hydrophobic or positively charged	Nematode MIFs	May interact with residue 109 also faces the interior of the trimer
		hydrophobic	Animal sequences	
Polymorphic residues lining the substrate binding sites	**3**	Hydrophobic or Aromatic residues	Animal	Near Pro2 may interact with substrates
		Cys	Plant, Plasmodium and Giardia	Free thiol may interact with substrate
	26	Hydrophobic or neutral	Animal and protozoan MIFs	
		neutral	Plant MIFs	
	42	Positively charged	Vertebrate DDTs Nematode MIF2s	Faces residue 4
		Aromatic	Most other sequences	

	55	hydrophobic	DDTs	
		No residue preference	Nematode MIF2s	
		Aromatic	Most other sequences	
	65	Ala	Most Animal and plant MIFs	Involved in thiol-oxioreductase substrate specificity
		No preference	Most nematode MIF2s	
	66	Leu or Val	Vertebrate and many nematode MIF1s	Involved in thiol-oxioreductase substrate specificity
		No preference	All other sequences	
	105	aromatic	Vertebrate, ecdysozoan , plant, *G. intestinalis* and *E. tenella* MIFs	
		hydrophobic	DDTs, nematode MIF2s, Apicomplexan and *P. sojae* MIFs	
	107	Polar or negatively charged	Vertebrate and most ecdysozoan MIFs	
		Positively charged	DDTs, most nematode MIF2s and Plant MIFs	
		Small hydrophobic	Apicomplexan and *P. sojae* MIFs	
	118	Aromatic	Plant MIFs	
		Hydrophobic	Animal MIFs	
	120	Aromatic	Most Animal and many plant MIFs	
		Positively charged	DDTs and some nematode MIF2s	
	126	Aromatic	Vertebrate MIF1s, ecdysozoa MIF1s, most plant MIFs	
		Hydrophobic	Vertebrate DDTs, nematode MIF2s, apicomplexa and Apicomplexan and *P. sojae* MIFs	

Table 5.2.0.2 The identified group specific synapomorphies and polymorphic residues lining substrate binding pocket. The residue position in the alignment, the group or species in which polymorphisms have been identified and any inferred or experimentally verified structure-function is listed.

Very little structure-function data has been accumulated outside of the active site residues so it is difficult to guess what functions other conserved residues might have. However, examination the MIF alignment has identified twenty-three residues that are conserved in the majority of sequences. Three of these residues line the substrate binding pocket, Presumably many of them are involved in maintaining tertiary structure. Eleven group specific synapomorphic residues were also identified. Many of these are found near residues lining the substrate binding pocket.

With the exception of CAE_MIF4, the catalytic proline has been conserved in all isolated sequences indicating the tautomerase activity will be a universal feature of MIFs. The conservation of this activity indicates it may be closely linked to the biological functions of MIF genes.

Conversely, the CXXC motif (residues 64-67 in the alignment) is limited to the metazoan MIFs and is absent from the protozoan and plant sequences. Some of the protozoan sequences to have cysteine residues at positions 64 and/or 65 which may be an ancestral state for the residues at these positions. However, whether the protozoan MIFs possess an oxioreductase activity comparable to the vertebrate MIFs is still to be determined. Interestingly most of the plant and protozoan MIFs have a cysteine (Cys3) immediately after the catalytic proline. The presence of a free-thiol close to the proline might radically change its activity or the substrate specificity of the enzyme.

In the metazoan MIFs there are a number of polymorphisms in the residues lining the substrate binding pockets. In the vertebrate MIFs and DDTs substitutions of aromatic residues for positively charged or smaller hydrophobic residues is believed to give the DDTs a preference for larger, less hydrophobic substrates. Interestingly some of these polymorphisms are shared by many of the nematode MIFs.

There are at least two examples of group specific synapomorphies in which changes occur in residues that may interact (4 and 42, 55 and 102). Whether these reciprocal changes in residues represent compensatory changes or the acquisition of novel functions remain to be elucidated.

5.3: Phylogenetic analysis of the protein and cDNA sequences of the MIF gene family

Phylogenetic analyses were performed using the protein and cDNA alignments excluding four partial sequences (MIF2_MELJA, MIF1_ANCCA, MIF2_ANCCA and LS1_TOXGO). Only characters 1-128 in the protein and 1-390 in the cDNA alignment were used in the subsequent analysis, excluding long C-terminal extensions found in several sequences (MIF1_CIOIN, MIF2_CIOIN, MIF2_CAEEL, MIF3_CAEEL, MIF2_STRSE, MIF2_ONCVO, MIF2_BRUMA, LS1_PHYSO, LS2_GOSAR and LS2_MESCR). Phylogenetic analyses were performed using distance, maximum parsimony and Bayesian methods (reviewed in (Huelsenbeck *et. al.*, 2001; Swofford *et. al.*, 1996)).

For the protein alignment distance searches using the neighbor joining (NJ) algorithm and distance parameters of total character difference were used to construct a single tree which was then tested using NJ bootstrap analysis (10,000 replicates, see figure 5.3.1).

An initial uncorrected NJ tree was used to build a likelihood model for the cDNA sequence alignment. A general time reversible model was used to estimate the actual base substitution rates. Base frequencies were calculated to remove base composition bias and the gamma distribution was calculated to determine if the rate of base substitutions was equal across all sites. This corrected matrix was then used as the evolutionary model used in the NJ analysis of the cDNA alignment. The 'corrected' NJ tree was tested by bootstrap analysis (10,000 replicates, see figure 5.3.2).

For both the protein and cDNA datasets a heuristic maximum parsimony (MP) search was performed using stepwise addition of taxa and 10 replicates. The search of the protein alignment yielded two islands containing 196 trees with a length of 2426. The search of the cDNA alignment yielded two islands containing 15056 trees with a length of 1856. A consensus tree was built from the trees saved during each search. This tree was tested by bootstrap analysis (10,000 replicates, using a heuristic algorithm, see figures 5.3.3 and 5.3.4). Both the distance and MP analyses were performed using PAUP* 4.0b8 (Sinauer Associates Inc., Sunderland, Mass).

The Bayesian analysis of Markov chain Monte Carlo estimations of maximum likelihood trees were carried out using Mr. Bayes 2.0 (Huelsenbeck and Ronquist, 2001). A Dayhoff amino acid substitution model was used for the protein searches. The analysis was run with four chains for 1,000,000 generations and the best tree saved every 100 generations. A consensus tree was generated from the saved trees excluding those found before the likelihood values stabilize (the "burn-in" generations, see figures 5.3.5 and 5.3.7). After initial examination of the consensus trees containing all of the MIF sequences several of the long branch taxa

(LS1_PLABU, MIF1_GIAIN, LS1_EMETE, LS1_PLAFA, LS1_PLAYO and LS1_PHYSO) were removed to examine whether these rapidly evolving sequences were contributing to the large polytomy at the base of the animal clade (see figures 5.3.6 and 5.3.8).

CHMI_ESCCO was chosen as the outgroup and each of the MIF sequences in the analysis were assigned to a group *a priori* depending on their origin and similarity to reference sequences. Table 5.3.9 shows a summary of the results of the phylogenetic analysis.

Metazoans:

- 🟪 Vertebrate MIFs
- 🟪 Vertebrate DDTs
- 🟦 Ecdysozoan MIF1s
- 🟨 Deuterostome MIF2s
- 🟧 Nematode MIF2s

Higher Plants:

- 🟩 Gymnosperms: Conifers
- 🟩 Angiosperms: Dicots
- 🟩 Angiosperms: Monocots

Unicellular Eukaryote and CHMI:

- 🟥 Apicomplexa
- ⬛ Other Unicellular Eukaryotes and CHMI

Figure 5.3.0 The color codes used to differentiate sequences from different groups of organisms within the animal sequences the vertebrate MIF1, nematode MIF1, vertebrate DDTs, and other non-vertebrate MIFs have been given different colors.

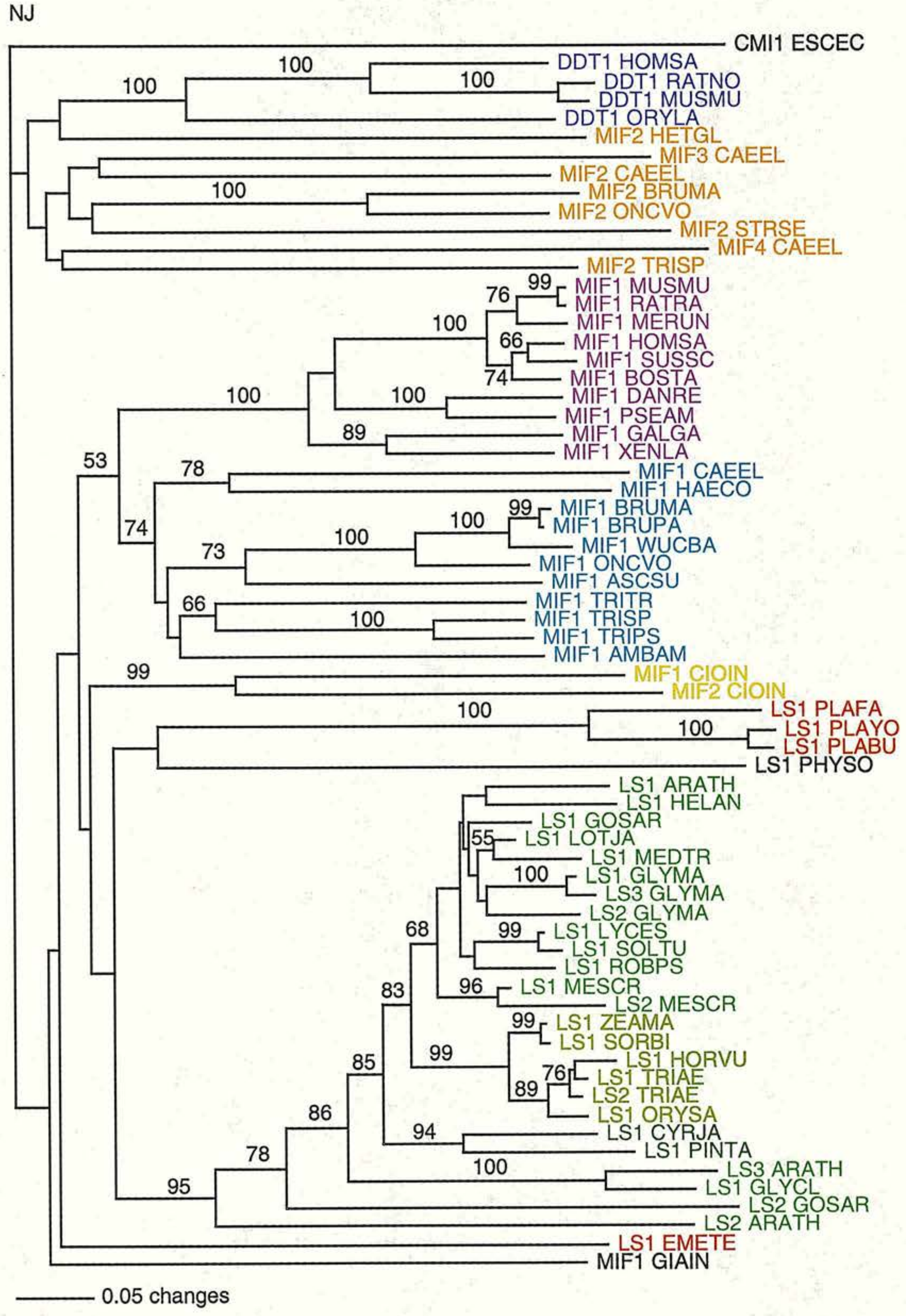


Figure 5.3.1 Rooted phylogram showing the results of NJ search of the protein alignment. Clades supported by bootstrap values >50% are indicated by bootstrap values placed at the base of the node.

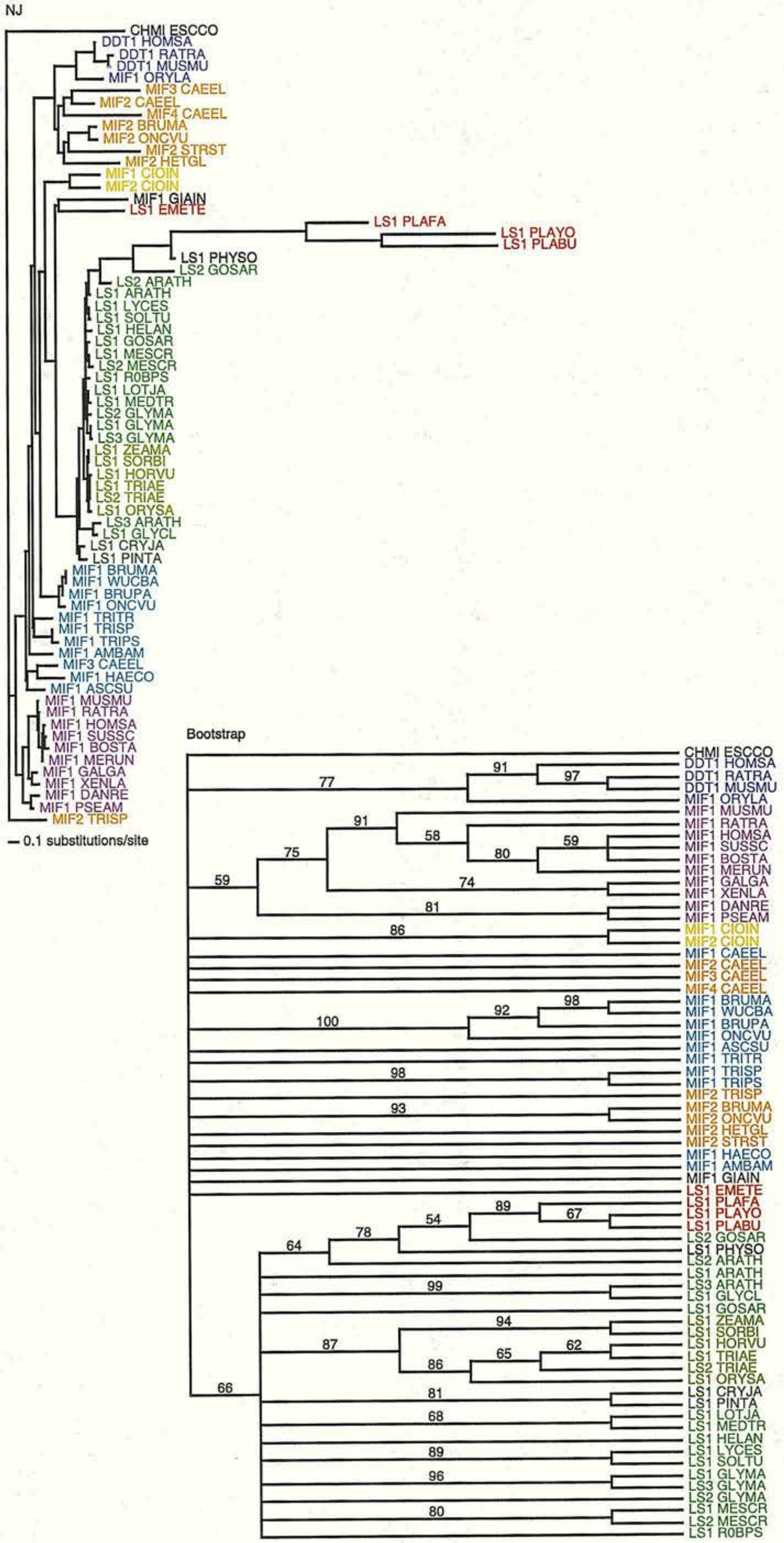


Figure 5.3.2 Rooted phylogram showing the results of corrected NJ search of the cDNA alignment. Clades supported by bootstrap values >50% are indicated by bootstrap values placed at the base of the node.

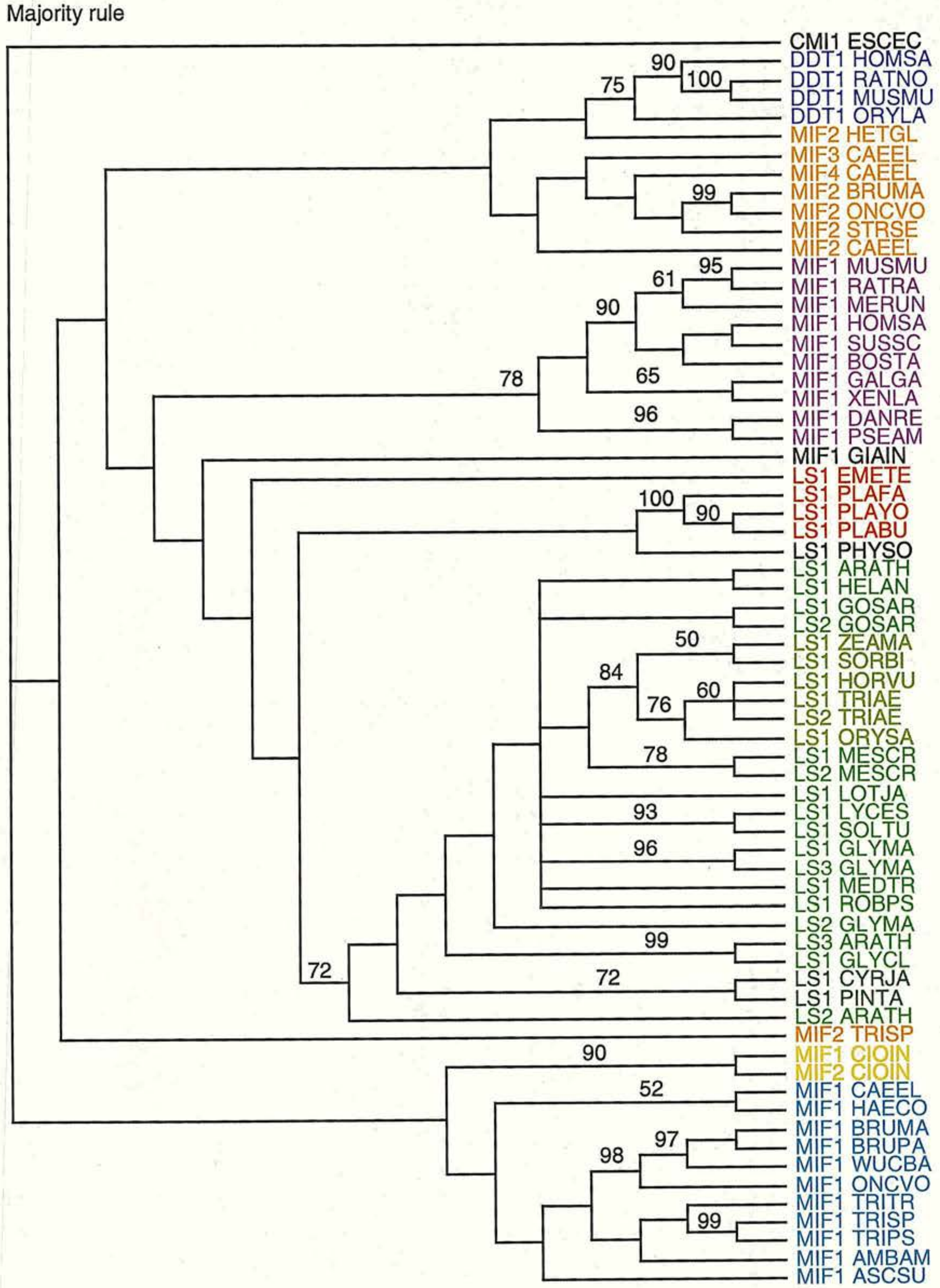


Figure 5.3.3 Rooted cladogram showing the consensus of the 196 trees found in the MP analysis of the MIF protein sequence alignment. Clades supported by bootstrap values >50% are indicated by bootstrap values placed at the base of the node.

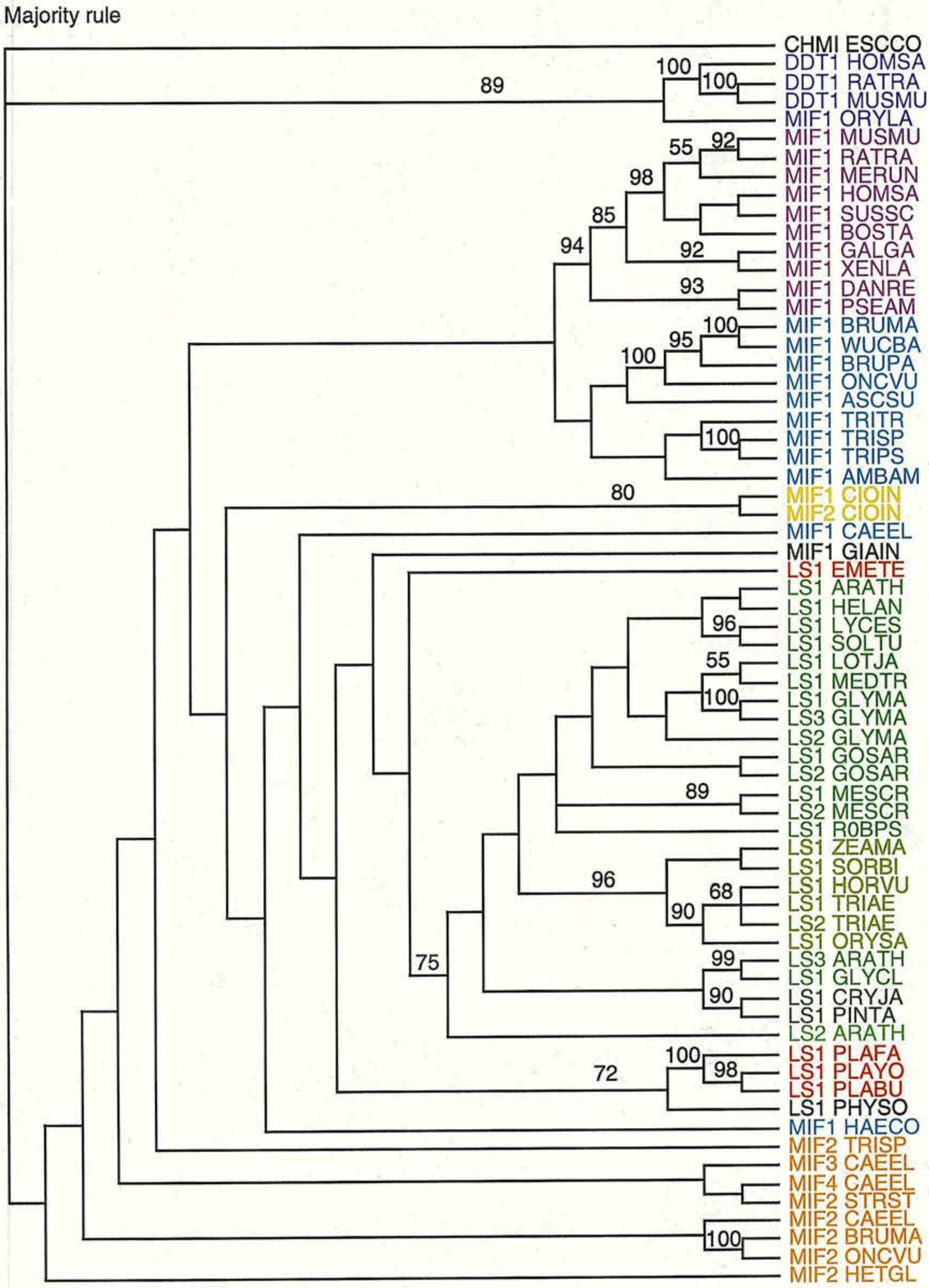


Figure 5.3.4 Rooted cladogram showing the consensus of the 15056 trees found in the MP analysis of the MIF protein sequence alignment. Clades supported by bootstrap values >50% are indicated by bootstrap values placed at the base of the node.

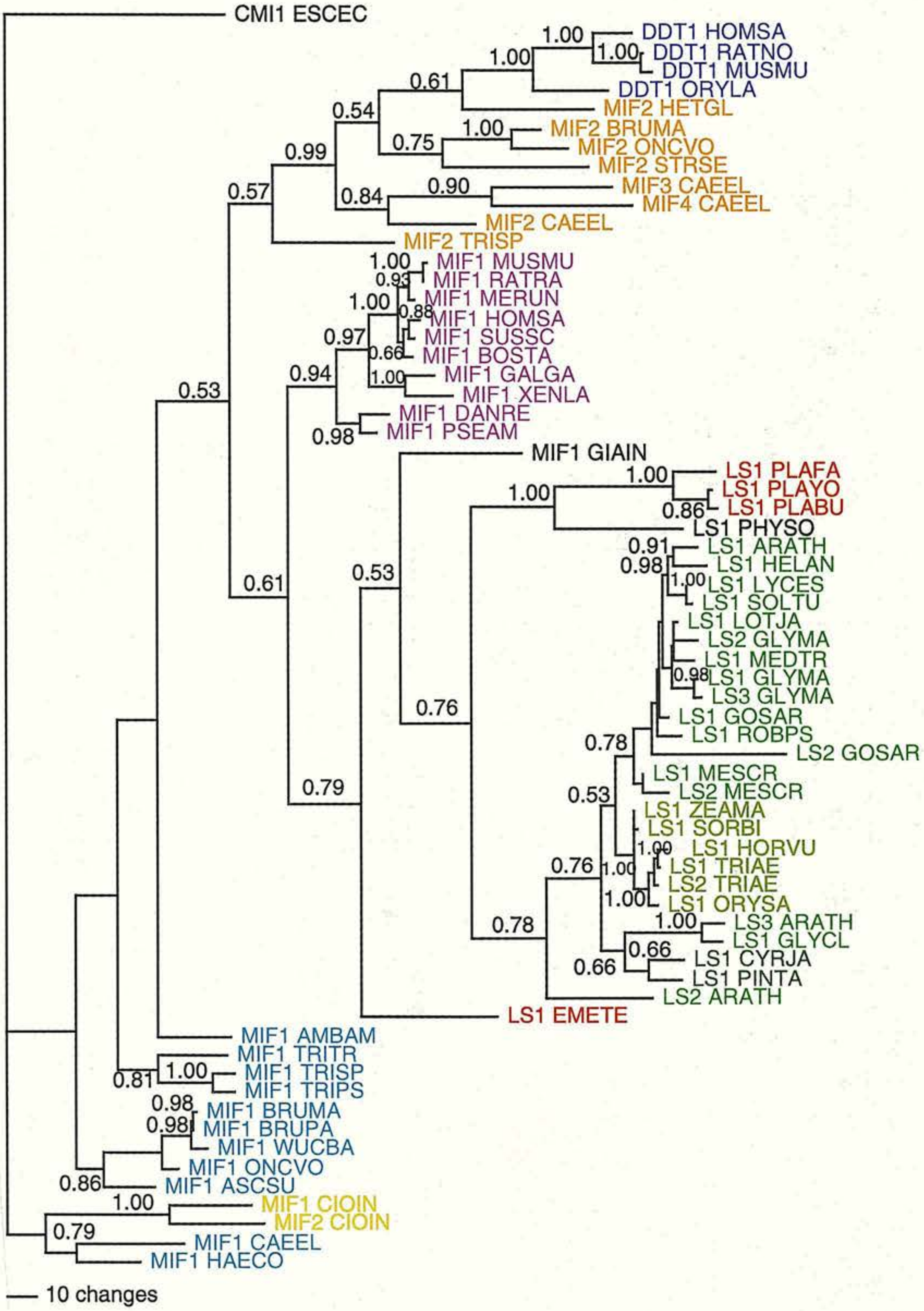


Figure 5.3.5 Rooted phylogram shows a consensus tree summarizing the results of the Bayesian analysis of the MIF protein alignment. The consensus tree was built from 999,625 best trees saved during the analysis.The posterior probability that the clade is correct is shown at the base of each node if it is > 0.50.

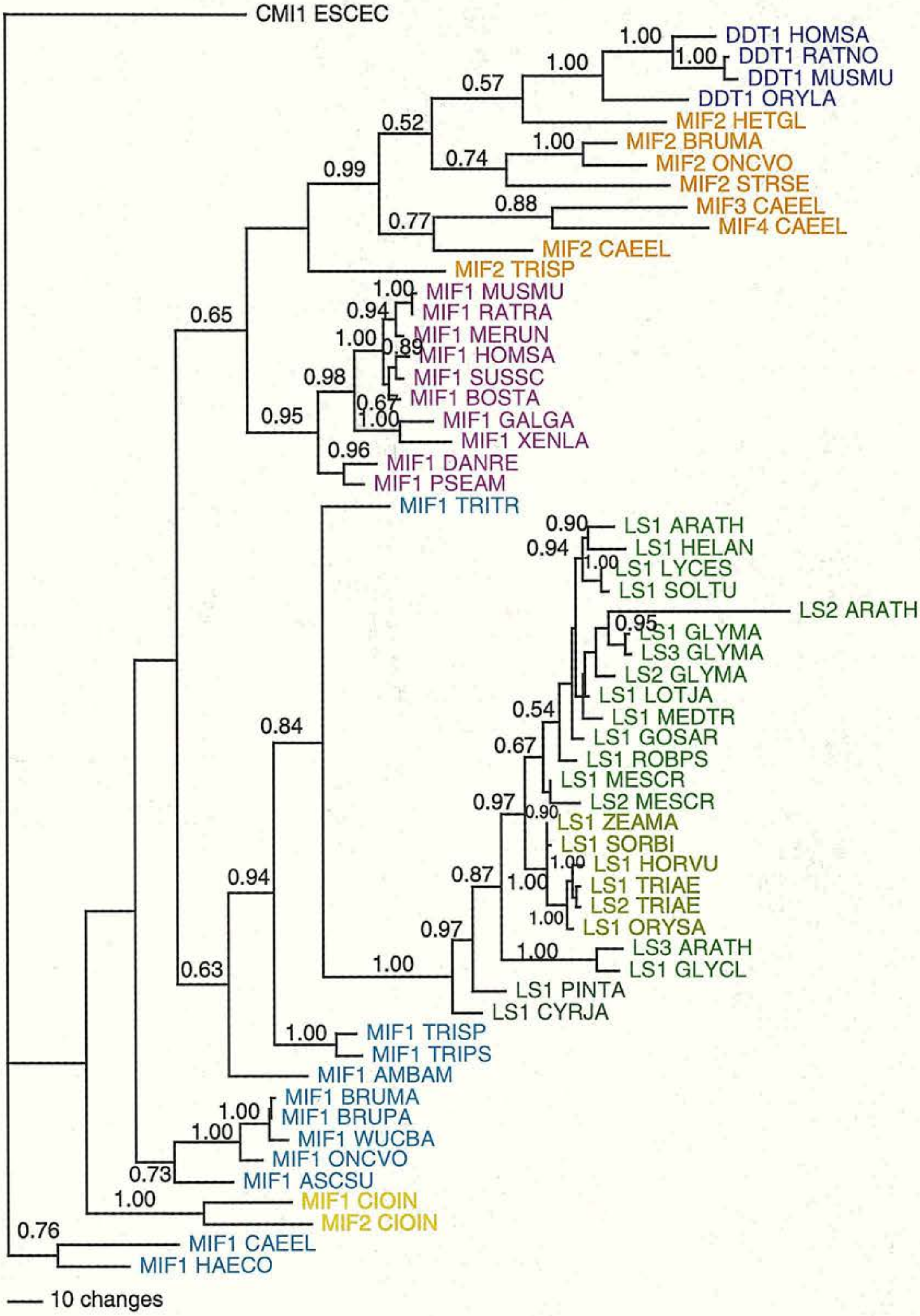


Figure 5.3.6 Rooted phylogram shows a consensus tree summarizing the results of the Bayesian analysis of the MIF protein alignment excluding the long branch protozoan sequences. The consensus tree was built from 999,037 best trees saved during the analysis.The posterior probability that the clade is correct is shown at the base of each node if it is > 0.50.

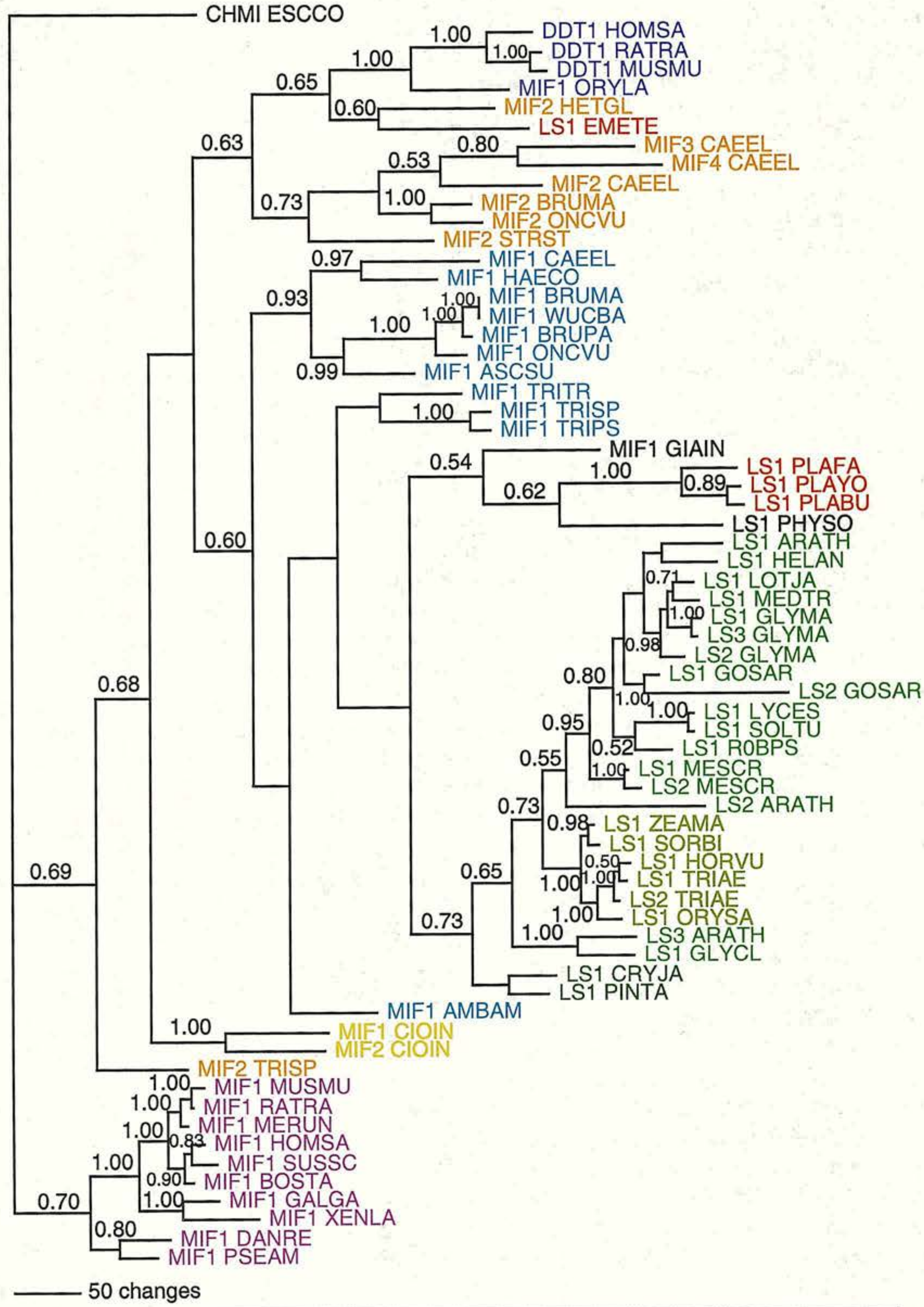


Figure 5.3.7 Rooted phylogram shows a consensus tree summarizing the results of the Bayesian analysis of the MIF cDNA alignment. The consensus tree was built from 998,630 best trees saved during the analysis.The posterior probability that the clade is correct is shown at the base of each node if it is > 0.50.

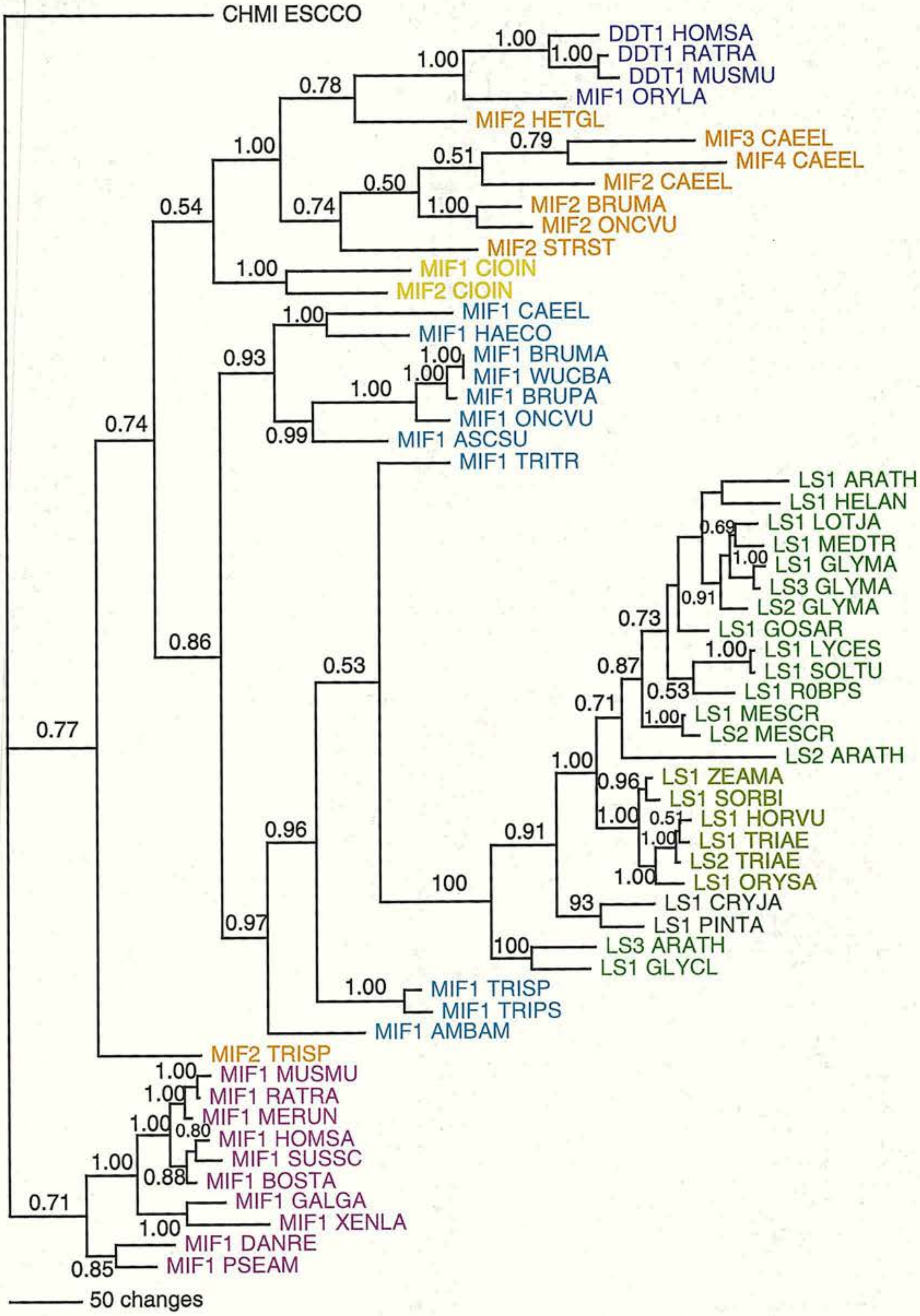


Figure 5.3.8 Rooted phylogram shows a consensus tree summarizing the results of the Bayesian analysis of the MIF cDNA alignment excluding the long branch protozoan sequences. The consensus tree was built from 998,690 best trees saved during the analysis.The posterior probability that the clade is correct is shown at the base of each node if it is > 0.50.

Clades	NJ aa	NJ cor. cDNA	MP aa	MP cDNA	MCMC aa	MCMC aa -LBT	MCMC cDNA	MCMC cDNA - LBT
Plant								
Plants separate from animals	X	X	X	X	X	X	X	X
Monocot	X	X	X	X	X	X	X	X
Gymnosperm	X	X	X	X	X	X	X	X
Plasmodium and *P. sojae*		X		X	X	NA	X	NA
Plasmodium and Plants		X			X	NA		NA
Animal								
Vertebrate MIFs	X	X	X	X	X	X	X	X
Vertebrate DDTs	X	X	X	X	X	X	X	X
Ecdysozoan MIF1	X				X	X	X	X
Secernentean MIF1s	X						X	X
Nematode MIF2					X	X	X	X
C. elegans MIF2,3,4					X	X		
Vertebrate DDTs with nematode MIF2					X	X	X	X
Ecdysozoan MIF1s basal to other animal MIF clades						X		X

Table 5.3.9 Summary of the results of the phylogenetic analyses of the MIF protein and cDNA sequences. The topology of the trees is broken down into clades. Clades are supported by high bootstrap scores (> 65%) or posterior probability values (> 0.65) are marked with an X. -LBT: excluding long branch protozoan sequences. NA: not applicable

The overall topology of trees found in the phylogenetic analyses of the MIF sequences were very similar. However, nodes at the base of the animal portion of the trees were generally not supported by NJ and MP bootstrap analyses. The plant sequences were robustly placed in a single clade. Within this clade the monocot and gymnosperm sequences were placed in separate groups from the dicot sequences. The gymnosperm sequences were always placed near the base of the plant clade. However, in some of the analyses several of the dicot sequences with long branch lengths superceded them at the base of the clade.

The placement of the protozoan sequences is more problematic with the *Plasmodium sp.* and *P. sojae* sequences being placed with the plant sequences in some analyses or rooting from the polytomy at the node separating the plant and animal groups and the *G. intestinalis* and *E. tenella* sequence. Interestingly, the *E. tenella* sequence was never placed within the *Plasmodium sp.* and *P. sojae* clade. This result is unexpected as *E. tenella* is also an Apicomplexan.

In the animal section of the tree the vertebrate sequences were consistently placed in two clades which separated the MIF and DDT sequences. The results of the NJ protein and MCMC searches indicate that nematode, *C. intestinalis* and *A. americanum* can be divided into at least two major clades. The first clade contains the nematode MIF2 sequences. The second clade (ecdysozoan MIFs) is a loosely associated set of sequences which root from a large polytomy at the node separating the plant and animal sequences. This group includes the nematode and the *A. americanum* MIF1s. The relationship of the sequences within this clade remains largely unresolved although several small clades of closely related sequences have been identified. The results of the MCMC searches place the vertebrate DDTs deep within the nematode MIF2 clade (see figures 5.3.5-8) The position of the *C. intestinalis* MIFs varies between analyses and it is unclear whether they belong to the ecdysozoan or MIF2/DDT groups.

The NJ analysis of the protein sequences and the MCMC analysis of the cDNA sequences have grouped the nematode MIF1s belonging to Secernentean species into a single clade. This indicates they may represent an orthologous set of genes. Similarly the MCMC analysis of the protein sequences placed *C. elegans* MIF2, 3 and 4 in a single clade with *Ce*-MIF-2 rooting at the base of the clade. This indicates *Ce*-MIF-3 and 4 may have arisen from an ancestral gene duplication of *Ce*-MIF-2.

How these major clades should be rooted in relation to each is not entirely clear. In the searches that contain all the sequences the major clades branch from a polytomy at the base of the animal clade. However, the MCMC searches which do not include the protozoan sequences indicate that the ecdysozoan MIFs (or several members of this group) root at the base of the animal portion of the tree with Secernentean MIFs, vertebrate MIF1s and the MIF2/DDT clades branching off from these sequences.

One of the most interesting features of the MIF phylogenetic trees is the observed differences in the rate of change in sequences belonging to different groups (as reflected in branch lengths in trees rooted with *Ec*-CMI-1). For instance the vertebrate DDT sequences have much longer branches than the vertebrate MIFs indicating that they are evolving much more quickly. A similar comparison of the plant and animal sequences show that the plant sequences have remained relatively static when compared to the animal sequences. These differences in relative rates may reflect functional constraints that are holding sequences constant in the groups that show high conservation or a burst of changes that reflect the acquisition of new functionalities in the rapidly evolving groups.

5.6: Evolution of Intron sequences in the MIF gene family

Intron placement and movement within genes families can provide another layer of information that can complement analysis of the gene sequences (see figure 5.6.1).

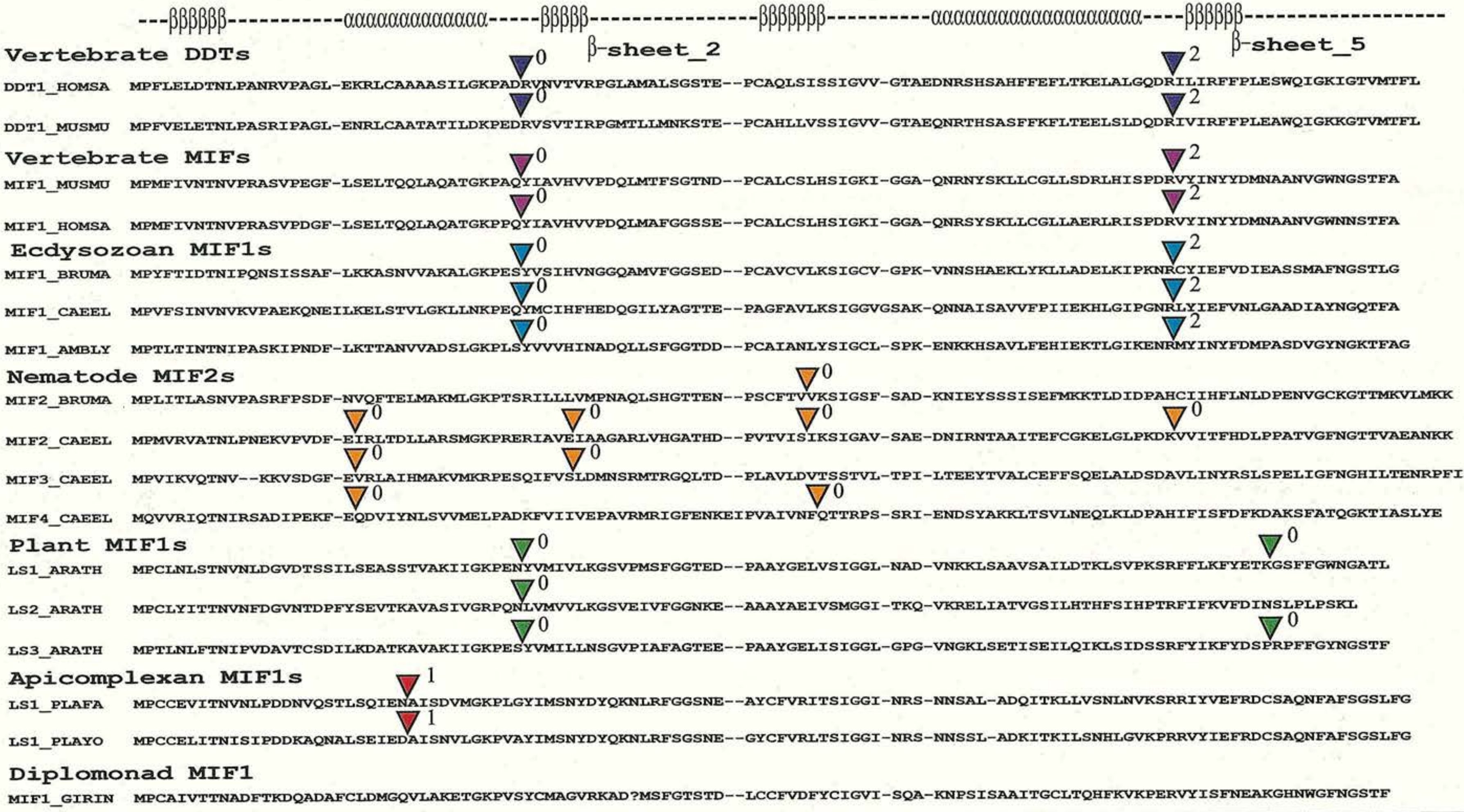

---ββββββ-----------ααααααααααααα-----βββββ---------------βββββββ----------ααααααααααααααααααα----ββββββ------------------
Vertebrate DDTs
β-sheet_2
β-sheet_5
0
2
DDT1_HOMSA MPFLELDTNLPANRVPAGL-EKRLCAAAASILGKPADRVNVTVRPGLAMALSGSTE--PCAQLSISSIGVV-GTAEDNRSHSAHFFEFLTKELALGQDRILIRFFPLESWQIGKIGTVMTFL
0
2
DDT1_MUSMU MPFVELETNLPASRIPAGL-ENRLCAATATILDKPEDRVSVTIRPGMTLLMNKSTE--PCAHLLVSSIGVV-GTAEQNRTHSASFFKFLTEELSLDQDRIVIRFFPLEAWQIGKKGTVMTFL
Vertebrate MIFs
0
2
MIF1_MUSMU MPMFIVNTNVPRASVPEGF-LSELTQQLAQATGKPAQYIAVHVVPDQLMTFSGTND--PCALCSLHSIGKI-GGA-QNRNYSKLLCGLLSDRLHISPDRVYINYYDMNAANVGWNGSTFA
0
2
MIF1_HOMSA MPMFIVNTNVPRASVPDGF-LSELTQQLAQATGKPPQYIAVHVVPDQLMAFGGSSE--PCALCSLHSIGKI-GGA-QNRSYSKLLCGLLAERLRISPDRVYINYYDMNAANVGWNNSTFA
Ecdysozoan MIF1s
0
2
MIF1_BRUMA MPYFTIDTNIPQNSISSAF-LKKASNVVAKALGKPESYVSIHVNGGQAMVFGGSED--PCAVCVLKSIGCV-GPK-VNNSHAEKLYKLLADELKIPKNRCYIEFVDIEASSMAFNGSTLG
0
2
MIF1_CAEEL MPVFSINVNVKVPAEKQNEILKELSTVLGKLLNKPEQYMCIHFHEDQGILYAGTTE--PAGFAVLKSIGGVGSAK-QNNAISAVVFPIIEKHLGIPGNRLYIEFVNLGAADIAYNGQTFA
0
2
MIF1_AMBLY MPTLTINTNIPASKIPNDF-LKTTANVVADSLGKPLSYVVVHINADQLLSFGGTDD--PCAIANLYSIGCL-SPK-ENKKHSAVLFEHIEKTLGIKENRMYINYFDMPASDVGYNGKTFAG
Nematode MIF2s
0
MIF2_BRUMA MPLITLASNVPASRFPSDF-NVQFTELMAKMLGKPTSRILLLVMPNAQLSHGTTEN--PSCFTVVKSIGSF-SAD-KNIEYSSSISEFMKKTLDIDPAHCIIHFLNLDPENVGCKGTTMKVLMKK
0
0
0
0
MIF2_CAEEL MPMVRVATNLPNEKVPVDF-EIRLTDLLARSMGKPRERIAVEIAAGARLVHGATHD--PVTVISIKSIGAV-SAE-DNIRNTAAITEFCGKELGLPKDKVVITFHDLPPATVGFNGTTVAEANKK
0
0
MIF3_CAEEL MPVIKVQTNV--KKVSDGF-EVRLAIHMAKVMKRPESQIFVSLDMNSRMTRGQLTD--PLAVLDVTSSTVL-TPI-LTEEYTVALCEFFSQELALDSDAVLINYRSLSPELIGFNGHILTENRPFI*
0
0
MIF4_CAEEL MQVVRIQTNIRSADIPEKF-EQDVIYNLSVVMELPADKFVIIVEPAVRMRIGFENKEIPVAIVNFQTTRPS-SRI-ENDSYAKKLTSVLNEQLKLDPAHIFISFDFKDAKSFATQGKTIASLYE
Plant MIF1s
0
0
LS1_ARATH MPCLNLSTNVNLDGVDTSSILSEASSTVAKIIGKPENYVMIVLKGSVPMSFGGTED--PAAYGELVSIGGL-NAD-VNKKLSAAVSAILDTKLSVPKSRFFLKFYETKGSFFGWNGATL
0
LS2_ARATH MPCLYITTNVNFDGVNTDPFYSEVTKAVASIVGRPQNLVMVVLKGSVEIVFGGNKE--AAAYAEIVSMGGI-TKQ-VKRELIATVGSILHTHFSIHPTRFIFKVFDINSLPLPSKL
0
0
LS3_ARATH MPTLNLFTNIPVDAVTCSDILKDATKAVAKIIGKPESYVMILLNSGVPIAFAGTEE--PAAYGELISIGGL-GPG-VNGKLSETISEILQIKLSIDSSRFYIKFYDSPRPFFGYNGSTF
Apicomplexan MIF1s
1
LS1_PLAFA MPCCEVITNVNLPDDNVQSTLSQIENAISDVMGKPLGYIMSNYDYQKNLRFGGSNE--AYCFVRITSIGGI-NRS-NNSAL-ADQITKLLVSNLNVKSRRIYVEFRDCSAQNFAFSGSLFG
1
LS1_PLAYO MPCCELITNISIPDDKAQNALSEIEDAISNVLGKPVAYIMSNYDYQKNLRFSGSNE--GYCFVRLTSIGGI-NRS-NNSSL-ADKITKILSNHLGVKPRRVYIEFRDCSAQNFAFSGSLFG
Diplomonad MIF1
MIF1_GIRIN MPCAIVTTNADFTKDQADAFCLDMGQVLAKETGKPVSYCMAGVRKAD?MSFGTSTD--LCCFVDFYCIGVI-SQA-KNPSISAAITGCLTQHFKVKPERVYISFNEAKGHNWGFNGSTF

Figure 5.6.1 Comparison of the intron positions found in MIF genes. The genomic sequence of HOMSA_DOT1 (AF058293), MUSMU_DOT1 (AF068199), HOMSA_MIF1 (L39357), MUSMU_MIF1 (L19686)), BRUMA_MIF1 (AF002699), CAEEL_MIF1 (AL132860), AMBLY_MIF1 (AF289543), BRUMA_MIF2, CAEEL_MIF2 (Z78012), CAEEL_MIF3 (AC084197), CAEEL_MIF4 (AC084197), ARATH_LS1 (AL161946), ARATH_LS2 (AL132968), ARATH_LS3 (BA000025), PLAFA_LS1, PLAYO_LS1, and GIAIN_MIF1 are shown as they were aligned . The intron positions and phase are shown with colored triangles and numbers. * The C-terminal extension of CAEEL_MIF3 which does not contain any introns has been removed. GIAIN_MIF1 does not contain any introns within its genomic sequence. The secondary structure of the *H. sapiens* MIF1 is shown at the top of the diagram. α indicates alpha-helical and β indicates beta-sheet structures. The sequences have been broken up into groupings derived from the phylogenetic analysis.

Genomic sequence was available for seventeen MIF genes, from nine different species. Each gene had between zero and four introns, with the majority of the genes having two introns. Comparison of the intron positions showed that several are shared by the MIF groups. The vertebrate DDTs, vertebrate MIFs, ecdysozoan MIF1s and the plant MIFs all share a common first intron position with phase zero. The vertebrate DDTs, vertebrate MIFs and ecdysozoan MIF1s share a common second phase zero intron position with a phase two. A unique second intron was found in two plant sequences in the C-terminal portion of the molecule. The nematode MIF2 sequences have a different set of intron positions relative to the other metazoan MIF genes. There is a great deal of variability in the conservation of these introns in different nematode sequences (figure 5.6.1). Interestingly, the fourth intron of CAEEL_MIF2 shares the same position as the second intron of the vertebrate DDTs, vertebrate MIFs and ecdysozoan MIF1s. However they do not share the same phase. The apicomplexan MIFs both have a single uniquely positioned intron at the N-terminus of the gene with a phase of one, while the *Giardia* MIF does not have any introns.

5.7: Discussion

Despite having been discovered almost 40 years, ago MIFs remain a very enigmatic set of molecules. In vertebrates the MIF gene family is an extremely important group of regulatory molecules which function in a variety of different contexts. In the mammalian immune system MIF functions as a signaling molecule that controls the initiation and scope of some types of immune responses. MIF and its interactions with glucocorticoids is particularly important in the initiation of the innate immune response after exposure to bacterial products (Froidevaux *et. al.*, 2001). The mechanisms through which MIF exerts its activities still remain largely unresolved. Unlike most cytokines, MIF as well as being a signaling molecule has two enzymatic activities (Kleemann *et. al.*, 1998a; Kleemann *et. al.*, 1998b; Sugimoto *et. al.*, 1999). These enzymatic activities are important for some but not all of MIF's effects on cells. The functions of MIFs in other organisms remain almost completely unexplored. Several nematode MIFs have been isolated from a variety of parasitic species in contexts that indicate they have both endogenous and exogenous functions. The endogenous functions of these MIFs is currently being tested in the model nematode *C. elegans* whose genome contains four MIF genes. A preliminary study indicates that nematode MIFs may have roles in embryogenesis, tissue differentiation and stress responses although gene knock out by RNAi has not revealed any phenotype (Marson *et. al.*, 2001). The functions of MIFs secreted into hosts by parasitic species are also being explored. Nematode MIFs have been shown to have effects on mammalian immune cells in both *in vivo* and *in vitro* assays. It is speculated that they are involved in the manipulation of the host immune system by the parasite, however definitive links between these molecules and effects on host immune system are still to be established (Falcone *et. al.*, 2001).

5.7.1 The distribution of the MIF gene family

One of the major goals of this study was to survey of the distribution of the MIF gene family. While MIFs have been cloned and characterized from several vertebrate and nematode species a preliminary survey of the whole genome and EST datasets revealed that there were many MIF-like gene sequences in organisms from which they have not been previously identified (see Figure 4.7.1.1). Sequences have been identified in many of the major eukaryote groups including the Diplomonadida, Apicomplexa, Heterokonts, Viridiplantae and Metazoa. No MIF like sequences were identified from the Euglenoidea or Fungi. Both of these groups (particularly the fungi) have whole or partial genome sequence available from several species. Until more whole genome sequences become publicly available it will remain unclear if this gene family has been completely lost from those groups. Searching of the sequenced portion of the *D. melanogaster* and *Anopheles gambiae* genomes has failed to identify any MIF genes. This indicates that within the metazoa there may be some species that have lost their MIF genes. The distribution of MIFs through so many evolutionarily distant phyla supports the possibility that MIF-like genes are likely to be found in most eukaryote genomes. None of the isolated MIFs show significant similarity to the structurally similar bacterial genes 4-OT or CHMI and thus it is possible that MIFs arose after the separation of eukaryotes from the bacterial lineages.

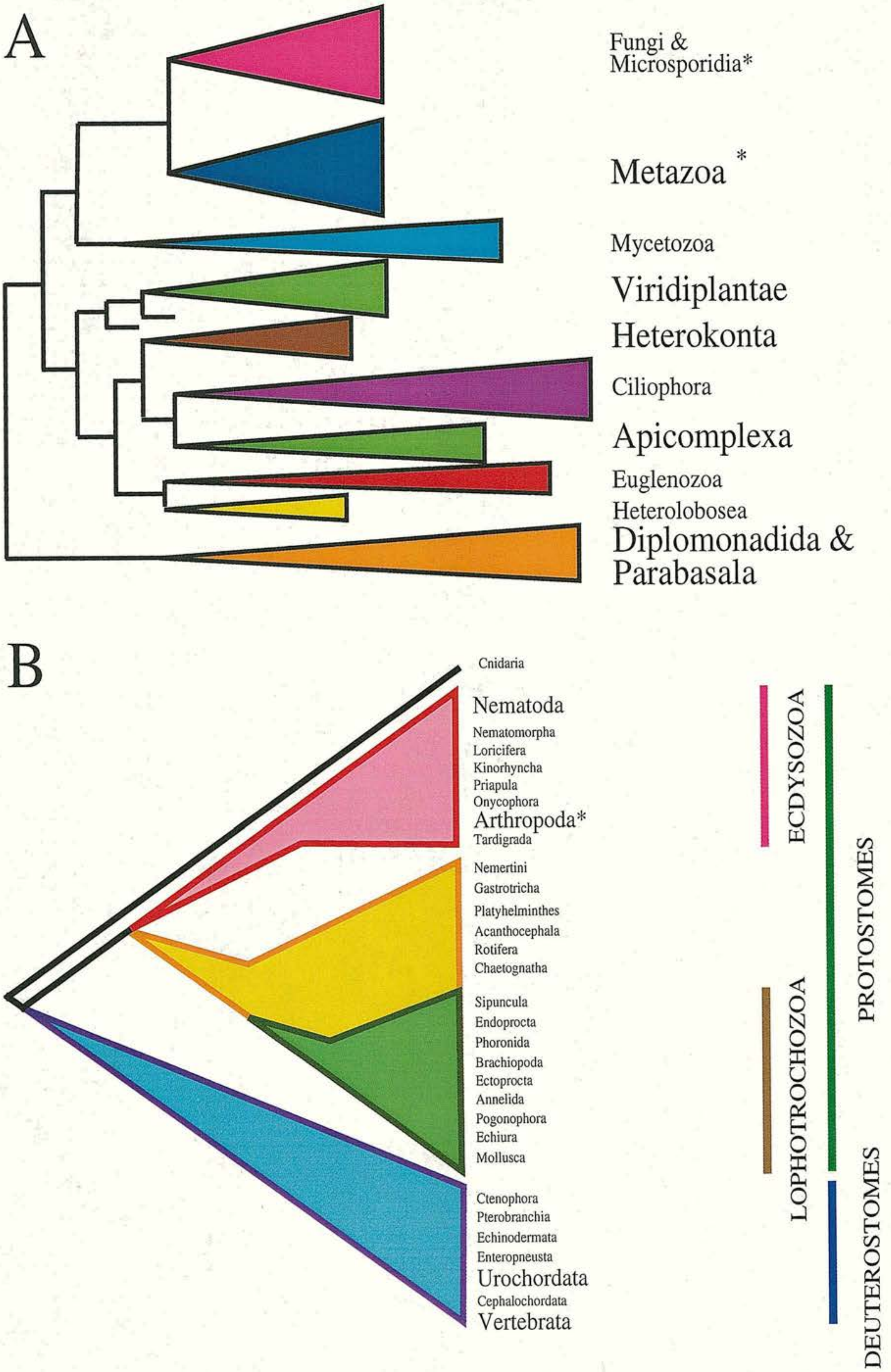


Figure 5.7.1.1 Distribution of the identified MIF gene family members across eukaryotes. Those groups from which MIF-like sequences have been identified are marked in bold. **A:** the phylogenetic distribution of MIF sequences through the eukaryotes adapted from Baldauf *et. al.* 2000. **B:** the phylogenetic distribution of MIF sequences through the metazoa adapted from Blaxter 1998. * indicates that searching of whole genome sequence from a species within this group has failed to yield any MIF gene family members. Arthropods: *Drosophila melanogaster* and *Anopheles gambiae* Fungi: *Candida albicans, Saccharomyces cerevisiae* and *Saccharomyces pombe* Microsporidian: *Encephalitozoon cuniculi.*

5.7.2 Analysis of the aligned MIFsequences

Analysis of the aligned MIF sequences has identified several interesting features. First the catalytic proline required for MIFs tautomerase activity is conserved in almost all MIF sequences identified, while the CXXC active site which is necessary for thiol-oxioreductase activity seen in some vertebrate MIFs is confined to a limited number of metazoan sequences. This indicates that the thiol-oxioreductase activity may be an innovation seen exclusively in metazoans.

When residues lining the substrate binding site are examined, those which vary between groups have given some clue as to what types of substrates the different enzymes might prefer. The substrate binding sites of the vertebrate MIFs, ecdysozoan MIF1s and plant MIFs are rich in hydrophobic and aromatic residues indicating a potential preference for smaller hydrophobic substrates. Conversely, in the vertebrate DDTs and nematode MIF2s many of aromatic residues have been replaced with smaller hydrophobic or positively charged residues indicating these enzymes could accommodate larger, more polar substrates. The protozoan sequences show much more variability in these residues than the plant and animal MIFs. MIFs from *G. intestinalis* and *E. tenella* tend to share more residues with the animal sequences while the *Plasmodium sp.* and *P. sojae* sequences share more residues with the plant sequences. However, until enzymatically active recombinant MIFs and a panel of suitable substrates are available it is impossible to test exactly how these substitutions have affected the activities of the enzyme.

5.7.3 Phylogenetic analysis of the MIF gene family

The results of these phylogenetic analyses of the MIF genes has given some clues as to how this family has evolved. Three different phylogenetic methods, distance (NJ), parsimony (MP) and Bayesian (MCMC) were used in the analysis of the dataset. All three techniques gave similar results that can be summarized as follows (Figure 5.7.2.1). The metazoan and plant sequences reproducibly partition into separate clades indicating that MIF genes have been duplicated independently in both lineages. Within the metazoan clade four groups are seen vertebrate MIFs, vertebrate DDTs, ecdysozoan MIF1s and nematode MIF2s. Depending on the analysis, the inferred relationship of these groups to each other is variable. However, there is some indication that the ecdysozoan MIF1s may represent a basal lineage from which the other three groups have evolved. The topology of the trees within the vertebrate MIF1 and DDT groups are well supported and consistent with previously published data on the evolution of the vertebrate species. In general there is very little resolution within the ecdysozoan MIF1s with most of the sequences rooting from a polytomy which separates the plant and metazoan sequences. However, the results of several analyses support the Secernentean MIF1 (SC) group which contains sequences originating from species placed in nematode clades III and V (Blaxter *et. al.*, 1998). If this group is genuine then this would indicate most likely orthologue for the filarial MIF1s in the *C. elegans* genome is *Ce*-MIF-1. One of the most interesting features of the SC group is the presence of the CXXC motif in the filarial and ascaris sequences (clade III) and its absence in the clade V nematodes *C. elegans* and *H. contortus*. Have the clade V nematodes lost the CXXC motif or have the clade III nematodes acquired it by convergent evolution with the vertebrate sequences? Only one other MIF sequence retains a CXXC motif *Ci*-MIF-1. However the *Ciona* MIF1 motif is divergent from vertebrate and nematode motifs (Cys-Ala-Cys-Cys versus Cys-Ala-Leu/Ile-Cys). How the presence of the extra cysteine residue will affect the thiol-oxioreductase activity of the enzyme is unknown.

Interestingly, the results of the MCMC analyses place the vertebrate DDTs deep within the nematode MIF2 group. Whether this placement is a genuine reflection of the evolution of the vertebrate DDTs from a protostome MIF2 ancestor or whether other factors such as the convergent evolution of the nematode MIF2 and vertebrate DDT sequences are obscuring the proper placement of these sequences has not been resolved.

None of the phylogenetic techniques used could resolve the placement of the protozoan sequences. The *P. sojae* MIF was consistently placed with the *Plasmodium sp.* sequences. Contrary to expectation the *E. tenella* MIF was never placed with the other apicomplexan MIFs. It is possible that there is more than one MIF gene in the apicomplexan genome and that the sequences present in the analysis do not represent orthologues gene sets. However, ninety-nine percent of the *P. falciparium* genome sequence is currently available in the public databases which makes it unlikely that a second MIF will be identified in its genome. Another possibility is that selective pressures have pushed the sequences of the apicomplexan MIFs away from each other. Perhaps the *Plasmodium sp* and *P. sojae* MIFs are involved in similar physiological processes which has driven them to evolve plant-like features which are absent from the *E. tenella* MIF.

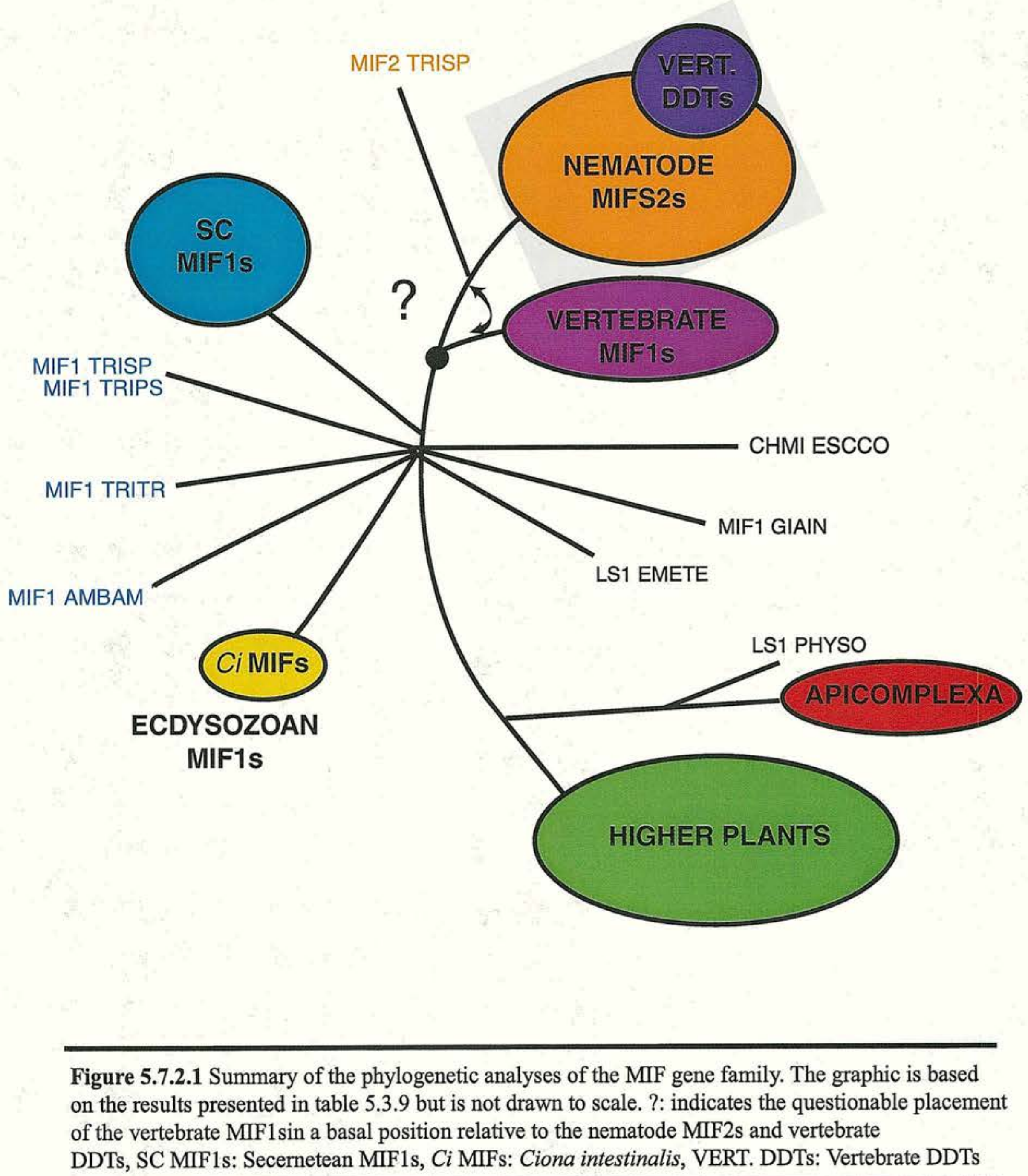


Figure 5.7.2.1 Summary of the phylogenetic analyses of the MIF gene family. The graphic is based on the results presented in table 5.3.9 but is not drawn to scale. ?: indicates the questionable placement of the vertebrate MIF1sin a basal position relative to the nematode MIF2s and vertebrate DDTs, SC MIF1s: Secernetean MIF1s, *Ci* MIFs: *Ciona intestinalis*, VERT. DDTs: Vertebrate DDTs

5.7.3 The evolution of MIF introns and general structural features

Several very important features are obvious from the examination of the placement of the MIF introns. The first is the conservation of an intron position and phase across the vertebrate DDTs, vertebrate MIFs, ecdysozoan MIF1s and the plant MIFs. Comparison of the sequences with MIF tertiary structures indicates that this intron lies before the start of β-sheet 2. The second is the conservation of a second intron position (and phase) within the vertebrate DDTs, vertebrate MIFs and ecdysozoan MIF1s. Comparison of the sequences with MIF tertiary structures indicates that this intron lies before the start of β-sheet 5. It has been speculated that the MIF gene arose from a fusion of two β-α-β monomers which would have previously functioned as a hexamer (i.e similar to the bacterial enzyme 4-OT). The placement of these introns gives some support to the idea that an ancestral β-α-intron-β gene may have undergone a duplication event which lead to the current β-α-intron-β–β-α-intron-β MIF gene structure. A second structural feature that supports the idea of an ancestral MIF with a hexameric structure is the placement of the active site residues. Both the Pro2 and CXXC motif are found before the β-sheet that would have begun each subunit. Interestingly, a highly conserved proline residue is found immediately before the CXXC motif. Could this residue be the proline that would have served as the catalytic site in the duplicated segment of the ancestral MIF? While this second proline has not been linked to the tautomerase activity of MIF it conservation indicates it must still serve an important function.

The evolutionary processes influencing intron movement are not well understood. The heterogeneity of intron positions in the nematode MIF2s has made it difficult to compare them to the other nematode MIFs. It is possible intron movement has been constrained in the vertebrates but not the nematodes. Characterization of the genomic sequences of MIFs from a other non-vertebrates such as *C. intestinalis* may help resolve whether these intron positions are nematode specific.

5.8 Conclusions

During the course of this study several important findings have been made. MIF genes have been identified in species from most of the major eukaryote groups, suggesting an ancient origin for this gene family. Phylogenetic analysis indicates that within the metazoa at least two gene families are conserved in both the vertebrate and non-vertebrate species. Three of these groups have common intron positions one of which is shared with the plant MIFs. The intron data along with some of the phylogenetic analyses suggests that the ecdysozoan MIF1s may represent an ancestral group of the metazoan MIFs. However, several problems have complicated this study. The short length of the MIF gene and the low level of conservation of the MIF sequences made robust phylogenetic studies difficult. Also several major questions still remain to be answered. Do the phylogenies derived from the analysis of the protein sequences reflect the true relationships of the MIF genes or has convergent evolution of important features such as the active site residues or substrate binding pockets obscured the true relationships of these proteins? Why do there appear to be accelerated rates of evolution in some MIF groups and what can this tell us about how their functions are evolving? Where do the protozoan sequences fit into the MIF phylogeny? As MIF sequences and functional studies become available from additional species particularly plants, free-living protozoa and other non-vertebrate metazoa the some of these questions and relationship of the different MIF families to each other may be elucidated.

Chapter 6

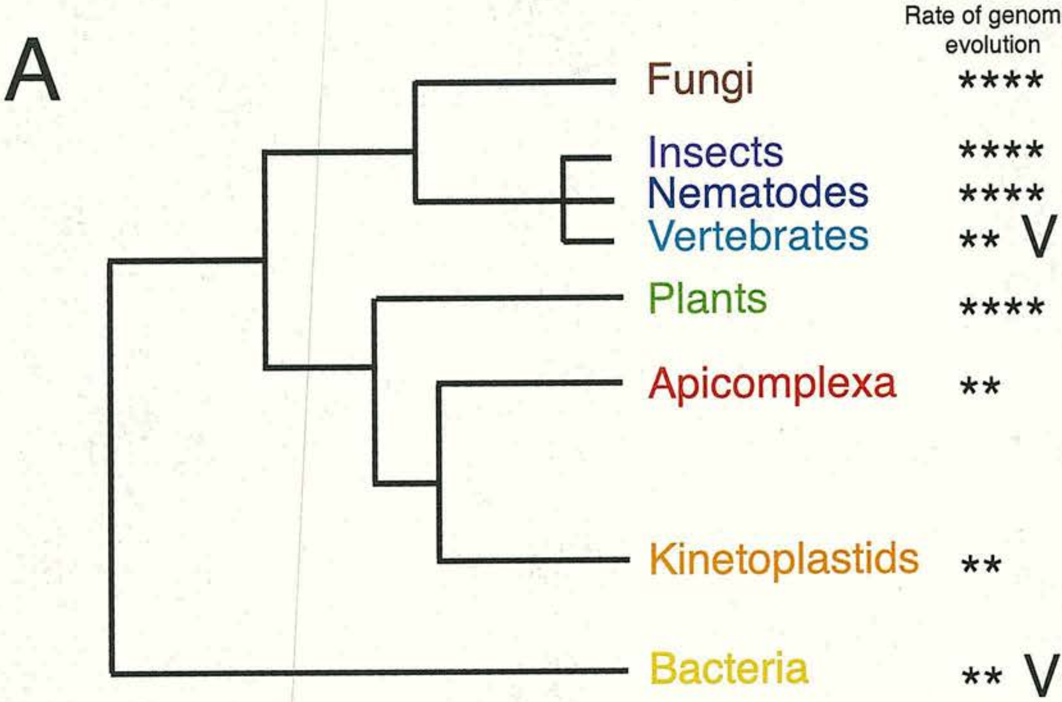
Conservation of linkage and conservation of synteny between nematode genomes

6.0 Introduction

All genomes encode a set of conserved genes that are shared with other organisms. The arrangement of these genes on the chromosomes is determined by a mixture of stochastic rearrangement and functional constraint selecting for linkage and synteny. Rearrangement events will tend to randomize gene order over time and therefore conservation of gene order (synteny) probably reflects either recent shared ancestry or functional constraint. Analysis of these patterns of the evolution in genomes can identify what forces shape their size, composition and organization. Very little is known about how nematode genomes have evolved. Current studies extend between two closely related species: *Caenorhabditis elegans* and *Caenorhabditis briggsae*. The goal of this study was to expand this survey and analyze sections of a genome from a nematode distantly related to the rhabditids. Comparison of the two genomes will help elucidate how the genomes of these organisms are changing and the rate with which these changes are taking place.

6.01 Linkage conservations and conservation of synteny between genomes

Some gene clusters have been conserved because of functional constraint. In metazoa these include the histone cluster (Hentschel and Birnstiel, 1981), the Hox cluster (Ferrier and Holland, 2001), the immunoglobulin cluster (Litman *et. al.*, 1993) and the MHC cluster (Ohta *et. al.*, 2000). Conservation of the order of functionally unrelated genes has been found in the comparison of many genomes. Does this conservation imply hitherto uncharacterized functional constraints that are keeping genes clustered? The processes underlying the dynamics of genome rearrangements are still being defined. However, comparisons of different organisms indicate the mechanics and tempo of these changes may vary dramatically between groups. Figure 6.0.1.1 summarizes what has been observed when the genomes of different groups organisms have been compared.



B

Group	Observed mode of genome evolution
Bacteria	Horizontal transfer, large inversions, genome reduction
Kinetoplastids	Diversity in subtelomeric regions, variable size in non-coding repeats, but overall conservation of coding portions of chromosomes
Apicomplexa	Diversity in subtelomeric regions but overall conservation of coding portions of chromosomes
Fungi	Duplication of genome segments, inversions and occasional chromosomal translocations
Plants	Frequent polypoloidy, frequent inversions, very little conservation between distantly related species
Insects	Large paracentric inversions but rough conservation of linkage groups, chromosome fissions and fusions
Nematodes	Conservation of linkage groups but frequent breaks in synteny
Vertebrates	Conservation large segments , translocation of large segments, chromosome fissions and fusions

Figure 6.0.1.1 Summary of the rate and mode observed in evolution of genome. **A:** a phylogentic tree showing the relationship of the different groups of organsims and the rate of evolution seen when their genomes are compared. ****: high rate **: low rate and V: indicates variable rates in different species examined. **B:** Table summarizing the mode of genome evolution within each groups.

The comparison of bacterial genomes has shown a surprising level of conservation of gene order between many related species. The evolution of these genomes appears to be driven by three major forces; horizontal transfer of DNA from other species or strains via mobile genetic elements (Brown *et. al.*, 2001; Edwards *et. al.*, 2002; Koonin *et. al.*, 2001), large inversions which are often centered around the site of replication termination (Eisen *et. al.*, 2000; Suyama and Bork, 2001) and the reduction of the genome complement in permissive environments (Cole *et. al.*, 2001; Mira *et. al.*, 2001). Local rearrangement of genes is very unusual and generally occur in conjunction with the inactivation of a member of the gene pair. The organization of genes in operonic structures is believed to be a major force constraining the movement of bacterial genes, but the dynamics of operon genesis and destruction have not yet been elucidated.

Similar but less extensive studies in unicellular eukaryotes indicate that different groups show individual modes of genome evolution. Studies of the genomes of parasitic protozoa belonging to the groups *Apicomplexa* and *Kinetoplastida* have demonstrated that, while genes and gene order of core sections of chromosomes have remained relatively static between closely and distantly related species, there are dramatic differences in the sub-telomeric sections of chromosomes (Bowman *et. al.*, 1999; Carlton *et. al.*, 1998; del Portillo *et. al.*, 2001; El-Sayed *et. al.*, 2000; Gardner *et. al.*, 1998; Janssen *et. al.*, 2001; Myler *et. al.*, 1999; Myler *et. al.*, 2000; Nomura *et. al.*, 2001; Ravel *et. al.*, 1999; Tchavtchitch *et. al.*, 2001). In both groups these regions contain arrays of species specific gene families involved in the biology of infection. These observations are surprising as early karyotype analysis of the chromosomes of some groups such as the trypanosomids indicated that hyper-plasticity within the chromosomes is common. In *T. brucei* one example of a polymorphic chromosome has been shown to result from variation in copies of repetitive elements flanking the conserved gene containing sections (El-Sayed *et. al.*, 2000). Interestingly, kinetoplastids also organize most of their genes in operons. Could these operons be constraining gene rearrangement in their genomes as they are believed to in bacteria?

A rather different story emerges from the study of hemiascomycete yeasts. When the genomes of *Saccharomyces cerevisiae* and *Candida albicans* (150-300 Myr divergence) were compared, over 1000 breaks in gene order were identified (Fischer *et. al.*, 2001; Llorente *et. al.*, 2000). Many of these breaks were shown to be inversions occurring around duplicated genomic segments. However, subsequent loss of redundant genome sections often obscured some of the break points. These duplicated genomic segments were also implicated in the rare interchromsomal translocations found in the genomes. A similar pattern was seen when segments of the genomes of two distantly related (200 Myr) filamentous fungi *Magnaporthe grisea* and *Neurospora crassa* are compared. Microsyntenic regions were rearranged by frequent inversions of gene clusters (Hamer *et. al.*, 2001). Comparison of these segments with the hemiascomycete yeast genomes showed no conservation of synteny and genes were randomly distributed between different linkage groups. So unlike the apicomplexa and kinetoplastids, the chromosomes of these organisms appear to be relatively plastic with frequent rearrangements, often facilitated by the duplication of chromosome segments.

Among the multicellular organisms extensive studies of plant, vertebrate and insect genomes have demonstrated that, as in the fungi, duplication of chromosome segments and rearrangement of chromosomes by inversions are a common occurrence.

During the analysis of several plant genomes, comparisons of monocots and dicots failed to find more than small sections of microsynteny (Liu *et. al.*, 2001a; Paterson *et. al.*, 2000). One complicating factor is the frequent polyploidization of plant genomes. Defining orthologous chromosomal segments is often extremely difficult, as uneven recombination and loss between duplicated chromosome pairs can quickly randomize the gene complement. However, when more closely related species were compared syntenic segments corresponding to orthologous linkage groups were identified indicating that, despite wholesale rearrangement, the gene complement of many chromosomes remained relatively constant (Grant *et. al.*, 2000; Ku *et. al.*, 2000; Lagercrantz, 1998).

In the vertebrates extensive analysis of mammal and teleost genomes has revealed a surprising conservation of synteny between distantly related species (~450

Myr divergence) in discrete chromosomal sections. Karyotype analysis and chromosomal painting of representative taxa from all the major groups of mammals and several outlying species such as chickens and zebrafish have shown that vertebrate chromosomes are mosaic structures with large orthologous segments conserved through the whole phylum (O'Brien *et. al.*, 1999; Postlethwait *et. al.*, 2000). However, these segments have been rearranged between species through translocation of large chromosomal segments and chromosome fission and fusion events. The tempo of chromosome rearrangement does not appear to be constant in all groups and some lineages such as the rodents show more extensive reshuffling than outlying taxa such as zebrafish. It is still being determined how static the arrangements of genes within these segments are. In zebrafish more then half of the 500 genes analyzed in a radiation hybrid survey showed conserved order when compared to the *H. sapiens* physical map (Barbazuk *et. al.*, 2000; Postlethwait *et. al.*, 2000; Woods *et. al.*, 2000). This suggests that, within these large blocks, conserved microsynteny will be found. Sequencing of several sections of the pufferfish genome supports this finding but indicates that inversions may be the predominant mechanism moving genes within these segments (Brunner *et. al.*, 1999; Davidson *et. al.*, 2000).

In situ analysis of the chromosomes of several insects (*Drosophila sp.* and *Aedes gambiae*) indicates that a similar pattern of genome evolution is occurring in the diptera (Fulton *et. al.*, 2001; Ranz *et. al.*, 2001). Large orthologous linkage groups could be identified which contained similar gene complements. However, when the incidence of rearrangements were measured and compared, the dipterans showed remarkably elevated incidences of paracentric rearrangements relative to the rates seen in vertebrates and plants (Ranz *et. al.*, 2001). Interestingly, these analyses also demonstrate uneven rates of intrachromsomal inversions and interchromsomal translocations are found in different chromosomes and chromosome segments (Fulton *et. al.*, 2001). For instance the X chromosomes of *D. melanogaster.* and *A. gambiae* showed a higher incidence of interchromosomal translocations than the 3R arms. Within the X chromosome the translocations occurred much more frequently at the tip of the chromosome than at the center. This indicates that even in a high background of internal recombination some hotspots occur which may dispose chromosomes or segments of chromosomes to movement within the genome. The forces controlling the location of these hotspots have not yet been defined although there is some evidence implicating transposable elements or their preferred integration sites as possible factors influencing the likelihood of a paracentric inversion occurring in a particular region (Caceres *et. al.*, 2001; Caceres *et. al.*, 1999).

6.0.2 Linkage conservations and conservation of synteny between nematode genomes

Very little is known about how nematode genomes have evolved. The limited number of karyotype analyses that have been performed have shown that, like other multicellular eukaryotes, nematodes are diverse in the number of autosomes and sex chromosomes found in closely and distantly related species (Albertson *et. al.*, 1979; Barabashova, 1974; Goldstein and Moens, 1976; Goldstein and Slaton, 1982; Goldstein and Triantaphyllou, 1979; Goldstein and Triantaphyllou, 1980; Goldstein and Triantaphyllou, 1981; Mutafova, 1976; Mutafova *et. al.*, 1982; Sakaguchi *et. al.*, 1983; Triantaphyllou and Moncol, 1977; Vassilev and Mutafova, 1974). Also an unusual process called chromatin diminution causes massive restructuring of somatic chromosomes in several species (Muller *et. al.*, 1996; Triantaphyllou and Moncol, 1977). The genomes of two closely related free-living nematode species *C. elegans* and *C. briggsa*e have been the focus of extensive comparative genetic studies (Kuwabara and Shah, 1994; Thacker *et. al.*, 1999). The forthcoming availability of the *C. elegans* and *C. briggsae* genome sequences will provide a wealth of comparative data including the first measurement of the rate and mode of genome evolution between two nematode species. One study has analyzed fragments (8 Mb in ~ 40kb sections) of sequence from the unfinished *C. briggsae* genome and found extensive rearrangements relative to the *C. elegans* genome, with one break in synteny estimated in every ~8.5 kb (Kent and Zahler, 2000). Syntenic fragments had a bimodal length distribution, with higher rates of rearrangement (and shorter fragments) mapping to the chromosome arms, while the chromosome centers had lower rates. Because the incidence of rearrangements were estimated from partial genome sequence they may be inflated due to virtual breaks resulting from clone ends. However, given that *C. elegans* and *C. briggsae* are believed to be closely related (25- 50 Myr divergence(Thomas and Wilson, 1991)) it is possible the genomes of nematodes may be evolving at an even faster the rate then seen in the diptera. If the rhabditid genomes are evolving at a high rate then it is unclear whether rapid genome evolution is unique to this group (as is seen in some mammals) or whether it extends through the phylum Nematoda. It would also suggest that the transferability of gene positional information which has been so useful in the comparative analysis of rhabditid genomes to more distantly related nematodes will be very limited. To address this question a third nematode genome outside of the rhabditid group needs to be surveyed. The data and resourses generated by the *B. malayi* genome project provides an ideal test bed for these questions. Molecular clock analysis of phylogenies derived from cytochrome c and globin genes yielded an estimated time of divergence between *C. elegans*, *C. briggsae* and the *B. malayi* of 300-500 Myr (Vanfleteren *et. al.*, 1994). Two approaches were taken to examine the conservation of genomic arrangement between the two distantly related nematodes. The first is a detailed analysis of the genes contained in the genomic regions around the *B. malayi* macrophage migration inhibitory factor 1 (*mif-1*) (Pastrana *et. al.*, 1998) and *Bm-mif-2* (Zang *et. al.* 2002 *in press*) loci which encode *B. malayi* homologues of a human cytokines. The second used the *B. malayi* BAC end sequences data generated by J. Daub, Edinburgh University and the PSU, Sanger Institute to ascertain if the observations made in the analysis of the *mif* genomic regions were consistent with what is observed in other sections of the genome.

6.1 Isolation and Characterization of *Bm-mif-1*locus

A probe for *Bm-mif-1* was synthesized by PCR amplification of the full length cDNA using vector primers T3 and T7 from the *B. malayi* EST MB2SLJ1E04T3 as a template (Genbank accession AA216506). The PCR product was cleaned using a Microcon-100 (Millipore) column and the vector portions of the PCR product removed by digestion with restriction enzymes *Xho*1 and *Eco*R1 (New England Biolabs). The digested PCR product was cleaned with a Microcon-100 and 500 ng of DNA random prime labeled with biotin using the Phototope random primed labeling kit (New England Biolabs). The *Bm-mif-1* hybridized to high density nylon filter arrays containing 4,608 bacterial artificial chromosomes (BACs) in the pBeloBACII vector with *B. malayi* genomic DNA inserts (Guiliano *et. al.*, 1999) and detected with Streptavidin/ Alkaline Phosphatase conjugate and CPD* (Phototope Detection Kit , NEB). Figure 6.1.1A shows the luminograph of the detected nylon filter. Isolated hybridization-positive BACs were PCR tested for the presence of the *Bm-mif-1* gene using the gene specific primers Bm-MIF-1.F1a and Bm-MIF-1.R1a with standard PCR reaction conditions using Taq polymerase (AGS-Gold, Hybaid) and isolated BAC colony boils as template (see figure 6.1.1B). Two BACs BMBAC101P19 and BMBAC102O03 were PCR positive for *Bm-mif-1*. BAC DNA was isolated by Midi preparation (Qiagen) and the ends of both clones sequenced with the vector primers T7 and SP6.

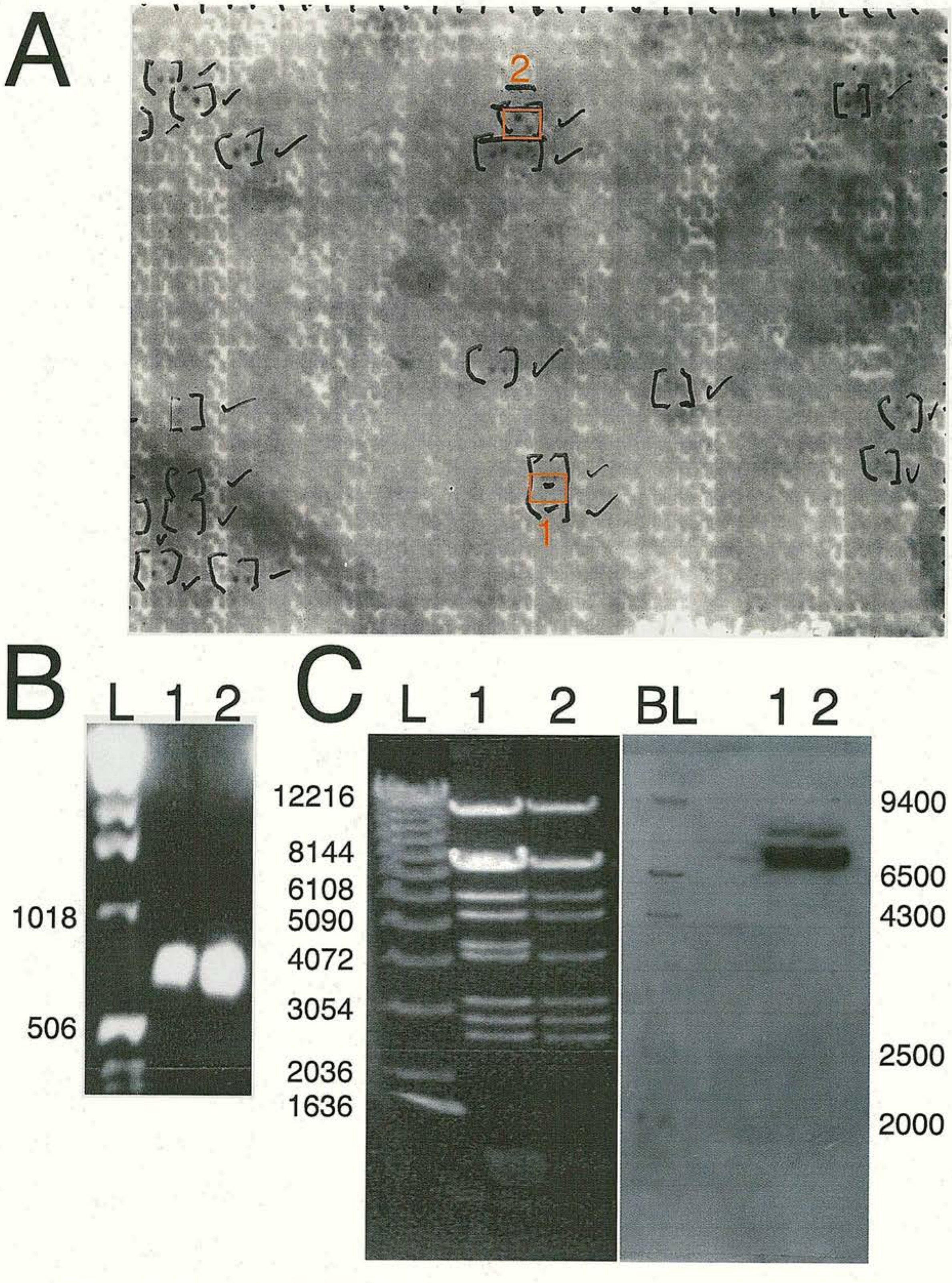


Figure 6.1.1 Hybridization of *Bm-mif-1* to BAC filter **A:** Luminograph of the high density BAC filter hybridized with the *Bm-mif-1* probe. Positive clones are highlighted and numbered in orange (1: BMBAC101P19, 2: BMBAC102O03). **B:** Hybridization positive BACs were tested for the presence of *Bm-mif-1*by PCR with specific primers 1: BMBAC101P19 and 2: BMBAC102O03 are shown. L1: Gibco 1kb ladder (GibcoBRL). **C:** The BACs were digested with restriction enzymes E: *EcoR*1 and H: *Hind*III and run on a 0.7% agarose gel. The gel was then blotted and hybridized with the *Bm-mif-1* probe. BL2: biotin labeled lambda digested with *Hind*III.

The size of the BAC inserts was determined by digestion of prepared BAC DNA with *Hind*III. The analysis of the *Hind*III fragment sizes indicated that the BMBAC101P19 insert is approximately 65 kb and is larger than the BMBAC02O03 insert by at least 4.5 kb (See figure 6.1.1B). The *Hind*III digested BAC DNA was southern blotted and hybridized with the biotin labeled *Bm-mif-1* probe. A 7.5 kb HindIII fragment containing the Bm-mif-1 gene was identified. The luminograph shown in figure 6.1.1B indicates that *Bm-mif-1* is located on a 7.5 kb fragment. BMBAC101P19 was submitted to the Sanger Centre Pathogen Sequencing Unit (PSU) for sequencing.

6.2 Identification of BMBAC101L03

To further expand the *Bm-mif-1*contig, the sequence from the T7 end of the BMBAC101P19 insert was used to design specific PCR primers BMBAC101P19.T7.F1 and BMBAC101P19.T7.R1. The reverse primer BMBAC101P19.T7.R1 and a specific vector primer containing two biotin moieties on the 5' end (2BiotinBACF3; NEB Organic Synthesis Unit) were used to synthesize a biotin labeled PCR product, using the standard reaction conditions. The PCR product was gel purified (UltraFree-DA, Amicon Bioseparations) and hybridized to the *B. malayi*18,000 BAC filter using the standard hybridization and detection protocol (see figure 6.2.1A) (Foster *et. al.*, 2001). Five BACs were isolated (BMBAC101L03, BMBAC213O19, BMBAC218E06, BMBAC230D12, BMBAC235B06) and the presence of the T7 end of BMBAC101P19 tested by PCR (see figure 6.2.1B). BAC DNA was isolated by Midi preparation (Qiagen)and both ends of the clone sequenced with the vector primers T7 and SP6.

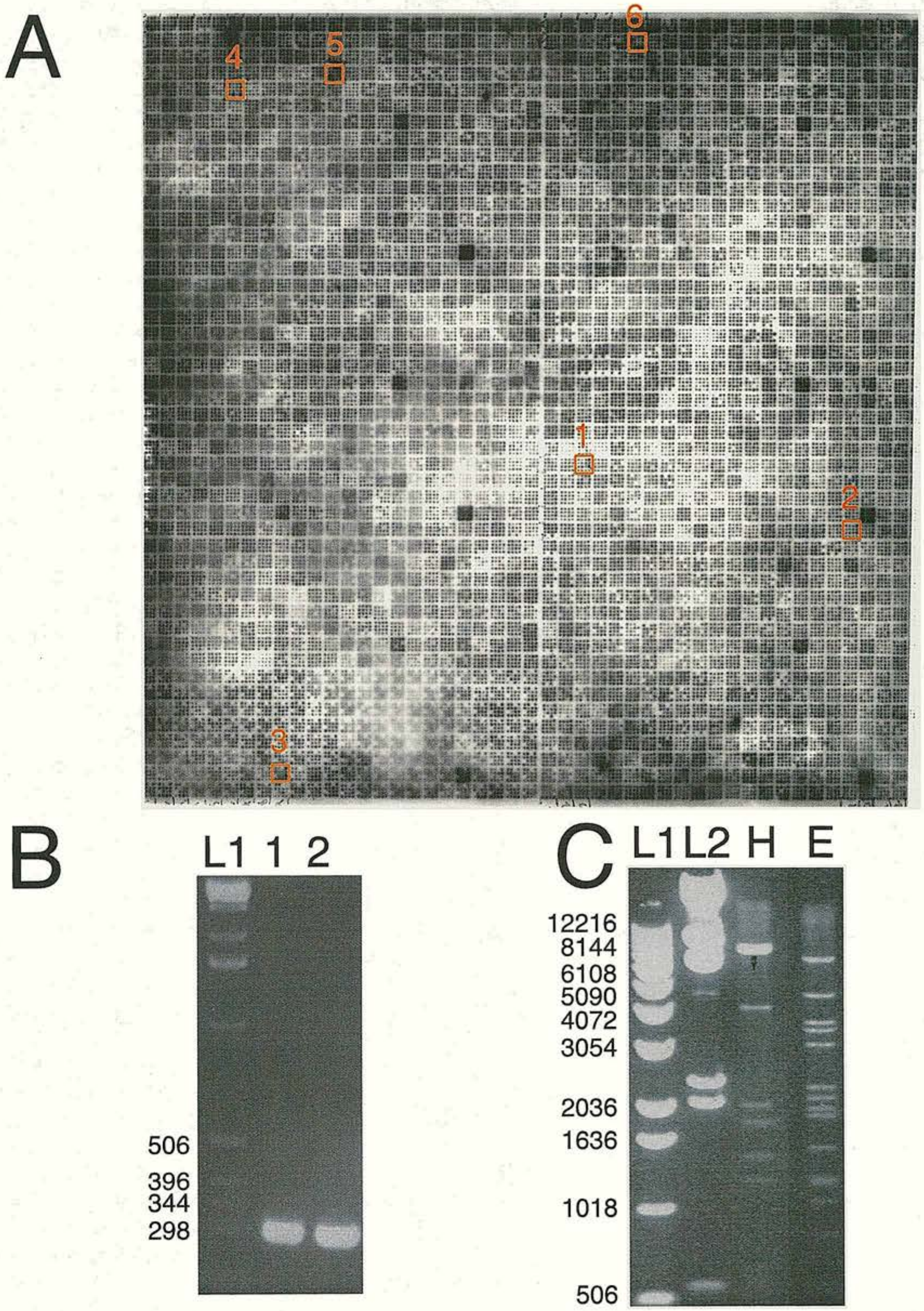


Figure 6.2.1 Hybridization of the T7 end of BMBAC101P19 to B. malayi BAC filter **A:** Luminograph of the high density BAC filter hybridized with the BMBAC101P19.T7 probe. Positive clones are highlighted and numbered in orange (1: BMBAC101L03, 2: BMBAC101P19, 3: BMBAC213O19, 4: BMBAC218E06, 5: BMBAC230D12, 5: BMBAC235B06). **B:** Hybridization positive BACs were tested for the presence of BMBAC101P19.T7 fragment by PCR with specific primers. The results for 1: BMBAC101P19 and 2: BMBAC101L03 are shown. L1: Gibco 1kb ladder (GibcoBRL). **C:** BMBAC101L03 was digested with restriction enzymes E: *Eco*R1 and H: *Hind*III and run on a 0.7% agarose gel. L2: Lambda digested with *Hind*III.

The amount of overlap between the clones and BMBAC101P19 was determined by comparing the two sequences with Pustell DNA Matrix in the MacVector package (Oxford Molecular) and summarized in table 6.2.2.

BAC clone	End	bp overlap BMBAC101P19	Size of clone insert in kb
BMBAC101L03	T7	10,637	35
BMBAC213O19	SP6	39,467	50
BMBAC218E06	T7	24,997	40
BMBAC230D12	T7	45,356	60
BMBAC235B06	SP6	58,132	70

Table 6.2.2 Summary of the overlap of identified BACs The ends of the BAC clones identified by the BMBAC101P19 T7 end probe hybridization were sequenced and compared to BMBAC101P19 and the bp overlap determined. This was compared to the estimated size of the insert determined by PFG analysis (data not shown)

BMBAC101L03 showed the least sequence overlap and the size of its insert was determined by digestion of prepared BAC DNA with *Eco*R1 and *Hind*III (see figure 6.2.1C). The restriction digests indicated that BMBAC101L03 has an insert of approximately 27 kb. To determine the 17 kb of additional sequence, BMBAC101L03 was submitted to the PSU, Sanger Institute for sequencing.

6.3 Sequencing of the *Bm-mif-1* locus: preparation, subcloning and sequencing of BACs at the PSU

The two BACs were sequenced at the PSU (shot sequencing and assembly performed by (Hall N, Clark L.N., Corton C.H. and Barrell B.G.) using a two-stage strategy involving random sequencing of sub-cloned DNA followed by directed sequencing to resolve problem areas. In the first stage, DNA prepared from BACs clones was shattered by sonification and fragments of 1.4-2 kb were cloned into pUC18. The DNA from randomly selected clones was sequenced with dye-terminator chemistry and analyzed on ABI3700 automated sequencers. Each BAC was sequenced to a depth of seven fold sequence coverage. Contig assembly was performed using Phrap (Phil Greene, Washington University Genome Sequencing Center, unpublished). Manual base calling and finishing was carried out using Gap4 software (http://www.mrclmb.cam.ac.uk/pubseq/manual/gap4_unix_1.html). Gaps and low quality regions of the sequence were resolved by techniques such as primer walking, PCR and re-sequencing clones under conditions giving increased read lengths.

6.4 Sequence analysis of the BAC inserts: Prediction of protein coding genes

The finished sequences of BMBAC101P19 and BMBAC101L03 were 64,685 bp and 28,757 bp respectively with an overlap of 10,637bps. The contiguated sequence of BMBAC101P19 and BMBAC101L03 (82,805 bp) was compared to all public databases including the GenBank nonredundant (nucleic acid and protein), EST (dbEST), *C. elegans* protein (Wormpep) databases and to the *B. malayi* clustered EST database (see chapter 3) using BLAST (Altschul *et. al.*, 1990; Altschul *et. al.*, 1997). GeneFinder (Phil Greene and LaDeana Hillier, Washington University Genome Sequencing Center, unpublished software) was trained with the 162 publicly available *B. malayi* gene sequences and used to predict genes on the contiguated BMBAC101P19/ BMBAC101L03 sequence. The sequence was annotated within the Artemis workbench (Rutherford *et. al.*, 2000). Predicted protein sequences were compared to the Pfam database (Bateman *et. al.*, 2000) and potential cellular localization examined using PSORTII (Nakai and Kanehisa, 1992). The sequence contig along with all of its annotation is available in Genbank (Accession AL606837).

6.5 Verification of the Genes Predicted from BMBAC101P19

To evaluate the accuracy of gene prediction in the 83 kb genomic fragment, specific primers spanning the protein coding regions of the genes contained on BMBAC101P19 were designed and PCR was performed on oligo (dT) primed *B. malayi* mixed adult first strand cDNA. Each predicted cDNA had at least two primer sets designed to it. A list of all primers designed and tested is listed in the appendix with the primers being named gene_number.primer_orientation # (i.e. 01P19.2.F1). The primer sets were designed to break the cDNAs into sections (300-400 bp) which should be easily amplified. Once primers were tested and found to amplify cDNA fragments primers spanning larger sections of some the cDNAs (1000-2000 kb) were used to produce PCR products which were then sequenced. PCR on first strand cDNA was performed with AGS-Gold Taq (Hybaid) using the standard conditions as described above using the gene-specific primers. An example of a typical RT-PCR which amplified the cDNA of a predicted gene can be seen in figure 6.5.1A.

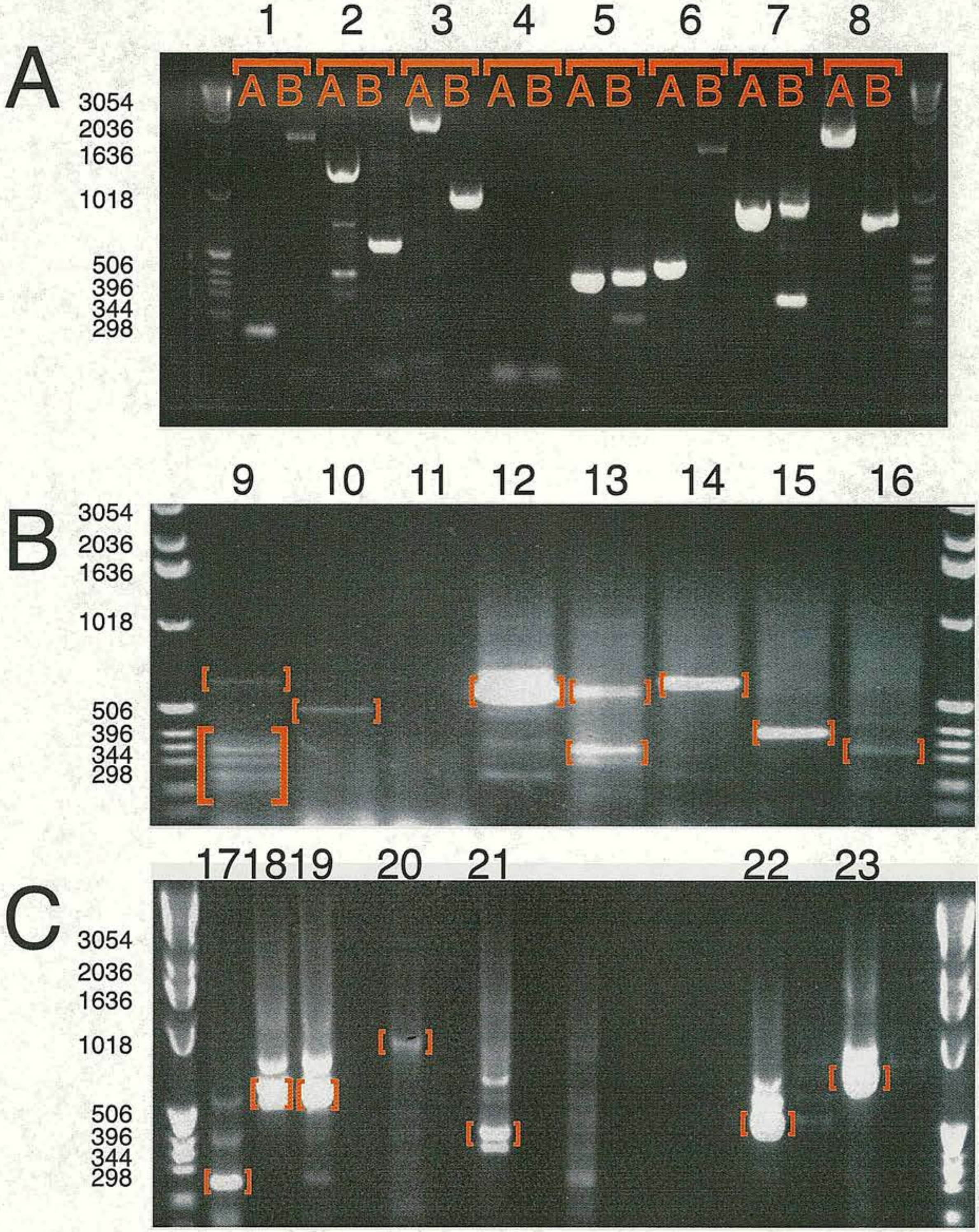


Figure 6.5.1A-C Isolation of cDNA predicted from the *B. malayi* BMBAC101L03 and BMBAC101P19 **A:** An example amplification of cDNA fragments by PCR from oligo dT primed cDNA. A: genomic DNA positive control using BMBAC101P19 as template. B: oligo dT primed cDNA as template. 1: 01P19.5F1, 01P19.5R1; 2: 01P19.5F1, 01P19.5R2; 3: 01P19.5F1, 01P19.5R3; 4: 01P19.5F1, 01P19.5R4; 5: 01P19.5F1, 01P19.5R5; 6: 01P19.5F2, 01P19.5R1; 7: 01P19.5F2, 01P19.5R2; 8: 01P19.5F2, 01P19.5R3; **B:** An example amplification of 5' RACE fragments by PCR with SL1 and gene specific primers from oligo dT primed cDNA. The numbers indicate which gene specific reverse primer was used in conjunction with SL1 in the PCR. 9: 01P19.4.R1; 10: 01P19.6.R1; 11: 01P19.6.R2; 12: 01P19.2.R5; 13: 01P19.2.R1;14: 01P19.3.R3; 15: 01P19.3.R5; 16: positive control PCR for BMBAC101P19.3 using primers 01P19.3.F2 and 01P19.3.R3. **C:** An example amplification of 3' RACE fragments by PCR with GeneRACER 3'primer and gene specific forward primers from tagged oligo dT primed cDNA. The numbers indicate which gene specific reverse primer was used in conjunction with GeneRACER 3' primer in the PCR. 17: 01P19.2.F14; 18: 01P19.6.F2; 19: 01P19.2.F12; 20: 01P19.3.F2; 21: 01P19.3.F3; 22: 01P19.5.F4; 23: 01P19.6.F1. The orange brackets indicate bands that were purified, cloned, and sequenced.

To recover the 3' ends of the transcripts, first strand cDNA was synthesized which contained the GeneRacer 3' RACE primer 5' to the oligo dT. PCRs were then performed with gene specific forward primers and the GeneRacer 3' RACE primer. An example of a typical 3'RACE PCR which amplified the 3' ends of predicted cDNAs can be seen in figure 6.5.1C. To retrieve the 5' ends of cDNAs the pan-nematode SL1 sequence (Blaxter and Liu, 1996) and specific reverse primers were used to PCR from oligo dT primed cDNA. To retrieve the 5' ends of BMBAC101P19.2 which has a very large cDNA and BMBAC101P19.3 and BMBAC101P19.7 which are not abundantly expressed, the template first strand cDNA was synthesized with the gene specific reverse primers (01P19.2.R5, 01P19.3.R3 and 01P19.7.R7). An example of a typical 5'RACE PCR that amplified the 5' ends of predicted cDNAs using the SL1 sequence and specific reverse primers can be seen in figure 6.5.1B. If the primary RACE PCR did not yield strong products or presented very complex mixtures of PCR products secondary PCRs were performed using gene specific nested primers and 2% of the primary PCR product. PCR products whose sizes were consistent with those of the predicted genes or were very abundant were selected for sequencing. The PCR products were cloned in to pCR4-TOPO and sequenced using the vector primers M13L and M13R or gene specific primers.

6.6 Analysis of theBMBAC101L03/BMBAC101P19 contig

6.6.1 General Characteristics of **Brugia malayi** *genomic DNA*

The insert of BMBAC101L03 and BMBAC101P19 were determined to be 28,757bp and 64,685 bp in length with 10,637bp of overlapping sequence (see figure 6.6.1.2). The contiguated sequence of the inserts of BMBAC101L03 and BMBAC101P19 is 82,805 bp with an AT content of 68%. The exonic DNA has an average AT content of 59.9% while the intergenic and intronic DNA have a higher AT content of 69 and 71.4% respectively. The average predicted gene size (ATG - stop) is 3.2 kb (ranging from 0.5 -20kb). The average distance between genes was 3.1 kb (range of 300 bp to 10.5 kb), giving an average gene density of one gene per 6.9 kb. This is lower than *C. elegans* which has an average of one gene per 5 kb (The *C. elegans*

Sequencing Consortium, 1998). Analysis of the *C. elegans* orthologues showed that they had a higher gene density with an average of one gene per 3.2 kb. There was an average of 9.3 introns per gene in the *B. malayi* sequences, with an average intron length of 316 bp (range of 48-2767 bp). The *C. elegans* orthologues of these genes have an average of 5.5 introns per gene with an average size of 142 bp (range of 46-1260bp).

Gene/Region	# of exons	length cDNA (bp)	%AT exonic DNA	# introns	length of intronic DNA (bp)	%AT intronic DNA	Gene size	length intergenic region (bp)	%AT intergenic region
BMBAC101P19.7[1]	8	1039	63.1	7	1185	76.2	2224	-	-
intergenic	-	-	-	-	-	-		4995	75.5
BMBAC101P19.6	5	804	61.9	4	1714	70.7	2518	-	-
intergenic	-	-	-	-	-	-		308	67.2
BMBAC101P19.5	19	2679	60.5	18	7141	70.4	9820	-	-
BMBAC101P19.4	2	446	59.8	1	155	70.3	601	-	-
intergenic	-	-	-	-	-	-		3182	64.6
BMBAC101P19.2	38	5955	59.8	37	14157	68.5	20112	-	-
intergenic	-	-	-	-	-	-		3868	71.5
BMBAC101P19.1	3	535	59.4	2	754	69.2	1289	-	-
intergenic	-	-	-	-	-	-		10498	67.1
BMBAC101P19.3	10	1182	60.5	9	2842	70.6	4024	-	-
intergenic	-	-	-	-	-	-		2872	70.4
BMBAC101L03.5	7	918	59.8	6	2090	71.5	3008	-	-
intergenic	-	-	-	-	-	-		862	67.4
BMBAC101L03.4	3	630	59.4	2	862	71.1	1492	-	-
intergenic	-	-	-	-	-	-		2065	68.3
BMBAC101L03.3	9	1239	59.3	8	1962	73.2	3201	-	-
intergenic	-	-	-	-	-	-		1503	70.6
BMBAC101L03.2	7	693	56.7	6	1081	72.0	1774	-	-
intergenic	-	-	-	-	-	-		1039	67.9
BMBAC101L03.1[1]	8	1340	59.7	7	1746	74.1	3086	-	-
average	***10.3***	***1508***	***59.9***	***9.3***	***3275***	***71.4***	***4783***	***3119***	***69***

Table 6.6.1.1 General features of BMBAC101L03/BMBAC101P19 contig. [1] Gene fragments (see text). The table summarizes the number of exons, length of the predicted cDNAs, number of introns, the length of the intronic portions of genes, the gene size, the size of intergenic regions separating genes and the %AT of the different sections DNA. Calculations of average number of exons per gene, length of cDNAs, number of introns, total length of introns and gene size did not include BMBAC101P19.7 and BMBAC101L03.1 because they represented partial gene fragments.

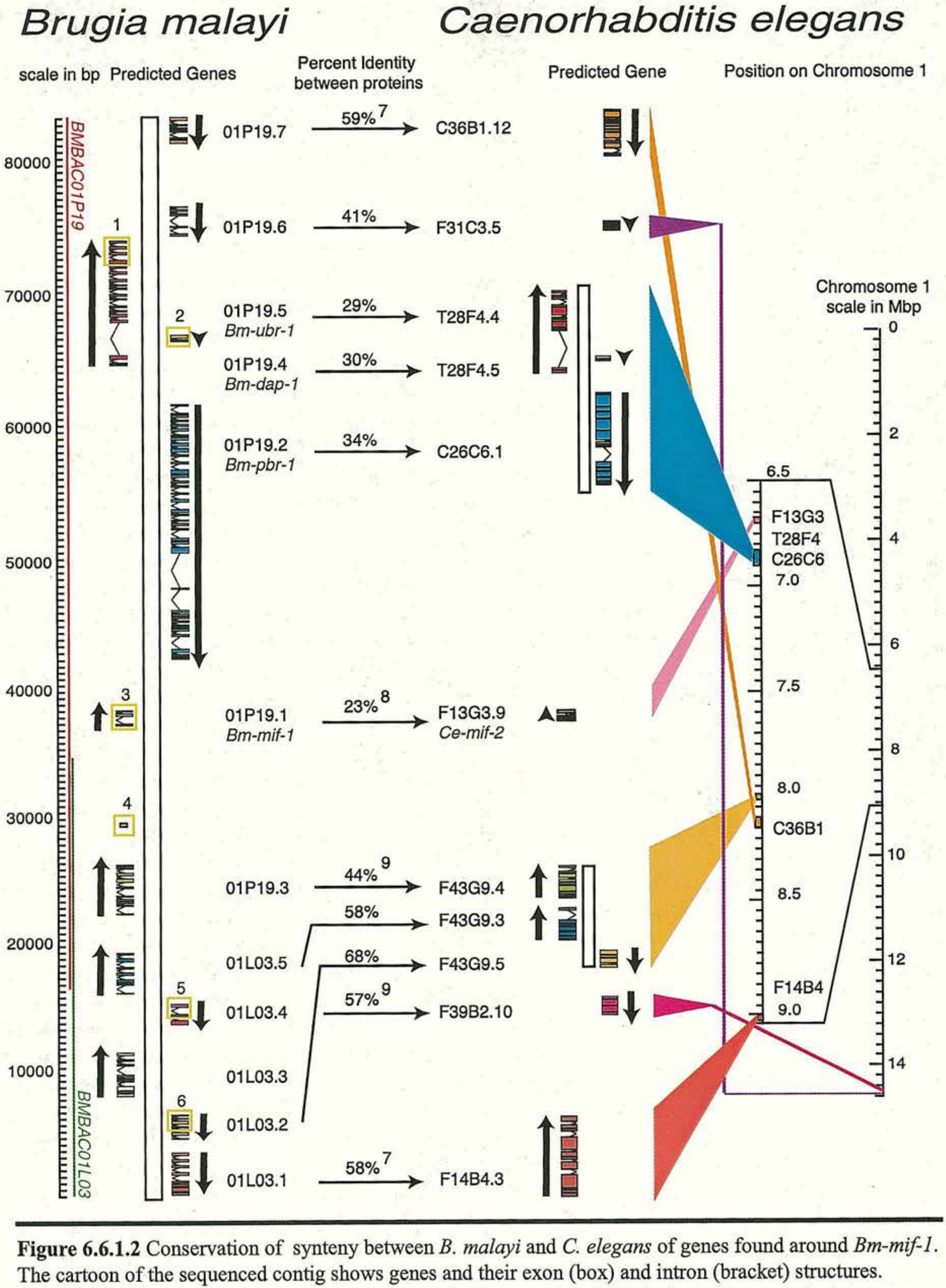


Figure 6.6.1.2 Conservation of synteny between *B. malayi* and *C. elegans* of genes found around *Bm-mif-1*. The cartoon of the sequenced contig shows genes and their exon (box) and intron (bracket) structures. Genes on the left are transcribed from the bottom to the top, and to the right from the top to the bottom. **1** Segment is identical to *B. malayi* EST cluster BMC03169 (Blaxter *et. al.* 2001). **2** ORF has high similarity to *O.volvulus* EST cluster OVC02481 (Lizotte-Waniewski *et. al.* 2000). **3** Segment is identical to *B. malayi* EST cluster BMC00238. **4** Segment is identical to *B. malayi* EST clusters BMC02055 and BMC01932, however, no ORF can be identified, so it is not believed to represent protein-coding sequence. **5** Segment is identical to *B. malayi* EST cluster BMC06334. **6** Segment is identical to *B. malayi* EST cluster BMC00400.**7** BMBAC101L03.1 and BMBAC101P19.7 are gene fragments, the percent identity calculation was based the alignable portion of the *C. elegans* orthologue. **8** *Ce*-F13G3.9 is a MIF homologue found on *C. elegans* chromosome I in close proximity to other synteny genes. However, phylogenetic analysis indicates that *Ce*-F13G3.9 is not the orthologue of *Bm-mif-1* (see chapter 5). **9** The percent identity was calculated for BMBAC101P19.3 and BMBAC101L03.4 only within the PWWP or dnaJ domains as these are the only alignable portions of the molecules.

6.6.2 Identified Genes and Evaluation of Gene Prediction

Of the twelve *B. malayi* genes identified seven were confirmed by RT-PCR. A summary of the differences between the predicted and the experimentally confirmed cDNA sequences is presented in table 6.6.2.0.1.

Gene	Observed differences between predicted and experimentally confirmed cDNA sequences
BMBAC101P19.2	- Failed to predict 142 bp exon 23. - Small splice site misprediction on the donor site of intron 24. 1 Small splice site misprediction on the acceptor site of intron 28.
BMBAC101P19.3	1.0 Predicted additional exons after exon 7 which were not confirmed by RT-PCR. • Failed to predict alternative start site in exon 6.
BMBAC101P19.4	• small splice site misprediction on the donor site of intron 1.
BMBAC101P19.5 BMBAC101P19.8	• Predicted BMBAC101P19.5 and BMBAC101P19.8 to be separate genes based on similarities to genes predicted in the *C. elegans* genome. RT-PCR confirms that these ORFs represent a single gene with a large intron separating the two segments in both species.
BMBAC101P19.6	• No differences.
BMBAC101P19.7	• Small splice site misprediction on the acceptor site of intron 2. • Small splice site misprediction on the donor site of intron 5. • Failed to predict additional exon 7 which is found on BMBAC101P19.7b cDNA.

Table 6.6.2.0.1 Summary of the differences observed between the predicted and experimentally confirmed cDNA sequences.

Alternatively spliced transcripts were identified for four of the cDNAs. In general the differences between the predicted and experimental cDNAs involved either slightly mispredicted splice sites or the omission of small exons. Eleven of the twelve predicted genes had *C. elegans* homologues. Salient features of some of the predicted genes are summarized in table 6.6.2.0.2 and are discussed below.

***B. malayi* ORF**	**Predicted cDNA length in bp**	**Predicted peptide length**	**Numbe r of Introns**	***C. elegans* orthologue**	**% Identity with *C. elegans* orthologue**	**Number of introns in C. elegans orthologue**	**Number of shared intron positions with *C. elegans* orthologue**	**Putative Identity**
BMBAC01L03.1	1340[1]	446[1]	7[1]	*Ce*-F14B4.3	58[2]	3[3]	3	N-terminal fragment of the β subunit of RNA polymerase I
BMBAC01L03.2	693	230	6	*Ce*-F43G9.5	68	3	1	Pre-mRNA cleavage factor
BMBAC01L03.3	1239	412	8	-	-	-	-	Contains LON-ATP dependent serine protease domain
BMBAC01L03.4	630	209	2	*Ce*-F39B2.10	57[4]	3	1	Contains dnaJ domain
BMBAC01L03.5	918	305	6	*Ce*-F43G9.3	58	6	2	mitochondrial carrier protein
BMBAC01P19.1 (*Bm-mif-1*)	535	115	2	*Ce*-Y56A3A.3	41	2	2	Macrophage migration inhibitory factor homologue
BMBAC01P19.2a/b (*Bm-pbr-1*)	5955/5748	1934/1865	37/35	*Ce*-C26C6.1	34	14	9	Polybromo domain protein, BAF180 homologue
BMBAC01P19.3 a/b	1182/919	367/283	9/7	*Ce*-F43G9.4	44[5]	8	2	Contains PWWP domain
BMBAC01P19.4 (*Bm-dap-1*)	446	111	1	*Ce*-T28F4.5	30	1	1	Homologue of mammalian death associated protein DAP-1
BMBAC01P19.5a/b (*Bm-ubr-1*)	2679/2602	847/821	18/17	*Ce*-T28F4.4	27	12	5	
BMBAC01P19.6	804	190	4	*Ce*-F31C3.5	41	1	1	Conserved protein of unknown function
BMBAC01P19.7a/b	1039/932[1]	274/298[1]	6/7[1]	*Ce*-C36B1.12	60[6]	3[3]	2	C-terminal fragment of a novel transmembrane protein

Table 6.6.2.0.2 Genes predicted on the BMBAC101L03/BMBAC101P19 contig. [1] Gene fragments (see text). [2] BMBAC01L03.1 gene fragment aligned with the N-terminal 450 aa of CeF14B4.3. [3] Number of introns in the aligned portion of the *C. elegans* orthologue. [4] Only the dnaJ domains of BMBAC01L03.4 and *Ce*-F39B2.10 were aligned. [5] Only the PWWP domains of BMBAC01P19.3and *Ce*-F43G9.4 were aligned. [6] The gene fragment of BMBAC01P19.7 aligned with the C-terminal 380 aa of *Ce*-C36B1.12.

6.6.2.1 BMBAC101L03.3

BMBAC101L03.3 contains two protein domains, an N-terminal LON ATP dependent serine protease domain (PF02190) and an anonymous C-terminal protein domain (PFB021704). Proteins predicted from the *H. sapiens*, *M. musculus*, *D. melanogaster* and *A. thaliana* genome (accession; XP_0421219, NP_067424, AE003685 and AAC42255.1) share this domain architecture (see figure 6.6.2.1.1). However, *C. elegans* has two proteins which show similarity to the individual protein domains (*Ce*-R08B4.3 18% identity to the C-terminal protein domain, *Ce*-M18.6 27%identity to the N-terminal LON domain). There are no predicted proteins in the *C. elegans* genome sequence that retain both domains. This gene was either lost or split into separate genes, in the *C. elegans* genome, but has been retained in *B. malayi*.

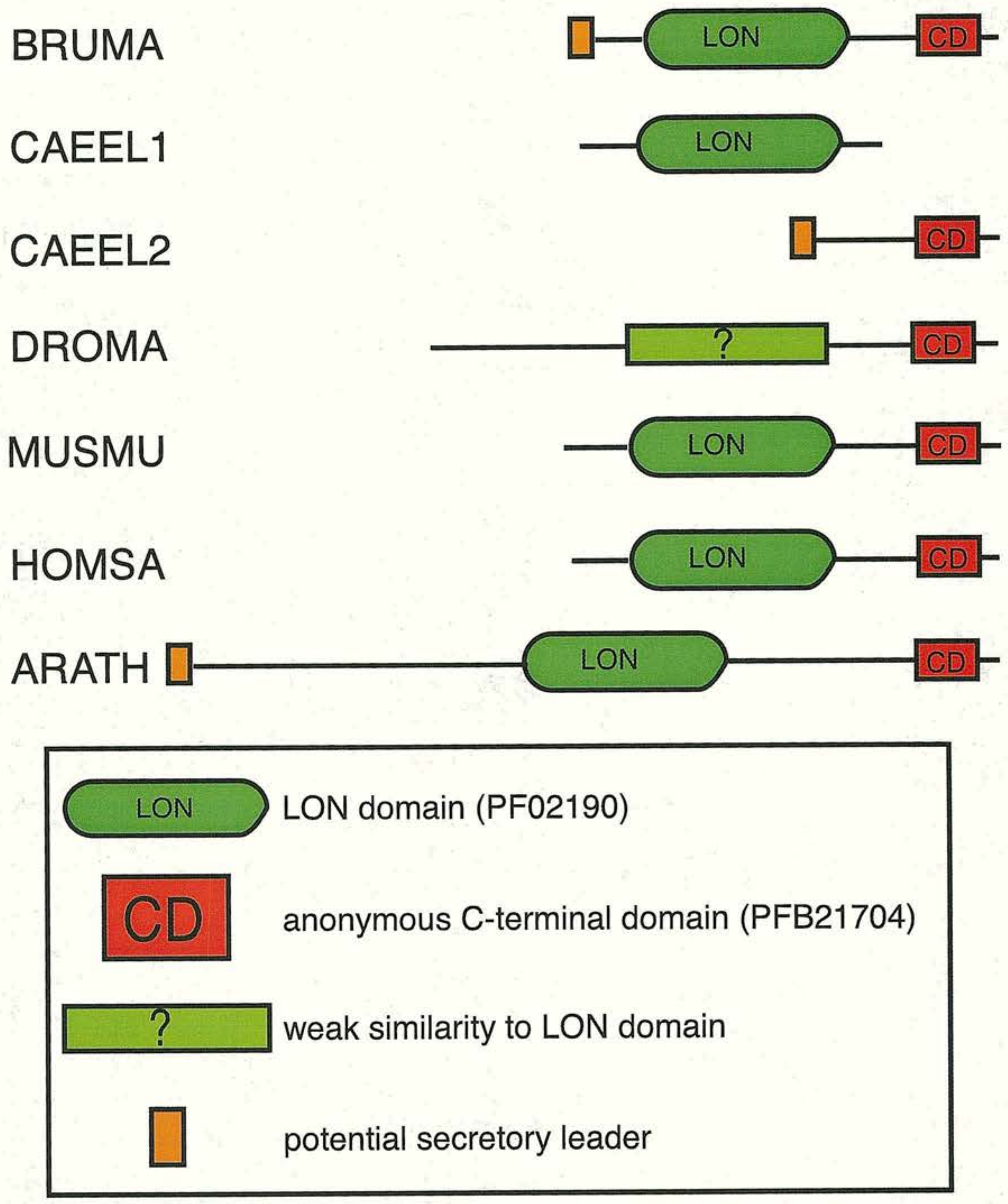

BRUMA
LON
CD
CAEEL1
LON
CAEEL2
CD
DROMA
?
CD
MUSMU
LON
CD
HOMSA
LON
CD
ARATH
LON
CD
LON
LON domain (PF02190)
CD
anonymous C-terminal domain (PFB21704)
?
weak similarity to LON domain
potential secretory leader

Figure 6.6.2.1.1 Comparison of the domain architecture of BMBAC101L03.3 and similar proteins found in other species. BMBAC101L03.3 was searched against GenBank and Pfam (Bateman *et. al.* 2000) and those proteins found with similar domain architecture analyzed. A schematic representation of the five identified proteins and their domains is shown. PSORTII was used to determine which proteins contain potential secretory leader. BRUMA: BMBAC101L03.4, CAEEL1: M18.6, CAEEL2: R08B4.3, DROMA: AE003685, HOMSA: XP_0421219, MUSMU: NP_067424.

6.6.2.2 BMBAC101L03.4

Searches against public databases show that the N-terminus of the BMBAC101L03.4 has a dnaJ (N-terminal) domain (PF00684). This dnaJ domain shows similarity to 41 predicted *C. elegans* proteins. The remainder of the molecule does not appear to be conserved. The dnaJ domain in BMBAC101L03.4 shows highest similarity (82%) to the dnaJ domain in the *C. elegans* predicted protein F39B2.10. Both proteins have the dnaJ domain at their N-terminus and share a common position of the first intron in this region (see figure 6.6.2.2.1). We therefore suggest that *Ce*-F39B2.10 is the closest homologue to BMBAC101L03.4 in the *C. elegans* genome.

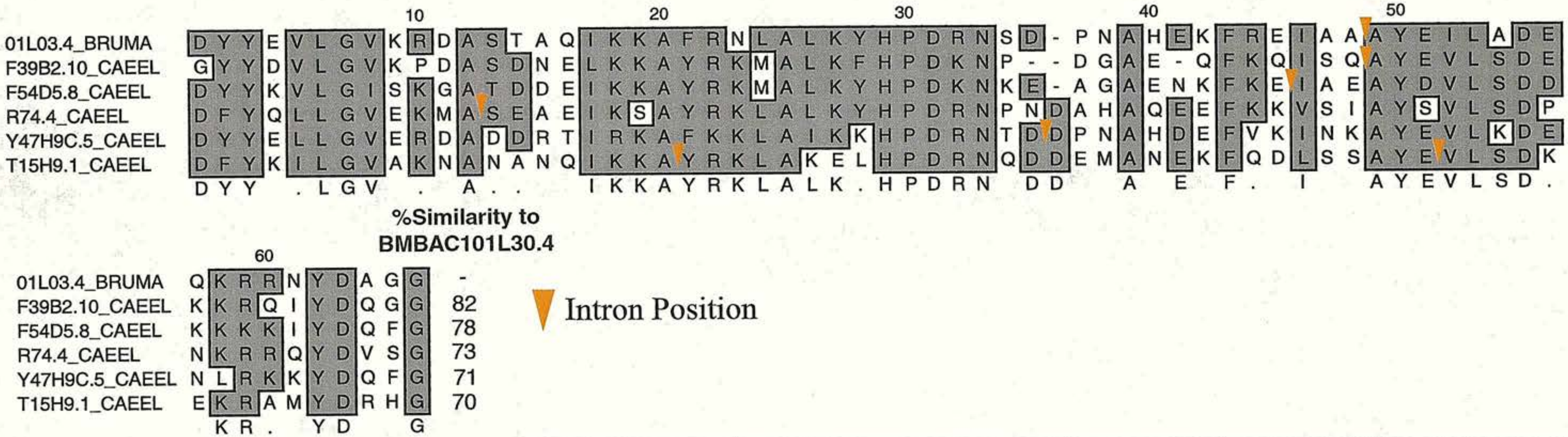
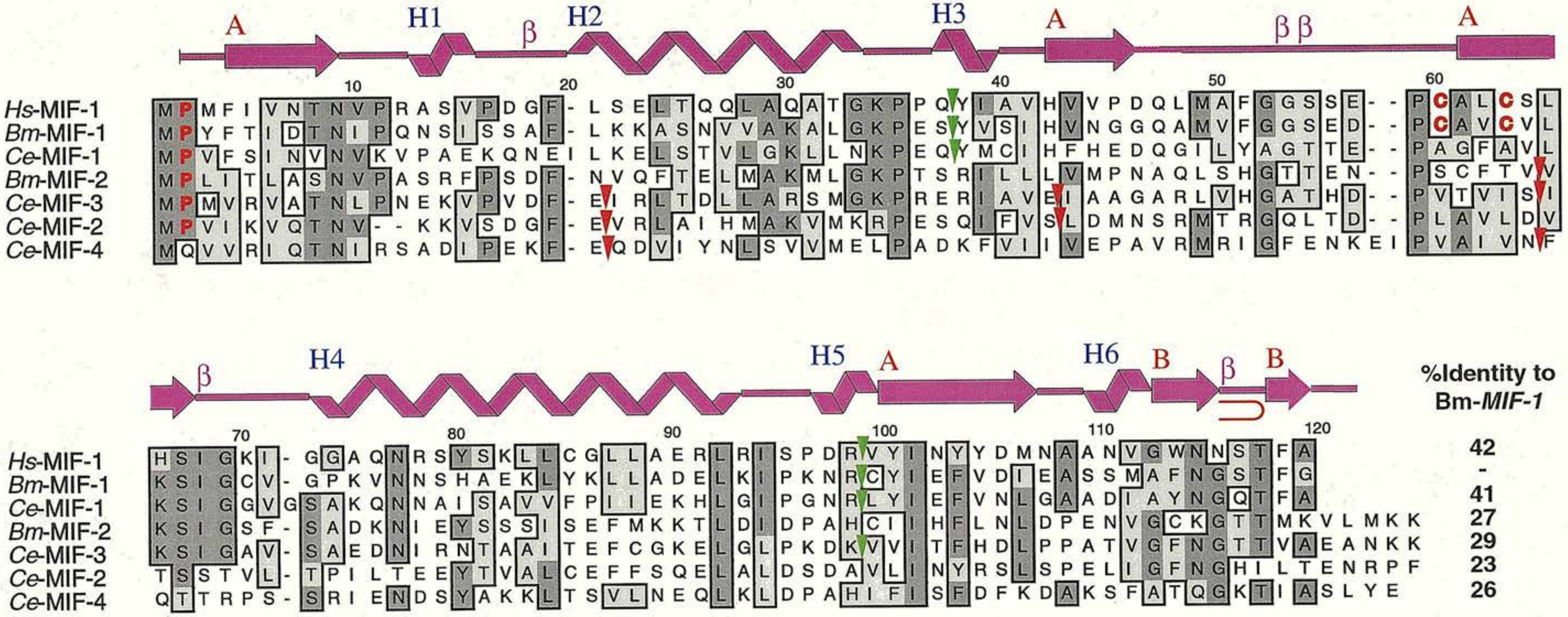

10
20
30
40
50
60
01L03.4_BRUMA D Y Y E V L G V K R D A S T A Q I K K A F R N L A L K Y H P D R N S D - P N A H E K F R E I A A A Y E I L A D E
F39B2.10_CAEEL G Y Y D V L G V K P D A S D N E L K K A Y R K M A L K F H P D K N P - - D G A E - Q F K Q I S Q A Y E V L S D E
F54D5.8_CAEEL D Y Y K V L G I S K G A T D D E I K K A Y R K M A L K Y H P D K N K E - A G A E N K F K E I A E A Y D V L S D D
R74.4_CAEEL D F Y Q L L G V E K M A S E A E I K S A Y R K L A L K Y H P D R N P N D A H A Q E E F K K V S I A Y S V L S D P
Y47H9C.5_CAEEL D Y Y E L L G V E R D A D D R T I R K A F K K L A I K K H P D R N T D D P N A H D E F V K I N K A Y E V L K D E
T15H9.1_CAEEL D F Y K I L G V A K N A N A N Q I K K A Y R K L A K E L H P D R N Q D D E M A N E K F Q D L S S A Y E V L S D K
D Y Y . L G V . A . . I K K A Y R K L A L K . H P D R N D D A E F . I A Y E V L S D .
%Similarity to BMBAC101L30.4
01L03.4_BRUMA Q K R R N Y D A G G -
F39B2.10_CAEEL K K R Q I Y D Q G G 82
F54D5.8_CAEEL K K K K I Y D Q F G 78
R74.4_CAEEL N K R R Q Y D V S G 73
Y47H9C.5_CAEEL N L R K K Y D Q F G 71
T15H9.1_CAEEL E K R A M Y D R H G 70
K R . Y D G
Intron Position

Figure 6.6.2.2.1 Alignment of the conserved dnaJ domain of BMBAC101L03.4 with five dnaJ domains from *C. elegans* proteins F39B2.10 (T21991), F54D5.8 (T22648), R74.4 (T24254), Y47H9C.5 (T26967), and T15H9.1 (T24938). The dnaJ domains were identified using Pfam (Bateman, *et. al.* 2000) and aligned. The percent similarity each *C. elegan* sequence has with BMBAC101L03.4 is shown. Intron positions are indicated with the orange triangles.

6.6.2.3 BMBAC101P19.1

BMBAC101P19.1 encodes *Bm-mif-1* (Pastrana *et. al.*, 1998). *Bm*-MIF-1 shows similarity to small proteins from plants, metazoa and protozoa, but only the mammalian MIF homologues have been studied in detail. Mammalian MIF is a cytokine involved in inflammation, growth and differentiation of immune cells (Nishihira, 2000). In *C. elegans* there are four MIF-like genes: *Ce-mif-1*(Y56A3A.3), *Ce-mif-2*(C52E4.2), *Ce-mif-3* (F13G3.9) and *Ce-mif-4* (Y73B6BL.13) (The *C. elegans* Sequencing Consortium, 1998). *Bm*-MIF-1 is most similar to *Ce*-MIF-1 (41% identity) which is located on *C. elegans* chromosome III. Comparison of *Bm*-MIF-1 with the four *C. elegans* MIFs, a second *B. malayi* MIF like molecule (*Bm*-MIF-2) and human MIF-1 (see figure 6.6.2.3.1) revealed that *Bm*-MIF-1 and *Ce*-MIF-3 share two intron/exon boundaries with each other and the vertebrate MIFs. *Bm*-MIF-1 (and other closely related filarial MIF-1 orthologues) also contain the CXXC motif which is critical for the thiol-oxioreductase activities of vertebrate MIFs (Kleemann *et. al.*, 2000b). None of the *C. elegans* MIF homologues contain this motif. Other residues important for MIF function are highlighted in figure 6.6.2.3.1. A fuller analysis of the MIF gene family is presented in chapter 3.


A
H1
β
H2
H3
A
β β
A
10
20
30
40
50
60
Hs-MIF-1 M P M F I V N T N V P R A S V P D G F - L S E L T Q Q L A Q A T G K P P Q Y I A V H V V P D Q L M A F G G S S E - - P C A L C S L
Bm-MIF-1 M P Y F T I D T N I P Q N S I S S A F - L K K A S N V V A K A L G K P E S Y V S I H V N G G Q A M V F G G S E D - - P C A V C V L
Ce-MIF-1 M P V F S I N V N V K V P A E K Q N E I L K E L S T V L G K L L N K P E Q Y M C I H F H E D Q G I L Y A G T T E - - P A G F A V L
Bm-MIF-2 M P L I T L A S N V P A S R F P S D F - N V Q F T E L M A K M L G K P T S R I L L L V M P N A Q L S H G T T E N - - P S C F T V V
Ce-MIF-3 M P M V R V A T N L P N E K V P V D F - E I R L T D L L A R S M G K P R E R I A V E I A A G A R L V H G A T H D - - P V T V I S I
Ce-MIF-2 M P V I K V Q T N V - - K K V S D G F - E V R L A I H M A K V M K R P E S Q I F V S L D M N S R M T R G Q L T D - - P L A V L D V
Ce-MIF-4 M Q V V R I Q T N I R S A D I P E K F - E Q D V I Y N L S V V M E L P A D K F V I I V E P A V R M R I G F E N K E I P V A I V N F
β
H4
H5
A
H6
B
β
B
%Identity to Bm-MIF-1
70
80
90
100
110
120
Hs-MIF-1 H S I G K I - G G A Q N R S Y S K L L C G L L A E R L R I S P D R V Y I N Y Y D M N A A N V G W N N S T F A 42
Bm-MIF-1 K S I G C V - G P K V N N S H A E K L Y K L L A D E L K I P K N R C Y I E F V D I E A S S M A F N G S T F G -
Ce-MIF-1 K S I G G V G S A K Q N N A I S A V V F P I I E K H L G I P G N R L Y I E F V N L G A A D I A Y N G Q T F A 41
Bm-MIF-2 K S I G S F - S A D K N I E Y S S S I S E F M K K T L D I D P A H C I I H F L N L D P E N V G C K G T T M K V L M K K 27
Ce-MIF-3 K S I G A V - S A E D N I R N T A A I T E F C G K E L G L P K D K V V I T F H D L P P A T V G F N G T T V A E A N K K 29
Ce-MIF-2 T S S T V L - T P I L T E E Y T V A L C E F F S Q E L A L D S D A V L I N Y R S L S P E L I G F N G H I L T E N R P F 23
Ce-MIF-4 Q T T R P S - S R I E N D S Y A K K L T S V L N E Q L K L D P A H I F I S F D F K D A K S F A T Q G K T I A S L Y E 26

C Cysteine residues required for thiol-oxioreductase activity
P N-terminal proline required for tautomerase activity
▼ Conserved introns of verebrate and nematode MIFs
▼ Conserved introns of nematode MIFs

Figure 6.6.2.3.1 *Bm*-MIF-1 (accession AAC82502) was aligned with the human MIF-1 (accession AAA21814), the *C. elegans* MIF homologues *Ce*-MIF-1(accession Z78012), *Ce*-MIF-2 (accession Z71259), *Ce*-MIF-3 (accession AL132860), *Ce*-MIF-4 (accession AC084197), and a second *B. malayi* MIF homologue *Bm-mif-2* (accession AAF91074) The cartoon shows the secondary structure of human MIF-1 (taken from the pdb summary of 1MIF structure on the CATH web site http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html) is aligned above the sequences. The active site proline and cysteine residues are marked in red and the conserved intron positions marked in red and green triangles. The identity of the protein sequences to *Bm*-MIF-1 is shown at the end of the alignment.

6.6.2.4 BMBAC101P19.2

Two splice variants were identified for BMBAC 01P19.2. BMBAC101P19.2b is missing the 25th and 26th exons of BMBAC101P19.2a. Predicted proteins for both transcripts show high levels of similarity to large multidomain proteins from humans (BAF180, accession AAG34760; (Xue *et. al.*, 2000)), chickens (accession JC5056; (Nicolas and Goodwin, 1996)), *D. melanogaster* (CG11375, accession AAF56339) and *C. elegans* (C26C6.1, T19481)(see figure 6.6.2.4.1). These proteins share six bromodomains (PGF00439), two BAH (Bromo Adjacent Homology, PF01426) domains, an HMG (High Mobility Group, PF00505) box and an anonymous C-terminal domain (PFB031551).

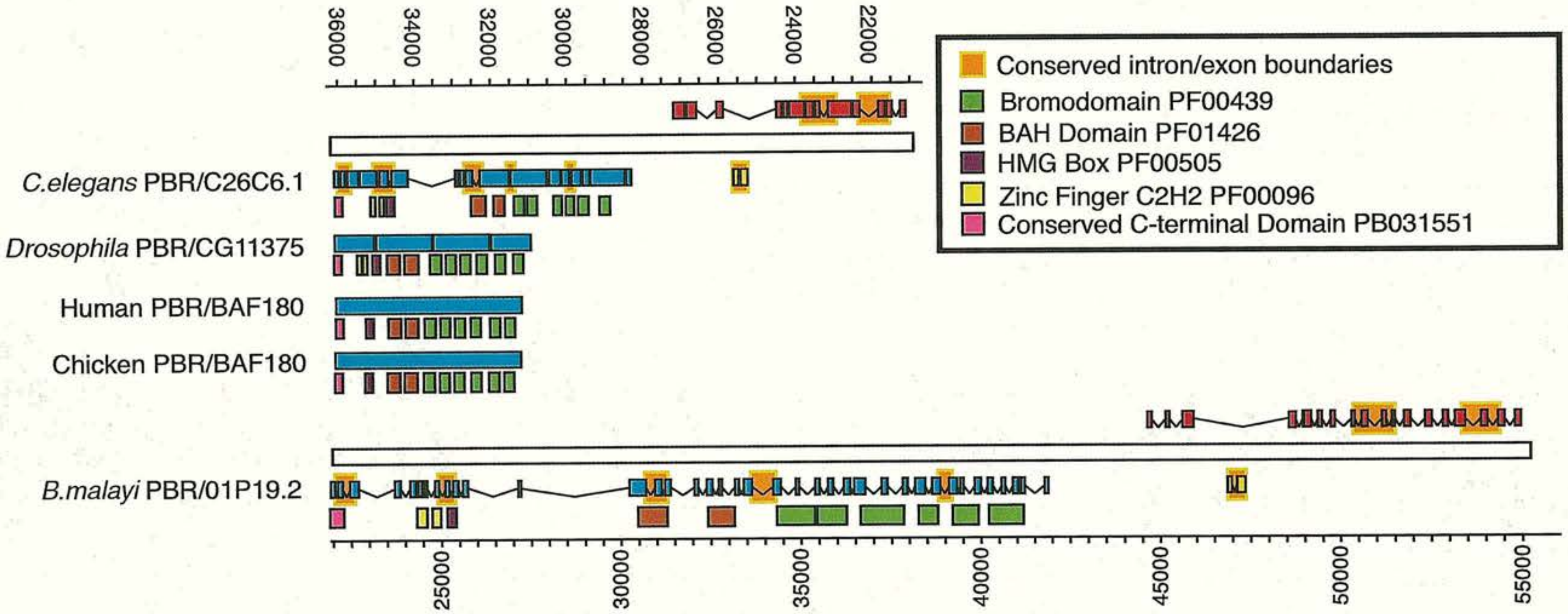


Figure 6.6.2.4.1 Comparison of the genomic organization of the pbr synteny cluster in *C. elegans* and *B. malayi* and the domain structure of the PBR homologues in *Drosophila melanogaster*, *Gallus gallus*, and *Homo sapiens*. Intron/exon boundaries that are conserved between both nematodes are shaded. Protein domains identified by searching Pfam are also indicated.

The *B. malayi*, *C. elegans* and *D. melanogaster* poly-bromodomain proteins (*pbr-1*) also have two zinc fingers (see figure 6.6.2.4.1). Bromodomains interact with acetylated lysine in histone complexes, while HMG boxes are found in chromatin proteins that bind to single stranded DNA and unwind double stranded DNA. This indicates this protein family may be involved in chromatin remodeling complexes. The human homologue of these proteins, BAF180, has been shown to localize to the kinetochores of mitotic chromosomes (Xue *et. al.*, 2000). None of the vertebrate BAF180 homologues have the zinc fingers found in the nematode and fly proteins. Zn fingers are known to be involved in binding to DNA so the nematode and arthropod *pbrs* may have different or additional functions. The *B. malayi* and *C. elegans* PBR-1 homologues show 33% identity (50% similarity) to each other.

6.6.2.5 BMBAC101P19.3

Two splice variants of BMBAC101P19.3 were isolated. BMBAC101P19.3b begins in exon three of BMBAC101P19.3a and is 5' transpliced to SL1. There is only one homologue of BMBAC101P19.3a in any organism, hypothetical protein F43G9.4 from *C. elegans* (46% identity in the N-terminal 100 aa). The N-termini of both BMBAC101P19.3a and *Ce*-F43G9.4 contain PWWP domains (PF00855). PWWP domains have been identified in a number of proteins with nuclear location that play roles in cell growth and differentiation (Stec *et. al.*, 2000; Stec *et. al.*, 1998). Psort profiling of BMBAC101P19.3 and *Ce*-F43G9.4 indicate that both proteins are likely to have nuclear localizations. BMBAC101P19.3b is missing exons one and two and thus does not have the PWWP domain. The predicted protein sequence of BMBAC101P19.3b has no similarity to any other sequences in the public databases.

6.6.2.6 BMBAC101P19.4

The protein encoded by BMBAC101P19.4 is homologous to the *C. elegans* protein T28F4.5 (30% identity). Using these two sequences as seeds, iterative searches against the GenBank non-redundant peptide database using PSI-blast (Altschul and Koonin, 1998) indicated that they belong to a group of small peptides which include the human DAP-1 protein (death associated protein). DAP-1 is a nuclear protein and is a positive regulator of IFN-γ induced apoptosis in HeLa cells (Deiss *et. al.*, 1995).

6.6.2.7 BMBAC101P19.5

Two splice variants for this gene were also isolated. BMBAC101P19.5b is missing the 11th exon of BMBAC101P19.5a. Only one homologue was found in any organism, hypothetical protein T28F4.4 from *C. elegans* (27% identity, 45% similarity). Psort profiling of BMBAC101P19.5 and *Ce*-T28F4.4 indicate that both proteins may have a nuclear localization. Because of its relative proximity to the *pbr-1* proteins BMBAC101P19.5 has been named *Bm-ubr-1* (upstream of *pbr-1*).

6.6.2.8 BMBAC101P19.7

BMBAC101P19.7 is the 3' fragment of a gene and has two splice variants. The C-terminal extension on BMBAC101P19.7b is the result of the splicing of an additional exon to an internal splice acceptor site in the 6th exon. Searches against PfamB revealed that both proteins have a anonymous domain (PFB005417) found in a variety of predicted proteins from other species. All of these proteins (including BMBAC101P19.7) are predicted to have multiple transmembrane segments, several of which span the Pfam domain. Two proteins in the predicted *C. elegans* genome contain this domain (C36B1.12 and T05E11.5). BMBAC101P19.7a/b are most similar to *Ce*-C36B1.12 (60% identity).

6.6.2.9Linkage conservation and conservation of synteny between the genomes of B. malayi and C. elegans

With the exception of BMBAC101L03.3 we have found orthologues in the *C. elegans* genome sequence for all of the genes predicted from the BMBAC101L03/BMBAC101P19 contig (see figure 6.6.1.1). All of the *C. elegans* genes except for *Ce*-Y56A3A.3 are found on chromosome I. One of the other MIF like genes in the *C. elegans* genome *Ce*-F13G3.9 (23% identity and 44% similarity) is found on *C. elegans* chromosome I in close proximity to the orthologues of *B. malayi* genes BMBAC101P19.2, 4 and 5. Eight of these ten *C. elegans* orthologues lie within a 2.3 Mb region in the center of chromosome I (6.7-9 Mb). The homologues/orthologues of the other two genes (BMBAC0L03.4 and

BMBAC101P19.6) are found at the distal tip of chromosome I around 14.2 Mb. This suggests that the within this section the composition of the genome (i.e the gene content) has been conserved between the two organisms.

Two groups of three genes in close proximity are conserved between the two genomes. In the first, *Ce*-F43G9.5 and *C*e-F43G9.3 are divergently transcribed from a 631bp intergenic region. *Ce*-F43G9.3 is followed by *Ce*-F43G9.4 in the same transcriptional orientation with 501 bp separating the genes. In *B. malayi* this microsynteny is conserved, except that two genes BMBAC101L03.3 and BMBAC101L03.4 are found in between BMBAC101L03.2 and BMBAC101L03.5.

In the second cluster (see figure 6.6.2.4.1) two large genes *Ce*-C26C6.1 and *Ce*-T28F4.4 are also divergently transcribed from a 5,274 bp intergenic region. A small third gene, *Ce*-T28F4.5 is found in the large third intron of *Ce*-T28F4.4 on the same strand and transcriptional orientation as *Ce*-C26C6.1. In *B. malayi* this microsynteny is conserved with BMBAC101P19.2 and BMBAC101P19.5 divergently transcribed and BMABAC01P19.4 sitting in the large third intron of BMBAC101P19.5 on the same strand and transcriptional orientation as BMBAC101P19.2.

While there has been extensive rearrangement of the order of the genes between the two genomes, when compared to the corresponding *C. elegans* genes ten of the *B. malayi* genes are in the same relative transcriptional orientation as their *C. elegans* orthologues.

6.7 Isolation and Characterization of *Bm-mif-2*locus

A probe for *Bm-mif-2* was synthesized by PCR amplification of the full length cDNA using vector primers T3 and T7 the *B. malayi* EST MBAFCZ1E9T3 as a template (Genbank accession AA257577). The PCR product was cleaned using a Microcon-100 (Millipore) and the vector portions of the PCR product removed by digestion with restriction enzymes *Xho*1 and *Eco*R1. The digested PCR product was cleaned with a Microcon-100 and 500 ng of DNA random prime labeled with biotin using the Phototope random primed labeling kit (New England Biolabs). The *Bm-mif-2* probe was hybridized to the high density nylon filter arrays containing 18,000 BACs as described above. Figure 6.7.1A shows the luminograph of the detected nylon filter.

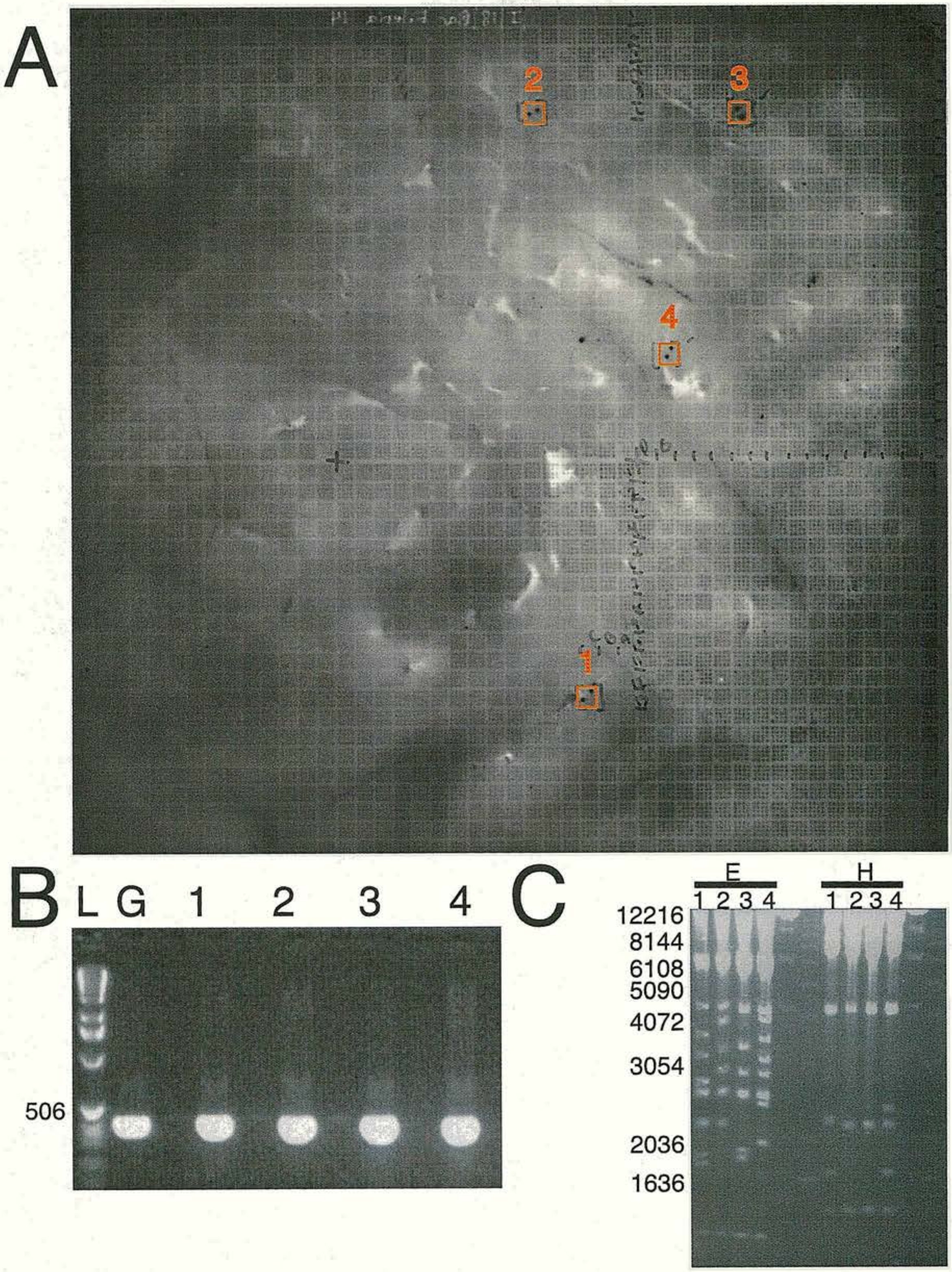


Figure 6.7.1 Hybridization of *Bm-mif-2* to BAC filter **A:** Luminograph of the high density BAC filter hybridized with the *Bm-mif-2* probe. Positive clones are highlighted and numbered in orange (1: BMBAC111C13, 2: BMBAC228F06, 3: BMBAC239K06, 4: BMBAC245O19). **B:** Hybridization positive BACs were tested for the presence of *Bm-mif-2* by PCR with gene specific primers the numbers correspond to the clones numbers listed above, G is a positive control amplified from prepared *B. malayi* genomic DNA. L: Gibco 1kb ladder (GibcoBRL). **C:** The BACs were digested with restriction enzymes E: *EcoR*1 and H: *Hind*III and run on a 0.7% agarose gel. The numbers correspond to the clones numbers listed above.

Four isolated hybridization-positive BACs (BMBAC111C13, BMBAC228F06, BMBAC239K06 and BMBAC245O19) were PCR tested for the presence of the *Bm-mif-2* gene using the gene specific primers Bm-MIF-2a/b.F4 and Bm-MIF-2.Z1E09.R1 with standard PCR reaction conditions and isolated BAC colony boils as template (see figure 6.7.1B). All four BACs were PCR positive for *Bm-mif-2*. BAC DNA was isolated by Midi preparation (Qiagen) and their ends of both clones sequenced with the vector primers. These end sequences were compared against each other and the public databases and specific primers designed and synthesized. The features of the end sequences of the four BACs are summarized in table 6.7.2.

B. malayi BAC end	Hits to *B. malayi* sequences	Hits to other filarial sequences	Hits to proteins sequences and *C. elegans* ORFs (chromosome)
BMBAC111C13.SP6	none	OVC00425	none
BMBAC111C13.T7	none	none	none
BMBAC228F06.SP6	none	none	none
BMBAC228F06.T7	Hha 1 repeat	none	none
BMBAC239K06.SP6	Related to BMC04100	none	acyl carrier protein *Ce*-F37C12.3 (III)
BMBAC239K06.T7	none	OVC00425	none
BMBAC245O19.SP6	BMC12236 BMC00207 BMC12237	OVC01395	ribosomal protein S14 *Ce*-F37C12.9 (III)
BMBAC245O19.T7	none	none	*Ce*-B0464.9 (III)

Table 6.7.2 The *B. malayi* BACs spanning the *Bm-mif-2* locus were compared to the public databases and hits to *B. malayi* or other filarial EST clusters as well as hits to proteins and *C. elegans* ORFs listed. The chromosome on which the *C. elegans* ORF is found is also noted.

Because the SP6 end of BMBAC111C13 and the T7 end of BMBAC239K06 were found to be overlapping sequences, only one set of primers was synthesized for these sequences (BMBAC11C13.SP6.F1 and BMBAC11C13.SP6.R1). Because the T7 end of BMBAC28F06 was exclusively HhaI repeat sequence (McReynolds *et. al.*, 1986) no primers were designed to this sequence. However, subsequent PCR tests with Hha repeat primers indicated that none of the other BACs min the *mif-2* contig contain the repeat (data not shown). It is possible that BMBAC228F06 is a chimeric clone with a fragment of an Hha 1 repeat array fused to the T7 end of the clone. To help determine the extent of the overlap of each of the clones PCR was performed with the end specific primers using BAC colony boils as templates. The PCR and end sequence data indicates that the four BACs extensively overlap with the T7 ends of BMBAC11C13 and BMBAC245O19 lying on the end of the contig (see figure 6.7.3B). The size of the BAC inserts were determined by digestion of prepared BAC DNA with *Hind*III and *Eco*R1 (see figure 6.7.1C). The restriction fragments indicate that the insert of BMBAC111C13 is ~45 kb, BMBAC228F06 is ~38 kb, BMBAC239K06 is ~43 kb and BMBAC245O19 is ~46kb. Because BMBAC111C13 and BMBAC245O19 are predicted to be at the ends of the *Bm-mif-2* contig and their inserts are extensively overlapping both BACs inserts will be shot gun sequenced together at the PSU, Sanger Institute.

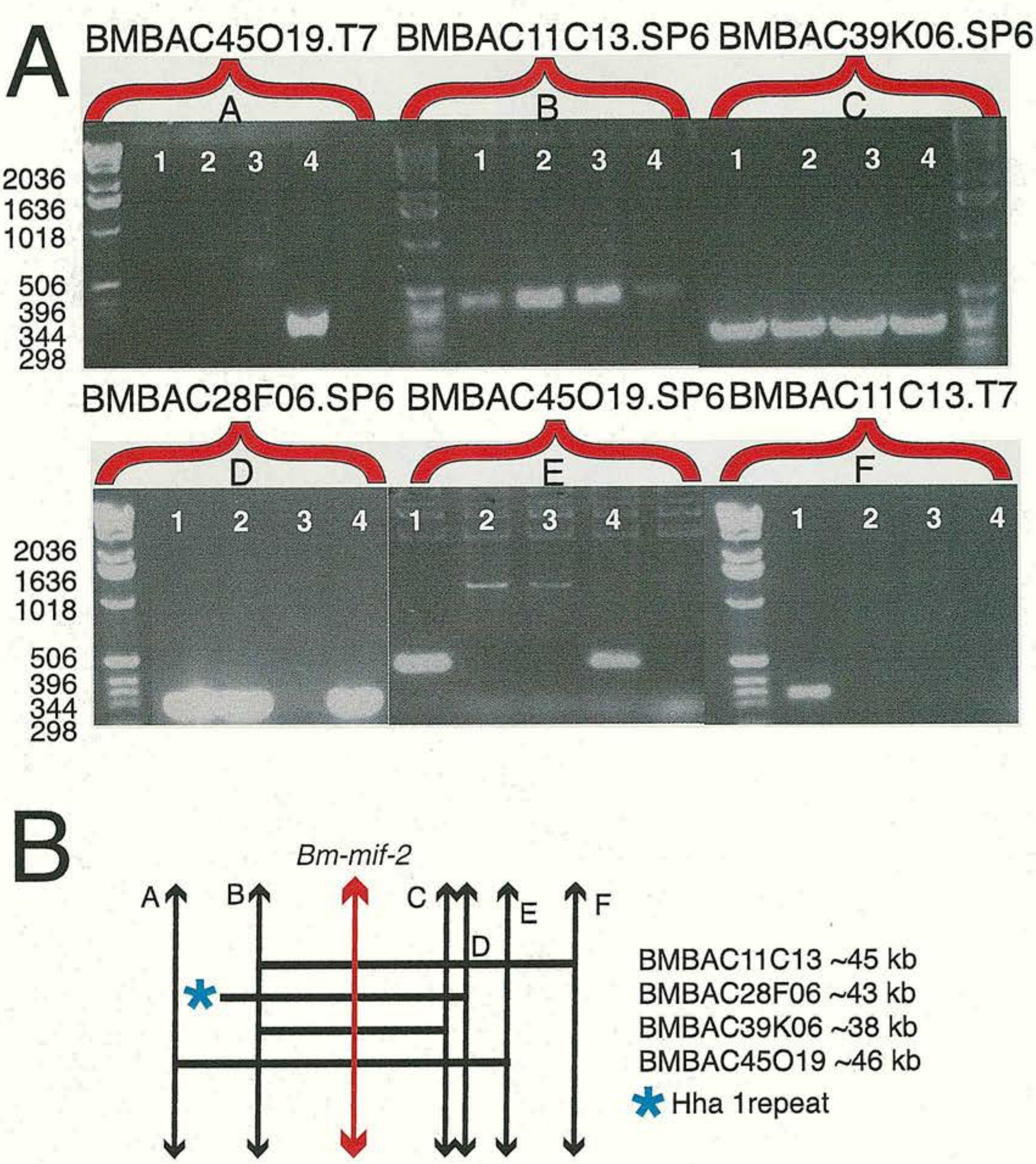


Figure 6.7.3A and **B** Determination of the overlap of *Bm-mif-2* containing BAC inserts **A:** PCR of the BACs in the *Bm-mif-2* contig with primers specific for the end sequences. The primer pairs used in the PCR are listed above the gel photo. The samples are loaded in the following order 1: BMBAC111C13; 2: BMBAC228F06; 3: BMBAC239K06; 4: BMBAC245O19. **B:** A cartoon summarizing the PCR results showing the extent of overlap between the different clones in the contig. The black arrows indicates PCR data was used to infer the overlap while the red arrow indicates hybridization data. The letters next to the black arrows indicate which primer pair was used (see A).

6.8 Preliminary analysis of the Bm-mif-2 contig

While full sequence of the *Bm-mif-2* contig is not yet available several interesting features have already been observed. Several of the BAC ends have matches to sequences in the public databases. BMBAC111C13.SP6 and BMBAC239K06.T7 has similarities to an *O. volvulus* EST cluster OVC00425. However, neither the BAC end sequences or OVC00425 has any similarities to any other sequences in Genbank. BMBAC239K06.SP6 has similarities to an acyl-carrier protein (*Ce*-F37C12.3) from *C. elegans*. BMBAC245O19.SP6 has similarities to *C. elegans rps-14* (*Ce*-F37C12.9) which also is found on the same cosmid. Examination of the *C. elegans* genome sequence reveals that *Ce*-F37C12.9 and *Ce*-F37C12.3 are divergently transcribed from a 1,150 bp intergenic region. In between the two genes a third *gene rpl-36* (*Ce*-F37C12.4) is found in the same transcriptional orientation as *Ce*-F37C12.3. The close proximity of *Ce*-F37C12.4 to *Ce*-F37C12.3 indicates that the two genes may be in an operon. PCR with primers to the *B. malayi rpl-36* has confirmed that this syntenic unit is conserved between the two species (data presented in chapter 7). Like the *Bm-mif-1* locus, the orthologue of the gene that initially identified the BACs in this contig (*Bm-mif-2*) is found on a different chromosome (*C. elegans* chromosome II) than the orthologues of other genes identified in the contig which are found on *C. elegans* chromosome III.

6.9 Analysis of *B. malay*i BAC end sequences

The synteny data from the *Bm-mif-1* and *Bm-mif-2* loci may not be representative of the whole *B. malayi* genome. To further expand this study the BAC end sequences deposited in Genbank (generated by the PSU and J. Daub ICAPB, Edinburgh University) were screened for clones which had matches on both ends to *C. elegans* proteins. The 5927 BAC end sequences were compared to Wormpep 43 with BLAST (Altschul *et. al.*, 1997). Thirty-six had significant matches (cut off $\leq e^{-8}$) on both the SP6 and T7 ends. The chromosomal location of the *C. elegans* protein with the best match was determined. Twenty of the thirty-six (55%) GSSs had matches to proteins on the same chromosome. If it is assumed that genes are randomly distributed between the autosomes in both nematodes than a basal level of intrachromosomal synteny of 20% would be expected. The rate observed is a much higher rate than would be expected to occur because of random segregation of both genes ($p<0.001$, Chi^2 test performed in Minitab, Minitab Inc). Like the data collected from the *Bm-mif-1* and *Bm-mif-2* loci the genes were often separated by large distances (average distance 4.3 Mb) on the *C. elegans* chromosomes (see figure 6.9.1A). While there was no general pattern of distribution within the chromosomes it is interesting to note that none of the syntenic sequences originated from *C. elegans* chromosome V. There also did not appear to be any general pattern to the distribution of the interchromosomal rearrangements (see figure 6.9.1B). Unlike the intrachromosomal rearrangements, there were several examples of interchromosomal rearrangements that involved genes with similarities to proteins from *C. elegans* chromosome V.

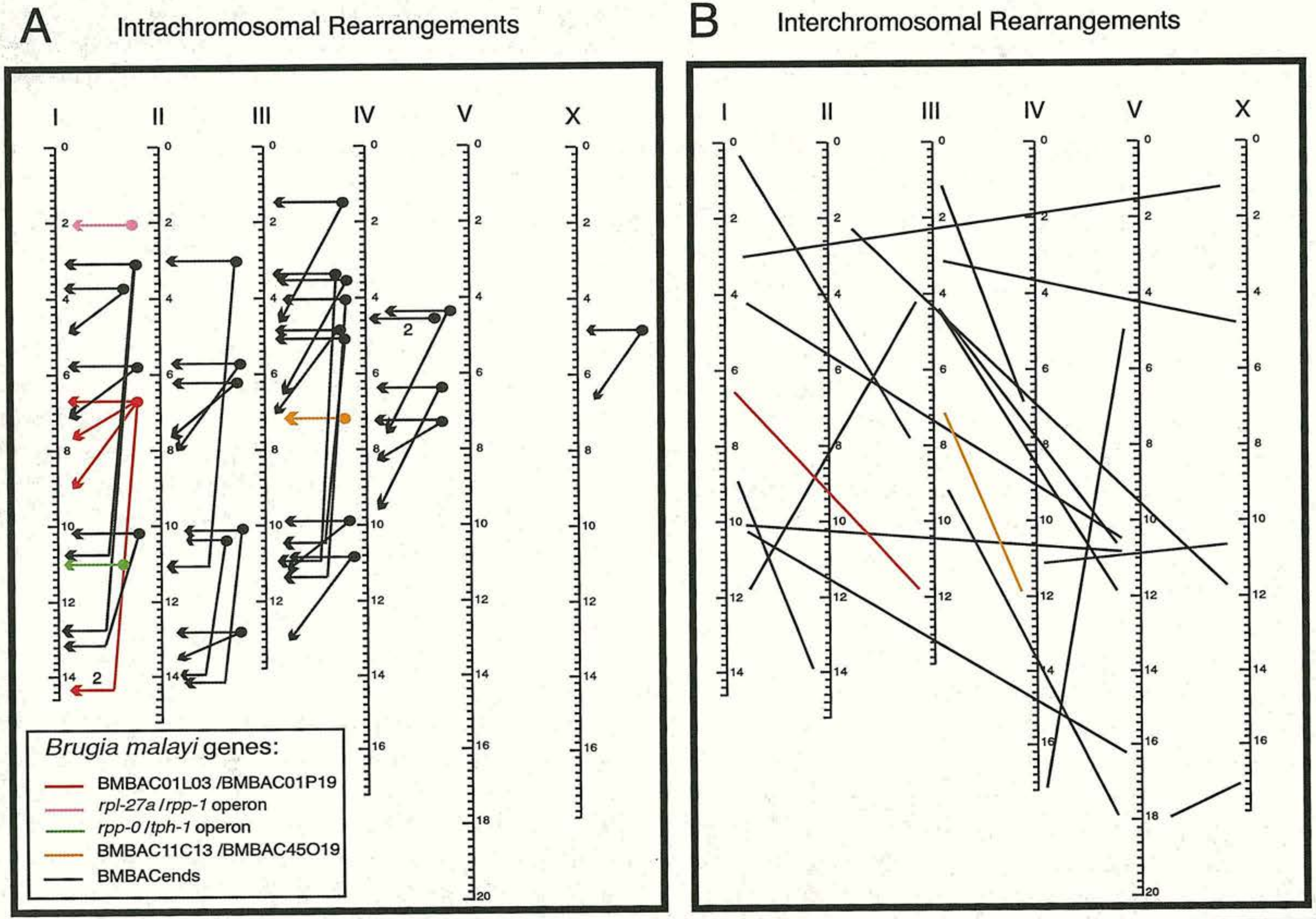


Figure 6.9.1 Genes identified in BAC end sequences show conservation of synteny between the genomes of *B. malayi* and *C. elegans* In the graphics the relative positions of the genes identified in *B. malayi* BAC end sequences have been mapped on the *C. elegans* chromosomes. **A:** those segments which show intrachromosomal rearrangements between the species while **B:** those segments which show interchromosomal rearrangements between the species.

6.10 Discussion

Brugia malayi is a parasitic nematode distantly related to the free-living model nematode *C. elegans* (Blaxter *et. al.*, 1998; Vanfleteren *et. al.*, 1994). Analysis of 83 kb of genomic DNA flanking the *B. malayi mif-1* locus has identified a 'fractured' conservation of microsynteny and conservation of linkage between the two nematode genomes. Twelve protein coding genes were predicted and eleven of these had orthologues in the *C. elegans* genome. The orthologues of ten of these genes are found on *C. elegans* chromosome I and eight of these genes are found in a 2.3 Mb segment. The other two are found at the distal tip of chromosome I. Some of these genes have remained tightly linked in the same or slightly modified relative transcriptional orientations. Preliminary analysis of the *Bm-mif-2* contig yields a similar result with a local cluster of 3 genes being conserved between the two species. The *C. elegans* orthologue of the fourth gene identified on the contig is found on the same chromosome as the cluster and is separated from the cluster by similar distances as the long range syntenic genes in the *Bm-mif-1* contig. These observations are reinforced by the analysis of the random BAC end sequences which indicates that genes shared between the two genomes are much more likely to share common linkage groups.

6.10.1 General features of B. malayi genomic DNA

This 83 kb fragment of *B. malayi* genomic DNA represents the largest contiguated portion of sequenced genomic DNA from a non-rhabditid nematode determined to date. Like *C. elegans, B. malayi* is thought to have a gene complement of ~20,000 genes (Blaxter *et. al.*, 2001) in a relatively compact genome (~100 Mb) (McReynolds *et. al.*, 1986). A large proportion (~60%) of genes identified in the *B. malayi* EST dataset (23,000 sequences) have no *C. elegans* homologue (Blaxter *et. al.*, 2001) using a BLAST search probability cutoff of $\leq e^{-8}$, supporting the idea that *B. malayi* has a reasonably diverse gene complement that is comparable to *C. elegans*. Notably, *C. elegans* orthologues could be identified for eleven of the twelve *B. malayi* genes identified in the BAC sequences. Some of these orthologous pairs had low pairwise identity, but were confirmed by shared intron positions. The criteria used to identify homologues in global searches with the ESTs would not have detected these pairs (probability values of e^{-10}) and thus the 'true' proportion of *B.*

malayi unique genes is likely to be much less than 60%. In the sequenced segment, an average of one gene has been predicted per 6.9 kb. The *B. malayi* genes were found to have larger and more numerous introns than the *C. elegans* genes (2.2 times longer and 1.7 times more frequent). Despite the difficulties presented by AT rich genomes with intron rich genes the gene predictions made by GeneFinder proved relatively accurate. If the BMBAC101L03/BMBAC101P19 contig is representative of a central region of an autosome this suggests that the *B. malayi* genome may be larger than estimated. The genome size should be reassessed.

6.10.2 Conservation of microsynteny

Synteny between the genomes of closely related eukaryotic organisms has been demonstrated in many taxa. However, it is only relatively recently that examples of conservation of microsynteny, that do not involve genes that are known to be functionally related (Brunner *et. al.*, 1999; Hamer *et. al.*, 2001). Most of the genes we have shown to have retained microsynteny between *C. elegans* and *B. malayi* do not fall into any clear functional categories when their homology data is compared. However, all the genes in the second microsynteny cluster (BMBAC101P19.2, .4 and .5) are predicted to have nuclear localization and two of the genes in the third microsyntenic cluster are ribosomal proteins (*rps-14* and *rpl-36*/ acyl-carrier protein operon).

It is also possible that the genes are not functionally linked but have promoters or *cis*-acting elements embedded within them which are required for the proper function of their synteny partners. This might account for conservation of the retention of the transcriptional orientation of the *pbr-1*, *dap-1*, *ubr-1, rps-14*, *rpl-36*/ acyl-carrier protein operon orthologues. Huynen *et. al.* have shown that coregulation of gene pairs in fungi increases the chances of that pair being retained in other fungal genomes (Huynen *et. al.*, 2001). This does not explain why the genes in the other example of microsynteny (BMBAC101L03.2, BMBAC101L03.5 and BMBAC101P19.3) have retained their arrangements, as the intergenic DNA between BMBAC101L03.2 and BMBAC101L03.5 has been interrupted by two other genes. Many genes in *C. elegans* are cotranscribed in operons (Blumenthal and Steward, 1997) and this could constrain synteny breakage. It is possible that BMBAC101L03.5 and BMBAC101P19.3 are in a transcriptional operon. The *C.*

elegans orthologues of BMBAC101L03.5 and BMBAC101P19.3 are separated by 501bp which is within the intergenic distance normally seen in *C. elegans* operons (Blumenthal and Steward, 1997). However, BMBAC101L03.5 and BMBAC101P19.3 are separated by 2.8 kb, which is outside the range of operonic intergenic spacing seen in *C. elegans*.

The function of many *C. elegans* genes have been investigated by RNA-mediated interference (RNAi) (Fraser *et. al.*, 2000; Maeda *et. al.*, 2001). In two of the clusters one gene has an RNAi phenotype, lethality (*Ce*-F43G9.5 the BMBAC101L03.2 homologue, (Maeda *et. al.*, 2001)) or altered post-embryonic morphology (*Ce*-C26C6.1 the BMBAC101P19.2 homologue, (Fraser *et. al.*, 2000)). The third cluster contains two ribosomal proteins, which are generally required for proper ribosome function. Therefore it is likely that the gene clusters are conserved because removing other members would be enough to interfere with the functions of the essential genes.

6.10.3 Long range conservation of linkage groups:

Many exceptions to the conservation of linkage have been identified in the study: the *Bm-mif-1*/*Ce-mif-1* and *Bm-mif-2*/*Ce-mif-2* orthologous pairs and the sixteen examples found in the BAC end sequence data. However, all of the other *C. elegans* orthologues of the *B. malayi* genes are found in the BMBAC101L03/BMBAC101P19 contig and *Bm-mif-3* contig are found on *C. elegans* chromosomes I or III. Another *C. elegans* MIF homologue, *Ce-mif-3*, is found on chromosme I in close proximity to the genes in the *pbr-1* synteny cluster suggesting that a gene conversion event may have obscured synteny of this gene.

The data yielded from the BAC sequence contigs and BAC end sequences indicate that genes in the *B. malayi* genome have undergone extensive rearrangement relative to the rhabditid genomes. While several gene clusters have been conserved most of the genes have been radically reorganized. This data supports the observations of the mutable order of genes observed in other protosomes. However, while it is apparent that the rate of evolution in these genomes is relatively high it is unclear if the mechanisms rearranging the genes are similar. The observed rearrangments could result from large inversions or interchromsomal translocations. It is also unclear why the gene complement within chromosomes would be constrained. In the metazoa the only examples of extensive long range synteny between the genomes of distantly related species (>300 Myr divergence) have been identified in vertebrates (teleost fish and humans (Grant *et. al.*, 2000; Ku *et. al.*, 2000)). In these organisms the movement of genes within the genomes appear to be constrained. Large sections of the genome are mosaicly arranged in different species. In mammals interchromosomal exchanges are rare events though some lineages, such as rodents, show elevated rates. As more protostome and basal deuterostome genomes are mapped and sequenced we will be able to determine how common this phenomenon rapid genome reshuffling is in other metazoa and if some taxa show slower rates of genome rearrangment then the nematodes and dipterans. One puzzling feature of the nematode and dipteran data is the infrequency of interchromosomal translocations. If there are no functional constraints linking most of their genes why would there be a bias in the types of rearrangements that occur? It is possible that rearrangments within chromosomes are more easily accomplished than interchromosomal translocations. Mechanistically this may be because they require fewer DNA breaks than interchromosomal translocations and the nuclear scaffold may hold local chromosomal regions in closer association. Because the data from the *C. brggsae* analysis is composed of small fragments it is difficult to use the *C. elegans* and *C. briggsae* comparison to establish what is occurring through the whole linkage groups in the rabditids. However, the imminent release of the whole genome sequence of *C. briggsae* will allow these comparisons to be completed. The data from the *Bm-mif-1* and *Bm-mif-2* contigs and the BAC end sequence data does not reveal any general patterns of gene rearrangement in nematodes. The *C. elegans* orthologues of the genes found on the BMABC01L03/BMBAC101P19 contig do tend to be found in a local segment of chromosome I (nine of eleven genes are in 2.3 Mb or 16% of the chromosome). This observation is supported by the preliminary data found in the analysis of the *Bm-mif-2* contig and BAC end sequence data and suggests that local rearrangements in nematodes intrachromosomal inversions/rearrangements have occurred more frequently than long-range intrachromosomal, or interchromosomal rearrangements.

6.10.4 Conclusions:

Because the sample size of the *B. malayi* genomic DNA we have surveyed is relatively small only tentative conclusions about the global extent of conservation of linkage and microsynteny between these two genomes can be drawn. However, the portions analyzed have conclusively demonstrated that such structures do exist even between these very distantly related species. The results yielded by the BAC end sequence dataset indicate that intrachromosomal rearrangements are more likely to be found than interchromosomal translocations. However despite the observed conservation the high rate of rearrangement of genes within chromosomes makes it unlikely that the positional information of genes in the rhabditid genomes will be useful in finding orthologous genes in the *B. malayi* genome. As the BAC end sequence dataset is expanded and large portions of the *B. malayi* genome are sequenced it may become possible to define what factors drive the rate of these exchanges and what maintains tight linkage between functionally unrelated genes.

Chapter 7

Operons and the resolution of polycistronic transcripts in nematodes

7.0 Introduction

Intermolecular ligation of RNA molecules, *trans*-splicing is a process that has been show to occur in all eukaryotes tested. These range from single celled protozoa such as trypanosomes and ciliates, to complex multicellular organisms such as nematodes, insects, mammals and plants (Been and Cech, 1986; Blaxter and Liu, 1996; Conklin *et. al.*, 1991; Dorn *et. al.*, 2001; Eul *et. al.*, 1995; Murphy *et. al.*, 1986; Perry and Agabian, 1991; Sutton and Boothroyd, 1986). The ubiquitous nature of this process indicates it must have essential functions in these groups. One common *trans*-splicing reaction is the addition of a mini-exon called a spliced leader (SL) to the 5' end of messenger RNAs. SLs have been isolated from a variety of protozoa and animals. Figure 7.0.1 shows a cartoon of the phylogenetic relationships of eukaryotic life adapted from Bauldauf *et. al.* ((Baldauf *et. al.*, 2000)). A portion of the tree, which includes the animal section of the crown group has been magnified (adapted from Blaxter 1998 (Blaxter, 1998)). Those groups from which SL have been isolated are indicated. To date only two of the major groups have been found to utilize SLs: the Eugleniods and the Metazoa. Within the eugleniods all members surveyed have been found to add SLs to the 5' end of their mRNAs. These include the free living protozoa *Euglena* and the parasitic trypansomatids (Frantz *et. al.*, 2000; Perry and Agabian, 1991). Within the Metazoa, SL usage has been found in several groups including the cnidarians (Stover and Steele, 2001), which are thought to be at the base of the metazoan group, indicating that this process may have evolved early in metazoan evolution. In the protostomes, nematodes and platyhelminths have been found to utilize SLs (Blaxter and Liu, 1996; Davis *et. al.*, 1995; Vandenberghe *et. al.*, 2001). Within the nematodes, one SL, SL1 has been found in all species tested, with only one reported base change in any species surveyed (Koltai *et. al.*, 1997). Unlike the nematodes the platyhelminths show diversity between species in the SL mini-exon sequences (Davis, 1997). To date in the deuterostomes only one group, the urochordates, have been shown to utilize SLs (Vandenberghe *et. al.*, 2001). Within both the protostomes and deuterostomes several species have been shown not to utilize SLs including *Drosophilia melanogaster*, *Homo sapiens* and *Mus musculus*. This indicates that if SL *trans*-splicing to mRNAs was a process that evolved early in metazoan evolution then it has been lost by several important lineages. However it has been shown that SL *trans*-splicing can be reconstituted in mammalian systems by expression of the nematode SL1 snRNA indicating that the machinery necessary for *trans*-splicing is still intact (Bruzik and Maniatis, 1992).

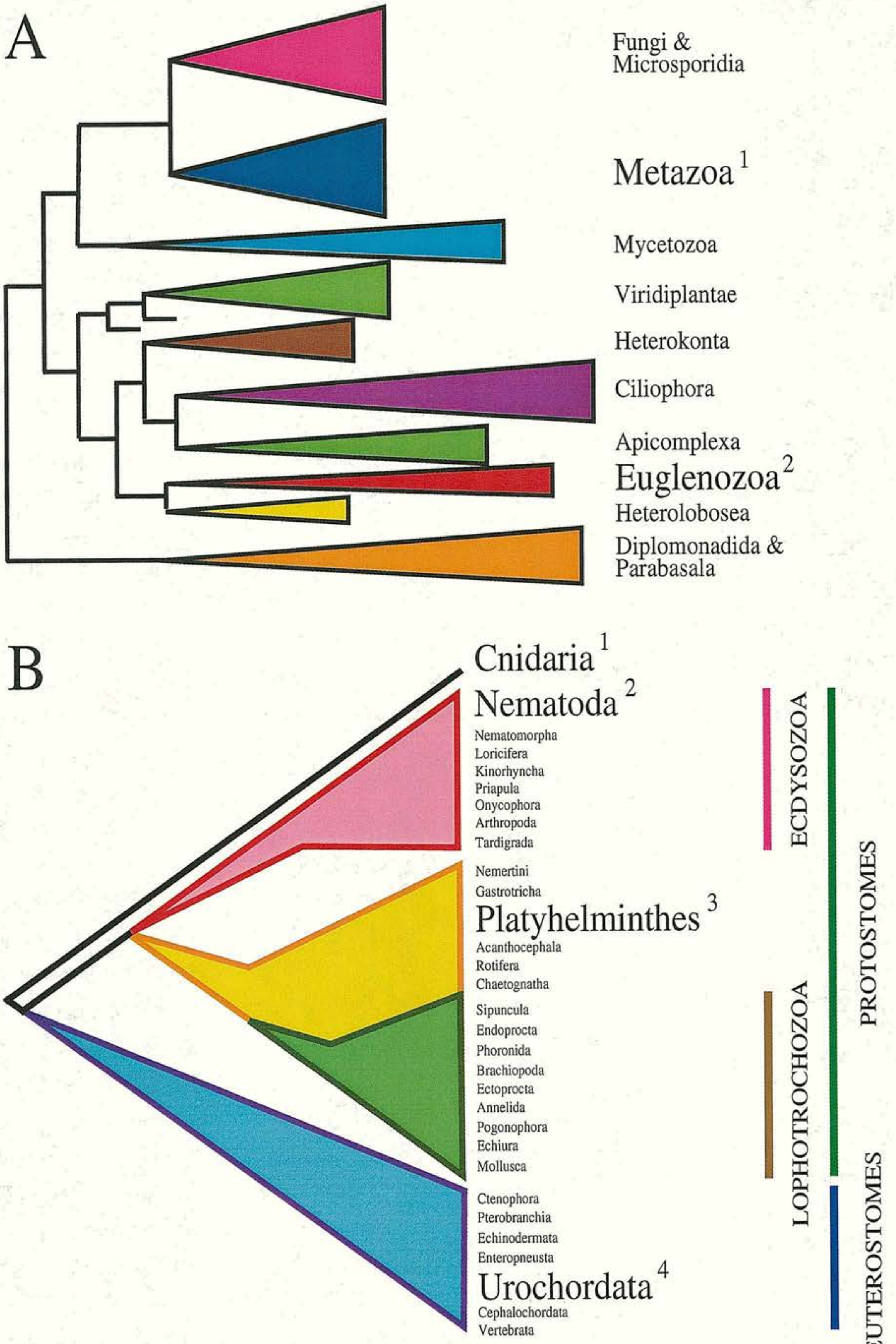
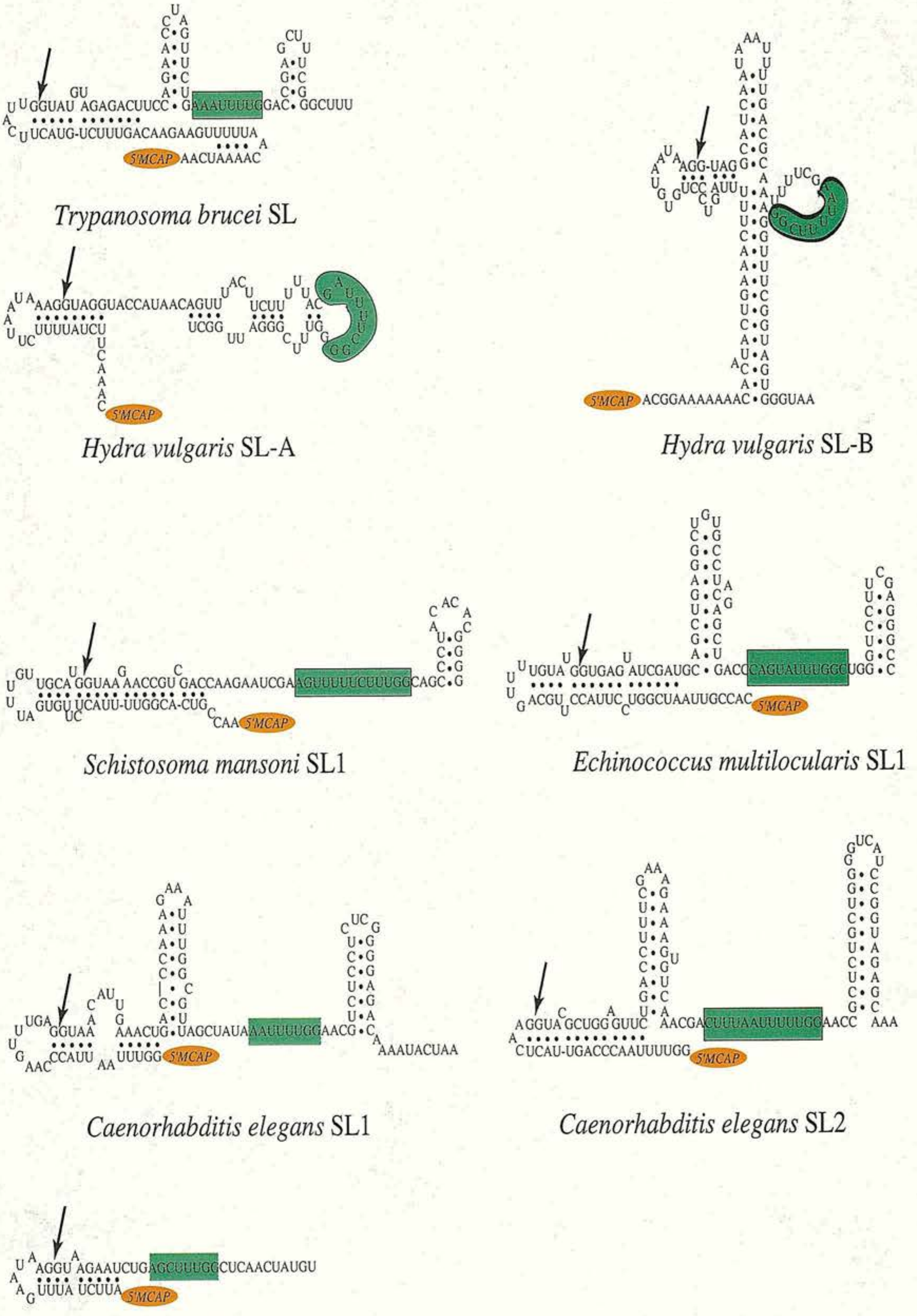


Figure 7.0.1A Phylogentic distribution of SL usage across the eukaryotes. Those groups utilizing SLs are in bold. A: the phylogenetic distribution of SLs through the eukaryotes adapted from Baldauf et. al. 2000. **1** (see references below) **2** Frantz, *et. al.* 2000 and Perry, *et. al.* 1991).
Figure 7.0.1B Phylogenetic distribution of SLs through the metazoa adapted from Blaxter 1998. **1** Stover *et. al.* 2001, **2** Blaxter *et. al.* 1996, **3** Davis *et. al.* 1995, **4** Vandenberghe *et. al.* 2001.

Genomic analysis of SL genes have shown they exist in tandem arrays (Blaxter and Liu, 1996). The only known exception to this observation is the SL2 gene family from *C. elegans*. In some organisms (eugleniods, nematodes and cnidarians) these arrays are associated with the 5S ribosomal genes (Blaxter and Liu, 1996; Frantz *et. al.*, 2000; Stover and Steele, 2001). While the primary sequences of these SLs are not conserved the overall gene structure as well as the predicted and experimentally verified secondary structure of the SL RNAs have several common features. All SL genes have two sections, the mini-exon (the portion spliced onto the 5' end of the mRNAs) which ranges from 18-41 nts in length is found at the 5' end of the gene. The intron portion of the gene follows the mini-exon and ranges from 44-128 nts in length. This portion of the RNA binds to the splicoseomal components and shows more sequence heterogeneity than the mini-exon. With the exception of *Ciona intestinalis*, all SL RNAs described to date have at least two stem loops which are believed to be important for interaction with the snRNAs which make up the splicing complex (Vandenberghe *et. al.*, 2001). The SL intron sequences contain a predicted Sm binding site, which is essential for SL function in several systems (Maroney *et. al.*, 1990; Sturm and Campbell, 1999). All SLs tested have also been shown to have methylated cap structures at their 5' ends, although the composition of these caps differs between organisms. Figure 7.0.2 shows the secondary structures of a selected set of SLs. *In vivo* and *in vitro* structure function studies in both trypanosomatids and nematodes have identified portions of the molecule which are important for the *trans*-splicing reaction. These studies have highlighted several important differences between the nematode and trypansome *trans*-splicing systems. Within the trypanosomatids it has been found that while the composition of the mini-exon is flexible its length must remain constant to maintain *trans*-splicing activity. *In vitro* and *in vivo* studies have shown that the sequence of the nematode SL mini-exon can be replaced or the majority of the leader deleted and functional *trans*-splicing still maintained (Lucke *et. al.*, 1996; Maroney *et. al.*, 1991). However, in both groups the Sm binding region and the second and third stem loops are important for association with the splicing complex as well as the *trans*-splicing reaction.



Trypanosoma brucei SL

Hydra vulgaris SL-A

Hydra vulgaris SL-B

Schistosoma mansoni SL1

Echinococcus multilocularis SL1

Caenorhabditis elegans SL1

Caenorhabditis elegans SL2

Ciona intestinalis SL1

Figure 7.0.2 The secondary structures of SLs representative of every major group. The location of the methy-cap structures at the 5' end of the molecules is shown with orange circles. The locations of the binding sites for the Sm protein are shown with green boxes. The junction of the mini-exon and intron is indicated with an arrow.

The mechanisms underlying the *trans*-splicing process also appear to be conserved between different groups of organisms. Mechanistically the *trans*-splicing process is very similar to *cis*-splicing which removes introns from internal segments of the mRNA. Like *cis*-splicing, in *trans*-splicing the donor (SL-snRNA) has a 5' splice site consensus (beginning in GT) and the acceptor (the target mRNA) has a 3' splice acceptor site (ending in AG) (Hannon *et. al.*, 1990; Laird, 1989). Many of the snRNAs which are required for *cis*-splicing (U2, U4, U5 and U6) are also required for the *trans*-splicing reaction (Palfi *et. al.*, 1994; Xu *et. al.*, 2000). The SL-snRNA is believed to replace the U1-snRNA in the splicing complex. In *trans*-splicing the reaction proceeds with the creation of a branched intermediate molecule which is similar to the lariat intermediate found in *cis*-splicing reactions.

In vivo and *in vitro* experiments indicate that SLs have important roles in the processing and translation of mRNAs. The presence of the SL close to the initiating methionine raises the possibility that SLs interact with the ribosome. Studies with extracts from *Typanosoma brucei* (Euglenozoa) and *Ascaris suum* (Nematoda) indicate that the presence of the SL may stimulate translation (Maroney *et. al.*, 1995; Moreno *et. al.*, 1991). Interestingly in both trypanosomes and nematodes SLs have been shown to play an important role in RNA processing by separating polycistronic transcripts (Zorio *et. al.*, 1994). These polycistrons are similar to bacterial operons in that they are sets of genes in the same transcriptional orientation driven by a common promoter. Unlike bacterial operons, in trypanosomatid and nematode operons translation of proteins does not occur by binding of the ribosomes directly to the polycistronic mRNA. Instead the polycistronic mRNA is processed and the individual cistrons separated by trans-splicing of the SL to the 5' end and polyadenylation of the 3' end of each transcript (see figure 7.0.3).

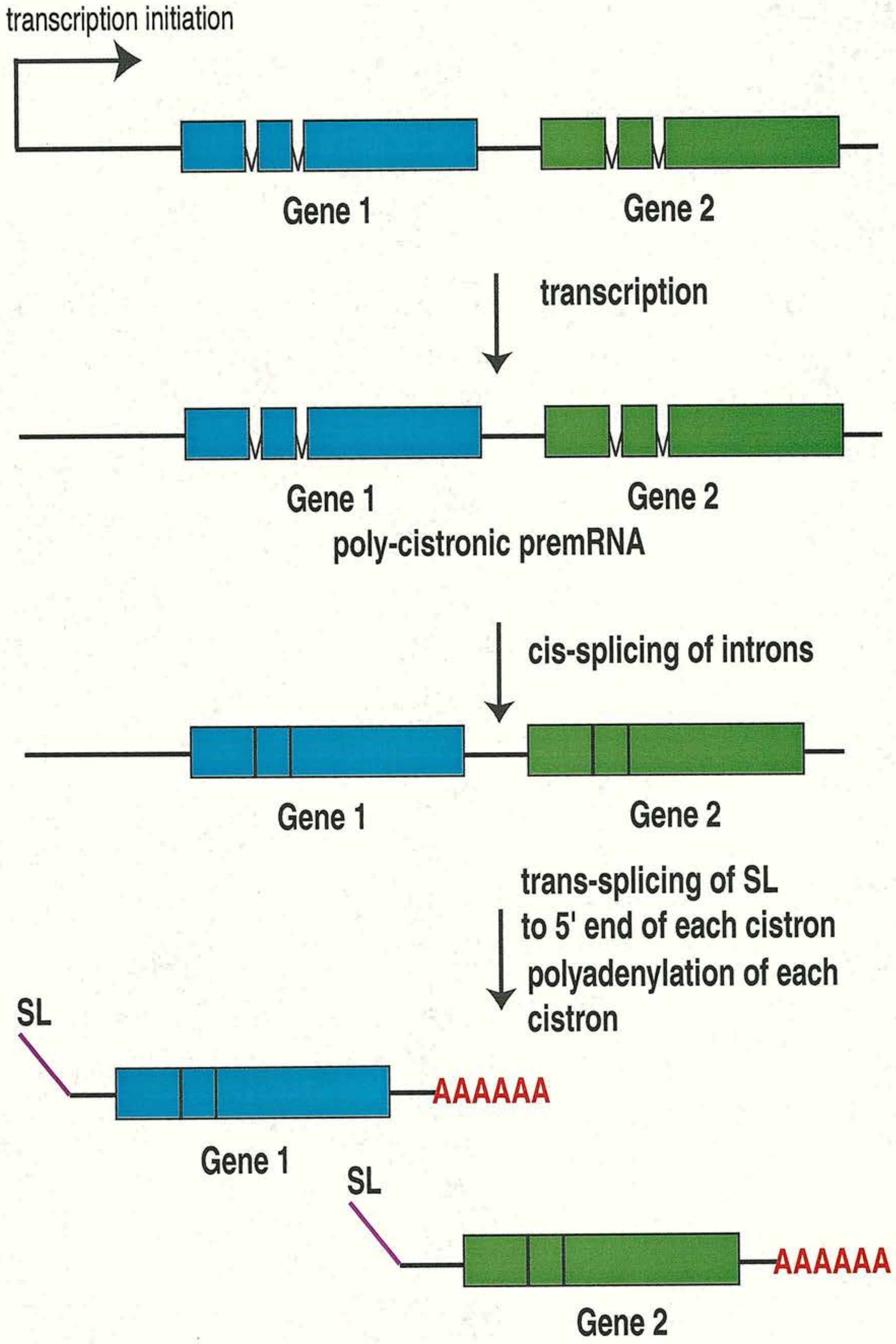


Figure 7.0.3 A schematic representation of the events occurring during the processing of polycistronic mRNAs from transcription of the polycistron and separation of the individual cistrons by *trans*-splicing and polyadenylation.

Also unlike bacterial operons, which generally contain genes which function in common biological processes, most operons found in these eukaryotes do not appear to contain genes with obvious biological interactions, although some exceptions to this rule have been published (Combes *et. al.*, 2000; Hough *et. al.*, 1999; Page, 1997; Treinin *et. al.*, 1998). The size of operons in trypanosomatids and nematodes indicate they play very different roles in the organization of the genes they contain. In trypanosomes most protein coding genes are organized in operons (>10 genes) which span large portions of the chromosome. The genes in these operons share a single promoter and transcription initiation site indicating that they are probably all transcribed at the same time and rate (Myler *et. al.*, 2001; Myler *et. al.*, 2000). This suggests that mRNA stability or translation may be more important mechanisms of control of gene activity in these organisms. This had already been observed in studies of individual genes (gp63, hsp83; (Brittingham *et. al.*, 2001; Zilka *et. al.*, 2001)) In nematodes, operons are much smaller, usually containing between two to three genes which could indicate that in nematodes gene expression is an important mechanism for controlling gene function (Blumenthal and Steward, 1997). Also the processing events that separate the cistrons occur at different times. In trypanosomes the *trans*-splicing of the SL to the 5' end of the cistron appears to occur before polyadenylation of the transcript (Sturm *et. al.*, 1999). In nematodes both processes appear to occur simultaneously with removal of introns (*cis*-splicing) (Spieth *et. al.*, 1993).

7.1 Nematode operons

In nematodes operons have only been extensively studied in the free-living nematode *Caenorhabditis elegans*. Analysis of the available *C. elegans* genome sequence indicates that approximately 13% of the genes are organized in ~850 operons. These operons contain an average of 2.6 genes (ranging from 2-7 genes) which are each separated by an average of 126 bp (Blumental *et. al.* 2001 unpublished). These operons show some unique features, which indicate the processes leading to the resolution of nematode polycistronic transcripts are different from those found in protozoa. Nematode operons were initially discovered after the isolation of a second nematode SL (SL2) at the 5' end of the *Ce-gpd-3* gene (glyceraldehyde-3-phosphate-dehydrogenase)(Huang and Hirsh, 1989). Subsequent analysis indicated that *Ce-gpd-3* received SL2 exclusively and that SL2 was *trans*-spliced to the 5' end of a variety of mRNAs (Spieth *et. al.*, 1993). Cloning and sequencing of the genomic region around the *Ce-gpd-3* gene revealed that it lay downstream of two other genes *Ce-mai-1* (mitochondrial atpase inhibitor) and *Ce-gpd-2* in the same transcriptional orientation (Huang *et. al.*, 1989). The intergenic spaces between these three genes were small, 245 bp and 309 bp respectively. Analysis of other SL2 containing mRNAs revealed they resided in similar gene clusters. RT-PCR analysis of the *Ce-mai-1/gpd-2/gpd-3* operon revealed that the three genes were transcribed as a single polycistronic transcript (Spieth *et. al.*, 1993). Experiments using transgenic *C. elegans* carrying *gpd-2* and *gpd-3* downstream of a heat shock promoter showed that *Ce-gpd-3* expression was dependent on heat shock indicating that it did not possess its own promoter (Spieth *et. al.*, 1993). The specificity of SL2 *trans*-splicing to *Ce-gpd-3* in this construct was dependent on the location of the promoter being upstream of the first gene (Spieth *et. al.*, 1993). When the *Ce-gpd-2* polyadenylation site was mutated, polycistronic mRNAs accumulated indicating that the maturation of the downstream mRNA is dependent on the maturation of the upstream mRNA (Spieth *et. al.*, 1993). Finally SL2 *trans*-splicing to genes that normally receive SL1 could be engineered by inserting these genes into constructs that placed them downstream of *Ce-gpd-2* and the intergenic region that separated *Ce-gpd-2* from *Ce-gpd-3*. This body of evidence indicated that SL2 *trans*-splicing was a process that was linked to the processing of genes downstream in operons. Since then several other studies have shown that while SL2-like SLs are the primary SLs added to downstream genes in operons, a small percentage of the transcripts receive SL1 (Hough *et. al.*, 1999). Low levels of SL2 addition to genes that are not in operons, or are the first gene in operons, has recently been reported (Hough *et. al.*, 1999). The functional interchangeability of SL1 and SL2 is also supported by studies with *Ce-rrs-1* (e2482) mutants which lack the 5S/SL1 ribosomal cluster. Normally a homozygous *rrs-1* mutant has an embyronic lethal phenotype. However embryonic lethality can be partially rescued (the larva completes embryogensis) by transformation with plasmid constructs containing tandem arrays of SL1 or SL2 genes (Evans and Blumenthal, 2000; Ferguson *et. al.*, 1996). Evans *et. al.* and Huang *et. al.* have established some of the mechanisms which make SL2 *trans*-splicing specific to genes down stream in operons (Evans *et.*

al., 2001; Huang *et. al.*, 2001). In a systematic screen for regulatory sequences in the intergenic region of *Ce-gpd-2* and *gpd-3* only a U-rich region 29bp downstream of the *Ce-gpd-2* polyadenylation site was necessary for specifying SL2 *trans*-splicing. This region was also functional when put in a heterologous context indicating that no other sequence elements within the *Ce-gpd-3* gene are required. Huang *et. al.* suggested that this U-rich region could be a binding sequence for the cleavage stimulation factor (CstF) complex and that successful binding of the polyadenylation complex could be required for SL2 *trans*-splicing (Huang *et. al.*, 2001). Evans *et. al.* have since shown that the SL2 snRNP specifically co-precipitates with CstF-64, a subunit of the CstF complex (Evans *et. al.*, 2001). Mutation studies indicate that while stem loop II is important for snRNP complex formation and *trans*-splicing, stem loop III confers operonic *trans*-splicing specificity and CstF-64 binding (Evans *et. al.*, 2001). SL2 stem loop III mutants are still capable of *rrs-1* rescue indicating that they can still function as SL donors *in vivo* (Evans *et. al.*, 2001). This data provides the first clues as to the mechanisms underlying the specificity of SL2 *trans*-splicing. The coupling of the SL2 snRNP to the polyadenylation complex may allow it to out-compete SL1 *trans*-splicing to the splice acceptors in the intergenic region and perhaps make the whole process of polyadenylation and *trans*-splicing more efficient.

Analysis of the *C. elegans* genome has revealed that it utilizes different types of operons, some of which control mRNA levels by differential transcription of individual cistrons. The canonical operon exemplified by the *Ce-mai-1/gpd-2/gpd-3* operon (type I) is the most common form. These operons generally contain two to six genes separated by small intergenic segments, usually ranging from 50 to 500 bp (Blumenthal and Steward, 1997). The first gene in the operon usually receives SL1. *Ce-mai-1* is an exception to this rule and does not receive any SL (Spieth *et. al.*, 1993). The genes downstream in the operon receive a mixture of SL1 and SL2 with SL2 usually being predominant. Again the *Ce-mai-1* operon is an exception to this rule with *Ce-gpd-2* receiving a mixture of SL1 and SL2 and *Ce-gpd-3* receiving SL2 exclusively. In another form of operon (type 2) exemplified by the *Ce-cyt-1/ced-9* operon there is no intergenic segment and the site of polyadenylation of *Ce-cyt-1* is adjacent to the SL acceptor site of *Ce-ced-9*. In this operon *Ce-ced-9* is *trans*-spliced to SL1 exclusively (Williams *et. al.*, 1999). Mutational studies of the polyadenylation site and SL acceptor site indicate that both processes are occurring simultaneously and in competition which could indicate that the generation of either transcript could be mutually exclusive. This form of transcriptional control has also been observed in polycistrons like *Ce-unc-17/cha-1* which share a single 5' untranslated exon but are otherwise not overlapping. *Ce-unc-17* is contained in the first intron of *cha-1* and alternative splicing of the first exon either to *Ce-unc-17* or *Ce-cha-1* ensures that both species will not be simultaneously generated from the same mRNA (Rand, 1989). In a third type of polycistron the *Ce-lir-2/lir-1/lin-26* operon actually contains two polycistrons *Ce-lir-2/lir-1A,B,C* or *Ce-lir-1D,E,F/lin-26*. GFP transcriptional fusions indicate that while *Ce-lir-2/lir-1A,B,C* are ubiquitously expressed *Ce-lir-1D,E,F/lin-26* is expressed in nonneuronal ectodermal tissue (Dufourcq *et. al.*, 1999). This expression in driven by a promoter embedded in the large (9kb) first intron of *lir-1*. So sections of single polycistrons can be differentially expressed if secondary promoter elements and transcription initiation sites are embedded within the individual cistrons. Figure 7.1.1 illustrates the different types of operons identified in the *C. elegans* genome

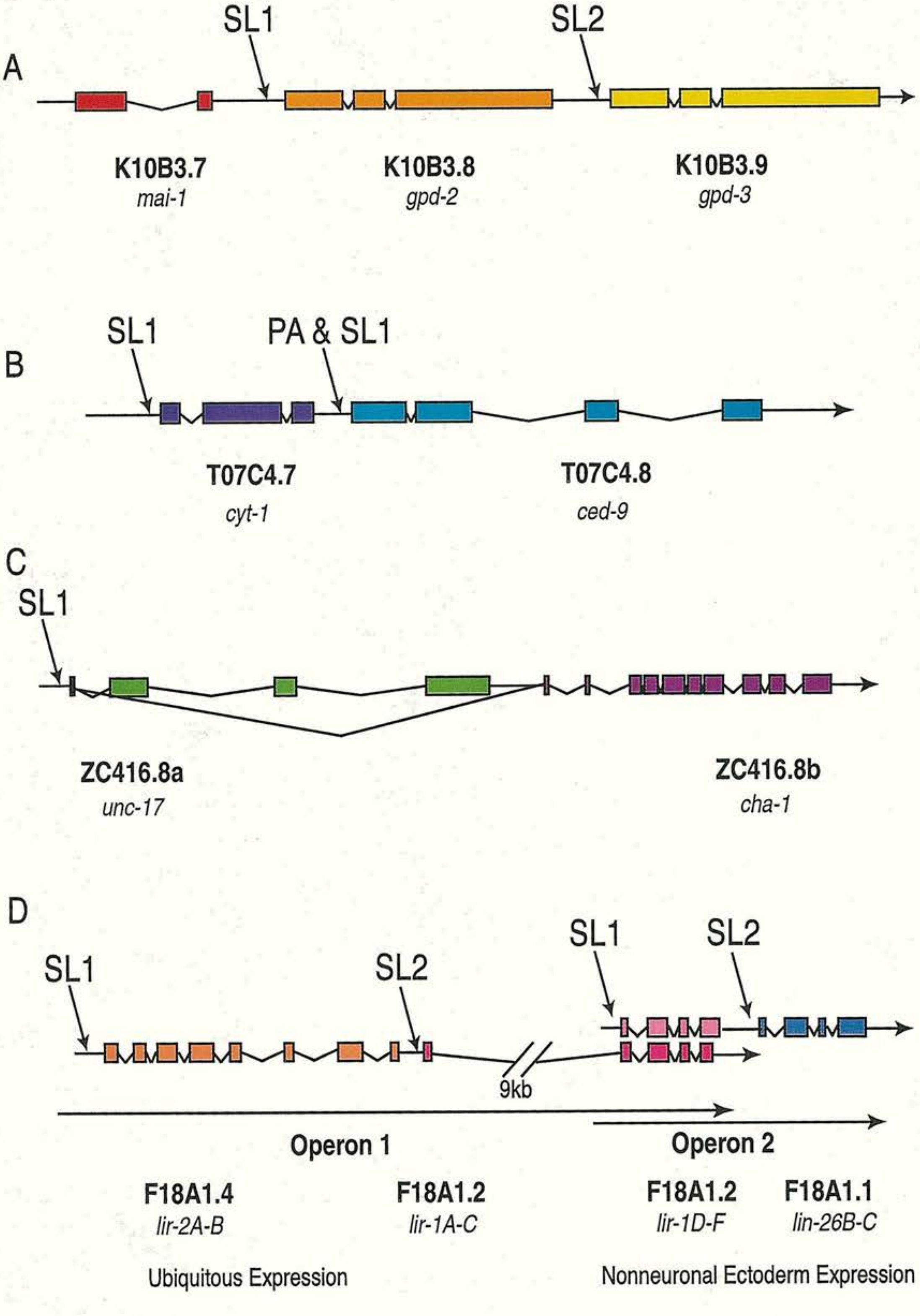


Figure 7.1.1 The different varieties of operons identified in the *C. elegans* genome. **A:** Type I operon *Ce-mai-1/gpd-2/gpd-3* operon, **B:** Type II operon *Ce-cyt-1/ced-9*, **C:** *Ce-unc-17/cha-1*, **D:** Type III operon *Ce-lir-2/lir-1/lin-26*.

7.2 Operons outside of *C. elegans*

Outside phylum nematoda there has been no definitive proof that any other metazoa organize their genes in operons. Dicistronic transcripts have been described from *Drosophila*, mammals and plants. However, their constituent open reading frames are not separated before translation (Andrews *et. al.*, 1996; Brogna and Ashburner, 1997; Garcia-Rios *et. al.*, 1997; Girardet *et. al.*, 1996; Szabo *et. al.*, 1994; Walker *et. al.*, 1996). Davis and Hodgson have described a potential operon from the parasitic trematodes *Schistosoma mansoni* and *Fasciola hepatica* (Davis and Hodgson, 1997). Two genes, ubiquitin binding protein (UbCRBP) and enolase were found to be the same transcriptional orientation with only 54 bp separating them. Isolation of the 5' ends of both genes showed that ubiquitin binding protein was not *trans*-spliced, while enolase receives the *S. mansoni* SL (Davis and Hodgson, 1997). Analysis of the mRNAs of both transcripts by RT-PCR and RNAse protection assay showed that pre-mRNA transcripts which spanned both UbCRBP, the intergenic region and enolase could be identified (Davis and Hodgson, 1997). However because of the lack of a transgenic system it could not be verified that a single common promoter drove the expression of the two genes and it could not be determined whether production of individual UbCRBP and enolase mRNA was mutually exclusive.

Within the nematodes operons have been identified in two other nematode species, *C. briggsae* and *Oscheius sp.*CEW1 (Evans *et. al.*, 1997; Kuwabara and Shah, 1994). Both of these nematodes are Rhabditidae and are closely related to *C. elegans* with an estimated 25 and 100 Myr divergence respectively (Fitch *et. al.*, 1995). Analysis of the unfinished *C. briggsae* genomic sequence has shown that there is conservation of synteny across large sections of the genome and structures such as operons appear to be highly conserved. *C. briggsae* also appears to utilize a similar repertoire of SL2-like SLs to *C. elegans* (Kuwabara and Shah, 1994). A single operon, the ribosomal protein *rpl-27a/rpp-1* operon, has been described which is conserved between *Oscheius sp.*CEW1 and *C. elegans* (Evans *et. al.*, 1997). Examination of the 5' ends of the *Os-rpp-1* transcript revealed that *Oscheius sp.*CEW1 also utilized a set of SL2-like SLs, although these were different from those identified in *C. elegans* or *C. briggsae* (Evans *et. al.*, 1997). Recently Redmond and Knox reported the isolation of an SL2-like SL from the parasitic strongylid *Haemonchus contortus* (Redmond and Knox, 2001). However they could not establish that this SL was utilized in the context of operon resolution. No systematic survey has yet been undertaken to see how common this form of gene organization is and whether the utilization of SL2-like SLs are utilized across the phylum *Nematoda*. To address this issue a selected set of operonic genes from *C. elegans* were tested for conservation of synteny in the genome of *Brugia malayi*. Three genes sets were found which were conserved between these two distantly related nematodes and the 5' ends of the transcripts of the second gene in the *rpl-27a* / *rpp-1* operon were isolated to determine if *B. malayi* also utilizes an SL2-like SL to resolve polycistronic transcripts.

7.3 Strategy for isolating operon candidates

Approximately 850 operons (13% of gene pairs) are predicted in the *C. elegans* genome sequence (Blumenthal *et. al.* 2001 unpublished). PCR screening this large dataset in other nematode genomes was not feasible, so several classes of genes and gene families were examined to select groups of operons that could be tested. Analysis of the ribosomal protein genes from *C. elegans* revealed that they occur in operons more frequently than expected. Of the 133 *C. elegans* ribosomal proteins examined 66 (49%) were found to be in operons. This is a statistically higher rate than would be expected ($p < 0.001$) based on the observed frequency of operons in the *C. elegans* genome (Chi^2 test, performed in Minitab, Minitab Inc). Because ribosomal protein sequences are highly conserved and abundantly expressed orthologous genes were easily identified by comparison of the *C. elegans* protein sequences to the nematode datasets in dbEST (Boguski *et. al.*, 1993) and NEMBASE (Parkinson *et. al.*, 2001). Tables 7.3.1A and B lists the full complement of *C. elegans* cytoplasmic ribosomal proteins and their *B. malayi* orthogues. Tables 7.3.2A and B list a partial complement of the *C. elegans* mitochondrial ribosomal proteins. Because these proteins are not as well conserved between different taxa, the characterized yeast and vertebrate protein sequences were used to probe the *C. elegans* genome sequence. Identified proteins were then compared to *B. malayi* EST dataset. From the 65 ribosomal protein operons identified in *C. elegans* twelve had adjacent orthologous gene pairs which had representative sequences in the *B. malayi*

EST dataset. Table 7.3.3 shows those operon candidates that were tested for conservation of synteny in *B. malayi* by PCR.

Large subunit cytoplasmic ribosomal proteins

protein	*H. sapiens*	*D. melanogaster*	*C. elegans*	Operonic in *C. elegans*	*B. malayi*
L1, L4	P36578	CG8195	B0041.5 (I)	No	-
L2, L8	XP_005130.1	AAF47659.1	B0250.1 (V)	No	BMC01536
	-	AAF47659.1	B0250.7 (V)	No	BMC00336
L3	XP_039345.1	AAF54609.1	F13B10.2 (II)	No	BMC00329
L5	XP_028341.1	AAG22457.1	F54C9.5 (II)	Yes	BMC01632
L6	XP_050941.1	CG11522	R151.3 (III)	Yes	BMC00250
L7	XP_035492.1	AAF52868.1	F53G12.10 (I)	Yes	BMC00673
L7a	XP_035105.1	AAF46169.1	Y24D9A.4 (IV)	No	BMC00621
L9	XP_047490.1	AAF53049.1	R13A5.8 (III)	No	BMC02343
L10	XP_018268.1	AAF45349.1	K11H12.2 (IV)	No	BMC00151
L10a	XP_017704.1	CG7283	Y71F9AL.13 A (V)	Yes	BMC00576
L11	P39026	AAF57560.1	F07D10.1 (X)	No	BMC00827
	-	-	T22F3.4 (V)	No	-
L12	XP_033467.1	CG3195	JC8.3 (IV)	No	BMC01716
L13	XP_047464.1	AAF52842.1	C32E8.2 (I)	Yes	BMC00623
L13a	XP_027886.1	CG1475	M01F1.2 (III)	Yes	BMC00137
L14	XP_044190.1	AAF50393.1	C04F12.4 (I)	No	BMC00082
L15	XP_048417.1	AAF45440.1	F10B5.1 (II)	Yes	BMC00009
L17, L23	XP_028962.1	AAF46914.1	B0336.10 (III)	Yes	BMC00275
L18	XP_049965.1	CG8615	Y45F10D.12 (IV)	Yes	BMC00862
L18a	XP_038594.1	AAF57838.1	E04A4.8 (IV)	No	BMC00060
L19	XP_008294.3	AAF47305.1	C09D4.5 (I)	Yes	BMC02351
L21	XP_033917.1	CG12775	C14B9.7 (III)	No	BMC01506
L22	XP_030989.1	AAF45546.1	C27A2.2 (II)	No	BMC03219
L23	XP_012891.1	AAF47545.1	F55D10.2 (X)	Yes	BMC00175
L23a	XP_017356.1	AAF47545.1	F52B5.6 (I)	No	BMC00175
L24	XP_015463.1	CG9282	D1007.12 (I)	Yes	BMC00120
L26	XP_016869.1	CG6846	F28C6.7 (II)	Yes	BMC00059
L27	XP_032124.1	CG4759	C53H9.1 (I)	Yes	BMC03199
L27a	XP_016869.1	AAF51006.2	Y37E3.8A (I)	Yes	BMC01540
L28	XP_035923.1	CG12740	R11D1.8 (V)	No	BMC01157
L29	XP_011055.1	AAF46708.1	B0513.3 (IV)	Yes	BMC00067
L30	XP_046141.1	CG6764	C03D6.8 (I)	No	BMC02758
L31, L41	XP_033301.1	CG1821	W09C5.6A/B (I)	No	BMC01624
L32	XP_003054.3	AAF57001.1	T24B8.1 (II)	No	BMC00221
L34	XP_034712.1	CG6090	C42C1.14 (IV)	No	BMC00060

L35	XP_044796.1	CG4111	ZK652.4 (III)	No	BMC00205
L35a	X52966	CG2099	F10E7.7 (II)	No	BMC00179
L36	XP_044614.1	AAF45531.1	F37C12.4 (III)	Yes	BMC00060
L36a	XP_052671.1	CG7424	C09H10.2 (II)	No	BMC01314
L37	XP_017770.1	CG9091	W01D2.1 (II)	No	BMC00060
	XP_017770.1	CG9091	C54C6.1 (III)	No	BMC00060
L37a	NP_000989	CG5827	Y48B6A.2 (II)	Yes	BMC01949
L38	P23411	CG18001	C06B8.8 (V)	Yes	BMC00157
L39	XP_010359.3	AAF47154.1	C26F1.9 (V)	No	BMC00809
L40	XP_009284.3	AAF51034.1	ZK1010.1 (III) *ubq-2*	Yes	BMC00134

Figure 7.3.1A Table listing the large subunit ribosomal proteins identified in the genomes of *H. sapiens*, *D. melanogaster*, *C. elegans* and the EST dataset of *B. malayi*. A GenBank accession number is given for each *H. sapiens* and *D. melanogaster* gene. The cosmid ORF number is given for each *C. elegans* gene. A cluster number is given for the *B. malayi* gene. A listing of the ESTs in each cluster can be found in NEMBASE (Parkinson *et. al.*, 2001) (http://nema.cap.ed.ac.uk). The chromosome on which the *C. elegans* protein is found is listed next the ORF number. It is also indicated if the *C. elegans* ribosomal protein gene is in an operon.

Small subunit cytoplasmic ribosomal proteins

protein	*H. sapiens*	*D. melanogaster*	*C. elegans*	Operonic in *C. elegans*	*B. malayi*
S2	XP_043619	AAF45638.1	B0393.1 (III)	No	BMC00957
S2/S5	XM_034464	AAF52822.1	C49H3.11 (IV)	Yes	BMC01616
S3	XP_035076.1	AAF56129.1	C23G10.3 (III)	Yes	BMC00419
S3a	XP_037456.1	AAF59372.1	F56F3.5 (II)	No	BMC01505
S4	XP_044025.1	AE003539	Y43B11AR.4 (IV)	Yes	BMC00375
S5	XP_034265	CG7014	T05E11.1 (IV)	Yes	BMC00499
S6	XP_048310.1	AAF46288.1	Y71A12B.1 (I)	No	BMC00011
S7	XP_012638.5	CG1883	ZC434.2 (I)	No	BMC00849
S8	XP_046554.1	CG7808	F42C5.8 (IV)	No	BMC00738
S9	XP_050590.1	AAF50249.1	F40F8.10 (II)	No	BMC01891
S10	XP_043285.1	CG14206	D1007.6 (I)	Yes	BMC00256
S11	P04643	AAF50249.1	F40F11.1 (IV)	No	-
S12	XP_017626.1	AAF49851.1	F54E7.2 (III)	Yes	BMC00188
S13	XP_047325.1	CG4263	C16A3.8 (III)	No	-
S14	XP_042550.1	AAF46297.1	F37C12.9 (III)	No	BMC00207
S15	XP_047576.1	CG8332	F36A2.6 (I)	No	BMC04325
S15a	XP_027366.1	CG2033	F53A3.3 (III)	No	BMC00121
S16	XP_046112.1	CG4046	T01C3.6 (V)	Yes	BMC00243
S17	XP_007615.3	AAF50272.1	T08B2.10	Yes	BMC00267
S18	XP_016854.1	AAF57491.1	Y57G11C.16 (IV)	No	BMC00161
S19	XP_008876.1	AAF48633.1	T05F1.3 (I)	No	BMC00176
S20, S22	XP_031816.1	AAF55809.1	Y105E8A.16 (I)	Yes	BMC00239
S21	XP_009693.3	AAF51191.1	F37C12.11 (III)	No	BMC00282
S23	XP_004020.1	CG8415	F28D1.7 (IV)	No	BMC00817
S24	XP_039577.1	CG3751	T07A9.11 (IV)	No	BMC00096
S24e	XP_039578.1	-	T26G10.3 (III)	No	-
S25	XP_051497.1	AAF54605.1	K02B2.5 (IV)	Yes	BMC00290
S26	XP_049421.1	AAF53666.1	F39B2.6 (I)	No	BMC00253
	-	-	C03H5.f (II)	No	-
S27	P42677	CG8338	F56F3.5 (III)	No	BMC01505
	XP_045145.1	CG10423	F56E10.4 (V)	No	BMC04343
S27a	XP_017513.2	AAF52941.1	F34H10.1 (X)	No	BMC03367

	-	-	K08C9.7 (I)	No	-
S28	XP_006026.2	CG2998	Y41D4B.5 (IV)	No	BMC00209
S29	XP_052669.1	CG8495	B0412.4 (III)	Yes	BMC00288
S30	XP_006522.3	CG15697	C26F1.4 (V)	Yes	BMC00146
P0	XP_017620.1	AAF51807.1	F25H2.10 (I)	Yes	BMC00405
P1	XP_035388.1	AAF51499.1	Y37E3.7 (I)	Yes	BMC00166
P2	M17887	R6FFP2	C37A2.7 (I)	Yes	BMC00278

Figure 7.3.1B Table listing the small subunit ribosomal proteins identified in the genomes of *H. sapiens*, *D. melanogaster*, *C. elegans* and the EST dataset of *B. malayi*. A GenBank accession number is given for each *H. sapiens* and *D. melanogaster* gene. The cosmid ORF number is given for each *C. elegans* gene. A cluster number is given for the *B. malayi* gene. A listing of the ESTs in each cluster can be found in NEMBASE (Parkinson *et. al.*, 2001) (http://nema.cap.ed.ac.uk). The chromosome on which the *C. elegans* protein is found is listed next the ORF number. It is also indicated if the *C. elegans* ribosomal protein gene is in an operon.

Mitochondrial large subunit ribosomal proteins

protein	*H. sapiens*	*D. melanogaster*	*C. elegans*	Operonic in *C. elegans*	*B. malayi*
L1	AI155727	CG7494	F33D4.5 (IV)	Yes	-
L2	AI313184	CG7636	F56B3.8 (IV)	Yes	-
L2a	AW960439	CG6547	Y48E1B.5 (II)	Yes	-
L3	P09001	CG8288	C26E6.6 (III)	Yes	BMC08589
L3a	AI074410	CG15871	Y34D9A.1 (I)	No	BMC00143
L4	AI087088	AAF45007	T23B12.2 (V)	Yes	BMC11493
L5	AL353983	CG17166	C47D12.6 (II)	-	-
L7/12A	P52815	CG5012	W09D10.3 (III)	No	-
L7a			Y48A6B.3 (III)	Yes	BMC00938
L9	AI359339	CG4923	B0205.11 (I)	Yes	BMC02815
L10	AW249086	CG11488	K01C8.6 (II)	No	-
L11	AI188527	CG3351	B0303.15 (III)	No	-
L13	AB049640	CG10603	F13G3.7 (II)	Yes	BMC07811
L14	AA642134	CAB63504	F45E12.4 (II)	Yes	BMC11522
L15	AAI28556	CG5219	Y92H12BR.8 (I)	No	BMC06711
L16	AA780023	CAA15945	T04A8.11 (III)	Yes	-
L17	AI141700	CG13880	Y54E10A.7 (I)	Yes	-
L19	P49406	CG8039	Y119C1B.4 (I)	Yes	-
L20	AI368972	CG11258	Y48C3A.1 (II)	No	-
L22	AA772054	CG4742	Y39A1A.6 (III)	No	-
L22	AI760300	CG5242	F54C4.1 (III)	No	-
L23	Z49254	CG1320	T08B2.8 (I)	Yes	-
L24	AI361046	CG8849	F59A3.3 (I)	Yes	-
L27	W81261	-	-	-	-
L30	AA772463	CG7038	W04B5.4 (III)	Yes	BMC05770
L31	N46796	CG12921	ZC410.7A/B (I)	Yes	BMC06932
L32	AA166925	CG12220	C30C11.1 (III)	No	-
L33	AF047440	CG3712	-	-	
L34	AA988598	-	-	-	
L36	AA454962	CG7528	W02A11.4A/	No	BMC02741

			B (I)		
L36a	AI075733	CG5479	C25A1.13 (I)	No	-
L41	XP_088427	CG12954	B0432.3 (II)	Yes	BMC01857

Figure 7.3.2A Table listing the mitochondrial large subunit ribosomal proteins identified in the genomes of *H. sapiens*, *D. melanogaster*, *C. elegans* and the EST dataset of *B. malayi*. A GenBank accession number is given for each *H. sapiens* and *D. melanogaster* gene. The cosmid ORF number is given for each *C. elegans* gene. A cluster number is given for the *B. malayi* gene. A listing of the ESTs in each cluster can be found in NEMBASE (Parkinson *et. al.*, 2001) (http://nema.cap.ed.ac.uk). The chromsome on which the *C. elegans* protein is found is (Parkinson *et. al.*, 2001) next the ORF number. It is also indicated if the *C. elegans* ribosomal protein gene is in an operon

Small subunit mitochondrial ribosomal proteins

protein	*H. sapiens*	*D. melanogaster*	*C. elegans*	Operonic in *C. elegans*	*B. malayi*
S2	AA421679	CG2937	T23B12.3	Yes	BMC06012
S4			C48B6.2	No	-
S5	AI660862		E02A10.1	Yes	BMC03177 BMC05967
S6	W07026	CG15016	R12E2.12	Yes	-
S7	D55616	CG5108	Y57G11C.4	No	-
S11	AI554856	CG5184	W04D2.5	Yes	BMC08678
S12	Y11681	P10735	T03D8.2	Yes	-
S14	HS262D12	CG12211	T01E8.6	Yes	-
S15	AI147651	CG4207	-	-	-
S16	AA305994	CG8338	F56D1.3	Yes	-
S17	AI309197	CG4326	C05D11.10A	Yes	BMC08489
S18a	NP_054765	CG10757	T13H5.5	No	BMC02782
S18b	AI083656	CG11744	T14B4.2	Yes	BMC07498
S21	AI066648	-	F29B9.10	Yes	-
S22	AA631191	CG12261	C14A4.14	Yes	BMC07370
S25	NM_022497	CG14413	Y55F3AM.1	No	BMC01309
S26	NM_030811	CG7354	C34E10.11	Yes	-
S34	NM_023936	CG13037	M88.2	No	-
S35	AF070663	CG5497	Y43F8C.8	Yes	-
DAP3	CAA58535.1	CG3633	C14A4.3	Yes	-

Figure 7.3.2B Table listing the mitochondrial small subunit ribosomal proteins identified in the genomes of *H. sapiens*, *D. melanogaster*, *C. elegans* and the EST dataset of *B. malayi*. A GenBank accession number is given for each *H. sapiens* and *D. melanogaster* gene. The cosmid ORF number is given for each *C. elegans* gene. A cluster number is given for the *B. malayi* gene. A listing of the ESTs in each cluster can be found in NEMBASE (Parkinson *et. al.*, 2001) (http://nema.cap.ed.ac.uk). The chromosome on which the *C. elegans* protein is found is listed next the ORF number. It is also indicated if the *C. elegans* ribosomal protein gene is in an operon.

Potential Candidate Operon	Gene 1	Gene 2	Intergenic Separation (bp)
rpp-0/tph-1	*Ce*-F25H2.10 *rpp-0*	*Ce*-F25H2.11 translationally controlled tumor protein *tph-1*	170
fibrillarin/*rps-16*	*Ce*-T01C3.7 fibrillarin	*Ce*-T01C3.6 *rps-16*	230
rpl-26/suf-1	*Ce*-F28C6.7 *rpl-26*	*Ce*-F28C6.6 polyadenylation factor *suf-1*	313
rpl-5/F54C9.6	*Ce*-F54C9.5 *rpl-5*	*Ce*-F54C9.6 ATP binding protein	136
rpl-27a/rpp-1	*Ce*-Y37E3.8 *rpl-27a*	*Ce*-Y37E3.7 *rpp-1*	207
rpl-36/ayc-1	*Ce*-F37C12.4 *rpl-36*	*Ce*-F37C12.3 acyl-carrier protein *ayc-1*	112
rps-12/F54E7.1	*Ce*-F54E7.2 *rps-12*	*Ce*-F54E7.1 conserved membrane protein of unknown function	349
rps-25/inx-7	*Ce*-K02B2.5 *rps-25*	*Ce*-K02B2.4 *inx-7/* innexin	503
mrpl-4/mrps-2	*Ce*-T23B12.3 *mrps-2*	*Ce*-T23B12.2 *mrpl-4*	238
mrpl-9/B0205.3	*Ce*- B0205.11 *mrpl-9*	*Ce*-B0205.3 26S proteosome subunit *rpn-10*	247
C01F6.8/*mrpl-31*	*Ce*-C01F6.8 Nucleotide-sensitive chloride conductance regulator	*Ce*- ZC410.7A/B *mrpl-31*	261
mrpl-41/B0432.4	*Ce*- B0432.3 *mrpl-41*	*Ce*-B0432.2 ThiJ/PfpI family protein	333

Figure 7.3.3 Operon candidates identified in *C. elegans* that were tested for conservation of synteny between operonic gene pairs in *B. malayi*. The intergenic separation between *the C. elegans* genes is also listed

7.4 Testing operon candidates by PCR

EST clones representative of each of the operon gene candidates from *B. malayi* were excised *in vivo* using standard protocols (Stratagene) and the 5' and 3' ends of the isolated pBluescript phagemids sequenced with the vector primers T3 and T7 or M13L and M13R. Gene specific primer pairs were synthesized for each gene. To determine if the operon was conserved between *C. elegans* and *B. malayi* three PCRs were performed using these primers *operon_gene-1*.F1/*operon_gene-1*.R1, *operon_gene-2*.F1/*operon_gene-2*.R1, *operon_gene-1*.F1/*operon_gene-2*.R1. The first two PCRs verified that each of the gene-specific primer pairs yielded products from genomic DNA while the third PCR tested the conservation of the operon. The PCRs were performed with Long Range Taq (Stratagene) or AGS-gold (Hybaid) according to the manufacturer's instructions using 300-400 ng of *B. malayi* genomic DNA as template. Figure 7.4.1 shows the results of these PCRs.

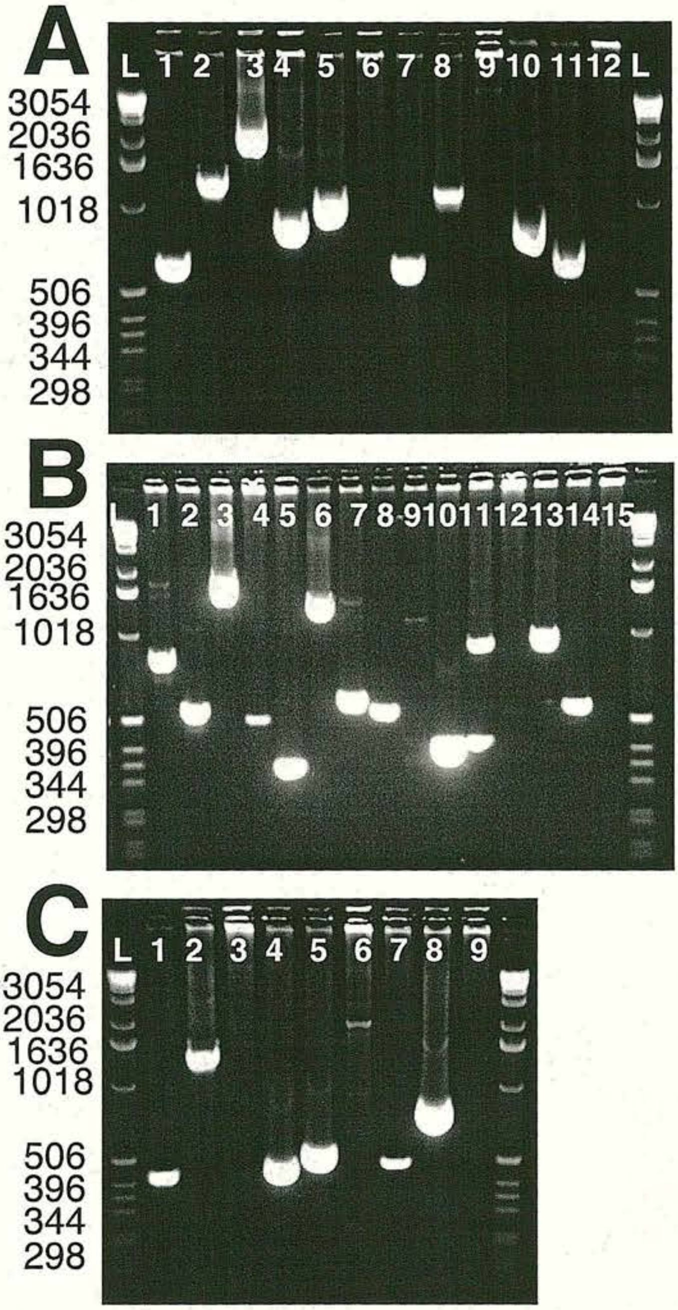


Figure 7.4.1 Testing potential operons from *B. malayi* by PCR from genomic DNA
A: 1: Bm-rpp0.F1/ Bm-rpp0.R1, 2: Bm-tph1.F1/Bm-tph1.R1, 3: Bm-rpp0.F1/Bm-tph-1.R1
4: Bm-fib1.F1/Bm-fib1.R1, 5: Bm-rps16.F1/Bm-rps-16.R1, 6: Bm-fib1.F1/Bm-rps16R1,
7: Bm-rpl26.F1/Bm-rpl26.R1, 8: Bm-suf1.F1/Bm-suf-1.R1, 9: Bm-rpl26.F1/Bm-suf1.R1,
10: Bm-rpl5.F1/Bm-rpl5.R1, 11: Bm- F54C9.6.F1/Bm- F54C9.6.R1, 12:Bm-rpl5.F1/Bm- F54C9.6.R1.
B: 1: Bm-rpl27a.F1/ Bm-rpl27a.R1, 2: Bm-rpp1.F1/Bm-rpp1.R1, 3: Bm-rpl27a.F1/ Bm-rpp1.R1,
4: Bm-rpl36.F1/ Bm-rpl36.R1, 5: Bm-ayc-1.F1/ Bm-ayc-1.R1, 6: Bm-rpl36.F1/Bm-ayc-1.R1,
7: Bm- rps12.F1/Bm- rps12.R1, 8: Bm- F54E7.1.F1/Bm- F54E7.1.R1, 9: Bm- rps12.F1/
Bm- F54E7.1.R1, 10: Bm-rps25.F1/ Bm-rps25.R1, 11: Bm- K02B2.4.F1/Bm- K02B2.4.R1,
12: Bm-rps25.F1/Bm- K02B2.4.R1, 13: Bm- mrpl4.F1/Bm- mrpl4.R1, 14: Bm- mrps2.F1/
Bm- mrps2.R1, 15: Bm- mrpl4.F1/ Bm- mrps2.R1.
C: 1: Bm- mrpl9.F1/Bm- mrpl9.R1, 2: Bm- B0205.3.F1/Bm- B0205.3.R1, 3: Bm- mrpl9.F1/
Bm- B0205.3.R1, 4: Bm- C01F6.8.F1/ Bm- C01F6.8.R1, 5: Bm- mrpl31.F1/ Bm- mrpl31.R1,
6: Bm- C01F6.8.F1/Bm-mrpl31.R1, 7: Bm-mrpl41.F1/ Bm-mrpl41.R1, 8: Bm-B0432.4.F1/
Bm-B0432.4.R1, 9: Bm-mrpl41.F1/ Bm-B0432.4.R1.

Three of the tested operon candidates *Bm-rpp-0/Bm-tph*-1, *Bm-rpl-27a/Bm-rpp-1* and *Bm-rpl-36/Bm-ayc-1* gave positive PCR results. In addition based on data yielded from the mapping of the *Bm-mif-2* locus an additional set of PCRs were performed to determine if position of *Bm-rps-14* upstream of *Bm-rpl-36/Bm-ayc-1* operon was conserved (data not shown).

To determine whether these operons are conserved across the *Secernentea*, two additional representative members of clade V and clade IV *Pristionchus pacificus* and *Strongyloides ratti* were surveyed for the operonic structures conserved between *B. malayi* and *C. elegans*. Specific primers designed for the genes in *rpp-0/tph-1*, *rpl-27a/rpp-1* and *rpl-36/ayc-1* operons. These primers were based on sequences found in their EST datasets available in GenBank. Genomic DNA was prepared from *P. pacificus* and *S. ratti* free living adults were collected from the fecal cultures of infected rats and washed in PBS (kind gift of Dr. Mark Viney, Bristol University). PCRs were performed as described above and figure 7.4.2 shows the resulting products.

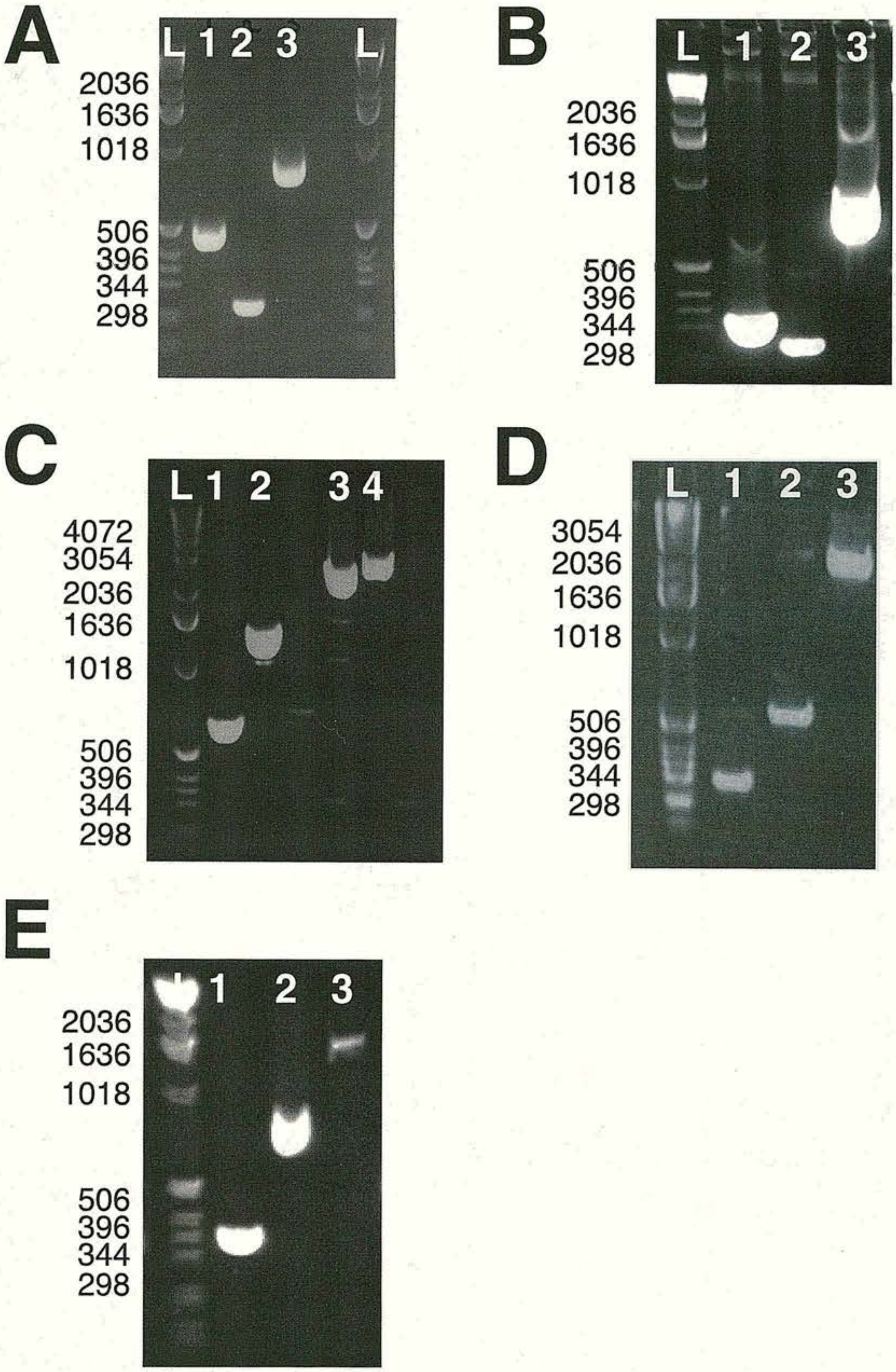
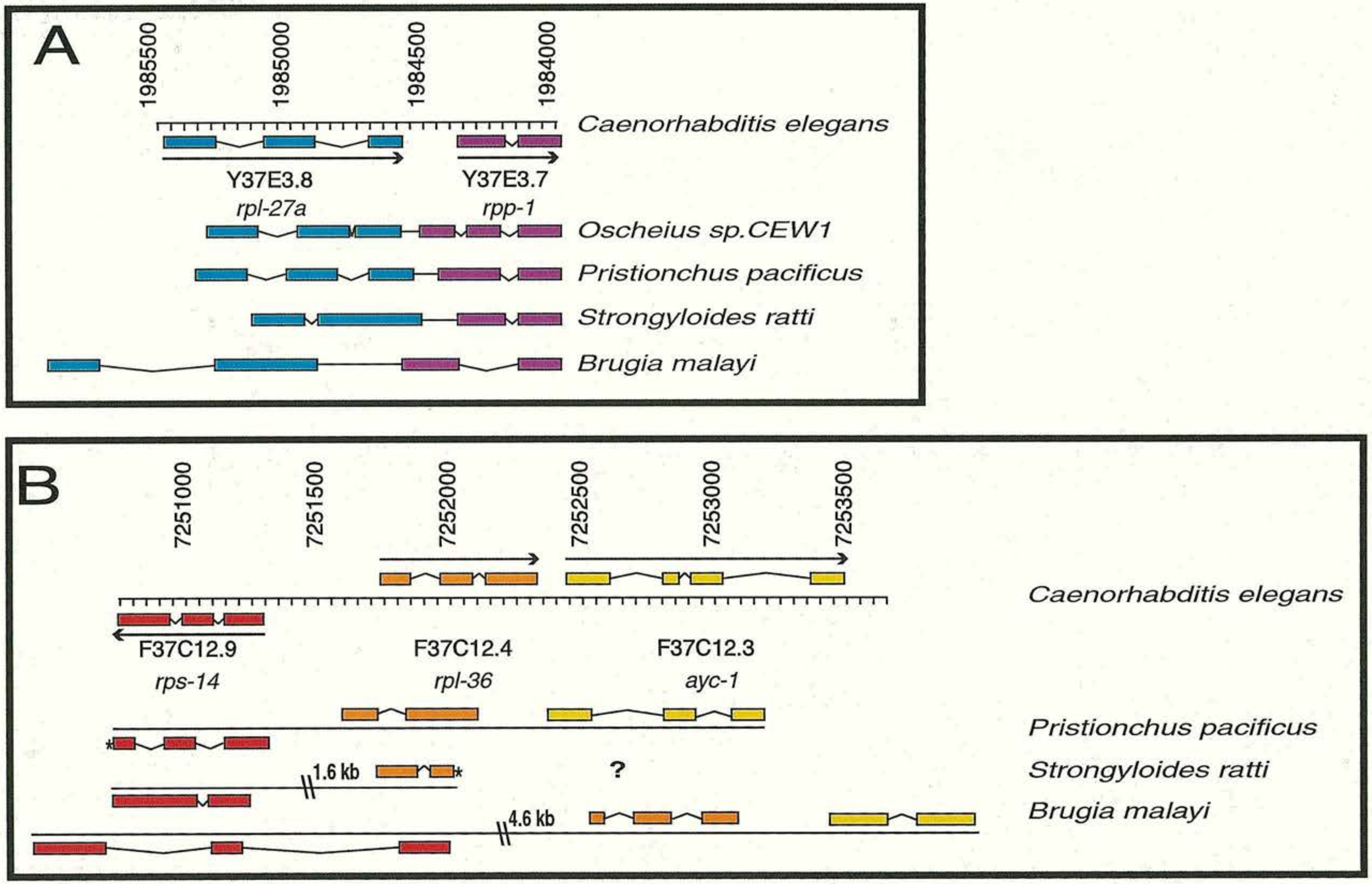


Figure 7.4.2 PCR of the rpl-27a/rpp-1 and rpl-36/ayc-1 rpp-0/tph-1operons from *P.pacificus* and *S. ratti*. **A:** 1: *Pp*-rpL27aF1/*Pp*-rpL27aR1, 2:*Pp*-rpP1F1/*Pp*-rpP1R1, 3:*Pp*-rpL27aF1/*Pp*-rpP1R1; **B:** 1: *Sr*-rpL27a.F1/*Sr*-rpL27a.R1, 2: *Sr*-rpP1.F1/*Sr*-rpP1.R1, 3: *Sr*-rpL27a.F1/*Sr*-rpP1.R1; **C:** 1:*Pp*-rpS14.F1/ *Pp*-rpS14.R1, 2:*Pp*-rpL36.F1/*Pp*-rpL36.R1, 3: *Pp*-rpS14.F1/*Pp*-rpL36.R1, 4: *Pp*-rpL36.F1/*Pp*-ayc-1.R1; **D:** 1:*Sr*-rpS14.F1/*Sr*-rpS14.R1, 2:*Sr*-rpL36.F1/*Sr*-rpL36.R1, 3: *Sr*-rpS14.F1/*Sr*-rpL36.R1; **E:** 1: *Pp*-rpP0.F1/*Pp*-rpP0.R1, 2: *Pp*-tph1.F1/*Pp*-tph-1.R1, 3: *Pp*-tph1.F1/*Pp*-tph-1.R1

7.5 Isolation and sequencing of potential operons

The PCR products *Bm-rpl-27a*.F1/*Bm-rpp-1*.R1, *Pp-rpl-27a*.F1/*Pp-rpp-1*.R1, *Sr-rpl-27a*.F1/*Sr-rpp-1*.R1, *Bm-rpl-36*.F1/*Bm*-F37C12.3.R1, *Sr-rps-14*.F1/*Sr-rpl-36*.R1, *Pp-rpl-36*.F1/*Pp*-F37C12.3.R1, *Pp-rps-14*.F1/*Pp-rpl-36*.R1, *Bm-rpp-0*.F1/*Bm-tph-1*.R1 and *Pp-rpp-0*.F1/*Pp-tph-1*.R1 were T- cloned into pCR4.0 (Invitrogen) and fully sequenced with the gene specific primers and the vector primers M13L and M13R. Figure 7.5.1 shows a graphical representation of the genomic fragments amplified and tables 7.5.2A-C summarizes the features of each gene in the three different species surveyed.


A
1985500
1985000
1984500
1984000
Caenorhabditis elegans
Y37E3.8
rpl-27a
Y37E3.7
rpp-1
Oscheius sp.CEW1
Pristionchus pacificus
Strongyloides ratti
Brugia malayi
B
7251000
7251500
7252000
7252500
7253000
7253500
Caenorhabditis elegans
F37C12.9
rps-14
F37C12.4
rpl-36
F37C12.3
ayc-1
Pristionchus pacificus
*
1.6 kb
*
?
Strongyloides ratti
Brugia malayi
4.6 kb

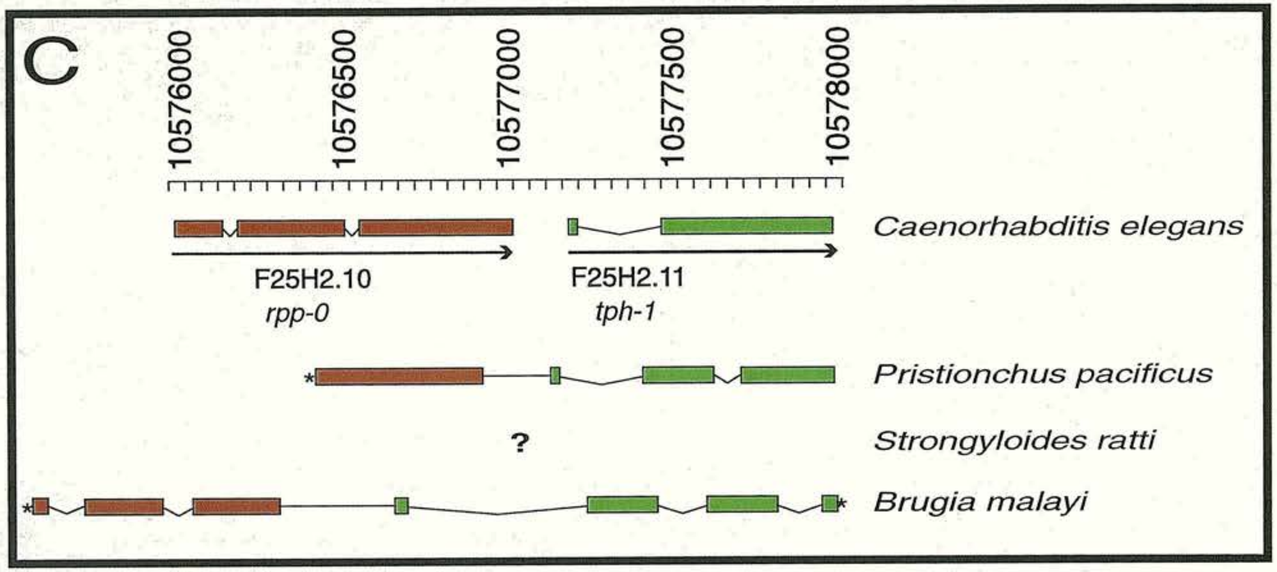


Figure 7.5.1 The sequenced fragments of genomic DNA of the three operons isolated from four nematode species, with genes indicated by exons (box) and intron (bracket) structures. The numbers at the top show the relative position of the *C. elegans* operons on their respective chromosomes and all graphics are drawn to scale based on these numbers. **A:** the *rpl-27a / rpp-1* operon, **B:** the *rpl-36 / ayc-1* operon with the conserved *rps-14* gene in the opposite strand with a diverging transcriptional orientation. **C:** the *rpp-0 / tph-1* operon. * indicates the cloned fragments that do not contain the entire coding portion of the gene.

Table 7.5.2A POP-1 (*rpl-27a* / *rpp-1*)

Species	*rpl2-7a* gene/ gene fragment size in bp	*rpl-27a* intron sizes in bp	Intergenic region size in bp	*rpp-1*gene/ gene fragment size in bp	*rpp-1* intron sizes in bp
C. elegans	883	62 215	208	384	50
*Oscheius sp.*CEW1	595	81 6	89	543	39 68
P. pacificus	630	74 96	89	211	69
S. ratti	547	48 53	163	400	43
B. malayi	873	425	323	583	212

Table 7.5.2B POP-2 (*rpl-36*/ *ayc-1*)

Species	*rpl-36* gene/ gene fragment size in bp	*rpl-36* intron sizes in bp	Intergenic region size in bp	*ayc-1* gene/gene fragment size in bp	*ayc-1* intron sizes in bp
C. elegans	590	113 51	113	1050	201 46 336
P. pacificus	378	101	436	436	287 137
S. ratti	378	43	?	?	?
B. malayi	566	104 92	405	572	103

Table 7.5.2C POP-3 (*rpp-0*/ *tph-1*)

Species	*rpp-0* gene/ gene fragment size in bp	*rpp-0* intron sizes in bp	Intergenic region size in bp	*tph-1* gene/gene fragment size in bp	*tph-1* intron sizes in bp
C. elegans	939	47 46	170	547	256
P. pacificus	550	-	179	512	274 75
S. ratti	?	?	?	?	?
B. malayi	482	150 101	543	514	520 150 177

Tables 7.5.2A-C Characteristics of the cloned genomic fragments of the *rpl-27a/rpp-1*, *rpl-36/ayc-1* and *rpp-0/tph-1* operons.

7.6 Survey of spliced leaders utilized on operonic genes

To test the composition of the SLs found at the 5' ends of the genes found in the operonic structures 5' RACE was performed on *Bm-rpp-1* and *Pp-rpp-1.* RACE cDNA was synthesized using poly-A selected mRNA and the Invitrogen GeneRacer Kit, which utilizes a cap selection methodology to selects for full length transcripts. PCR was performed on 5 μL first strand RACE cDNA with the 5' GeneRACER primer and gene-specific reverse primers (*Bm*-rpp1.R4 and *Pp*-rpp1.R1). Figure 7.6.1 shows PCR of the operon genes from 5' RACE cDNA.

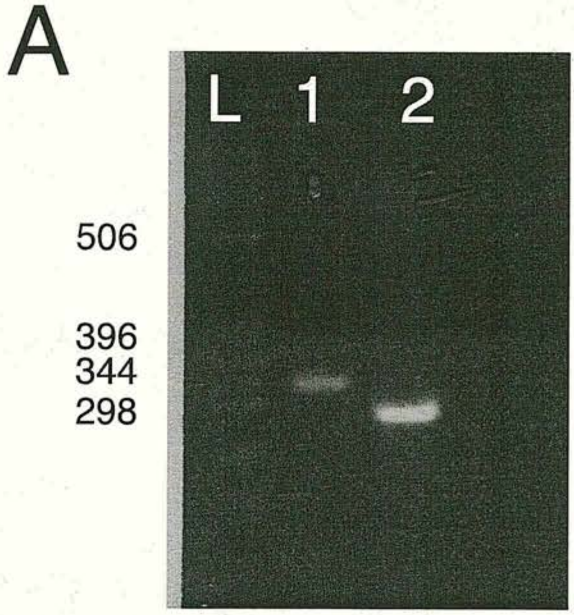


Figure 7.6.1 The PCR amplification of the 5' RACE fragments of the operon genes. A: 1: *Bm-rpp-1* and 2: *Pp-rpp-1*

The resulting PCR products were T-cloned into pCR4.0. Individual colonies were picked and the insert isolated by PCR with the vector primers M13L and M13R. The PCR products were then treated with exonuclease I and shrimp alkaline phosphatase (Amersham) to remove excess primer and dNTPs. 4µL of treated PCR product was used as template for sequencing with a gene specific reverse primer. Tables 7.6.2, 7.6.3A and 7.6.3B show the results of the sequencing of the isolated 5' RACE cDNA fragments.

Nematode Species	Operon	Gene	# isolated SL-1 RACE ends (% total dataset)	# isolated SL-2 RACE ends (% total dataset)
P. pacificus	*rpl-27a/rpp-1*	*rpp-1*	8 (5%)	163 (95%)
B.malayi	*rpl-27a/rpp-1*	*rpp-1*	126 (100%)	0

Table 7.6.2 Number of RACE fragments isolated containing SL-1 or SL-2 like spliced leaders for each surveyed ribosomal protein operon gene.

Table 7.6.3A Spliced leader sequences isolated from *B. malayi* RACE ends of *Bm-rpp-1*

Gene	SL1-like sequences	Number of clones	Percent of dataset	Found in other species
	GGTTT-AATTACCCAAGTTTGAG	115	91%	all
	ATATATCGACGGTTTAATTACCCAAGTTTGAG	006	4%	
	GGGTTT-AATTACCCAAGTTTGAG	001		
α	GGTTTTAATTACCCAAGTTTGAG	001		*P. pacificus*
	GGTTT-AATCACCCAAGTTTGAG	001		
	GGTTT-AACTACCCAAGTTTGAG	001		
	GTTT-AATTACCCAAGTTTGAG	001		

Table 7.6.3B Spliced leader sequences isolated from *P. pacificus* RACE ends *Pp-rpp-1*

Gene	SL2-like sequences	Number of clones	Percent of dataset	Found in other species
δ	GGTTTTTAACCCAGTATCTCAAG	001		
α	GGTTTTT-ACCCAGTATCTCAAG	021	12%	*H. contortus*
ε	GGTTTTT-ACTCAGTATCTCAAG	001		
	GGGTTTTT-ACCCAGTATCTCAAG	001		
ϕ	GGTCTTT-ACCCAGTATCTCAAG	001		
β	GGTTTT-AACCCAGTATCTCAAG	106	62%	
γ	GGTTTT-AACCCGGTATCTCAAG	001		
η	GGTTTT-AACCCAGTATCTTAAG	001		
ι	GGTTTT-GACCCAGTATCTCAAG	001		
φ	GTTTTATACCCAGTATCTCAAG	001		
κ	GGTTT-ATACCCAGTATCTCAAG	023	13%	
χ	GGGTTT-ATACCCAGTATCTCAAG	001		

λ	GGTTT-AAACCCAGTATCTCAAG	002	1%	
	SL1-like sequences	Number of clones	Percent of dataset	Found in other species
	GGTTT-AATTACCCAAGTTTGAG	007	4%	all
α	GGTTTTAATTACCCAAGTTTGAG	001		*B. malayi*

Tables 7.6.3A and **B** The diversity of the SL sequences found on the ends of the 5' RACE cDNA fragments. The table lists the different SL1 and SL2-like sequences found in the race survey. Each variant of the SL1 and SL2-like sequences used in subsequent phylogenetic analyses has been given a gene name. The sequence, number of isolated clones, the percent of dataset (when >1 clone was found) and whether the sequence has been found in other nematode species is shown.

While a diverse set of SL2-like sequences were found to be *trans*-spliced to the 5' end of *Pp-rpp-1*only SL1 was found on the 5' ends of *Bm-rpp-1*.

7.7 Primer extension of *Bm-rpp-1* and isolation of processing intermediate mRNAs for *Bm-rpl-27a/Bm-rpp-1*

Primer extension was performed on *Bm-rpp-1* to verify that it contains a single 5' splice acceptor site and to assess whether there are polymorphisms in the length of the transcripts. Polymorphisms could indicate the use of alternative SLs. Briefly the gene specific primer *Bm*-rpp1.R4 was labeled with T4 kinase and P^{32}-γATP. The primer extension was performed on 10 μg of *B. malayi* mixed adult total RNA using AMV-RT. The extension reaction was electrophoresed on a polyacrylamide gel with an M13 sequencing ladder. The gel was then visualized using X-ray film. Figure 7.7.1A shows the results of the primer extension. Only one product was detected at 248 bp which is consistent with the SL1-containing transcript found in the 5' RACE survey.

To test if *B m-rpl-27a/Bm-rpp-1* are transcribed as a single polycistronic transcript, RT-PCR was performed on DNAse treated *B. malayi* mixed adult total RNA using MMLV-RT (Stratagene), AGS-gold Taq (Hybaid) and the gene specific primers *Bm*-rpl27a.F1 and *Bm*-rpp1.R4 using standard protocols. To test that none of the resulting PCR products were due to contaminating genomic DNA, a sham control was performed, where no MMLV-RT was added to the RT reaction. Figure 7.7.1b shows the results of the RT-PCR with three major products detected. These three PCR products were isolated by gel extraction, cleaned and concentrated with a Microcon-100 and T-cloned into pCR4.0. Cloned PCR products were fully sequenced using the gene specific primers and the vector primers M13L and M13R. The sequencing verified that the three PCR products represented processing intermediates of the mRNA, the largest band being unprocessed, the middle having the *Bm-rpl-27a* intron removed and the smallest having both the *Bm-rpl-27a* and the *Bm-rpp-1* introns removed. All three processing intermediates had the intergenic region still connecting the two genes (see Figure 7.7.1B). While the results of the RT-PCR do not offer conclusive evidence of the order of processing events they establish that in *B. malayi cis*-splicing of the introns in the polycistron can occur before it is separated into individual cistrons. These results are consistent with the processing intermediates isolated from *C. elegans* operons (Spieth *et. al.*, 1993). However, it has not been possible to establish whether *cis*-splicing always occurs before the separation of the cistrons. Like the processing of conventional mRNAs *cis*-splicing of the polycistron appears to occur 5' to 3' with the introns contained in the upstream gene being removed before those in the downstream gene.

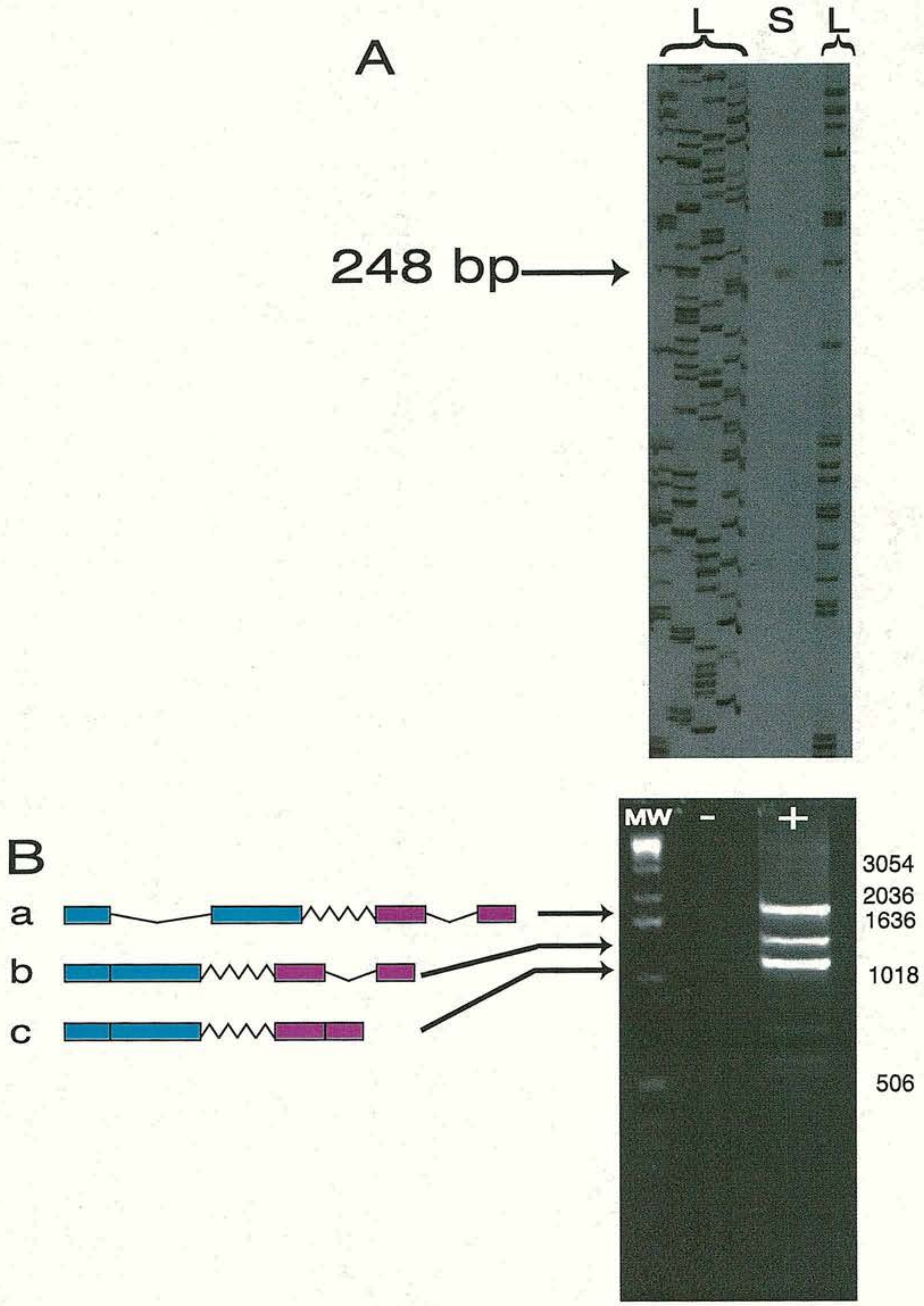


Figure 7.7.1A The primer extension product of *B. malayi rpp-1* showing that only one band can be observed at 248 bp which is consistent with the expected size of an SL1 *trans*-spliced cDNA. **L:** M13 sequencing ladder, **S:** primer extension sample.

Figure 7.7.1B Amplification of the unprocessed RNA intermediates of *rpl-27a* and *rpp-1* from B. malayi. **A:** no processing, introns in both genes present, **B:** processing intermediate with the *rpl-27a* intron removed, **C:** processing intermediate with the *rpl-27a* and *rpp-1* introns removed. **+:** reaction with RT added, **-:** sham reaction with no RT added

7.8 Diversity and genus-specific elaboration of SL2 gene families

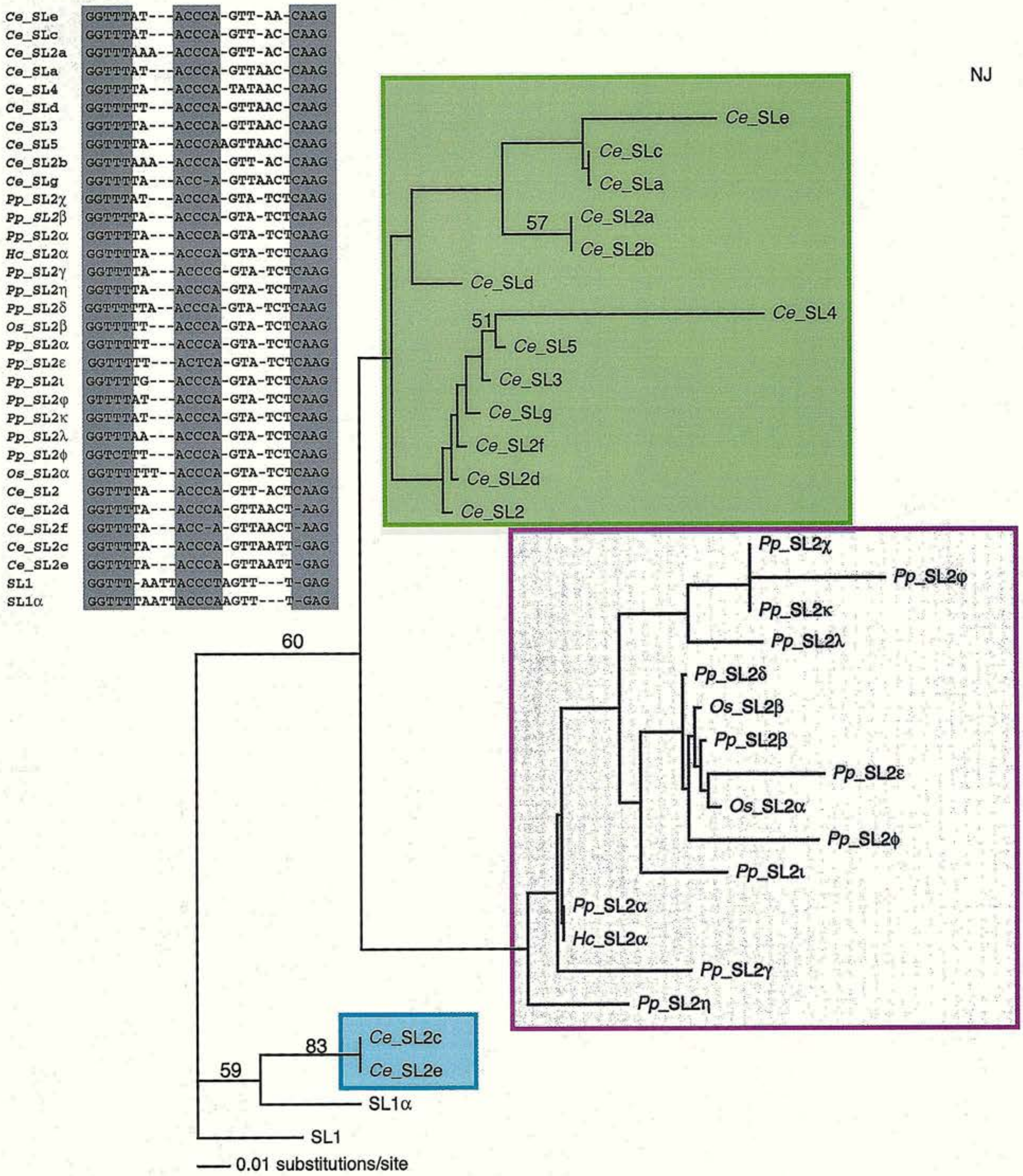
It is clear from the RACE survey that three SL sequences, *Pp*-SL2β, *Pp*-SL2α and *Pp*-SL2κ were the predominant species *trans*-spliced to the 5' end of *Pp-rpp-1* (see Tables 7.6.2 and 7.6.3). To determine whether *C. elegans* also utilized specific SL2-like sequences preferentially, the ESTs derived from a cap selected cDNA library for *Ce-rpl-27a*, *Ce-rpp-1*, *Ce-rpl-36*, *Ce-ayc-1*, *Ce-rpp-0* and *Ce-tph-1* were retrieved from GenBank and their 5' ends examined. Figure 7.8.1 summarizes this data and shows that, like *P. pacificus*, in *C. elegans* there are two sequences, SL2 and SL3, which comprise the majority of the SLs *trans*-spliced to the 5' end of these genes. In the *C. elegans* genome there are multiple copies of both of these spliced leader genes on chromosomes I, II and III. Whether the *Pp*-SL2β,α,κ are repeated in the *P. pacificus* genome is still to be determined. It is not clear whether the number of copies of a SL2 gene is the factor determining their abundance. It is possible they have promoters that drive their transcription at different rates. There also may be inherent structural differences between the different SL2s which makes some more efficiently *trans*-spliced (i.e. some associate with the polyadenylation complex more readily etc.). It has also not been established if all genes receive the same ratio of SL2 sequences. In *Meliodogyne javanica* variation has been observed in the levels SL1 and SL1M *trans*-spliced to the 5' ends of different genes (Koltai *et. al.*, 1997) so it is possible that there are differences in SL2 usage between downstream genes in different operons.

Observed differences in the abundance of SL sequences *trans*-spliced to the *C. elegans* genes found in the conserved operons

Protein	Total ESTs	SL1	SL2	SL2a	SL2b	SL2c	SL2d	SL3	SL5
Ce-rpl-27a	12	12 (100%)	-	-	-	-	-	-	-
Ce-rpp-1	54	1 (2%)	44 (81%)	5 (9%)	1 (2%)	1 (2%)	1 (2%)	-	1 (2%)
Ce-rpl-36	15	15 (100%)	-	-	-	-	-	-	-
Ce-ayc-1	1	-	1 (100%)	-	-	-	-	-	-
Ce-rpp-0	18	18 (100%)	-	-	-	-	-	-	-
Ce-tph-1	11	-	4 (36%)	-	-	-	2 (19%)	4 (36%)	1 (9%)

Table 7.8.1 Summary of the abundances of the various SLs seen at the 5' ends of the cap selected *C. elegans* ESTs. The number of full length ESTs for gene is shown along with the number of ESTs having a particular spliced leader and the percentage of the dataset it represents.

Phylogenetic analysis of the SL2-like SLs was suggests that the majority of *C. elegans* SLs are grouped separately from the other nematode SL2s. The nodes of the NJ tree which separates the *C. elegans* SL2s from the other nematode SL2s is not supported by bootstrap analysis which is not surprising considering the small size of the mini-exon and the overall homogeneity of the sequences in the alignment. When the full sequences of the SL genes are available more meaningful phylogenetic studies can be performed. If the NJ tree accurately reflects the evolution of the *C. elegans* SL2s it would indicate they have arisen monophyletically from a common ancestor and have spread through the genome. Figure 7.8.2 shows the phylogenetic analysis of the SL2 mini-exons. *Ce*-SL2c and *Ce*-SL2e do not group with these other *Ce*-SL2 sequences and appear to be basal to the other nematode SL2s. They share a 3' GAG with the SL1 instead of AAG which is conserved in the other SL2-like mini-exons. However, the rest of the molecule appears to be SL2-like indicating that they represent a possible intermediate between SL1 and SL2 mini-exons. Searches of the *C. elegans* genome sequence with the other nematode SL2 SLs have not identified any of these other genes. Why *C. elegans* (and presumably the other caenorhabditids) have lost these SLs and evolved a different family while they are conserved in other more distantly related groups like the strongylids is not clear.



```
Ce_SLe    GGTTTAT---ACCCA-GTT-AA-CAAG
Ce_SLc    GGTTTAT---ACCCA-GTT-AC-CAAG
Ce_SL2a   GGTTTAAA--ACCCA-GTT-AC-CAAG
Ce_SLa    GGTTTAT---ACCCA-GTTAAC-CAAG
Ce_SL4    GGTTTTA---ACCCA-TATAAC-CAAG
Ce_SLd    GGTTTTT---ACCCA-GTTAAC-CAAG
Ce_SL3    GGTTTTA---ACCCA-GTTAAC-CAAG
Ce_SL5    GGTTTTA---ACCCAAGTTAAC-CAAG
Ce_SL2b   GGTTTAAA--ACCCA-GTT-AC-CAAG
Ce_SLg    GGTTTTA---ACC-A-GTTAACTCAAG
Pp_SL2χ   GGTTTAT---ACCCA-GTA-TCTCAAG
Pp_SL2β   GGTTTTA---ACCCA-GTA-TCTCAAG
Pp_SL2α   GGTTTTA---ACCCA-GTA-TCTCAAG
Hc_SL2α   GGTTTTA---ACCCA-GTA-TCTCAAG
Pp_SL2γ   GGTTTTA---ACCCG-GTA-TCTCAAG
Pp_SL2η   GGTTTTA---ACCCA-GTA-TCTTAAG
Pp_SL2δ   GGTTTTTA--ACCCA-GTA-TCTCAAG
Os_SL2β   GGTTTTT---ACCCA-GTA-TCTCAAG
Pp_SL2α   GGTTTTT---ACCCA-GTA-TCTCAAG
Pp_SL2ε   GGTTTTT---ACTCA-GTA-TCTCAAG
Pp_SL2ι   GGTTTTG---ACCCA-GTA-TCTCAAG
Pp_SL2φ   GTTTTAT---ACCCA-GTA-TCTCAAG
Pp_SL2κ   GGTTTAT---ACCCA-GTA-TCTCAAG
Pp_SL2λ   GGTTTAA---ACCCA-GTA-TCTCAAG
Pp_SL2ϕ   GGTCTTT---ACCCA-GTA-TCTCAAG
Os_SL2α   GGTTTTTT--ACCCA-GTA-TCTCAAG
Ce_SL2    GGTTTTA---ACCCA-GTT-ACTCAAG
Ce_SL2d   GGTTTTA---ACCCA-GTTAACT-AAG
Ce_SL2f   GGTTTTA---ACC-A-GTTAACT-AAG
Ce_SL2c   GGTTTTA---ACCCA-GTTAATT-GAG
Ce_SL2e   GGTTTTA---ACCCA-GTTAATT-GAG
SL1       GGTTT-AATTACCCTAGTT---T-GAG
SL1α      GGTTTTAATTACCCAAGTT---T-GAG
```

Figure 7.8.2 Evolution of SL sequences. **A:** An alignment of the SL sequences is shown. Conserved portions of the molecule are shaded in grey. **B:** The phylogenetic analysis of SL sequence alignment (neighbor joining, total mean difference) was performed in PAUP*v4.08b (Sinauer Associates Inc.). Bootstrap analysis was performed on the NJ tree (10,000 replicates) and nodes with bootstrap support >50% are shown. The SL1 and SL1α sequences were selected as an outgroup. Additional SL2-like sequences that have not been previously described were identified by searching the *C. elegans* genome with the previously identified genes (Pigaga V. 2000 *honors thesis*, Edinburgh University).

7.9 Discussion

In metazoans the organization of genes in operons and resolution of polycistronic transcripts is a biological process that has so far only been characterized in the nematodes. While parallels can be drawn to the resolution of the large operons utilized by the eugleniods, when the mechanics of these processes are examined it is clear that SLs have evolved very different functions in the two groups of organisms. It is possible that SLs along with *cis*-and *trans*-splicing evolved early in the evolution of eukaryotes. Their absence from most eukaryotic taxa, including the groups considered the most primitive representatives, the Diplomonadida and Parabasala, lends strength to the view that SLs may have evolved independently in the Eugleniods and Metazoa. However, until more comprehensive surveys of the genomes and transcriptomes of these groups are finished the possibility remains that other large groups of eukaryotes will be found that utilize SLs. The question still remains why SLs are utilized at all. There are several possibilities: a) SLs evolved as convenient method of capping and thus stabilizing mRNAs, b) SLs evolved as a targeting signal that more efficiently directed mRNAs to the ribosomes or c) SLs evolved as a mechanism to help resolve polycistronic transcripts. Any of these explanations singly or in combination could account for their prevalence in these two groups. However, when mRNA processing in euglenoid and the metazoan species is compared the resolution of polycistronic transcripts appears to be common only in the two groups which *trans*-splice SLs to the majority of their mRNAs (Kinetoplastids 100% and Nematodes 70-90% of mRNA transcripts respectively). In metazoans such as the platyhelminths in which polycistrons have not yet been conclusively demonstrated, SLs do not appear be added to the majority of their mRNA transcripts (~30% in *S. mansoni*). The identification of SLs in cnidaria and urochordates indicates that if the origin of SL *trans*-splicing in the metazoa is monophyletic, then it evolved before the radiation of the deuterostomes and protostomes. However, within the different metazoan groups the presence and usage of SLs is highly variable. In some groups like the nematodes and platyhelminths, SLs have been isolated from taxa across the group, but in others such as the vertebrates or arthropods it appears wholly absent. Functional studies in trypanosomatids indicate *trans*-splicing precedes polyadenylation of the cistrons and may be the rate limiting step in the maturation of mRNAs. In *C. elegans* both of these processes appear to be coupled with the SL2 snRNP associating with the cleavage and polyadenylation complex. This indicates that these processes may occur simultaneously. Under conditions where the association of the polyadenylation complex is inhibited (i.e. mutation of the polyadenylation signal or removal of the polypyrimidine stretch in the intergenic region) resolution of the cistrons still occurs (Liu *et. al.*, 2001b; Williams *et. al.*, 1999). However, SL1 replaces SL2 as the major SL found on the downstream cistron (Liu *et. al.*, 2001b). Also the first cistron is polyadenylated either at the mutated polyadenylation signal site or the site of separation of the two cistrons by SL1 *trans*-splicing. While there do not appear to be any functional differences between the mini-exons of SL1 and SL2, the coupling of polyadenylation and *trans*-splicing could offer a number of advantages. First both processes would occur simultaneously to each pre-mRNA reducing the accumulation of processing intermediates. Second, because those mRNAs with caps and polyadenylation are more stable than those which are unprocessed, downstream cistrons that were not given SLs or polyadenylated relatively quickly could be targeted for degradation

7.9.1 SL Usage and Operons in the Nematodes

The one major question this study addressed was to determine if operons could be found in nematode groups that were distantly related to the rhabditina. By rationally selecting and testing operon candidates for conservation of synteny three operonic structures, *rpp-0/tph-1*, *rpl-27a/rpp-1* and *rpl-36/ayc-1*, have been identified that are conserved through nematode clades III, IV and V (Blaxter *et. al.*, 1998). Conservation of these operonic structures is not conclusive evidence of polycistronic transcription of the two genes. However, the close proximity of the syntenic genes (see table 7.5.2A-C) makes it unlikely that promoter elements are present in the intergenic region. Also, processing intermediates (polycistrons with introns removed) of the *rpl-27a/rpp-1* operon were isolated by RT-PCR from *B. malayi*. While this is evidence that the genes are transcribed as a single RNA and that *cis*-splicing of the polycistron occurs it does not establish if the polycistron is processed into two productive mRNAs or whether their maturation is mutually exclusive. When the SL composition of the downstream cistron *rpp-1* was surveyed it revealed that only SL1 was found on the 5' ends of the gene cloned from *B. malayi* (Clade V). However, when the *rpp-1* ortholog from a diplogasterid (Clade V) was surveyed it showed that like rhabditids *C. elegans* and *Oscheius sp.*CEW1, *P. pacificus trans*-splices SL2-like spliced leaders to the 5' end of *Pp-rpp-1*. Figure 7.9.1.1 summarizes these findings within the context of the whole phylum Nematoda. Several layers of data are still missing from this analysis: 1) Do clade IV nematodes utilize SL2-like sequences? 2) Are the operons isolated in this study conserved in clade I and II nematodes? If the operons are conserved what is the composition of the SLs *trans*-spliced to the 5' end of the downstream genes?

If only SL1 is found at the 5' ends of the downstream cistrons in the *S. ratti* operons this would indicate that use of a distinct set of SL sequences in the resolution of operons may be an innovation of the clade V nematodes (strongylids, rhabditids, diplogasterids). If the operons found in the clade III, IV and V nematodes are conserved through clade I and II it would suggest their origin may predate the radiation of the phylum.

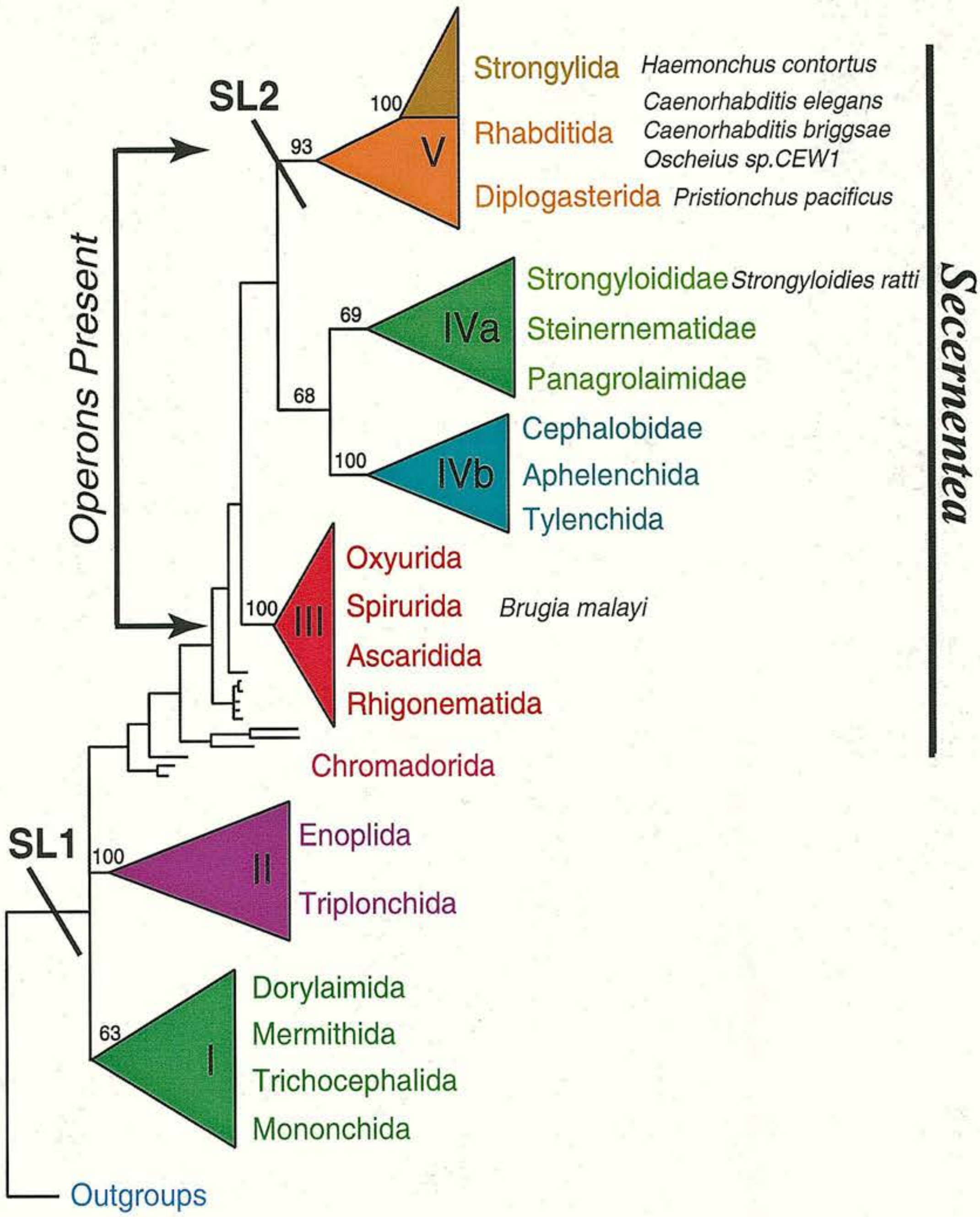


Figure 7.9.1.1 A cartoon showing SL usage and the evolution of the phylum Nematoda as derived from analysis of the 18S rRNA (adapted from Blaxter *et. al.* 1998). The nematode species that have been studied and whether they utilize SL2-like spliced leaders is indicated.

7.9.2 SL2 evolution in Clade V

SL2-like spliced leaders are a large group of diverse sequences. The 5' RACE survey of the *P. pacificus rpp-1* gene has highlighted some interesting features of this gene family and provided clues to how SL2s have evolved through clade V. Unlike SL1, which appears to be monomorphic, each of the species examined presented their own set of SL2 sequences. However, the mini-exon sequence of *Pp*-SL2α was found in two of the five species examined (*P. pacificus* and *H. contortus*). The sequence of the mini-exon of SL2β from *Oscheius sp.*CEW1differed from *Pp*/*Hc*-SL2α in only a single thymine insertion in the poly-thymine track in the 5' portion of the molecule. Figure 7.9.2.1 shows the evolutionary relationship of the clade V nematodes which SL2s have been identified (based on phylogenetic analysis of the SSU rRNA (Blaxter *et. al.*, 1998)). The usage of various SL2-like SLs is mapped onto the species. From this phylogeny it can be inferred that the presence of *Pp*/*Hc*-SL2α and the closely related *Os*-SL2β in all the three major groups (strongylid, rhabditid, diplogasterid) suggests that it may represent an ancestral SL2 sequence. The elaboration of another set of SL2-like sequences in *C. elegans* and *C. briggsae* may be a feature of that genus. If the *Pp*/*Hc*-SL2α is representative of ancestral sequence, why it would be lost from the caenorhabditids is unclear. SL2s are multicopy gene families in *C. elegans.* Some of these genes are clustered. If SL2s are also multi-copy and clustered in other nematodes deletions of portions of the genome containing these clusters could remove specific SL2 families. As long as the other SL2 sequences sufficiently compensated for the losses of the genes the deletions might not prove detrimental. Subsequent duplications and drift of the remaining SL2 genes could then repopulate the genome with a new family of SL2. To test this hypothesis additional nematodes within clade V, particularly those placed near the caenorhabditids would need to be surveyed. However, because the sequences of the SL2 genes are not well conserved outside the short mini-exon it would be difficult to show that an individual SL2 gene was missing from a particular genome other than by completely sequencing it.

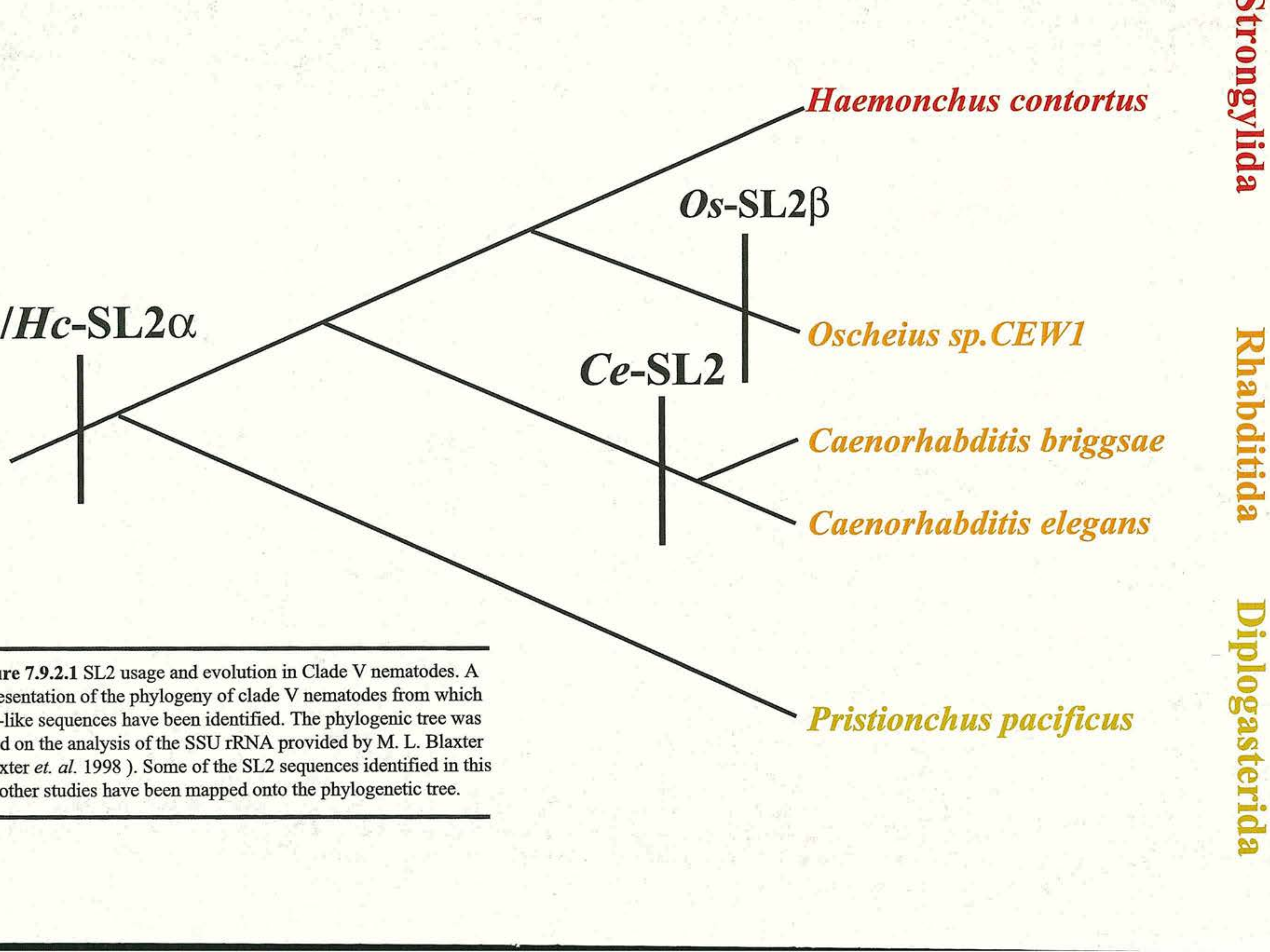


Figure 7.9.2.1 SL2 usage and evolution in Clade V nematodes. A representation of the phylogeny of clade V nematodes from which SL2-like sequences have been identified. The phylogenic tree was based on the analysis of the SSU rRNA provided by M. L. Blaxter (Blaxter *et. al.* 1998). Some of the SL2 sequences identified in this and other studies have been mapped onto the phylogenetic tree.

7.10 Conclusions:

During the course of this work two important finding have been made. First, operonic structures are a common feature in all nematodes within the Secernentea (clades III, IV and V). Second, utilization of SL2-like SLs in the resolution of these operons has only been found in clade V nematodes. While this survey does not cover the entire phylum there are still several conclusions that can be drawn. Operons are an ancient feature of the Nematoda and are conserved through large sections of the phylum. While it is unclear if the mechanisms involved in the resolution of these operons are the same across the phylum the lack of SL2-like SLs in the clade III nematode suggests that SL2 may be a relatively recent innovation. If the SL2 snRNPs from other clade V species interact with the polyadenylation complex this would be strong evidence supporting the idea that SL2-like sequences evolved as a mechanism to couple (and perhaps streamline) two important steps in the maturation process of polycistronic mRNA.

We do not know how common operons are in nematodes outside of the Caenorhabditids. However, the widespread use of operons as a form of gene organization in clade V may have pushed the development of SL2 by making it necessary to couple polyadenylation and *trans*-splicing reactions to more efficiently resolve polycistrons. Unfortunately lack of large EST datasets has prevented the expansion of this study to the other major nematode groups clades C, I and II. While SL1 is present in these nematodes it is unclear if these groups rely on SL1 usage to the same extent as the Secernentea. If these operons are conserved throughout the phylum and the separation of the downstream cistrons is a process reliant on *trans*-splicing this could indicate a very ancient origin for this form of genomic organization. It could also suggest some possible reasons for the widespread use of SLs in nematodes. One possibility is that the utilization of polycistrons as a form of genomic organization drove the reliance on SL usage in the nematodes. As this study is expanded and the use of operons and SL sequences determined in other species, it will become possible to map the evolution of these processes more accurately and extensively across the entire phylum.

Chapter 8

General Conclusions

The rationale behind these studies was to utilize the sequence datasets and other resources produced by the FGP to address several questions about the mode and tempo of the evolution of nematode genes and genomes. This sequence data also served as a platform to determine the conservation of a biological process that had until this point only been characterized in *C. elegans* and its close relatives.

The EST datasets from a human lymphatic and cutaneous filarial species were clustered into a non-redundant set of gene fragments using a custom built process designed and implemented during the course of these studies. These clusters were extensively analyzed to determine the extent of conservation between filarial genes and the datasets of other organisms. Interestingly, the filarial datasets had a much higher rate of novel sequences than the other nematode EST datasets or the proteins predicted from other animal genomes. Whether this high rate of novel sequences is reflective of a high rate of novel protein sequences in filaria or is an artifact of problems with the EST sequences (short reads, poor sequence etc.) has not been satisfactorily resolved. Many new nematode and parasite specific gene families were identified in this analysis and these will serve as a starting point for subsequent studies to identify novel vaccine candidates or drug targets. Analysis of abundant differentially expressed genes identified many novel genes and gene families that may play a role in biologies specific to particular lifecycle stages or serve as mediators of host-parasite interactions. Interestingly when the two filarial datasets were compared there was very little correlation in the identities of the hyper-abundant transcripts between the two species. The exception to this observation was the abundant differentially expressed transcripts from the infective L3. Several of these genes or gene families showed similar expression profiles (*alts*, *agys*, *vah-1* and *cpl-1*). The presence of these genes in both species at the same developmental time point indicates that they may share similar biological functions.

Within the EST datasets there were several gene families whose similarities to proteins in the public databases suggested they may play a role in mediating host-parasite interactions. One of these families showed similarity to the vertebrate cytokine family macrophage migration inhibitory factor (MIFs). Two MIF genes were identified in both *B. malayi* and *O. volvulus*. Comparisons to the public databases, other nematode EST datasets and partial genome sequences from a variety of organisms revealed that the MIFs form a diverse group of sequences that have representatives in animals, plants and several protozoan groups. To better understand how the filarial MIFs were related to MIFs from other nematodes and vertebrates as well as determine how this important gene family has evolved, MIF sequences were compared and analyzed with several molecular phylogenetic techniques. Comparison of the MIF proteins sequences revealed several interesting features and indicate not all MIFs share the same enzymatic activities and substrate specificities. Also some MIF families appear to be evolving at different rates indicating that different selection pressures may be shaping these gene families. Phylogenetic studies and examination of intron positions from available genomic sequences indicate that the animal MIFs form two distinct families. One of these families may represent an ancestral group but this observation is not supported by all the phylogenetic analyses. Comparison of sequences isolated from mammals and teleosts indicated that there is one functional gene from each family contained within the vertebrate genome. Within the nematodes the number of MIFs appear to be more variable with the *C. elegans* genome containing four MIF sequences. In the vertebrates the two MIF families appear to have distinct functions: whether this is also true of the nematodes MIFs is yet to be determined. One of the most intriguing aspects of the parasitic nematode MIFs is their potential function as immuno-modulators. MIFs are known to be secreted from several animal parasitic nematodes. However, in this study several representatives have been found in the EST datasets of plant parasitic nematodes. Are these secreted into the parasitized plant? Do they also have modulatory functions? Phylogenetic studies have indicated that both animal and plant parasitism have independently evolved several times in nematodes. Could MIF have been repeatedly recruited as a mediator of host-nematode interaction? Another intriguing finding in this study was the discovery of MIFs in the genome sequences of several parasitic protozoa. If these MIFs also have immunomodulatory functions this would indicate that MIFs have not only been repeatedly recruited by parasitic metazoa but also by parasites through the whole eukaryote phylum.

Very little is known about the forces that shape animal genomes. In vertebrates the composition and relative order of genes in large segments of linkage groups are conserved between evolutionarily distant species. Conversely, studies in insects have shown that while the composition of linkage groups in closely related species may be conserved, the order of genes on chromosomes is fluid, with genes being moved freely within and between chromosomes. This implies that either functional constraints are stabilizing chromosomes in vertebrates or some unknown process is causing increased incidence of rearrangements in insect genomes. Outside of the insects very little is known about the dynamics of genome evolution in non-vertebrates. To examine the mode and tempo of chromosome evolution in nematodes regions of genome flanking the two *B. malayi* MIF genes were examined. *C. elegans* orthologues of all of the other genes found in these segments of the *B. malayi* genome were found in common linkage groups. In the case of the genes surrounding *Bm-mif-1* the majority of their putative *C. elegans* orthologues were found in a 2.5 MB region in the center of chromosome I. The evolutionary distance between *C. elegans* and *B. malayi* is comparable to that separating mammals and teleost fish. The results indicate that like the insects the relative order of genes within nematodes genomes is relatively fluid and while the relative composition of linkage groups may be conserved their arrangement will vary greatly between species. When the full genome sequence of *C. briggsae* becomes available and larger portions of the *B. malayi* genome are sequenced it will be possible to robustly estimate the rate of these rearrangements in nematode species and compare them to the rates observed in other animals. Interestingly in both genome segments examples of conserved microsynteny could be found. All of these gene clusters contained a pair of genes that were divergently transcribed on opposite strands. Why would the configuration of these genes be conserved between two such evolutionarily distant species? Given the observation that most of the genes are rearranged randomly it is highly probable that the movement of these genes is functionally constrained. Perhaps promoter or enhancer elements contained within the intergenic regions separating the genes are binding them together? As more fine scale comparisons of insect genomes are preformed it will be come clear if this is a common occurrence in other non-vertebrates.

Operons and the resolution of polycistronic transcripts is an unusual process in metazoans that has only been shown to be a common form of gene organization in the nematode *C. elegans* and a few of its close relatives. The resolution of the polycistrons is dependent on *trans*-splicing reactions which cap the downstream mRNAs with a second distinct family of SL sequences (SL2s). It has recently been shown that this *trans*-splicing reaction is coupled to the polyadenylation of upstream genes via specific interactions of the SL2 RNA with the polyadenylation complex. Until this study it was not known how universal these process are in other nematodes. Using a directed approach several operonic structures conserved between *C. elegans, P. pacificus, S. ratti* and *B. malayi* were identified. Interestingly, when the 5' ends of cDNAs of a gene found downstream in a *B. malayi* operon was examined only SL1 was found. This raises the possibility that while operons may be a form of gene organization utilized by other nematodes the mechanisms by which polycistrons are resolved may not be conserved. It still remains to be determined if the *trans*-splicing reaction is coupled to polyadenylation in other nematodes. It is possible that this is an adaptation found exclusively in *C. elegans* and its relatives. How common operons are in *B. malayi* and other nematodes and whether they are also conserved outside the Secernetea remains to be determined. When this data is available it will be possible to more accurately assess how ancient nematode operons are and what forces lead to their genesis.

Appendix I

General Media and Solutions

Solution	Ingredient	Amount
LB:	Tyrptone	10g
	Yeast Extract	5g
	NaCl	5g
	ddH$_2$O to	1L
SOC	Tyrptone	20g
	Yeast Extract	5g
	NaCl	0.5g
	KCl	186 mg
	ddH$_2$O to	1L
	adjust pH to 7.0 with NaOH (~0.2 mL). After autoclaving add 10mL of sterile 1M MgCl2 and 10mL of 50% glucose	
2x YT	Tyrptone	16g
	Yeast Extract	8g
	NaCl	8g
	ddH$_2$O to	1L
1x TE pH8.0	10mM Tris-Cl (pH 8.0)	
	1mM EDTA	
5x TBE	Tris-Base	54g
	Boric Acid	27.5g
	50mM EDTA	20mL
	ddH$_2$O to	1L

Appendix II

Gene Name	Species	Common Name	Group	Origin	Genbank Ac.
CHMI_ESCCO	*Escherichia coli*	*E.coli*	Proteobacteria/ gamma subdivision	nr	Q05354
MIF1_BOSTA	*Bos taurus*	cow	Metazoa/Vertebrata	EST assembly	AF119571 AW463009 BF773948
MIF1_DANRE	*Danio rerio*	zebra fish	Metazoa/Vertebrata	EST assembly	AW279673 AW280038 AW280595 AW281307 AW281683
MIF1_GALGA	*Gallus gallus*	chicken	Metazoa/Vertebrata	nr	M95776
DDT1_HOMSA	*Homo sapien*	human	Metazoa/Vertebrata	nr	XM_037716
MIF1_HOMSA	*Homo sapien*	human	Metazoa/Vertebrata	nr	XM_037707
MIF1_MERUN	*Meriones unguiculatus*	jird	Metazoa/Vertebrata	nr	AF045740
DDT1_MUSMU	*Mus musculus*	mouse	Metazoa/Vertebrata	nr	6753617
MIF1_MUSMU	*Mus musculus*	mouse	Metazoa/Vertebrata	nr	AK013328
MIF1_ORYLA	*Oryzias latipes*	japanese medaka	Metazoa/Vertebrata	EST assembly	AV670637 AV670802
MIF1_PSEAM	*Pseudopleuronectes americanus*	winter flounder	Metazoa/Vertebrata	EST	AW013408
DDT1_RATNO	*Rattus norvegicus*	rat	Metazoa/Vertebrata	nr	Z36980
MIF1_RATNO	*Rattus norvegicus*	rat	Metazoa/Vertebrata	nr	U62326
MIF1_SUSSC	*Sus scrofa*	pig	Metazoa/Vertebrata	nr	AF176246
MIF1_XENLA	*Xenopus laevis*	african clawed frog	Metazoa/Vertebrata	EST	BE681403
MIF1_CIOIN	*Ciona intestinalis*	common tunicate	Metazoa/Urocordate	EST assembly	AV847874 AV895295 AV901255

MIF2_CIOIN	*Ciona intestinalis*	common tunicate	Metazoa/Urochordate	EST	AV873510
MIF1_AMBLY	*Amblyomma americanum*	tick	Metazoa/ Chelicerata	nr	AF126688
MIF1_ANCCA	*Ancylostoma caninum*	canine hookworm	Metazoa/Nematoda	EST	AW626839
MIF2_ANCCA	*Ancylostoma caninum*	canine hookworm	Metazoa/Nematoda	EST	AW181853
MIF1_ASCSU	*Ascaris suum*	pig roundworm	Metazoa/Nematoda	EST	BI593989
MIF1_BRUMA	*Brugia malayi*	lymphatic filarial nematode	Metazoa/Nematoda	nr	AF040629
MIF2_BRUMA	*Brugia malayi*	lymphatic filarial nematode	Metazoa/Nematoda	nr and EST	AY004865 AA257577
MIF1_BRUPA	*Brugia pahangi*	lymphatic filarial nematode	Metazoa/Nematoda	nr	AJ275477
MIF3_CAEEL	*Caenorhabditis elegans*	free-living nematode	Metazoa/Nematoda	nr	NM_059668
MIF2_CAEEL	*Caenorhabditis elegans*	free-living nematode	Metazoa/Nematoda	nr	NM_073602
MIF1_CAEEL	*Caenorhabditis elegans*	free-living nematode	Metazoa/Nematoda	nr	NM_067135
MIF4_CAEEL	*Caenorhabditis elegans*	free-living nematode	Metazoa/Nematoda	nr	NM_068567
MIF1_HAECO	*Haemonchus contortus*	barber's pole worm of sheep	Metazoa/Nematoda	EST	BI595478 BI595614
MIF2_HETGL	*Heterodera glycines*	soybean cyst nematode	Metazoa/Nematoda	EST	BF013613 BF013909
MIF2_MELJA	*Meloidogyne javanica*	root-knot nematode	Metazoa/Nematoda	EST	BG894322
MIF1_ONCVO	*Onchocerca volvulus*	cutaneous filarial nematode	Metazoa/Nematoda	nr	AF384027
MIF2_ONCVO	*Onchocerca volvulus*	cutaneous filarial nematode	Metazoa/Nematoda	nr	AF384028
MIF2_STRSE	*Strongyloides stercoralis*	human gut parasite	Metazoa/Nematoda	EST	BE028834 BE028850

					BE029714 BE223719 BE223780 BE223857 BE224537
MIF1_TRITR	*Trichuris trichiura*	human whip worm	Metazoa/Nematoda	nr	AJ237770
MIF1_TRISP	*Trichinella spiralis*	trichina muscle parasite	Metazoa/Nematoda	nr	AY050661
MIF2_TRISP	*Trichinella spiralis*	trichina muscle parasite	Metazoa/Nematoda	EST	BG354844
MIF1_TRIPS	*Trichinella pseudospiralis*	trichina muscle parasite	Metazoa/Nematoda	nr	AY050662
MIF1_WUCBA	*Wuchereria bancrofti*	lymphatic filarial nematode	Metazoa/Nematoda	nr	AF040629
LS1_PHYSO	*Phytophthora sojae*	soybean pathogen	Heterokonta/ Oomycetes	EST	BE582695
LS1_EIMTE	*Eimeria tenella*	chicken parasite	Apicomplexa	EST	AI755805 AI757530 BE027544
LS1_PLABE	*Plasmodium berghei*	murine malaria	Apicomplexa	EST	BF294177 BF294229 BF294865 BF295329 BF295726 BF295939 BF297076 BF297641 BF297969
LS1_PLAFA	*Plasmodium falciparum*	human malaria	Apicomplexa	HGS	NA[1]
LS1_PLAYO	*Plasmodium yoelii*	murine malaria	Apicomplexa	HGS	NA[1]

LS1_TOXGO	*Toxoplasma gondii*	feline gut parasite	Apicomplexa	EST	N81780 W63329
MIF1_GIAIN	*Giardia intestinalis*	gut parasite	Diplomonadida	HGS	AC056441
LS1_CRYJA	*Cryptomeria japonica*	Japanese cedar	Viridiplantae/ Coniferopsida	EST	AU085668
LS1_PINTA	*Pinus taeda*	loblolly pine	Viridiplantae/ Coniferopsida	EST	AW042584
LS1_HORVU	*Hordeum vulgare*	barley	Viridiplantae/ Liliopsida	EST	BI953781
LS1_ORYSA	*Oryza sativa*	rice	Viridiplantae/ Liliopsida	EST	BI799671
LS1_SORBI	*Sorghum bicolor*	common sorghum	Viridiplantae/ Liliopsida	EST	AW745430
LS1_TRIAE	*Triticum aestivum*	wheat	Viridiplantae/ Liliopsida	EST assembly	BE400433 BG907503 BG907504
LS2_TRIAE	*Triticum aestivum*	wheat	Viridiplantae/ Liliopsida	EST assembly	BE213336 BE430035
LS1_ZEAMA	*Zea mays*	corn	Viridiplantae/Liliopsida	EST	AW506633
LS1_ARATH	*Arabidopsis thaliana*	thale-cress	Viridiplantae/ Eudicotyledons	nr	AL161946
LS2_ARATH	*Arabidopsis thaliana*	thale-cress	Viridiplantae/ Eudicotyledons	nr	AL132968
LS3_ARATH	*Arabidopsis thaliana*	thale-cress	Viridiplantae/ Eudicotyledons	nr	AB023042
LS1_GLYCL	*Glycine clandestina*	soybean sp.	Viridiplantae/ Eudicotyledons	EST	BG838287
LS1_GLYMA	*Glycine max*	common soybean	Viridiplantae/ Eudicotyledons	EST	AW760171 AW704652 BE347192

					AW310908
LS2_GLYMA	*Glycine max*	common soybean	Viridiplantae/ Eudicotyledons	EST	AW099348 AW569088 AW508929 BE022510 AI437663 AW306529
LS3_GLYMA	*Glycine max*	common soybean	Viridiplantae/ Eudicotyledons	EST	AW397382 AW734204
LS1_GOSAR	*Gossypium hirsutum*	upland cotton	Viridiplantae/ Eudicotyledons	EST	AI726121
LS2_GOSAR	*Gossypium hirsutum*	upland cotton	Viridiplantae/ Eudicotyledons	EST	BF275087
LS1_HELAN	*Helianthus annuus*	common sunflower	Viridiplantae/ Eudicotyledons	EST	AJ412556
LS1_LOTJA	*Lotus japonicus*	common lotus	Viridiplantae/ Eudicotyledons	EST	AV416658
LS1_LYCES	*Lycopersicon esculentum*	tomato	Viridiplantae/ Eudicotyledons	EST	AW040389
LS1_MEDTR	*Medicago truncatula*	barrel medic	Viridiplantae/ Eudicotyledons	EST	BE202493
LS1_MESCR	*Mesembryanthemum crystallinum*	ice plant	Viridiplantae/ Eudicotyledons	EST	AI881581
LS2_MESCR	*Mesembryanthemum crystallinum*	ice plant	Viridiplantae/ Eudicotyledons	EST	AI822262
LS1_ROBPS	*Robinia pseudoacacia*		Viridiplantae/ Eudicotyledons	EST	BI679000
LS1_SOLTU	*Solanum tuberosum*	potato	Viridiplantae/ Eudicotyledons	EST assembly	BG590748 BG592304 BG889886

					BI432941 BI435097

Table 5.1.1 Summary of the characteristics of each MIF sequence isolated from GenBanks non-redundant or EST database. The gene name used in the multiple sequence alignment, the organism from which the sequence originates, taxonomic grouping of the species, the GenBank database the sequence was found in and the GenBank accession numbers of the sequence(s) are listed. nr: non-redundantnucleotide and protein database, EST: dbEST, HGS: high throughput genome sequence, assembly: indicates an assembly of two or more sequence was used to derive the amino acid and cDNA sequences used in the alignments. NA[1] sequences were found by using the blast facilities hosted at http://www.plasmodb.org to search the combined shotgun and assembled contigs of *Plasmodium falciparum* and *Plasmodium yoelii*. These sequences are not yet available in GenBank. MIF or DDT was used as an identifier when the sequences were of animal origin. Light sensitive transcript (LS) was used for those sequences of plant origin. This annotation was attached to a set of the plant ESTs derived from cDNA library of light stimulated *A. thaliana*embryos.

Appendix III

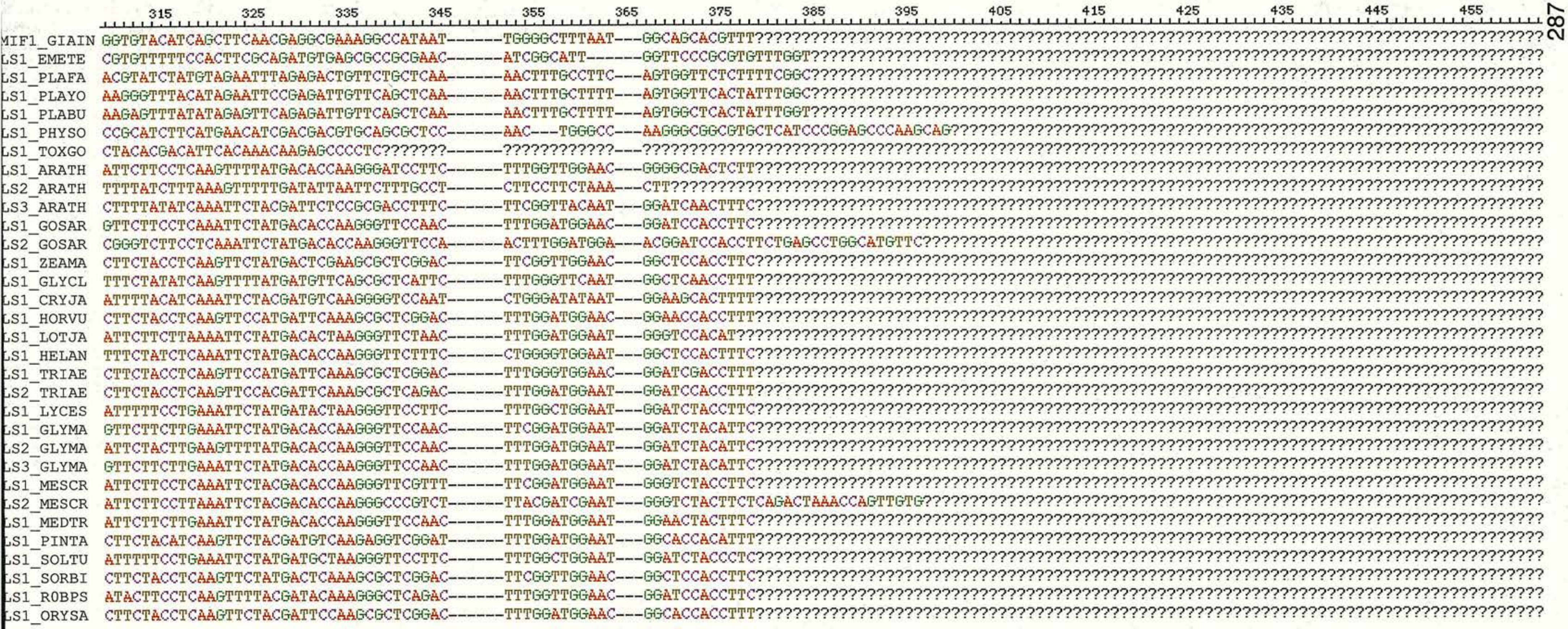
```
               5         15        25        35        45        55        65        75        85        95        105       115       125       135       145     1
CHMI_ESCCO ATGCCGCACTTTATCGTTGAATGCAGTGATAACATCCGCGAAGAAGCCGACCTGCCGGGGTTG---TTCGCCAAAGTGAATCCGACGCTGGCAGCCACGGGTATTTTTCCGCTGGCGGGTATTCGCAGCCGCGTGCATTGGGTCGATACCTGGC
DDT1_HOMSA ATGCCGTTCCTGGAGCTGGACACG------AATTTGCCCGCCAACCGAGTGCCCGCGGGGCTG---GAGAAACGACTCTGCGCCGCCGCTGCCTCCATCCTGGGCAAACCTGCGGAC------CGCGTGAACGTGACGGTACGGCCGGGCCTGG
DDT1_RATRA ATGCCGTTCGTTGAGTTGGAAACA------AACTTGCCGGCTAGCCGCATACCCGCAGGGCTG---GAGAACCGGTTGTGTGCGGCCACAGCCACCATCCTGGACAAACCCGAAGAC------CGCGTGAGCGTGACGATACGACCGGGCATGA
DDT1_MUSMU ATGCCATTCGTTGAGTTGGAAACA------AACTTGCCGGCTAGCCGCATACCCGCGGGGCTG---GAGAACCGGCTGTGTGCGGCCACAGCCACCATCCTGGACAAACCCGAAGAC------CGCGTGAGCGTTACGATACGACCTGGCATGA
MIF1_ORYLA ATGCCTTTCGTGGAGCTCCAGACC------AATTTACCTGGTTCCTCGTTTAATGAGGATTTT---TTGAAGAAGTTGTGTTCGTGCGTCGCTTCAACCTTAAGCAAACCAGAGGAG------AGGATGAACGTGGTGGTGAAACCTGGGCTGC
MIF1_MUSMU ATGCCTATGTTCATCGTGAACACC------AATGTTCCCCGCGCCTCCGTGCCAGAGGGGTTT---CTGTCGGAGCTCACCCAGCAGCTGGCGCAGGCCACCGGCAAGCCCGCACAG------TACATCGCAGTGCACGTGGTCCCGGACCAGC
MIF1_RATRA ATGCCTATGTTCATCGTGAACACC------AATGTTCCCCGCGCCTCCGTGCCAGAGGGGTTT---CTCTCCGAGCTCACCCAGCAGCTGGCGCAGGCCACCGGCAAGCCGGCACAG------TACATCGCAGTGCACGTGGTCCCGGACCAGC
MIF1_HOMSA ATGCCGATGTTCATCGTAAACACC------AACGTGCCCCGCGCCTCCGTGCCGGACGGGTTC---CTCTCCGAGCTCACCCAGCAGCTGGCGCAGGCCACCGGCAAGCCCCCCCAG------TACATCGCGGTGCACGTGGTCCCGGACCAGC
MIF1_MERUN ATGCCTATGTTCATCGTGAACACC------AACGTTCCCCGCTCCTCCGTGCCAGAGGGGCTT---CTCTCTGAGCTCACCCAGCAGCTGGCGCAGGCCACCGGCAAGCCGGCACAG------TACATCGCAGTGCACGTGGTCCCGGACCAGC
MIF1_SUSSC ATGCCGATGTTCGTGGTAAACACC------AACGTTCCCCGCGCCTCTGTGCCGGACGGGTTC---CTCTCCGAGCTGACTCAGCAGTTGGTGCAGGCCATGGGCAAGCCGGCGCAG------TACATCGCGGTGCACGTCGTCCCGGACCAGC
MIF1_BOSTA ATGCCGATGTTCGTGGTGAACACC------AACGTGCCCCGCGCCTCCGTGCCGGACGGGCTC---CTCTCCGAGCTCACGCAGCAGCTGGCGCAGGCCACGGGCAAGCCGGCACAG------TACATCGCGGTGCACGTGGTCCCAGACCAGC
MIF1_GALGA ATGCCTATGTTCACCATCCACACC------AACGTCTGCAAGGACGCCGTGCCCGACAGCCTG---CTGGGCGAGCTGACCCAGCAGCTGGCCAAGGCCACCGGCAAGCCCGCGCAG------TACATAGCCGTGCACATCGTACCTGATCAGA
MIF1_CIOIN ATGCCCCATCTATTTGTGAAAACA------AATGTTGCAAAAGACAAGTTGCCAAAGAGCATACTGCAGGATTTGACTAAGTTGGTGTCTTCAACAATTCCAAACAAGCCTGAGAAA------TATGTCTGTGTGACTGTGGTTCCTGATGTAT
MIF2_CIOIN ATGCCAGAAATTACGATACAAACC------AACGTTTCAAGCGACAAAATCGCTTCGGATTTA---CAAGAGATAGTGGTTGAACTTGTATCGCAGCATCTCAACAAACCGAAAGCA------AATATATGCGTCACTGTGTTAACAGATCTAT
MIF1_XENLA ATGCCTGTCTTCACCATCCGTACC------AACGTCTGCCGGGACTCCGTGCCCGATACCCTT---CTGTCCGATCTCACCAAGCAGCTGGCCAAGGCTACGGGCAAACCAGCTGAG------TACATTGCAATTCATATTGTGCCTGATCAAA
MIF1_DANRE ATGCCGATGTTTGTAGTGAACACA------AATGTTGCTAAAGACTCGGTTCCGGCGGAGCTG---CTGTCGGAGGCCACGCAGGAGCTCGCCAAAGCCATGGGCAAACCCCAGCAG------TACATCGCCGTACAGGTTGTGCCGGATCAGA
MIF1_PSEAM ATGCCGATGTTCGTGGTGAACACC------AACGTGGCCAAAGGCGACGTGCCCGCGGCGCTG---CTGTCCGAGGCCACCGAGGAGCTGGCCAAGGAAATGGGCAAACCTGCACAG------TATATCGCTGTGCACATCAACCCTGACCAAA
MIF1_CAEEL ATGCCCGTCATCAAAGTGCAAACA------AATGTC------AAAAAAGTATCCGATGGGTTC---GAGGTCCGACTTGCAATTCATATGGCGAAAGTGATGAAACGTCCTGAAAGC------CAGATATTCGTCTCGCTCGACATGAATTCTC
MIF2_CAEEL ATGCCGATGGTCAGAGTTGCGACG------AATCTACCAAATGAGAAGGTTCCTGTTGATTTT---GAGATTCGTCTCACCGATCTTCTTGCTCGATCAATGGGAAAACCAAGAGAG------AGAATTGCTGTCGAGATAGCCGCCGGTGCTC
MIF3_CAEEL ATGCCAGTTTTCTCCATCAATGTC------AACGTAAAGGTCCCTGCCGAGAAGCAGAATGAGATCTTGAAAGAGTTGTCAACTGTTCTCGGAAAGCTTCTCAACAAGCCGGAGCAG------TACATGTGTATTCACTTCCACGAGGATCAGG
MIF4_CAEEL ATGCAAGTTGTTCGAATTCAAACC------AACATCCGAAGTGCTGATATTCCAGAAAAGTTT---GAACAAGATGTGATCTACAATTTATCAGTTGTAATGGAATTACCAGCTGAT------AAATTCGTGATTATTGTTGAGCCGGCAGTTA
MIF1_BRUMA ATGCCATATTTTACGATTGATACC------AACAAACCACAGGATAGCATTTCAAGTGCTTTC---CTAAAGAAGGCACCAAATGTGGTTCCAAAAGCACTTGGAAAACCGGAAAGT------TATGTATCAATCCATGTGAATGGTGGACAAC
MIF1_BRUPA ATGCCATATTTTACGATTGATACA------AACATACCGCAGAATAGCATTTCGAGTGCTTTC---CTAAAGAAGGCATCAAATGTGGTTGCAAAAGCACTCGGAAAACCGGAAAGT------TATGTATCAATCCATGTGAATGGTGGACAAG
MIF1_WUCBA ATGCCATATTTTACGATTGATACC------AACAAACCACAGGATAGCATTTCAAGTGCTTTC---CTAAAGAAGGCACCAAATGTGGTTCCAAAAGCACTTGGAAAACCGGAAAGT------TATGTATCAATCCATGTGAATGGTGGACAAC
MIF1_ONCVU ATGCCTGCTTTTACGATCAATACA------AACATACCGCAGAGCAATGTTTCGGATGCGTTC---CTAAAGAAGGCATCAAGCACGGTTGCGAAAGCACTTGGAAAGCCGGAAAGT------TATGTGGCAATTCATGTGAACGGTGGACAAG
MIF1_ASCSU ATGCCNGTGCTCACTATCAACACG------AATGTGCCGTCAGACAAGGTTCCACAGGACTTC---CTCAAGAAGACCTCTGCGCTCGTTGCGAAGTCCCTCAGCAAGCCTGAGAGC------TATGTGGCAGTACGCGTGAACCCCGACCAGC
MIF1_TRITR ATGCCWATYTTYACRTTYWSNACG------AACGTTCCTTCTGAGAACATTTCCGTCGATTTC---CTGAAGAGCACAAGCAAGTTGATAGCCGGTATGCTCGGCAAACCAGAATCG------TACGTCGCAGTTCATATAAACGGTGGACAAA
MIF1_TRISP ATGCCTATCTTTACTCTTAATACA------AACATCAAAGCTACCGATGTTCCGTCGGACTTT---TTGTCCAGCACAAGCGCACTTGTTGGTAATATATTATCAAAACCAGGAAGT------TATGTAGCTGTGCACATCAACACAGATCAGC
MIF1_TRIPS ATGCCTATTTTTACGTTTAATACA------AACATCAAAGCTACCGATGTTCCGTCAGACTTT---TTGTCTAGCACAAGCGCACTGCTTGCTGATATATTATCTAAACCTGAAAGT------TATGTAGCTGTGCACCTCAACACAGATCAGC
MIF2_TRISP ATGCCAATTTTCACAATAATTACA------AAT------AAAAAAACTGCACCGAAAGATTTT---CACCGATTGCTAACAGATCTGTTGGCGGAATTGCTGAAAAAACCGAAAGAG------CTAGTGGTGGTTGATTTATTGCTTGATCAAA
MIF2_BRUMA ATGCCGCTGATAACTCTTGCTTCG------AACGTTCCCGCGAGTAGATTTCCGAGTGATTTC---AATGTTCAGTTCACGGAGTTAATGGCGAAAATGCTAGGAAAACCAACAAGT------CGGATACTTTTACTGGTAATGCCAAATGCAC
MIF2_ONCVU ATGCCGCTGATAACGCTCGCCTCG------AATGTTCTTGCAAGTGGATTTCCGACTGATTTC---AGTGTCCAATTCACGAAGTTAATGGCGGAATTGCTTGGGAAGCCAATAAGT------CGGATAACTCTATTGGTGACGCCAAGCGCAC
MIF2_HETGL ATGCCATTTATAAATCTTTGGACC------AACCTTCCCGAACTCAAATTGGACAACGAATTC---AGACGAACATTCCTCGCGACAGTGGCCAATTCAATGCAAAAGCCAGTGGAA------TCGACAGCACTAATTGTCCAGTCGGGACCAA
MIF2_MELJA ??????????????????GATCAT------TTCCCTCCTAAAATTTTTCATATTCCCCCATTT---CAGGGAGGGCTTCAGCCTCCCCTCGCTACCTCCGAAAAAATTAAATTCAAT------TTTTCCAAGGTTCTAGTTAATGCTGGGAATG
MIF2_STRST ATGCCATATGTTCGTTTGTTCTCT------AATTTGCCAGAAACATCTTTTACAGATGCTTTT---TGTACACAATTTACCGATTTATTAGCTGAAAAATTACATAAAGACAAATCA------AGAATTGTTATGCTTGTTCAACCACATACAA
MIF1_HAECO ATGCCGGTTTTCTCATTTCACACG------AACGTTGCCGCCTCCAAGGTTACACCTGATCTG---CTGAAACAGATCTCTGCACTTGTTGCTCGTATTTTGCACAAAAAAAAAAGT------TATGTTTGCGTGCACGTCGTGCCGGACCAGC
MIF1_ANCCA ????????????????????????------???????????????????????????????????????????????????????????????????????????????????????????------???????????????????????????????
MIF2_ANCCA ????????????????????????------????????????CAGAACAAAGTCACACCGGAC---CTGTTGAAGCAGATCTCCGAACTCGTCGCTCGTATTCTGCACAAACCCGAG------AGTTATGTTGCTGTCCATGTGGTTCCCGATC
MIF1_AMBAM ATGCCAACCCTTACAATTAACACG------AACATCCCCGCAAGCAAGATTCCAAATGACTTC---CTGAAGACTACTGCGAACGTCGTGGCTGACTCTCTGGGAAAGCCGCTTTCG------TATGTTGTGGTCCACATCAACGCCGATCAGC
```

```
              5        15        25        35        45        55        65        75        85        95       105       115       125       135       145     1
MIF1_GIAINATGCCTTGCGCNATTGTCACCACT------AATGCTGACTTCACCAAGGATCAGGCCGACGCGTTCTGCCTAGATATGGGCCAGGTTCTGGCTAAAGAAACCGGAAAGCCTGTGAGC------TATTGCATGGCGGGGGTTCGTAAAGCA---G
LS1_EMETE ATGCCACTGTGCCAGATCGTGTGC------AACACCCAAGTGGAGAGCGGCGCGGAGGAGGCCTTCCTCGCTGCTGTGGAATCAGGTCTCAGCAAAATCCTGGGCAAGCCCACCCAA------TACATCACAGTAACCCTCACTCGGGGCTCCG
LS1_PLAFA ATGCCTTGCTGTGAAGTAATAACA------AACGTAAACCTCCCTGATGATAATGTACAAAGTACTTTATCTCAAATAGAAAATGCAATTTCTGATGTTATGGGTAAACCACTTGGT------TATATTATGAGTAATTATGATTATCAAAAAA
LS1_PLAYO ATGCCTTGCTGCGAATTAATAACA------AACATTTCTATCCCTGACGATAAAGCTCAAAATGCGTTATCCGAAATAGAAGATGCTATATCTAACGTTTTAGGAAAACCAGTAGCA------TATATTATGAGCAACTATGATTATCAAAAAA
LS1_PLABU ATGCCGTGCTGTGAATTAATAACA------AACATTTCTATCCCTGACGATAAAGCTCAAAATACACTATCCGAAATAGAAGATGCTATATCTAATATTTTAGGGAAGCCAGTTGCA------TATATTATGAGCAACTATGATTATCAAAAAA
LS1_PHYSO ATGCCGAACGTGCAGGTGACGAGC------AACGTGCCGTCGAGCGGCGTGGACAAGGCCAAGGCCATGGCCGCCATCTCCAAGGGCGTGGCCACGGCGCTGGGCAAGTCGGAGCAG------GTGGTGATGGTGCACCTGAACCTGGACACGC
LS1_TOXGO ATGCCCAAGTGCATGATCTTTTGC------CCCGTCGCGGCGACCCGGCGCAACAGGACGCCCTCTTNAAGGACGCCGAAAAAGCCGTCGCAGAGCATCCTGGGGAAGCCTCTGAGC------TACGTCATGGTGGGATACTCGCAGACCGGGC
LS1_ARATH ATGCCGTGCCTCAACCTCTCCACC------AACGTTAACCTTGACGGCGTCGATACATCTTCCATTCTCTCGGAAGCTTCCTCCACCGTCGCGAAAATCATCGGCAAGCCTGAGAAC------TATGTGATGATTGTCTTGAAAGGCTCAGTGC
LS2_ARATH ATGCCTTGTCTTTACATTACAACA------AACGTCAATTTTGACGGCGTTAACACCGATCCGTTCTACTCCGAAGTCACCAAAGCCGTCGCTTCCATCGTCGGACGACCTCAAAAC------TTAGTGATGGTGGTGTTGAAGGGATCAGTAG
LS3_ARATH ATGCCCACTTTGAATCTCTTCACT------AACATACCAGTCGACGCCGTCACTTGCTCAGACATCCTCAAGGACGCCACTAAGGCCGTCGCTAAAATCATCGGCAAACCTGAATCC------TATGTGATGATACTGCTTAACAGTGGAGTGC
LS1_GOSAR ATGCCTTGCCTAAACCTTTCAACC------AACGTCAATCTCGACGGCGTCGACACCTCCGCCATCCTTTCTGAAGCCACCTCTTCCGTCGCTAAACTCATCGGCAAACCTGAGGCC------TATGTGATGATTGTGTTGAAGGGTTCAGTAC
LS2_GOSAR ATGCCTTGCCTAAACCTTTCAACC------AACGTCAATCTCGACGGGGTCGACACCTCCGCCATCCTTTCTGAAGCCACCTCTTCCGGCGCTAAACTCATCGGCAAACCTGGGGCC------TATGTGATGATTGGGTTGAAGGGTTCAATAC
LS1_ZEAMA ATGCCGTGCCTGAACGTGTCGACC------AACGTGAACCTGGAGGGGGTGGACACCTCCGCCATCCTCGCCGAAGCCTCCAAGTCCGTCGCCAACATCATCGGCAAGCCCGAGGCC------TACGTGATGGTTGTTCTCAAGGGTTCGGTGC
LS1_GLYCL ATGCCCACTTTGAATCTCTTCACA------AACGTTCCTGTCGACACCGTCGTTGCTTCTGACATTCTCAGAGATGCCACCAAAGCTGTTGCAAAAATCATCGGAAAACCCGAATCC------TATGTGATGATTTTGTTGAATGGGGGAGTGC
LS1_CRYJA ATGCCTTCTCTTAGCATTTCAACA------AATGTACCCCTGGATGGAGTCAACACATCGGGGATACTTTCACAAGCAAGCAAATCTGTTGCCCAGATAATCGGCAAGCCAGAAGCT------TATGTGATGGTGCAGCTGAAAGGATCGGTTG
LS1_HORVU ATGCCTTGCCTGAACGTGTCGACG------AACGTGAACCTGGAGGGGGTGGATACCTCCGCCGTCCTCGCCGACGCCTCCAGCACCGTCGCAACCATCATCGGCAAGCCGGAGGGC------TATGTGATGGTTGTTCTCAAGGGTTCAGTGC
LS1_LOTJA ATGCCGTGCCTTAACCTCTCCACC------AACGTCAACCTCGACGGCGTCGACACCTCTTCCATCCTCTCCGAAGCCACCTCCACCGTCGCCACCCTTATCGGAAAACCTGAGGCC------TATGTGATGATTGTCCTGAAAGGATCAGTAC
LS1_HELAN ATGCCGTGCCTGACTCTGTCAACC------AACGTCAACCTCGACGGACTTGACTCCTCCTCCATCCTTTCCGACGCCACCTCCGCCGTCGCCAAAATCATCGGAAAACCCGAAGCG------TATGTGATGATTGTGTTGAAGGGTTCTATAC
LS1_TRIAE ATGCCTTGCCTGAACGTGTCGACG------AACGTGAACCTGGAGGGGGTGGACACCTCCGCCGTCCTCGCCGACGCCTCCAGCACCGTCGCAACCATCATCGGCAAGCCGGAGAAC------TATGTGATGGTTGTTCTCAAGGGTTCAGTGC
LS2_TRIAE ATGCCTTGCCTGAACGTGTCGACG------AACGTGAACCTGGAGGGGGTGGACACCTCCGCCGTCCTCGCCGACGCCTCCAGCACCGTCGCAACCATCATCGGCAAGCCGGAGGCC------TATGTGATGGTTGTTCTCAAGGGTTCAGTGC
LS1_LYCES ATGCCGTGCCTTAACATTTCTACA------AATGTGAACTTGGAAGGAGTTGATACTTCCTCCGTACTTTCTGAAGCTACCTCCACTGTTGCCAAACTCATCGGCAAACCAGAAGCT------TATGTCATGATTGTGTTGAAGGGATCTGTTC
LS1_GLYMA ATGCCGTGCCTCAACCTCAGCACC------AACGTGTCCCTCGAAGGCGTCGACACCTCTTCCATCCTCGCCGAAGCCACCTCCTCCGTCGCCAACATCATCGGCAAACCCGAAGCC------TATGTGATGATTGTACTGAAAGGATCAGTAC
LS2_GLYMA ATGCCGTGCCTCAACCTCTCCACC------AACGTTAACCTAGACGGCGTCGACACCTCCTCCATTCTCTCCGAAGCCACCTCAACCGTCGCCAGCATCATCGGCAAACCCGAGGCC------TATGTGATGATTGTATTGAAAGGATCAGTGC
LS3_GLYMA ATGCCGTGCCTCAACCTCAACACC------AACGTGTCCCTCGACGGCGTCGACACCTCTTCCATCCTCGCCGAAGCCACCTCCTCCGTTGCCAACATCATCGGCAAACCCGCGGCC------TATGTGATGATTGTACTGAAAGGATCAGTAC
LS1_MESCR ATGCCGTGCCTGAACATTTCCACC------AACGTCAGCCTCGACGGCGTTGACACCTCCGCCATTCTTTCCGAGGCCACCTCCTCCGTCGCCAATATCATCGGCAAACCTGAGGCT------TATGTTATGGTTGTATTGAAGGGATCAGTGC
LS2_MESCR ATGCCGTGCCTGAACATTTCCACC------AACGTCAGCCTCGACGGCGTTGACACCTCCGCCATTCTTTCCGAGGCCACCTCCTCCGTCGCCAATATCATCGGCAAACCTGAGGCT------TATGTTATGGTTGTATTGAAGGGATCAGTGC
LS1_MEDTR ATGCCGTGCCTCAACCTCTCCACC------AACGTCAACCTCGAAGGTGTCGACACCTGTTCCATCCTCTCCGAAGCCACCTCCACCGTTGCTACTCTCATCGGCAAACCCGAATCC------TATGTGATGATTGTGCTCAAAGGATCAGTAC
LS1_PINTA ATGCCTTCCCTTAGTATTTCAACA------AATGTGTCTTTGGATGGGTTTAACACCTCCGAAATACTTTCAGAGACAAGCAAAAACGTTGCTAAGATCATTGGCAAGCCAGAAGCT------TATGTGATGGTGCAGCTGAAAGGGTCGGTAG
LS1_SOLTU ATGCCGTGCCTTAACATTTCTACA------AACGTGAACTTGGAAGGAGTTGATACTTCCTCCGTACTTTCTGAAGCTACCTCCACTGTTGCCAAACTCATCGGCAAACCAGAAGCT------TATGTCATGATTGTGTTGAAGGGATCTGTTC
LS1_SORBI ATGCCGTGCCTCAACGTGTCGACC------AACGTGAACCTGGAGGGGGTGGACACCTCCGTCATCCTCGCCGAAGCCTCCAAGTCCGTCGCTAACATCATCGGCAAGCCAGAGGCC------TACGTGATGGTTGTCCTCAAGGGTTCAGTGC
LS1_R0BPS ATGCCTTGCCTCACCATTTCAACC------AACGTGAGCCTCGACGGCGTCGACACTTCTACCATACTTTCCGAGGCTACATCCACTGTTGCCAAGCTCATCGGCAAACCAGAGGCC------TATGTCATGATTGTACTGAAGGGATCTGTAC
LS1_ORYSA ATGCCTTGCCTCAACGTGTCCACC------AACGTGAACCTCGACGGAGTGGACACCTCCGCCGTCCTCGCCGACGCATCCAAGACCGTCGCCACCATCATCGGCAAGCCCGAGGCC------TATGTGATGGTTGTTCTCAAGGGTTCAGTGC
```

```
           55        165       175       185       195       205       215       225       235       245       255       265       275       285       295       305
CHMI_ESCCO AGATGGCCGACGGGCAGCATGATTAT------GCCTCCGTGCATATGACGTTGAAA---ATCGGCGCAGGTCGCAGCCTGGAAAGCCGCCAGCAGGCGGGTGAAATGCTGTTTGAACTGATTAAAACGCACTTCGCCGCCCTGATGGAGAGCCG
DDT1_HOMSA CCATGGCGCTGAGCGGGTCCACCGAG------CCCTGCGCGCAGCTGTCCATCTCCTCCATCGGCGTAGTGGGCACCGCCGAGGACAACCGCAGCCACAGCGCCCACTTCTTTGAGTTTCTCACCAAGGAGCTAGCCCTGGGCCAG---GACCG
DDT1_RATRA CCTTGTTGATGAACAAATCCACAGAG------CCCTGCGCCCACCTCCTGATCTCTTCCATCGGTGTTGTGGGCACCGCGGAGCAGAACCGCAGCCACAGCTCCAGCTTCTTCAAGTTCCTCACCGAGGAGCTGTCCCTGGACCAG---GACAG
DDT1_MUSMU CCCTGTTGATGAACAAATCCACAGAG------CCTTGTGCTCACCTTCTGGTCTCTTCCATCGGGGTTGTGGGCACCGCGGAGCAGAACCGCACTCACAGCGCCAGCTTCTTCAAGTTCCTCACCGAGGAGCTGTCCCTGGACCAG---GACCG
MIF1_ORYLA CGATGCTGATGGCTGGCTCTTGCTCG------CCGTGCGTCATTCTATCGGTGGCGGCCATTTCTGTGACGGACTCTGCTGACAAGAACAAACAGCACAGTGCCAAGATCTTTGAGTTCCTAACTAAGGAGCTCAGTATGACTGAA---GACAG
MIF1_MUSMU TCATGACTTTTAGCGGCACGAACGAT------CCCTGCGCCCTCTGCAGCCTGCACAGCATCGGCAAGATCGGTGGTGCC---CAGAACCGCAACTACAGTAAGCTGCTGTGTGGCCTGCTGTCCGATCGCCTGCACATCAGCCCG---GACCG
MIF1_RATRA TCATGACTTTTAGTGGCACGAGCGAC------CCCTGCGCCCTCTGCAGCCTGCACAGCATCGGCAAGATCGGTGGCGCC---CAGAACCGCAACTACAGCAAGCTGCTGTGCGGCCTGCTGTCCGATCGCCTGCACATCAGCCCG---GACCG
MIF1_HOMSA TCATGGCCTTCGGCGGCTCCAGCGAG------CCGTGCGCGCTCTGCAGCCTGCACAGCATCGGCAAGATCGGCGGCGCG---CAGAACCGCTCCTACAGCAAGCTGCTGTGCGGCCTGCTGGCCGAGCGCCTGCGCATCAGCCCG---GACAG
MIF1_MERUN TCATGACTTTCAGCGGCTCGAGCGAC------CCCTGTGCCCTCTGCAGCCTGCACAGCATCGGCAAGATCGGCGGCGCC---CAGAACCGCACCTACAGCAAGCTGCTCTGCGGCCTGCTTGCCGATCGCCTGCGCATCAGCCCG---GACCG
MIF1_SUSSC TCATGGCCTTCGGAGGGTCCAGCGAG------CCGTGCGCCCTTTGCAGTCTGCATAGCATCGGCAAGATCGGCGGCGCG---CAGAACCGTTCCTACAGCAAGCTGCTGTGCGGCCTGCTGGCCGAACGCTTGCGCATCAGCCCG---GACAG
MIF1_BOSTA TCATGACCTTCGGGGGCTCCAGCGAG------CCCTGCGCGCTCTGCAGCCTGCACAGCATCGGCAAGATCGGCGGCGCG---CAGAACCGCTCCTACAGCAAGCTGCTGTGCGGCCTGCTGACCGAGCGCCTGCGCATTAGCCCG---GACAG
MIF1_GALGA TGATGTCCTTCGGGGGCTCCACGGAT------CCTTGCGCTCTCTGCAGCCTCTACAGCATTGGCAAGATTGGAGGGCAG---CAGAACAAGACCTACACCAAGCTCCTGTGCGATATGATTGCGAAGCACTTGCACGTGTCTGCA---GACAG
MIF1_CIOIN GGATGTCGTTTGGTGGGACTGAGGAG------CCCTGCGCTGCCGCTGTGCTTACTTCTATCAGTGATTTCAATGCAGAA---ACATGCACTACATATGCTGAAGCTATGCTTGGTGAGATATACAAGTTACTGGGAGTTGCACAA---GACAG
MIF2_CIOIN GGATGTCATTTGGTGAAAGTGAAGAG------CCGTGTGCTTGCTGTACAGTCACTTCTATTGTAGACTTCAACGCAGAA---ACCTGCGAAAAGCTTGCAGCGCTTTTGATGCCTCCTCTTCAAAAAGCGTTAGGAGTTTCTGGA---ACCCG
MIF1_XENLA TTATGAGTTTTGGTGATTCAACTGAT------CCTTGTGCCGTGTGCAGTCTGTGCAGCATTGGAAAGATAGGGGGGCCG---CAGAACAAAAGCTACACTAAGCTGCTCTGTGACATTCTGACTAAACAGCTGAACATCCCTGCT---AACAG
MIF1_DANRE TGATGATGTTCGGGGGAAAAGGAGAC------CCGTGTGCGCTCTGCTCGCTCACCAGCATCGGGAAGATCAGCGGCGCG---CAGAATAAACAATACTCCAAACTCCTCATGGGTTTACTCAACAAACACCTCGGCGTTTCTGCT---GACAG
MIF1_PSEAM TGATGATGTTCGGAGGAAAGGGAGAC------CCCTGTGCGCTCTGCTCCCTTCACAGCATCGGAAAGATCAGCGGCGCT---CAGAACAAGAAGTACTCTAAACTTCTGTGTGGTCTCCTCAACAAACACCTGGGCATCTCTCCT---GACAG
MIF1_CAEEL GAATGACGCGAGGACAATTGACTGAT------CCCTTAGCTGTTTTAGATGTCACATCATCCACAGTTCTTACTCCGATT---CTCACCGAAGAATACACGGTTGCTCTCTGTGAATTCTTCTCGCAGGAGCTGGCTCTAGATTCT---GACGC
MIF2_CAEEL GTCTAGTTCATGGGGCCACTCATGAT------CCTGTGACCGTAATTTCGATCAAATCAATCGGAGCAGTGTCTGCTGAA---GACAATATTCGCAATACGGCAGCCATCACTGAGTTTTGTGGAAAGGAATTGGGTCTTCCAAAA---GATAA
MIF3_CAEEL GAATCCTGTATGCTGGGACCACTGAG------CCAGCAGGCTTTGCCGTGTTGAAATCGATCGGCGGAGTTGGAAGCGCTAAGCAGAATAATGCGATTTCCGCCGTCGTTTTCCCGATTATTGAGAAACATCTTGGGATTCCAGGA---AATCG
MIF4_CAEEL GAATGAGAATTGGATTTGAGAATAAAGAAATTCCAGTTGCAATTGTAAATTTCCAAACAACCCGGCCTTCTTCTCGAATC---GAAAATGACTCATATGCAAAAAAATTGACATCAGTGCTCAATGAGCAGTTGAAGCTTGATCCC---GCCCA
MIF1_BRUMA CGATGGTGTTCGGAGGAAGTGAGGAC------CCCTGTCCTGTGTGTGTTTTAAAATCGATTGGTTGTGTTGGTCCTAAA---GTCAATAATTCGCACGCAGAGAAATTGTACAAATTGCTCGCTGATGAGCTGAAAATTCCGAAG---AATCG
MIF1_BRUPA CGATGGTGTTTGGAGGAAGTGAGGAT------CCCTGTGCTGTATGTGTCTTGAAATCGATTGGTTGTGTTGGTCCTAAA---GTCAATAATTCGCATGCAGAGAAATTGTACAAATTGCTTGCTGATGAGCTGAAAATTCCGAAG---AATCG
MIF1_WUCBA CGATGGTGTTCGGAGGAAGTGAGGAC------CCCTGTCCTGTGTGTGTTTTAAAATCGATTGGTTGTGTTGGTCCTAAA---GTCAATAATTCGCACGCAGAGAAATTGTACAAATTGCTCGCTGATGAGCTGAAAATTCCGAAG---AATCG
MIF1_ONCVU CGATGGTATTCGGTGGAAGTACTGAT------CCATGTGCTGTGTGTGTTTTAAAATCAATCGGTTGTGTTGGTCCTAAT---GTTAATAATTCGCACTCTGAAAAATTGTTCAAATTACTCGCTGATGAATTGAAAATTCCAAAA---AATCG
MIF1_ASCSU AAATGACCTTCGGAGGGAGCGCCGAT------CCATGCGCCGTATGCACCCTGGAATCGATTGGTGCTGTAGGAGGCAGC---CGCAACAACGCTCATGCCGAGAAGCTGTACAATCATCTCAACGAAACGCTTGGCATCCCCAAG---AACAG
MIF1_TRITR AGATAACGTTTGGTGGTACCGACGCC------CCAGCTGGCTTCGGGCAGTTGTTGTCGCTTGGCGGAGTTGGTGGCGAA---AAGAACCGTAGTCACTCCGCCAAACTATTCAAGCATTTAACTGACGGTCTCGGCATTCCTGGC---AACCG
MIF1_TRISP AGTTGTCGTTTGGCGGAAGTACAAAA------CCTGCTGCATTCGGTACTCTGATGTCGATTGGTGGAATAGAACCAAGC---AGAAATCGTGATCATTCGGCCAAACTGTTTGATCATCTTAACAAAAAATTGGGCATTCCAAAG---AATAG
MIF1_TRIPS AGTTAACATTTGGCGGAAATACAAGT------CCTGCTGCATTCGGTTCTTTGATGTCGATTGGTGGAATTGAAGCAAGC---AGAAATCGTGATCATTCGACAAAACTGTTTGATCATATCAACAAAAAATTGGGCATTCCAAAG---AATAG
MIF2_TRISP AAATGGAATTTGGCGGCGCTGATGAT------CCTTGTCTGATTGGCGTAGTTCGAGCGGTTGGAAGAATCAGTGCAGAA---GAAAATGCACAATATGCCGAAAGATTGAGTGAATTTCTACATCAGCAATTAGGCATTCTTCCA---CAACG
MIF2_BRUMA AATTATCACATGGCACCACTGAAAAT------CCATCATGTTTTACAGTGGTTAAATCAATTGGATCATTTTCGGCTGAT---AAAAATATCGAATATTCCTCATCAATATCAGAGTTCATGAAGAAAACATTGGATATCGATCCT---GCGCA
MIF2_ONCVU AGCTATCGCGTGGTGCCACACAAGAT------CCAACATGTCTCATTGTGATTAAATCGATCGGATCATTTTCGGCTGAT---AAAAACATCAAATACTCCGGGTCAATATCGGAATTCATCAAGAAAACATTGAATATTGATCCA---GCATA
MIF2_HETGL ACTGTTGCCAAATCGGAAGTGACCCTGCCGAACCAGCGATGATTCTGCAAATCAAATCAATTGGCTGTGTCAGCGCAGAT---GAAAATGTAATTCACTGTAAGAACATCAGCGAATTCGTCCAAAGTCGCCTTGGGATTCCGCCC---GACCG
MIF2_MELJA TTGGCTGTTTTGGAGGTTCTACCGAT------CCATTTATTTATGCCGAACTTCAATCAATTGGGGGATTTACTGATCCAAACAAAGTAACTGGAGAAATGACTAAACTTTTTACTGAACATTTTGGTGTTCCTGGCAGTCGAGTT---TATAT
MIF2_STRST TGATGTCATCTGGAGGTGTTCCTAAT---CAACCAAGTATCTGGATCGAAATAAATAATGTTGGACAACTTTCACCAAGA---CAAACTCAAGAATTATCACGAGATTTAACACATTTTGTCATGGAACAAACTACAGTACCAAGA---GAAAG
MIF1_HAECO ATATGATTTTCGCCGGGACCGACGAA------CCGTGTGGTGTTGGTGTGCTAAAATCGATCGGAGGTGTCGGTGGGAGT---AAGAATAATGAGCATGCCAAGGCCTTATTTGCTCTAATCAAGGATCACCTCGGCATCTCAGGG---AACAG
MIF1_ANCCA ??????????????????????????????????????????????????GAATCAATAGGTGCGCTCTCCGCGGATGAC---AATATTCGTCATACACAGAAAATAACACAATTTTGTCAGGATACTTTGAAATTGCCAAAGGAC---AAGGT
MIF2_ANCCA AGAAAATGACATTTGCTGGAACTGAC------GCGCCTTGTGGTATTGGTATCCTGAAGTCGATTGGAGGCGTTGGTGGC---TCGCAGAACAATAGCCATGCTAAGGCTCTGTTCGCGTTGATCAAAGACCATCTGGGAATCGAA---GGGAA
MIF1_AMBAM TGTTGTCATTCGGAGGCACTGATGAC------CCATGCGCTATTGCAAATCTGTACAGCATCGGCTGTCTGAGTCCAAAG---GAGAACAAAAAGCATTCAGCTGTTCTTTTTGAACACATTGAAAAGACCCTGGGCATCAAGGAA---AACAG
```

```
           55        165       175       185       195       205       215       225       235       245       255       265       275       285       295       305
MIF1_GIAIN ATATGTCTTTTGGAACATCTACTGAT------CTCTGCTGTTTTGTCGACTTCTACTGCATAGGTGTAATCTCCCAAGCC---AAGAATCCTTCTATTTCTGCCGCCATCACGGGCTGCCTTACCCAGCACTTCAAGGTGAAGCCA---GAGAG
LS1_EMETE  TCCGCCAC---AGTGGCAGCTGCGAT------CCTGCAGCTAGTGTTTCTGTGCATTCAATTGGGGGCATCAGCAGCCGC---ACCAACAATATGATTTGCGCAGAGGTCGCGGCTCTGTGCCAGCAACACCTGAAGGTGCCCGTA---GATCG
LS1_PLAFA  ACTTAAGATTTGGAGGTAGCAACGAA------GCTTATTGTTTTGTAAGAATAACAAGTATTGGAGGAATTAATAGGTCA---AATAATTCTGCTCTT---GCTGATCAAATAACGAAACTCCTTGTAAGCAACTTGAATGTAAAA---TCTAG
LS1_PLAYO  ACCTCCGATTTTCTGGAAGTAATGAA------GGATATTGTTTTGTTAGATTAACAAGTATTGGTGGAATAAATAGATCA---AATAATTCTTCATTG---GCAGATAAAATTACAAAAATTCTTTCCAATCATTTAGGTGTTAAA---CCAAG
LS1_PLABU  ACCTCCGATTTTCTGGAAGTAATGAA------GGATATTGCTTTGTTAGATTAACTAGTATTGGTGGAATAAATAGATCA---AATAATTCTTTACTA---GCAGATAAAATTACAAAAATCCTTTCCAATCATTTAAGTGTTAAA---CCAAG
LS1_PHYSO  CCATGCTCTTCCAGGCCTCGGACGCC------CCGTGCGCCATGATTCAGCTCAAGAGCATCGGCAAGGTGGACGCCCAG---CACAACCCGACGACG---GCGTCGATCCTGACGGAGACCGTGAGCCAGGAGCTGAACGTGCCC---AAGGA
LS1_TOXGO  AGATGCGTTTCGGCGGGACAGCGACC------CGTGTGCGTTCATTCGCGTTGCTTCCATTGGAGGCATCACCAGTTCCA---GAACTGCAAAATCGCCGCTGCTCTCTCCGCTGCATGCGAACNCACCTGGGCGTCCCCAAGAAC---CGCAT
LS1_ARATH  CTATGTCATTTGGCGGGACCGAGGAT------CCTGCAGCTTATGGTGAATTAGTTTCTATCGGTGGCCTTAATGCGGAT---GTGAACAAGAAGCTAAGCGCTGCTGTTTCCGCCATTCTTGAGACTAAGCTATCGGTGCCCAAG---TCTCG
LS2_ARATH  AGATAGTATTTGGAGGGAACAAAGAA------GCAGCTGCATATGCAGAGATTGTGTCAATGGGAGGCATCACCAAACAA---GTTAAGAGAGAACTCATAGCGACCGTTGGTTCTATTCTTCACACTCATTTTTCTATTCATCCC---ACTCG
LS3_ARATH  CCATTGCATTTGCCGGTACCGAGGAA------CCTGCTGCATATGGAGAATTGATATCTATTGGGGGATTAGGACCTGGC---GTAAACGGGAAGCTTAGCGAGACGATATCTGAGATTCTCCAAATTAAGCTCTCCATAGACAGC---TCTCG
LS1_GOSAR  CCATGTCATTTGGTGGTACTGAGCAG------CCAGCTGCCTATGGAGAACTGGTATCAATTGGGGGCCTTAATCCAGAT---ATGAACAAGAAACTGAGTGCTGCAATTGCTGCAATTCTTGAAACCAAGCTACAAGTGCCTAAG---TCACG
LS2_GOSAR  CCATGTCATTTGGTGGGACTGAGCAG------CCAGCTTGCTATGGAGAACTGGTATCAATTGGGGGCCTTAATCCAGAA---TATGAACAAGAAACTGAGTGCTTGCAATTGCTGGAATTCTTGAAACCAAGCTACAAGTGCCTA---AGTCA
LS1_ZEAMA  CTATGGCATTTGGAGGTACCCAGGAG------CCAGCAGCTTACGGTGAGCTGGTTTCCATCGGAGGCCTGAACCCTGAT---GTGAACAAGAAGCTTAGTGCTGGCATCGCTTCTATCCTGGAGTCAAAGCTGTCTGTTCCCAAG---TCACG
LS1_GLYCL  CTATTGCATTTGCTGGAACTGAAGAG------CCAGCTGCTTATGGAGAACTGATCTCGATTGGGGGCCTTGGTCCGAGT---GTCAATGGAAAACTGAGTTCTACAATTGCAGAAATTCTGCAAACTAAGCTATATATTGACAGT---TCCCG
LS1_CRYJA  CAATTTCGTTTGGGGGCACTGAGGAC------CCAGCAGCCTATGGGGAATTGGTGTCCATTGGAGGGCTCAGCGCTGAC---ACAAACAAAAAGCTCAGCGCAGCCATTTCCAGCATTCTCGAATCCACGCTCTCTGTACCAAAG---TCCCG
LS1_HORVU  CCATGGCATTCGGAGGTACCCAGGAG------CCTGCAGCCTATGGCGAGCTTGTTTCCATCGGAGGGCTGAACCCTGAT---GTGAACAAGAAGCTGAGCGCTGGCATTTCTTCCATCCTGGAGTCAAAGCTGTCCATCTCCAAG---TCCCG
LS1_LOTJA  CCGTGTCTTTTGGTGGAACTGAGCAA------CCAGCAGCTTATGGAGAATTGGTGTCCATCGGTGGTCTTAACCCTGAT---GTGAATAAGAAACTTAGTGCTGCAATTGCTTCAATTCTTGAAACCAAGTTGTCCGTGCCCAAG---TCACG
LS1_HELAN  CAATTTCTTTTGGTGGTAACGAACAG------CCAGCAGCCTACGGGGAATTAGTCTCAATCGGGGGGCTTAATGCTGAT---GTAAACAAGAAGTTGAGTGCTGCCGTTGCAGAAATTCTTGAGACCAAATTGTCCGTCCCTAAG---TCGCG
LS1_TRIAE  CCATGGCGTTCGGAGGTACCCAGGAG------CCTGCAGCCTACGGTGAGCTGGTTTCCATCGGAGGGCTGAACCCTGAT---GTGAACAAGAAGCTGAGTGCCGGCATTGCTTCCATCCTGGAGTCAAAGCTGTCCATCCCCAAG---TCCCG
LS2_TRIAE  CCATGGCATTTGGAGGTACCCAGGAG------CCTGCAGCCTACGGCGAGCTGGTTTCCATCGGAGGGCTGAACGCTGAT---GTGAACAAGAAGCTGAGCGCCGGCATTGCTTCCATCCTGGAGTCAAAGCTGTCCATCCCCAAG---TCCCG
LS1_LYCES  CCATGGCTTTCGGGGGTACAGAACAG------CCCGCTGCCTATGGAGAGTTGGTCTCCATTGGGGGTTTAAATGCTGAT---GTCAACAAGAAACTCAGTGCTGCAATAGCTGACATACTTGAGACTAAATTGTCTATTCCCAAA---TCTCG
LS1_GLYMA  CCATAGCTTTTGGTGGGAATGAGCAG------CCAGCAGCTTATGGAGAATTGGTTTCCATTGGTGGTCTTAACCCTAGT---GTAAACAAGGAACTTAGTGCTGCAATTGCTTCAATTCTGGAAACCAAATTGTCAGTGCCGAAG---TCGCG
LS2_GLYMA  CTATATCTCATGGTGGGAGTGAGCAG------CCAGCAGCTTATGGTGAATTGGTGTCCATTGGTGGTCTTAGCCCTGAC---GTGAACAAGAAACTTAGTGCTGGCATTGCTTCAATTCTCGAAAACAAGTTGTCTGTGCCAAAG---TCGCG
LS3_GLYMA  CCATAGCTTTTGGTGGGAATGAACAG------CCAGCAGCTTATGGAGAATTGGTGTCCATTGGTGGTCTTAACCCTAGT---GTAAACAAGGAACTTAGTGCTGCAATTGCTTCAATTCTGGAAACCAAATTGTCGGTGCCCAAG---TCACG
LS1_MESCR  CAATGGCATTTGGCGGAACTGAGCAG------CCTGCTGCATATGGTGAGCTGGTGTCCATTGGAGGTCTTAACCCGGAC---ACAAACAAGAAATTGAGTGCTGCCATCGCAGCAATTCTGGAGAGCAAATTGTCTGTTCCCAAA---TCACG
LS2_MESCR  CAATGGCATTTGGCGGAACTGAGCAG------CCTGCTGCATATGGCGAGCTGGTGTCCATTGGAGGTCTTAACCCGGAC---ACAAACAAGAAATTGAGTGCTGCCATCGCAGCAATTCTGGAGAGCACATCGCCTGTTCCCAAA---TCACG
LS1_MEDTR  CCATATCTTTTGGTGGGACTGAGCAG------GAAGCAGCTTATGGAGAATTGGTGTCCATCGGTGGTCTTAACCCTGAT---GTGAACAAGAAACTTAGTGCTGCAATTGCTGCAATTCTTGAAACCAAGTTGTCTGTTCCCAAA---ACTCG
LS1_PINTA  CAATTTCATTTGGGGGTACGGAGGAA------CCAGCAGCTTATGGGGAATTGGTGTCCATTGGAGGGTTGGGATCTGAT---ACCAACAAGAAGCTCAGCTCCGCCATTGCCAACGTACTGGAGACCAAGCTCGGTGTATCCAAG---TCACG
LS1_SOLTU  CCATGGCTTTCGGGGGTACAGAACAG------CCCGCTGCCTATGGAGAGTTGGTCTCCATTGGGGGATTAAATGCTGAT---GTCAATAAGAAACTCAGTGCTGCAATAGCTGACATACTTGAGACTAAATTGTCTATCCCCAAA---TCTCG
LS1_SORBI  CTATGGCATTTGGAGGTACCCAGGAG------CCAGCAGCGTACGGTGAGCTGGTTTCCATCGGAGGCCTGAACCCTGAC---GTGAATAAGAAGCTGAGTGCTGGCATTGCTTCCATCCTGGAGTCAAAGCTGTCTGTTCCCAAG---TCACG
LS1_R0BPS  CCATGGCATTTGGTGGGACTGAGCAG------CCTGCAGCATATGGTGAGCTGGTATCCATTGGTGGCCTGAATCCTAAT---GTGAACAAGAAACTTAGTGCTGCAATAGCTGAGATTCTCGAGACTAAGTTGTCCGTGCCCAAG---TCACG
LS1_ORYSA  CTATGGCATTTGGAGGTACCCAGGAG------CCTGCCGCTTATGGCGAGCTGGTTTCCATTGGTGGGCTGAACCCTGAT---GTCAACAAGAAGTTGAGTGCTGGCATTGCCTCTATCCTGGAGTCAAAGCTATCCATTCCCAAG---GGCCG
```

```
                315       325       335       345       355       365       375       385       395       405       415       425       435       445       455
CHMI_ESCCO CCTGCTGGCGTTGTCGTTTGAGATTGAAGAGCTGCATCCGACGCTGAATTTTAAACAAAACAACGTGCACGCATTGTTTAAG????????????????????????????????????????????????????????????????????????
DDT1_HOMSA GATACTTATCCGCTTTTTCCCCTTGGAGTCCTGGCAG------ATTGGCAAGATA---GGGACGGTCATGACTTTTTA????????????????????????????????????????????????????????????????????????????
DDT1_RATRA GATCATTATCCGATTCTTCCCCTTGGAGCCCTGGCAG------ATCGGAAAGAAA---GGAACTGTTATGACGTTTCTG???????????????????????????????????????????????????????????????????????????
DDT1_MUSMU GATCGTTATCCGCTTCTTCCCCTTGGAGGCTTGGCAG------ATCGGAAAGAAA---GGAACTGTCATGACATTTCTG???????????????????????????????????????????????????????????????????????????
MIF1_ORYLA GATTCTTATCAAGTTTGATGAGTTGCAGCCTCACCAA------GTCGGGAAGAAG---GGAACAGTTATGAGCTTCCTG???????????????????????????????????????????????????????????????????????????
MIF1_MUSMU GGTCTACATCAACTATTACGACATGAACGCTGCCAAC------GTGGGCTGGAAC---GGTTCCACCTTCGCT?????????????????????????????????????????????????????????????????????????????????
MIF1_RATRA GGTCTACATCAACTATTACGACATGAACGCAGCCAAC------GTGGGCTGGAAC---GGTTCCACCTTCGCT?????????????????????????????????????????????????????????????????????????????????
MIF1_HOMSA GGTCTACATCAACTATTACGACATGAACGCGGCCAAT------GTGGGCTGGAAC---AACTCCACCTTCGCC?????????????????????????????????????????????????????????????????????????????????
MIF1_MERUN GATCTACATCAACTATTACGACATGAACGCAGCCAAC------GTGGGCTGGAAC---GGTTCCACCTTCGCT?????????????????????????????????????????????????????????????????????????????????
MIF1_SUSSC GATCTACATCAACTACTACGACATGAACGCGGCCAAT------GTGGGCTGGAAC---GGC?????????????????????????????????????????????????????????????????????????????????????????????
MIF1_BOSTA GATCTACATCAACTTCTGCGACATGAACGCGGCCAAC------GTGGGCTGGAAC---????????????????????????????????????????????????????????????????????????????????????????????????
MIF1_GALGA GGTATACATCAATTACTTCGACATAAATGCTGCCAAC------GTGGGCTGGAAC---GGTTCCACCTTTGCA?????????????????????????????????????????????????????????????????????????????????
MIF1_CIOIN AATGTATCTTGAATTCCATGAAGCAACAAGGGAAACC------ATGGGGTATAAC---GGCACAACTTTCCACCAACTCGCAGCCAAGAAA???????????????????????????????????????????????????????????????
MIF2_CIOIN TTTTTACCTACAGTTTCACGAAATAACTGCAGGAATA------ATGGGCTTCCAG---GGTACAACTGTTAAAGTAGTTCGTGAAAGAAAGCAGAGTTCT??????????????????????????????????????????????????????
MIF1_XENLA AGTCTACATAAATTATTATGACCTTAATGCTGCAAAC------GTTGGCTGGAAT---GGATCTACCTTTGCC?????????????????????????????????????????????????????????????????????????????????
MIF1_DANRE GATCTATATAAACTTTGTTGACATGGATCCAGCCAAT------GTGGCCTGGAAC---AACAGCACCTTTGGA?????????????????????????????????????????????????????????????????????????????????
MIF1_PSEAM GATTTATATTAACTTTTTTGACATGGATGCAGCCAAC------GTAGCCTGGAAC---AACAGTACCTTTGCC?????????????????????????????????????????????????????????????????????????????????
MIF1_CAEEL AGTCCTCATTAATTATCGATCGCTAAGCCCGGAACTT------ATCGGATTCAAT---GGACATATTCTTACCGAAAATCGACCATTCATCTCTACAGATCGTGCTCGCTTCATTATTGGAGTTCTAGGCATTGCGTTTCTTGCATTCCTCCTT
MIF2_CAEEL GGTTGTTATCACATTCCACGACTTACCACCAGCTACG------GTAGGGTTCAAC---GGTACAACTGTCGCCGAAGCAAATAAGAAA??????????????????????????????????????????????????????????????????
MIF3_CAEEL CCTCTACATCGAATTCGTCAACCTCGGCGCCGCCGAC------ATCGCCTACAAC---GGTCAAACCTTCGCC?????????????????????????????????????????????????????????????????????????????????
MIF4_CAEEL TATTTTCATATCATTTGATTTTAAAGATGCAAAAAGT------TTTGCAACTCAA---GGGAAGACAATTGCTTCTTTGTATGAA?????????????????????????????????????????????????????????????????????
MIF1_BRUMA ATGCTACATCGAATCTGTGGATATCGAAGCTTCTTCA------ATGGCTTTTAAT---GGATCTACTTTTGGA?????????????????????????????????????????????????????????????????????????????????
MIF1_BRUPA ATGCTACATCGAATTTGTGGATATCGAAGCGTCTTCA------ATGGCTTTTAAT---GGATCTACTTTTGGA?????????????????????????????????????????????????????????????????????????????????
MIF1_WUCBA ATGCTACATCGAATCTGTGGATATCGAAGCTTCTTCA------ATGGCTTTTAAT---GGATCTACTTTTGGA?????????????????????????????????????????????????????????????????????????????????
MIF1_ONCVU ATGCTACATCGAATTTGTGAATATCGATGCGTCTACA------ATGGCTTTTAAT---GGATCTACTTTTGGA?????????????????????????????????????????????????????????????????????????????????
MIF1_ASCSU AATGTACATCAGCTTCGTCGACATTGATCCCACTACG------ATGGCGTATAAT---GGATCGACCTTCGCT?????????????????????????????????????????????????????????????????????????????????
MIF1_TRITR CATGTACATCAACTTCGTCGACATGCGTGGCAGCGAT------GTTGGATACAAT---GGGTCGACTTTC????????????????????????????????????????????????????????????????????????????????????
MIF1_TRISP AATGTATATCCATTTCGTCAATCTGAACGGAGACGAT------GTTGGTTGGAAC---GGTACTACATTCTGA?????????????????????????????????????????????????????????????????????????????????
MIF1_TRIPS AATGTATATTCATTTCGTCAATCTGCGCGGAAACGAT------GTTGGTTGGAAC---GGTACTACATTCTGA?????????????????????????????????????????????????????????????????????????????????
MIF2_TRISP AATGTACATACGGTACTTGAATATGGACGGCTTTTAC------GTTGGATGGAGT---GGCTGTCTGCGAGCG?????????????????????????????????????????????????????????????????????????????????
MIF2_BRUMA TTGTATCATTCATTTCTTAAATTTGGACCCAGAAAAT------GTTGGATGCAAG---GGAACAACAATGAAAGTGCTTATGAAGAAA??????????????????????????????????????????????????????????????????
MIF2_ONCVU CTGTATCATTCATTTCTTGGATTTGAATCCAGAAGAT------ATTGGATGCAAT---GGTACGACAATGAAAGAGCTGATGAAGAAA??????????????????????????????????????????????????????????????????
MIF2_HETGL CGTGATGATTCACTTTCAGTCTTTGGAGAAGCACGAG------GTCGGGAAAGGA---GGAACGACGGTGGAAAAGATGTGCCAG?????????????????????????????????????????????????????????????????????
MIF2_MELJA GAAATTGACTGGT---------CCTGATGCAAATAAA------TTCGCTCATAAT---GGAAAGACTTTTGCT?????????????????????????????????????????????????????????????????????????????????
MIF2_STRST TGTTTCAATACTTTATTTTGATATGTCACCTGATATG------GTCGCAAGAGGT---GGAATAACAATTGCTGAATCAATTGCTGGTTTAAAAT???????????????????????????????????????????????????????????
MIF1_HAECO GATGTACGTTGAATTCATCGACATCGGTGCTGCCGAC------ATTGCCTTCAAT---AGCAAGACCTTTGCT?????????????????????????????????????????????????????????????????????????????????
MIF1_ANCCA GATCATTACATATTTCGATCTGCAACCAATCCACGTT------GGTTTCAACGGC---ACCACTGTGGCAGCGGCAACGATGT???????????????????????????????????????????????????????????????????????
MIF2_ANCCA CAGGATGTATATCGAGTTCGTCGACATTGGAGCTAGC------GACATTGCCCAT---AATGGTAGAACATTTGCG??????????????????????????????????????????????????????????????????????????????
MIF1_AMBAM GATGTACATCAATTACTTCGACATGCCAGCAAGTGAT------GTTGGCTACAAC---GGAAAAACTTTTGCTGGC??????????????????????????????????????????????????????????????????????????????
```



```
                315       325       335       345       355       365       375       385       395       405       415       425       435       445       455
MIF1_GIAIN GGTGTACATCAGCTTCAACGAGGCGAAAGGCCATAAT------TGGGGCTTTAAT---GGCAGCACGTTT????????????????????????????????????????????????????????????????????????????????????
LS1_EMETE  CGTGTTTTTCCACTTCGCAGATGTGAGCGCCGCGAAC------ATCGGCATT------GGTTCCCGCGTGTTTGGT??????????????????????????????????????????????????????????????????????????????
LS1_PLAFA  ACGTATCTATGTAGAATTTAGAGACTGTTCTGCTCAA------AACTTTGCCTTC---AGTGGTTCTCTTTTCGGC??????????????????????????????????????????????????????????????????????????????
LS1_PLAYO  AAGGGTTTACATAGAATTCCGAGATTGTTCAGCTCAA------AACTTTGCTTTT---AGTGGTTCACTATTTGGC??????????????????????????????????????????????????????????????????????????????
LS1_PLABU  AAGAGTTTATATAGAGTTCAGAGATTGTTCAGCTCAA------AACTTTGCTTTT---AGTGGCTCACTATTTGGT??????????????????????????????????????????????????????????????????????????????
LS1_PHYSO  CCGCATCTTCATGAACATCGACGACGTGCAGCGCTCC------AAC---TGGGCC---AAGGGCGGCGTGCTCATCCCGGAGCCCAAGCAG???????????????????????????????????????????????????????????????
LS1_TOXGO  CTACACGACATTCACAAACAAGAGCCCCTC???????------????????????---????????????????????????????????????????????????????????????????????????????????????????????????
LS1_ARATH  ATTCTTCCTCAAGTTTTATGACACCAAGGGATCCTTC------TTTGGTTGGAAC---GGGGCGACTCTT????????????????????????????????????????????????????????????????????????????????????
LS2_ARATH  TTTTATCTTTAAAGTTTTTGATATTAATTCTTTGCCT------CTTCCTTCTAAA---CTT?????????????????????????????????????????????????????????????????????????????????????????????
LS3_ARATH  CTTTTATATCAAATTCTACGATTCTCCGCGACCTTTC------TTCGGTTACAAT---GGATCAACTTTC????????????????????????????????????????????????????????????????????????????????????
LS1_GOSAR  GTTCTTCCTCAAATTCTATGACACCAAGGGTTCCAAC------TTTGGATGGAAC---GGATCCACCTTC????????????????????????????????????????????????????????????????????????????????????
LS2_GOSAR  CGGGTCTTCCTCAAATTCTATGACACCAAGGGTTCCA------ACTTTGGATGGA---ACGGATCCACCTTCTGAGCCTGGCATGTTC??????????????????????????????????????????????????????????????????
LS1_ZEAMA  CTTCTACCTCAAGTTCTATGACTCGAAGCGCTCGGAC------TTCGGTTGGAAC---GGCTCCACCTTC????????????????????????????????????????????????????????????????????????????????????
LS1_GLYCL  TTTCTATATCAAGTTTTATGATGTTCAGCGCTCATTC------TTTGGGTTCAAT---GGCTCAACCTTT????????????????????????????????????????????????????????????????????????????????????
LS1_CRYJA  ATTTTACATCAAATTCTACGATGTCAAGGGGTCCAAT------CTGGGATATAAT---GGAAGCACTTTT????????????????????????????????????????????????????????????????????????????????????
LS1_HORVU  CTTCTACCTCAAGTTCCATGATTCAAAGCGCTCGGAC------TTTGGATGGAAC---GGAACCACCTTT????????????????????????????????????????????????????????????????????????????????????
LS1_LOTJA  ATTCTTCTTAAAATTCTATGACACTAAGGGTTCTAAC------TTTGGATGGAAT---GGGTCCACAT??????????????????????????????????????????????????????????????????????????????????????
LS1_HELAN  TTTCTATCTCAAATTCTATGACACCAAGGGTTCTTTC------CTGGGGTGGAAT---GGCTCCACTTTC????????????????????????????????????????????????????????????????????????????????????
LS1_TRIAE  CTTCTACCTCAAGTTCCATGATTCAAAGCGCTCGGAC------TTTGGGTGGAAC---GGATCGACCTTT????????????????????????????????????????????????????????????????????????????????????
LS2_TRIAE  CTTCTACCTCAAGTTCCACGATTCAAAGCGCTCAGAC------TTTGGATGGAAT---GGATCCACCTTT????????????????????????????????????????????????????????????????????????????????????
LS1_LYCES  ATTTTTCCTGAAATTCTATGATACTAAGGGTTCCTTC------TTTGGCTGGAAT---GGATCTACCTTC????????????????????????????????????????????????????????????????????????????????????
LS1_GLYMA  GTTCTTCTTGAAATTCTATGACACCAAGGGTTCCAAC------TTCGGATGGAAT---GGATCTACATTC????????????????????????????????????????????????????????????????????????????????????
LS2_GLYMA  ATTCTACTTGAAGTTTTATGACACCAAGGGTTCCAAC------TTTGGATGGAAT---GGATCTACATTC????????????????????????????????????????????????????????????????????????????????????
LS3_GLYMA  GTTCTTCTTGAAATTCTATGACACCAAGGGTTCCAAC------TTTGGATGGAAT---GGATCTACATTC????????????????????????????????????????????????????????????????????????????????????
LS1_MESCR  ATTCTTCCTCAAATTCTACGACACCAAGGGTTCGTTT------TTCGGATGGAAT---GGGTCTACCTTC????????????????????????????????????????????????????????????????????????????????????
LS2_MESCR  ATTCTTCCTTAAATTCTACGACACCAAGGGCCCGTCT------TTACGATCGAAT---GGGTCTACTTCTCAGACTAAACCAGTTGTG??????????????????????????????????????????????????????????????????
LS1_MEDTR  ATTCTTCTTGAAATTCTATGACACCAAGGGTTCCAAC------TTTGGATGGAAT---GGAACTACTTTC????????????????????????????????????????????????????????????????????????????????????
LS1_PINTA  CTTCTACATCAAGTTCTACGATGTCAAGAGGTCGGAT------TTTGGATGGAAT---GGCACCACATTT????????????????????????????????????????????????????????????????????????????????????
LS1_SOLTU  ATTTTTCCTGAAATTCTATGATGCTAAGGGTTCCTTC------TTTGGCTGGAAT---GGATCTACCCTC????????????????????????????????????????????????????????????????????????????????????
LS1_SORBI  CTTCTACCTCAAGTTCTATGACTCAAAGCGCTCGGAC------TTCGGTTGGAAC---GGCTCCACCTTC????????????????????????????????????????????????????????????????????????????????????
LS1_R0BPS  ATACTTCCTCAAGTTTTACGATACAAAGGGCTCAGAC------TTTGGTTGGAAC---GGATCCACCTTC????????????????????????????????????????????????????????????????????????????????????
LS1_ORYSA  CTTCTACCTCAAGTTCTACGATTCCAAGCGCTCGGAC------TTTGGATGGAAC---GGCACCACCTTT????????????????????????????????????????????????????????????????????????????????????
```

Figure 5.4.1 The MIF cDNA multiple sequence alignment. The bases in the alignment were colored using the default settings in Seqpup v0.6 (D.G. Gilbert ,Biology Dept., Indiana University). A: red, T: orange, G: green, C: blue. The sequences are named following table 5.1.1

Appendix IV

Primer Name	Primer Sequence 5'-3'
01P19.2F1	ATGAGTGGAGTGAAACGTAAA
01P19.2F2	CTGCAGATGTCCGTCTCTTGGTC
01P19.2F3	CGAATTGGGTTCAGTGATTAGT
01P19.2F4	AGGTTGGCAAGATCAAAGAC
01P19.2F5	ATTGACTTAACTACGATCAAGC
01P19.2F6	CGTTACAATGAACCGGAGAGTATG
01P19.2F7	TGCCCGTTGGTATACTGAGAAG
01P19.2F8	GAAGATAGCGATATATATGTTTG
01P19.2F9	AAGCACGAACCAGCACGGATG
01P19.2F10	ATCTCGGTAGATGATGATAGCAG
01P19.2F11	CTCTTCCACGTCTTCATCGACAT
01P19.2F12	AGCTATACGGCTCCATCTACCTC
01P19.2R1	CCGTGAGCTCCAAAGTACAAAC
01P19.2R2	GGATTGGGTTCATTACGAATTG
01P19.2R3	TGATCCCTCTATTTTCCGCAGC
01P19.2R4	GCATTCTTCGCGGAGTACCACG
01P19.2R5	GAGAATCGGAATGAGCATGATC
01P19.2R6	ACATATAACTTCCACTCTTTGC
01P19.2R7	TGGCTGTGCTGCTAACCACGTC
01P19.2R8	CATTCACATATCCATTTGCAAC
01P19.2R9	TTATAGTGGACCAGCACAATCC
01P19.3F1	AGCTGTGAAAAATTCAGTACCAAC
01P19.3F2	CAGCTGTCCACCACAAAAACTCCG
01P19.3F3	CCCGAGCGTTACCCATTGAGTG
01P19.3F4	TGATGCACAGCAGAGACGTAACGG
01P19.3F5	CGAGGATGCGGAACTGAAAGCAGC
01P19.3F6	GTTACTGTTCTGGAAGTGAAGC
01P19.3R1	GGTGTTGCAGACCATAGCGTGCC
01P19.3R2	AATGAGGTCGCATCCGTTTCTGG
01P19.3R3	CTGTACCACGCGCGTCTTCTCTG
01P19.3R4	TTGAAGTGCCATACGTAGTTCGC
01P19.3.R5	CAAGGTCACCCAGAACGCTCTG
01P19.4F1	ATGACTGAAAGCAAGCAGTGTGG
01P19.4.F2	AAGGTGGTCATGCACCAGCCG
01P19.4R1	GTTGACTTGTGAAGCTGTGGTATC
01P19.4.R2	TCTGCTTATCATGATAAGCGCG
01P19.5F1	GGTGTTGGTGTTTTGCTGGATTG
01P19.5F2	GGAAGTGGACTAAGCTACCTG
01P19.5F3	TCCATCAGATGTACTGTCGAC
01P19.5F4	TAGCTTACGAAAGTCCGCTCAG
01P19.5F5	GTCAGTACTTTGCTACACTAC
01P19.5R1	GTTGCTTTTGCCTCTTCATAC
01P19.5R2	GTGGTGACTGCGATGGACTCC
01P19.5R3	ACGGTGCATCAATATGAGAAG
01P19.5R4	CTGAAAATCCACTTCCAAAC
01P19.5R5	GGCACCGTGTCCACGAAGTGC
01P19.6F1	ATGACGCCGGAACAATGTGAATT
01P19.6F2	CCACCACAATGGTTATGTGTGGAG
01P19.6R1	AAGATCTTCTTTGGCATATTGTAC

01P19.6R2	GATCACTGCTGTTTTGACTGG
01P19.7F1	ATGAATGACATAACACAGTCAA
01P19.7F2	ACAGTTTTACTGACATGCATG
01P19.7F3	GAGAAGGAATTTCATCCAGTAATC
01P19.7.F4	GGGTGATATTATAGTACCTGG
01P19.7R1	AAGTAAAAAGTTGATTTGGTGG
01P19.7R2	CGACGAATGTTACAATAAGACC
01P19.7R3	TGAACAATCTGTTCCAGCTGC
01P19.7.R4	TCCATAAAGATAACGTGTTCG
01P19.8F1	ATGGTGCCTTCTTTATGCTCGGAGG
01P19.8R1	AGTTTGCTGCTTTGAGAGCCA
BAC01P19.T7.F1	GCAGCAAATGCTTATTTGTCTTG
BAC01P19.T7.R1	GTTTGGTGATTCATGTCCATGAGC
BAC11C13.SP6.F1	CTGACAATTAACCATTCATAC
BAC11C13.SP6.R1	TGATCAGCAAAACAGTAACCG
BAC11C13.T7.F1	TCATATCTCTGAAAGATTTGC
BAC11C13.T7.R1	AATTGAGAAGATAAGGAGTTG
BAC28F06.SP6.F1	GGAACTCGCAGAATTGAAGTC
BAC28F06.SP6.R1	CGAGTTATAATCGTTCGAATG
BAC39K06.SP6.F1	TATGACTGATTAGCTGGTGGA
BAC39K06.SP6.R1	ACGAGAGTCTACATCTCATTC
BAC45O19.SP6.F1	GCAACATCCTGTGCAGCCAGC
BAC45O19.SP6.R1	TGCTTATTTCATTACAACATC
BAC45O19.T7.F1	CTGGCATTGGATAGGCTGGAC
BAC45O19.T7.R1	GCAAAGCCTGCTAGCGTATCG
Bm-B0205.3.F2	ATGGTTCAGGAAAGTACGATG
Bm-B0205.3.R2	GCATTGATAAAATTTGATGCACC
Bm-B0432.2.R1	ACATATCGCAGCTATGTAACG
Bm-B0432.2.F1	GCTATGGTTATACTTGCTGAG
Bm-C01F6.8.R1	GAAAACGTAAAGCTAAATGTGC
Bm-C01F6.8.F1	CATGCTGTTTCACGGGATCCA
Bm-FIB-1.F1	TGCTATAAATGCCCACCATTTC
Bm-FIB-1.R1	CAACGGACGGTACTCAGCGACC
Bm-F10B5.2.F1	GCTTGACTCATCATGATTAAG
Bm-F10B5.2.R1	ACGTGGTGCCGTTTTGACGCTG
Bm-F37C12.3.F1	GGAACAAGAATTATACAAGATATG
Bm-F37C12.3.R1	CTCACTTTCCAGTTCTGAGATCTC
Bm-F54C9.6.F1	ACTGAACCATTTGTCGAACCA
Bm-F54C9.6.R1	TTGCCGAGTTTGACGTGCGCTG
Bm-F54E7.1.F1	CTTACAAATAGCAACATTTACG
Bm-F54E7.1.R1	AATAGAATACGAATACAGAATC
Bm-K02B2.4.F1	CCACCGAAAGTCTTCGTCGAAG
Bm-K02B2.4.R1	TTCACTCGTTCATGGAATCTT
Bm-MIF-1.F1a	ATGCCATATTTTACGATTGATAC
Bm-MIF-1.R1a	GAACACCATCGCTTGTCCACC
Bm-MIF-2a/b.F4	ATTGGATCATTTTCGGCTGAT
Bm-MIF-2.Z1E09.R1	CTTGTGAAGATTTATCAAACCAA
Bm-rpL5.F1	GTAGGATGGGATTCGTGAAGG
Bm-rpL5.R1	AACCAGTGGCATAAGCTGATGC
Bm-rpL10.F1	CGACGACCAGCAAGATGCTAC
Bm-rpL10.R1	CAACTCGTGCTACTAAGCCCTG

Bm-rpL26.F1	CGAGGACATTATAAAGGAAACG
Bm-rpL26.R1	GTATGTTTACCTTCAACAGGCC
Bm-rpL27a.F1	GGCTACATCAAAGAAAAAGACAAGG
Bm-rpL27a.R1	GCAACCAATACACAGGCACCAC
Bm-rpL36.F1	GGTCGCAGTTGAAGCAGTTGC
Bm-rpL36.R1	CGCTTTATAATCACTTATGAT
Bm-rpP0.F1	TACGGTCTTGTTGTGCGGCAGG
Bm-rpP0.R1	CTTAATCGAAAAGTCCGAATCC
Bm-rpP1.F1	ATGGCAAACCAAGAATTAGC
Bm-rpP1.R1	CAACTCCTTCCAAGGCCTTAGC
Bm-rpP1.R4	ACCACTACCAACACTTGAACT
Bm-rpS12.F2	CAACAGCAGCTGCAGCAGACG
Bm-rpS12.R2	TGCATAGCCCAATCCATTCAC
Bm-rpS14.F1	GGAAGGGTAAAGTTAAGGAGG
Bm-rpS14.R1	GAAGTGCGGATTGGGCTCCTG
Bm-rpS16.F1	GTTGTTACACAATCTGTGCAGG
Bm-rpS16.R1	CAAATTACGATCATACGATACC
Bm-rpS25.F1	AAAACATCAGCGAAAGCTGGG
Bm-rpS25.R1	CCTTCGTGCATCTTGTGTAGAC
Bm-mrpL4.F1	ATGTCATCAGCGCAGTTGGTTC
Bm-mrpL4.R1	GAATATATCTGGATGTAAACTG
Bm-mrpL9.R1	CAGCTTCATAAGATCTGGATG
Bm-mrpL9.F1	CTCAATGCTTTCTGTATCACT
Bm-mrpL31.R1	ACACTCGAACCCTGCTTCTGCG
Bm-mrpL31.F1	ATGTTTCGCTCTGCCCATCGA
Bm-mrpL41.F2	CACTCAATCCTCATTATTTCC
Bm-mrpL41.R2	GATACATTTCCTTAACATGAAC
Bm-mrpS2.F1	CCTATTCATTACGACAGAAAGG
Bm-mrpS2.R1	GTCATCATTTCCGGGAATCGG
Bm-suf-1.F1	TGAGTTCGCTTTCGATTGTAAG
Bm-suf-1.R1	GTTTCACAGTATTCCCATGTCT
Bm-tph-1.F1	CGATATGTTGATCTTCAAGGATG
Bm-tph-1.R1	TTGTTTTTCTTCAATGAGTGCCTCCTT
GeneRacer 3' RACE primer	GCTGTCAACGATACGCTACGTAACGGCATGACAGTG
Pp-F37C12.3.F1	CAGCTCACATTCAAGGAGGTTG
Pp-F37C12.3.R1	CTCTTCATTCAAACACGTCTTC
Pp-rpL27A.F1	GCTTAGAGGACACGTGTCCCACG
Pp-rpL27A.R1	GTGGGAGAAGAACTTGGCCTTCAC
Pp-rpL36.F1	CGCAAGGTTACCAAGCTCGAGG
Pp-rpL36.R1	TACGTTCTGCATCTCGTCACGC
Pp-rpP0.F1	ATGGGTCCTGAGAAGACCTCG
Pp-rpP0.R1	CGCGATTCCGAGCATGCTCTGG
Pp-rpP1.F1	GTGAGCTAAGGCTTAAAGCAACG
Pp-rpP1.R1	GGCGAAGAGACCAGGCCAGAAGGG
Pp-rpS14.F1	GCACGTAAGGGAAAGGTAAAGG
Pp-rpS14.R1	AAGAGCGCGGAGAGCAGACTGG
Pp-tph-1.F1	ATGCTGATCTACAAGGACGCG
Pp- tph-1.R1	CTGCTTCTCAATAATGAGAGC
Sr-rpL27A.R1	CGATCATTGGTTGACTTGGAAC
Sr-rpL27A.F1	GAGGTCGTGGTAATGCTGGAGG

Sr-rpL36.F1	GAAAAAGGTGTCCGTGTCACC
Sr-rpL36.R1	TGGATTTCATCACGCTTCTTC
Sr-rpP1.F1	ATGACTTCTACTCAAGAACTCG
Sr-rpP1.R1	GCAGATCCAGCTCCAGATCCAAT
Sr-rpS14.F1	CCGTAAAGGCAAAGTTCGTGAG
Sr-rpS14.R1	TTCAACAAACGCCATTTATGA
SL1	GGTTTAATTACCCAAGTTTGAG
SP6	CATACGATTTAGGTGACACTATAG
T3	ATTAACCCTCACTAAAGGGA
T7	TAATACGACTCACTATAGGGAGA
2BiotinBACF3	GAGTCGACCTGCAGGCATGC

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et. al.* (2000). The genome sequence of Drosophila melanogaster, Science *287*, 2185-95.

Albertson, D. G., Nwaorgu, O. C., and Sulston, J. E. (1979). Chromatin diminution and a chromosomal mechanism of sexual differentiation in Strongyloides papillosus, Chromosoma *75*, 75-87.

Allen, J. E., Daub, J., Guiliano, D., McDonnell, A., Lizotte-Waniewski, M., Taylor, D. W., and Blaxter, M. (2000). Analysis of genes expressed at the infective larval stage validates utility of Litomosoides sigmodontis as a murine model for filarial vaccine development, Infect Immun *68*, 5454-8.

Allen, J. E., and Loke, P. (2001). Divergent roles for macrophages in lymphatic filariasis, Parasite Immunol *23*, 345-52.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool, J Mol Biol *215*, 403-10.

Altschul, S. F., and Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases, Trends Biochem Sci *23*, 444-7.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res *25*, 3389-402.

Anant, S., Martin, S. A., Yu, H., MacGinnitie, A. J., Devaney, E., and Davidson, N. O. (1997). A cytidine deaminase expressed in the post-infective L3 stage of the filarial nematode, Brugia pahangi, has a novel RNA-binding activity, Mol Biochem Parasitol *88*, 105-14.

Anderson, R. C., and Bain, O. (1976). CIH keys to the nematode parasites of vertebrates No. 3. keys to genera of the order Spirurida Part 3. Diplotriaenoidea, Aproctoidea and Filarioidae. In CIH keys to the nematode parasites of vertebrates, R. C. Anderson, A. G. Chabaud, and S. Willmott, eds. (Slough, England, Commonwealth Agricultural Bureaux).

Andrews, J., Smith, M., Merakovsky, J., Coulson, M., Hannan, F., and Kelly, L. E. (1996). The stoned locus of Drosophila melanogaster produces a dicistronic transcript and encodes two distinct polypeptides, Genetics *143*, 1699-711.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data, Science *290*, 972-7.

Bandi, C., McCall, J. W., Genchi, C., Corona, S., Venco, L., and Sacchi, L. (1999). Effects of tetracycline on the filarial worms Brugia pahangi and Dirofilaria immitis and their bacterial endosymbionts Wolbachia, Int J Parasitol *29*, 357-64.

Bandi, C., Trees, A. J., and Brattig, N. W. (2001). Wolbachia in filarial nematodes: evolutionary aspects and implications for the pathogenesis and treatment of filarial diseases, Vet Parasitol *98*, 215-38.

Barabashova, V. N. (1974). [Karyotype characteristics of certain forms of stem eelworms of the collective species Ditylencus dipsaci], Parazitologiia *8*, 408-12.

Barbazuk, W. B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., Bedell, J. A., McPherson, J. D., and Johnson, S. L. (2000). The syntenic relationship of the zebrafish and human genomes, Genome Res *10*, 1351-8.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). The Pfam protein families database, Nucleic Acids Res *28*, 263-6.

Been, M. D., and Cech, T. R. (1986). One binding site determines sequence specificity of Tetrahymena pre- rRNA self-splicing, trans-splicing, and RNA enzyme activity, Cell *47*, 207-16.

Bendrat, K., Al-Abed, Y., Callaway, D. J., Peng, T., Calandra, T., Metz, C. N., and Bucala, R. (1997). Biochemical and mutational investigations of the enzymatic activity of macrophage migration inhibitory factor, Biochemistry *36*, 15356-62.

Bernhagen, J., Mitchell, R. A., Calandra, T., Voelter, W., Cerami, A., and Bucala, R. (1994). Purification, bioactivity, and secondary structure analysis of mouse and human macrophage migration inhibitory factor (MIF), Biochemistry *33*, 14144-55.

Bin, Z., Hawdon, J., Qiang, S., Hainan, R., Huiqing, Q., Wei, H., Shu-Hua, X., Tiehua, L., Xing, G., Zheng, F., and Hotez, P. (1999). Ancylostoma secreted protein 1 (ASP-1) homologues in human hookworms, Mol Biochem Parasitol *98*, 143-9.

Blaxter, M. (1998). Caenorhabditis elegans is a nematode, Science *282*, 2041-6.

Blaxter, M., Aslett, M., Guiliano, D., and Daub, J. (1999). Parasitic helminth genomics. Filarial Genome Project, Parasitology *118*, S39-51.

Blaxter, M., Daub, J., Guiliano, D., Parkinson, J., Whitton, C., and Project, F. G. (2002). The Brugia malayi genome project: Expressed Sequence tags and gene discovery, Transactions of the Royal Society of Tropical Medicine and Hygiene *96*, 1-11.

Blaxter, M., Daub, J., Guiliano, D. B., Parkinson, J., Whitton, C., and Project, F. G. (2001). The *Brugia malayi* genome project: expresed seqeunce tags and gene discovery, Transactions of the Royal Society of Tropical Medicine and Hygiene *in press*.

Blaxter, M., and Liu, L. (1996). Nematode spliced leaders--ubiquity, evolution and utility, Int J Parasitol *26*, 1025-33.

Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., Vanfleteren, J. R., Mackey, L. Y., Dorris, M., Frisse, L. M., *et. al.* (1998). A molecular evolutionary framework for the phylum *Nematoda*, Nature *392*, 71-5.

Blaxter, M. L., Raghavan, N., Ghosh, I., Guiliano, D., Lu, W., Williams, S. A., Slatko, B., and Scott, A. L. (1996). Genes expressed in Brugia malayi infective third stage larvae, Mol Biochem Parasitol *77*, 77-93.

Bloom, B. R., and Bennett, B. (1966). Mechanism of a reaction in vitro associated with delayed-type hypersensitivity, Science *153*, 80-2.

Blumenthal, T., and Steward, K. (1997). RNA Processing and Gene Structure. In *C.elegans* II, D. L. Riddle, T. Blumenthal, B. J. Meyer, and J. R. Priess, eds. (Plainview, New York, Cold Spring Harbor Laboratory Press), pp. 117-146.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). dbEST--database for "expressed sequence tags", Nat Genet *4*, 332-3.

Boguski, M. S., and Schuler, G. D. (1995). ESTablishing a human transcript map, Nat Genet *10*, 369-71.

Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M., Craig, A., Davies, R. M., Devlin, K., Feltwell, T., *et. al.* (1999). The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum, Nature *400*, 532-8.

Bozza, M., Satoskar, A. R., Lin, G., Lu, B., Humbles, A. A., Gerard, C., and David, J. R. (1999). Targeted disruption of migration inhibitory factor gene reveals its critical role in sepsis, J Exp Med *189*, 341-6.

Brattig, N. W., Buttner, D. W., and Hoerauf, A. (2001). Neutrophil accumulation around Onchocerca worms and chemotaxis of neutrophils are dependent on Wolbachia endobacteria, Microbes Infect *3*, 439-46.

Brittingham, A., Miller, M. A., Donelson, J. E., and Wilson, M. E. (2001). Regulation of GP63 mRNA stability in promastigotes of virulent and attenuated Leishmania chagasi, Mol Biochem Parasitol *112*, 51-9.

Brogna, S., and Ashburner, M. (1997). The Adh-related gene of Drosophila melanogaster is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms, Embo J *16*, 2023-31.

Brouqui, P., Fournier, P. E., and Raoult, D. (2001). Doxycycline and eradication of microfilaremia in patients with loiasis, Emerg Infect Dis 7, 604-5.

Brown, E. W., LeClerc, J. E., Kotewicz, M. L., and Cebula, T. A. (2001). Three R's of bacterial evolution: how replication, repair, and recombination frame the origin of species, Environ Mol Mutagen *38*, 248-60.

Brunner, B., Todt, T., Lenzner, S., Stout, K., Schulz, U., Ropers, H. H., and Kalscheuer, V. M. (1999). Genomic structure and comparative analysis of nine *Fugu* genes: conservation of synteny with human chromosome Xp22.2-p22.1, Genome Res *9*, 437-48.

Bruzik, J. P., and Maniatis, T. (1992). Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells, Nature *360*, 692-5.

Burke, J., Davison, D., and Hide, W. (1999). d2_cluster: a validated method for clustering EST and full-length cDNAsequences, Genome Res *9*, 1135-42.

Caceres, M., Puig, M., and Ruiz, A. (2001). Molecular characterization of two natural hotspots in the Drosophila buzzatii genome induced by transposon insertions, Genome Res *11*, 1353-64.

Caceres, M., Ranz, J. M., Barbadilla, A., Long, M., and Ruiz, A. (1999). Generation of a widespread Drosophila inversion by a transposable element, Science *285*, 415-8.

Calandra, T., Bernhagen, J., Metz, C. N., Spiegel, L. A., Bacher, M., Donnelly, T., Cerami, A., and Bucala, R. (1995). MIF as a glucocorticoid-induced modulator of cytokine production, Nature *377*, 68-71.

Carlton, J. M., Vinkenoog, R., Waters, A. P., and Walliker, D. (1998). Gene synteny in species of *Plasmodium*, Mol Biochem Parasitol *93*, 285-94.

Casiraghi, M., Anderson, T. J., Bandi, C., Bazzocchi, C., and Genchi, C. (2001). A phylogenetic analysis of filarial nematodes: comparison with the phylogeny of Wolbachia endosymbionts, Parasitology *122 Pt 1*, 93-103.

Chan, M. S. (1997). The global burden of intestinal nematode infections - fifty years on, Parasitology today *13*, 438-443.

Chandrashekar, R., Curtis, K. C., Ramzy, R. M., Liftis, F., Li, B. W., and Weil, G. J. (1994). Molecular cloning of Brugia malayi antigens for diagnosis of lymphatic filariasis, Mol Biochem Parasitol *64*, 261-71.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D., *et. al.* (2001). Massive gene decay in the leprosy bacillus, Nature *409*, 1007-11.

Combes, D., Fedon, Y., Grauso, M., Toutant, J. P., and Arpagaus, M. (2000). Four genes encode acetylcholinesterases in the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. cDNA sequences, genomic structures, mutations and in vivo expression, J Mol Biol *300*, 727-42.

Conklin, P. L., Wilson, R. K., and Hanson, M. R. (1991). Multiple trans-splicing events are required to produce a mature nad1 transcript in a plant mitochondrion, Genes Dev *5*, 1407-15.

consortium, T. C. e. g. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium, Science *282*, 2012-8.

Cookson, E., Blaxter, M. L., and Selkirk, M. E. (1992). Identification of the major soluble cuticular glycoprotein of lymphatic filarial nematode parasites (gp29) as a secretory homolog of glutathione peroxidase, Proc Natl Acad Sci U S A *89*, 5837-41.

Cross, H. F., Haarbrink, M., Egerton, G., Yazdanbakhsh, M., and Taylor, M. J. (2001). Severe reactions to filarial chemotherapy and release of Wolbachia endosymbionts into blood, Lancet *358*, 1873-5.

Daub, J., Loukas, A., Pritchard, D. I., and Blaxter, M. (2000). A survey of genes expressed in adults of the human hookworm, Necator americanus, Parasitology *120*, 171-84.

David, J. R. (1966). Delayed hypersensitivity in vitro: its mediation by cell-free substances formed by lymphoid cell-antigen interaction, Proc Natl Acad Sci U S A *56*, 72-7.

Davidson, H., Taylor, M. S., Doherty, A., Boyd, A. C., and Porteous, D. J. (2000). Genomic sequence analysis of *Fugu* rubripes CFTR and flanking genes in a 60 kb region conserving synteny with 800 kb of human chromosome 7, Genome Res *10*, 1194-203.

Davis, R. E. (1997). Surprising diversity and distribution of spliced leader RNAs in flatworms, Mol Biochem Parasitol *87*, 29-48.

Davis, R. E., Hardwick, C., Tavernier, P., Hodgson, S., and Singh, H. (1995). RNA trans-splicing in flatworms. Analysis of trans-spliced mRNAs and genes in the human parasite, Schistosoma mansoni, J Biol Chem *270*, 21813-9.

Davis, R. E., and Hodgson, S. (1997). Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in Schistosoma mansoni, Mol Biochem Parasitol *89*, 25-39.

Deiss, L. P., Feinstein, E., Berissi, H., Cohen, O., and Kimchi, A. (1995). Identification of a novel serine/threonine kinase and a novel 15-kD protein as potential mediators of the gamma interferon-induced cell death, Genes Dev *9*, 15-30.

del Portillo, H. A., Fernandez-Becerra, C., Bowman, S., Oliver, K., Preuss, M., Sanchez, C. P., Schneider, N. K., Villalobos, J. M., Rajandream, M. A., Harris, D., *et. al.* (2001). A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax, Nature *410*, 839-42.

Donelson, J. E., Duke, B. O., Moser, D., Zeng, W. L., Erondu, N. E., Lucius, R., Renz, A., Karam, M., and Flores, G. Z. (1988). Construction of Onchocerca volvulus cDNA libraries and partial characterization of the cDNA for a major antigen, Mol Biochem Parasitol *31*, 241-50.

Dorn, R., Reuter, G., and Loewendorf, A. (2001). Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila, Proc Natl Acad Sci U S A *98*, 9724-9.

Dufourcq, P., Chanal, P., Vicaire, S., Camut, E., Quintin, S., den Boer, B. G., Bosher, J. M., and Labouesse, M. (1999). lir-2, lir-1 and lin-26 encode a new class of zinc-finger proteins and are organized in two overlapping operons both in Caenorhabditis elegans and in Caenorhabditis briggsae, Genetics *152*, 221-35.

Edwards, R. A., Olsen, G. J., and Maloy, S. R. (2002). Comparative genomics of closely related salmonellae, Trends Microbiol *10*, 94-9.

Eisen, J. A., Heidelberg, J. F., White, O., and Salzberg, S. L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria, Genome Biol *1*.

El-Sayed, N. M., Hegde, P., Quackenbush, J., Melville, S. E., and Donelson, J. E. (2000). The African trypanosome genome, Int J Parasitol *30*, 329-45.

Erttmann, K. D., and Gallin, M. Y. (1996). Onchocerca volvulus: identification of cDNAs encoding a putative phosphatidyl-ethanolamine-binding protein and a putative partially processed mRNA precursor, Gene *174*, 203-7.

Eul, J., Graessmann, M., and Graessmann, A. (1995). Experimental evidence for RNA trans-splicing in mammalian cells, Embo J *14*, 3226-35.

Evans, D., and Blumenthal, T. (2000). trans splicing of polycistronic Caenorhabditis elegans pre-mRNAs: analysis of the SL2 RNA, Mol Cell Biol *20*, 6659-67.

Evans, D., Perez, I., MacMorris, M., Leake, D., Wilusz, C. J., and Blumenthal, T. (2001). A complex containing CstF-64 and the SL2 snRNP connects mRNA 3' end formation and trans-splicing in C. elegans operons, Genes Dev *15*, 2562-71.

Evans, D., Zorio, D., MacMorris, M., Winter, C. E., Lea, K., and Blumenthal, T. (1997). Operons and SL2 trans-splicing exist in nematodes outside the genus Caenorhabditis, Proc Natl Acad Sci U S A *94*, 9751-6.

Fadiel, A., Lithwick, S., Wanas, M. Q., and Cuticchia, A. J. (2001). Influence of intercodon and base frequencies on codon usage in filarial parasites, Genomics *74*, 197-210.

Falcone, F. H., Loke, P., Zang, X., MacDonald, A. S., Maizels, R. M., and Allen, J. E. (2001). A Brugia malayi homolog of macrophage migration inhibitory factor reveals an important link between macrophages and eosinophil recruitment during nematode infection, J Immunol *167*, 5348-54.

Ferguson, K. C., Heid, P. J., and Rothman, J. H. (1996). The SL1 trans-spliced leader RNA performs an essential embryonic function in Caenorhabditis elegans that can also be supplied by SL2 RNA, Genes Dev *10*, 1543-56.

Ferrier, D. E., and Holland, P. W. (2001). Ancient origin of the Hox gene cluster, Nat Rev Genet *2*, 33-8.

Fingerle-Rowson, G. R., and Bucala, R. (2001). Neuroendocrine properties of macrophage migration inhibitory factor (MIF), Immunol Cell Biol *79*, 368-75.

Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C., and Dujon, B. (2001). Evolution of gene order in the genomes of two related yeast species, Genome Res *11*, 2009-19.

Fitch, D. H., Bugaj-Gaweda, B., and Emmons, S. W. (1995). 18S ribosomal RNA gene phylogeny for some Rhabditidae related to Caenorhabditis, Mol Biol Evol *12*, 346-58.

Foster, J. M., Kamal, I. H., Daub, J., Swan, M. C., Ingram, J. R., Ganatra, M., Ware, J., Guiliano, D., Aboobaker, A., Moran, L., *et. al.* (2001). Hybridization to high-density filter arrays of a *Brugia malayi* BAC library with biotinylated oligonucleotides and PCR products, Biotechniques *30*, 1216-8, 1220, 1222 passim.

Frank, G. R., and Grieve, R. B. (1991). Metabolic labeling of Dirofilaria immitis third- and fourth-stage larvae and their excretory-secretory products, J Parasitol *77*, 950-6.

Frantz, C., Ebel, C., Paulus, F., and Imbault, P. (2000). Characterization of trans-splicing in Euglenoids, Curr Genet *37*, 349-55.

Fraser, A. G., Kamath, R. S., Zipperlen, P., Martinez-Campos, M., Sohrmann, M., and Ahringer, J. (2000). Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference, Nature *408*, 325-30.

Froidevaux, C., Roger, T., Martin, C., Glauser, M. P., and Calandra, T. (2001). Macrophage migration inhibitory factor and innate immune responses to bacterial infections, Crit Care Med *29*, S13-5.

Fukunishi, Y., and Hayashizaki, Y. (2002). Amino-acid translation program for full-length cDNA sequences with frame-shift error, Physiol Genomics *in press*.

Fulton, R. E., Salasek, M. L., DuTeau, N. M., and Black, W. C. T. (2001). SSCP analysis of cDNA markers provides a dense linkage map of the *Aedes aegypti* genome, Genetics *158*, 715-26.

Garcia-Rios, M., Fujita, T., LaRosa, P. C., Locy, R. D., Clithero, J. M., Bressan, R. A., and Csonka, L. N. (1997). Cloning of a polycistronic cDNA from tomato encoding gamma-glutamyl kinase and gamma-glutamyl phosphate reductase, Proc Natl Acad Sci U S A *94*, 8249-54.

Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., *et. al.* (1998). Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum, Science *282*, 1126-32.

Ghosh, I., Eisinger, S. W., Raghavan, N., and Scott, A. L. (1998). Thioredoxin peroxidases from Brugia malayi, Mol Biochem Parasitol *91*, 207-20.

Girardet, C., Walker, W. H., and Habener, J. F. (1996). An alternatively spliced polycistronic mRNA encoding cyclic adenosine 3',5'-monophosphate (cAMP)-responsive transcription factor CREB (cAMP response element-binding protein) in human testis extinguishes expression of an internally translated inhibitor CREB isoform, Mol Endocrinol *10*, 879-891.

Goldstein, P., and Moens, P. B. (1976). Karyotype analysis of Ascaris lumbricoides var. suum. Male and female pachytene nuclei by 3-D reconstruction from electron microscopy of serial sections, Chromosoma *58*, 101-11.

Goldstein, P., and Slaton, D. E. (1982). The synaptonemal complexes of caenorhabditis elegans: comparison of wild-type and mutant strains and pachytene karyotype analysis of wild- type, Chromosoma *84*, 585-97.

Goldstein, P., and Triantaphyllou, A. C. (1979). Karyotype analysis of the plant-parasitic nematode Heterodera glycines by electron microscopy. 1. The diploid, J Cell Sci *40*, 171-9.

Goldstein, P., and Triantaphyllou, A. C. (1980). Karyotype analysis of the plant-parasitic nematode Heterodera glycines by electron microscopy. II. The tetraploid and an aneuploid hybrid, J Cell Sci *43*, 225-37.

Goldstein, P., and Triantaphyllou, A. C. (1981). Pachytene karyotype analysis of tetraploid Meloidogyne hapla females by electron microscopy, Chromosoma *84*, 405-12.

Grant, D., Cregan, P., and Shoemaker, R. C. (2000). Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and Arabidopsis, Proc Natl Acad Sci U S A *97*, 4168-73.

Graul, R. C., and Sadee, W. (1997). Evolutionary relationships among proteins probed by an iterative neighborhood cluster analysis (INCA). Alignment of bacteriorhodopsins with the yeast sequence YRO2, Pharm Res *14*, 1533-41.

Gregory, W. F., Atmadja, A. K., Allen, J. E., and Maizels, R. M. (2000). The abundant larval transcript-1 and -2 genes of Brugia malayi encode stage-specific candidate vaccine antigens for filariasis, Infect Immun *68*, 4174-9.

Gregory, W. F., Blaxter, M. L., and Maizels, R. M. (1997). Differentially expressed, abundant trans-spliced cDNAs from larval Brugia malayi, Mol Biochem Parasitol *87*, 85-95.

Guiliano, D., Ganatra, M., Ware, J., Parrot, J., Daub, J., Moran, L., Brennecke, H., Foster, J. M., Supali, T., Blaxter, M.*, et. al.* (1999). Chemiluminescent detection of sequential DNA hybridizations to high- density, filter-arrayed cDNA libraries: a subtraction method for novel gene discovery, Biotechniques *27*, 146-52.

Hamer, L., Pan, H., Adachi, K., Orbach, M. J., Page, A., Ramamurthy, L., and Woessner, J. P. (2001). Regions of microsynteny in *Magnaporthe grisea* and *Neurospora crassa*, Fungal Genet Biol *33*, 137-43.

Hannon, G. J., Maroney, P. A., Denker, J. A., and Nilsen, T. W. (1990). Trans splicing of nematode pre-messenger RNA in vitro, Cell *61*, 1247-55.

Hawdon, J. M., Jones, B. F., Hoffman, D. R., and Hotez, P. J. (1996). Cloning and characterization of Ancylostoma-secreted protein. A novel protein associated with the transition to parasitism by infective hookworm larvae, J Biol Chem *271*, 6672-8.

Hawdon, J. M., Narasimhan, S., and Hotez, P. J. (1999). Ancylostoma secreted protein 2: cloning and characterization of a second member of a family of nematode secreted proteins from Ancylostoma caninum, Mol Biochem Parasitol *99*, 149-65.

Henkle, K. J., Liebau, E., Muller, S., Bergmann, B., and Walter, R. D. (1991). Characterization and molecular cloning of a Cu/Zn superoxide dismutase from the human parasite Onchocerca volvulus, Infect Immun *59*, 2063-9.

Hentschel, C. C., and Birnstiel, M. L. (1981). The organization and expression of histone gene families, Cell *25*, 301-13.

Herman, R. K., Albertson, D. G., and Brenner, S. (1976). Chromosome rearrangements in Caenorhabditis elegans, Genetics *83*, 91-105.

Hermanowski-Vosatka, A., Mundt, S. S., Ayala, J. M., Goyal, S., Hanlon, W. A., Czerwinski, R. M., Wright, S. D., and Whitman, C. P. (1999). Enzymatically inactive macrophage migration inhibitory factor inhibits monocyte chemotaxis and random migration, Biochemistry *38*, 12841-9.

Hermans, P. G., Hart, C. A., and Trees, A. J. (2001). In vitro activity of antimicrobial agents against the endosymbiont Wolbachia pipientis, J Antimicrob Chemother *47*, 659-63.

Hirai, H., Sakaguchi, Y., and Tada, I. (1985). Chromosomes of Onchocerca volvulus and O. gutturosa, Z Parasitenkd *71*, 135-9.

Hirai, H., Tada, I., Takahashi, H., Nwoke, B. E., and Ufomadu, G. O. (1987). Chromosomes of Onchocerca volvulus (Spirurida: Onchocercidae): a comparative study between Nigeria and Guatemala, J Helminthol *61*, 43-6.

Hoerauf, A., Volkmann, L., Hamelmann, C., Adjei, O., Autenrieth, I. B., Fleischer, B., and Buttner, D. W. (2000a). Endosymbiotic bacteria in worms as targets for a novel chemotherapy in filariasis, Lancet *355*, 1242-3.

Hoerauf, A., Volkmann, L., Nissen-Paehle, K., Schmetz, C., Autenrieth, I., Buttner, D. W., and Fleischer, B. (2000b). Targeting of Wolbachia endobacteria in Litomosoides sigmodontis: comparison of tetracyclines with chloramphenicol, macrolides and ciprofloxacin, Trop Med Int Health *5*, 275-9.

Hough, R. F., Lingam, A. T., and Bass, B. L. (1999). Caenorhabditis elegans mRNAs that encode a protein similar to ADARs derive from an operon containing six genes, Nucleic Acids Res *27*, 3424-32.

Hu, M., Chilton, N. B., and Gasser, R. B. (2002). The mitochondrial genomes of the human hookworms, Ancylostoma duodenale and Necator americanus (Nematoda: Secernentea), Int J Parasitol *32*, 145-58.

Huang, T., Kuersten, S., Deshpande, A. M., Spieth, J., MacMorris, M., and Blumenthal, T. (2001). Intercistronic region required for polycistronic pre-mRNA processing in Caenorhabditis elegans, Mol Cell Biol *21*, 1111-20.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program, Genome Res *9*, 868-77.

Huang, X. Y., Barrios, L. A., Vonkhorporn, P., Honda, S., Albertson, D. G., and Hecht, R. M. (1989). Genomic organization of the glyceraldehyde-3-phosphate dehydrogenase gene family of Caenorhabditis elegans, J Mol Biol *206*, 411-24.

Huang, X. Y., and Hirsh, D. (1989). A second trans-spliced RNA leader sequence in the nematode Caenorhabditis elegans, Proc Natl Acad Sci U S A *86*, 8640-4.

Hudson, J. D., Shoaibi, M. A., Maestro, R., Carnero, A., Hannon, G. J., and Beach, D. H. (1999). A proinflammatory cytokine inhibits p53 tumor suppressor activity, J Exp Med *190*, 1375-82.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics *17*, 754-5.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology, Science *294*, 2310-4.

Huynen, M. A., Snel, B., and Bork, P. (2001). Inversions and the dynamics of eukaryotic gene order, Trends Genet *17*, 304-6.

Janssen, C. S., Barrett, M. P., Lawson, D., Quail, M. A., Harris, D., Bowman, S., Phillips, R. S., and Turner, C. M. (2001). Gene discovery in Plasmodium chabaudi by genome survey sequencing, Mol Biochem Parasitol *113*, 251-60.

Jaworski, D. C., Jasinskas, A., Metz, C. N., Bucala, R., and Barbour, A. G. (2001). Identification and characterization of a homologue of the pro- inflammatory cytokine Macrophage Migration Inhibitory Factor in the tick, Amblyomma americanum, Insect Mol Biol *10*, 323-31.

Joseph, G. T., Huima, T., and Lustigman, S. (1998). Characterization of an Onchocerca volvulus L3-specific larval antigen, Ov-ALT-1, Mol Biochem Parasitol *96*, 177-83.

Jung, H., Kim, T., Chae, H. Z., Kim, K. T., and Ha, H. (2001). Regulation of macrophage migration inhibitory factor and thiol-specific antioxidant protein PAG by direct interaction, J Biol Chem *276*, 15504-10.

Juttner, S., Bernhagen, J., Metz, C. N., Rollinghoff, M., Bucala, R., and Gessner, A. (1998). Migration inhibitory factor induces killing of Leishmania major by macrophages: dependence on reactive nitrogen intermediates and endogenous TNF-alpha, J Immunol *161*, 2383-90.

Keddie, E. M., Higazi, T., and Unnasch, T. R. (1998). The mitochondrial genome of Onchocerca volvulus: sequence, structure and phylogenetic analysis, Mol Biochem Parasitol *95*, 111-27.

Kent, W. J., and Zahler, A. M. (2000). Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment, Genome Res *10*, 1115-25.

Kleemann, R., Hausser, A., Geiger, G., Mischke, R., Burger-Kentischer, A., Flieger, O., Johannes, F. J., Roger, T., Calandra, T., Kapurniotu, A., *et. al.* (2000a). Intracellular action of the cytokine MIF to modulate AP-1 activity and the cell cycle through Jab1, Nature *408*, 211-6.

Kleemann, R., Kapurniotu, A., Frank, R. W., Gessner, A., Mischke, R., Flieger, O., Juttner, S., Brunner, H., and Bernhagen, J. (1998a). Disulfide analysis reveals a role for macrophage migration inhibitory factor (MIF) as thiol-protein oxidoreductase, J Mol Biol *280*, 85-102.

Kleemann, R., Kapurniotu, A., Mischke, R., Held, J., and Bernhagen, J. (1999). Characterization of catalytic centre mutants of macrophage migration inhibitory factor (MIF) and comparison to Cys81Ser MIF, Eur J Biochem *261*, 753-66.

Kleemann, R., Mischke, R., Kapurniotu, A., Brunner, H., and Bernhagen, J. (1998b). Specific reduction of insulin disulfides by macrophage migration inhibitory factor (MIF) with glutathione and dihydrolipoamide: potential role in cellular redox processes, FEBS Lett *430*, 191-6.

Kleemann, R., Rorsman, H., Rosengren, E., Mischke, R., Mai, N. T., and Bernhagen, J. (2000b). Dissection of the enzymatic and immunologic functions of macrophage migration inhibitory factor. Full immunologic activity of N-terminally truncated mutants, Eur J Biochem *267*, 7183-93.

Koltai, H., Spiegel, Y., and Blaxter, M. L. (1997). Regulated use of an alternative spliced leader exon in the plant parasitic nematode Meloidogyne javanica, Mol Biochem Parasitol *86*, 107-10.

Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification, Annu Rev Microbiol *55*, 709-42.

Ku, H. M., Vision, T., Liu, J., and Tanksley, S. D. (2000). Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny, Proc Natl Acad Sci U S A *97*, 9121-6.

Kuwabara, P. E., and Shah, S. (1994). Cloning by synteny: identifying *C. briggsae* homologues of *C. elegans* genes, Nucleic Acids Res *22*, 4414-8.

Lagercrantz, U. (1998). Comparative mapping between Arabidopsis thaliana and Brassica nigra indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements, Genetics *150*, 1217-28.

Laird, P. W. (1989). Trans splicing in trypanosomes--archaism or adaptation?, Trends Genet *5*, 204-8.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et. al.* (2001). Initial sequencing and analysis of the human genome, Nature *409*, 860-921.

Langworthy, N. G., Renz, A., Mackenstedt, U., Henkle-Duhrsen, K., de Bronsvoort, M. B., Tanya, V. N., Donnelly, M. J., and Trees, A. J. (2000). Macrofilaricidal activity of tetracycline against the filarial nematode Onchocerca ochengi: elimination of Wolbachia precedes worm death and suggests a dependent relationship, Proc R Soc Lond B Biol Sci *267*, 1063-9.

Lavrov, D. V., and Brown, W. M. (2001). Trichinella spiralis mtDNA: a nematode mitochondrial genome that encodes a putative ATP8 and normally structured tRNAS and has a gene arrangement relatable to those of coelomate metazoans, Genetics *157*, 621-37.

Leonard, M. W., and Patient, R. K. (1996). Primer Extension Analysis of mRNA. In Methods in Molecular Biology, A. Harwood, ed. (Totowa, NJ, Humana Press Inc), pp. 137-145.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., and Quackenbush, J. (2000). An optimized protocol for analysis of EST sequences, Nucleic Acids Res *28*, 3657-65.

Litman, G. W., Rast, J. P., Shamblott, M. J., Haire, R. N., Hulst, M., Roess, W., Litman, R. T., Hinds-Frey, K. R., Zilch, A., and Amemiya, C. T. (1993). Phylogenetic diversification of immunoglobulin genes and the antibody repertoire, Mol Biol Evol *10*, 60-72.

Liu, D. Y., David, J. R., and Remold, H. G. (1982). Glycolipid affinity purification of migration inhibitory factor, Nature *296*, 78-80.

Liu, H., Sachidanandam, R., and Stein, L. (2001a). Comparative genomics between rice and Arabidopsis shows scant collinearity in gene order, Genome Res *11*, 2020-6.

Liu, Y., Huang, T., MacMorris, M., and Blumenthal, T. (2001b). Interplay between AAUAAA and the trans-splice site in processing of a Caenorhabditis elegans operon pre-mRNA, Rna *7*, 176-81.

Lizotte-Waniewski, M., Tawe, W., Guiliano, D. B., Lu, W., Liu, J., Williams, S. A., and Lustigman, S. (2000). Identification of potential vaccine and drug target candidates by expressed sequence tag analysis and immunoscreening of Onchocerca volvulus larval cDNA libraries, Infect Immun *68*, 3491-501.

Llorente, B., Malpertuy, A., Neuveglise, C., de Montigny, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., *et. al.* (2000). Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*, FEBS Lett *487*, 101-12.

Lu, W., Egerton, G. L., Bianco, A. E., and Williams, S. A. (1998). Thioredoxin peroxidase from Onchocerca volvulus: a major hydrogen peroxide detoxifying enzyme in filarial parasites, Mol Biochem Parasitol *91*, 221-35.

Lubetsky, J. B., Swope, M., Dealwis, C., Blake, P., and Lolis, E. (1999). Pro-1 of macrophage migration inhibitory factor functions as a catalytic base in the phenylpyruvate tautomerase activity, Biochemistry *38*, 7346-54.

Lucke, S., Xu, G. L., Palfi, Z., Cross, M., Bellofatto, V., and Bindereif, A. (1996). Spliced leader RNA of trypanosomes: in vivo mutational analysis reveals extensive and distinct requirements for trans splicing and cap4 formation, Embo J *15*, 4380-91.

Lustigman, S., Brotman, B., Huima, T., Prince, A. M., and McKerrow, J. H. (1992). Molecular cloning and characterization of onchocystatin, a cysteine proteinase inhibitor of Onchocerca volvulus, J Biol Chem *267*, 17339-46.

Lustigman, S., McKerrow, J. H., Shah, K., Lui, J., Huima, T., Hough, M., and Brotman, B. (1996). Cloning of a cysteine protease required for the molting of Onchocerca volvulus third stage larvae, J Biol Chem *271*, 30181-9.

Maeda, I., Kohara, Y., Yamamoto, M., and Sugimoto, A. (2001). Large-scale analysis of gene function in *Caenorhabditis elegans* by high- throughput RNAi, Curr Biol *11*, 171-6.

Maizels, R. M., Blaxter, M. L., and Scott, A. L. (2001). Immunological genomics of Brugia malayi: filarial genes implicated in immune evasion and protective immunity, Parasite Immunol *23*, 327-44.

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982). Molecular Cloning A Laboratory Manual (Cold Spring Harbor, New York, Cold Spring Harbor Laboratory).

Manoury, B., Gregory, W. F., Maizels, R. M., and Watts, C. (2001). Bm-CPI-2, a cystatin homolog secreted by the filarial parasite Brugia malayi, inhibits class II MHC-restricted antigen processing, Curr Biol *11*, 447-51.

Manson-Bahr, P. E. C., and Bell, D. R. (1987). Manson's Tropical Diseases, 19th edn (London, England, Bailliere Tindall).

Maroney, P. A., Denker, J. A., Darzynkiewicz, E., Laneve, R., and Nilsen, T. W. (1995). Most mRNAs in the nematode Ascaris lumbricoides are trans-spliced: a role for spliced leader addition in translational efficiency, Rna *1*, 714-23.

Maroney, P. A., Hannon, G. J., Denker, J. A., and Nilsen, T. W. (1990). The nematode spliced leader RNA participates in trans-splicing as an Sm snRNP, Embo J *9*, 3667-73.

Maroney, P. A., Hannon, G. J., Shambaugh, J. D., and Nilsen, T. W. (1991). Intramolecular base pairing between the nematode spliced leader and its 5' splice site is not essential for trans-splicing in vitro, Embo J *10*, 3869-75.

Marson, A. L., Tarr, D. E., and Scott, A. L. (2001). Macrophage migration inhibitory factor (mif) transcription is significantly elevated in Caenorhabditis elegans dauer larvae, Gene *278*, 53-62.

Martin, S. A., Hunter, S. J., Thompson, F. J., and Devaney, E. (1996). Stage specific gene expression in the post-infective L3 of the filarial nematode, Brugia pahangi, Mol Biochem Parasitol *79*, 109-12.

McCarter, J., Abad, P., Jones, J. T., and Bird, D. (2000). Papid gene discovery in plant parasitic nematodes via Expressed Sequence Tags, Nematology *2*, 719-731.

McReynolds, L. A., DeSimone, S. M., and Williams, S. A. (1986). Cloning and comparison of repeated DNA sequences from the human filarial parasite Brugia malayi and the animal parasite Brugia pahangi, Proc Natl Acad Sci U S A *83*, 797-801.

Meredith, S. E., Lando, G., Gbakima, A. A., Zimmerman, P. A., and Unnasch, T. R. (1991). Onchocerca volvulus: application of the polymerase chain reaction to identification and strain differentiation of the parasite, Exp Parasitol *73*, 335-44.

Michalski, M. L., and Weil, G. J. (1999). Gender-specific gene expression in Brugia malayi, Mol Biochem Parasitol *104*, 247-57.

Miller, R. T., Christoffels, A. G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T. R., and Hide, W. A. (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base, Genome Res *9*, 1143-55.

Mira, A., Ochman, H., and Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes, Trends Genet *17*, 589-96.

Mischke, R., Kleemann, R., Brunner, H., and Bernhagen, J. (1998). Cross-linking and mutational analysis of the oligomerization state of the cytokine macrophage migration inhibitory factor (MIF), FEBS Lett *427*, 85-90.

Mitchell, R. A., Metz, C. N., Peng, T., and Bucala, R. (1999). Sustained mitogen-activated protein kinase (MAPK) and cytoplasmic phospholipase A2 activation by macrophage migration inhibitory factor (MIF). Regulatory role in cell proliferation and glucocorticoid action, J Biol Chem *274*, 18100-6.

Molyneux, D. H. (1995). Onchocerciasis control in west africa: current status and future of the onchocerciasis control program, Parasitology Today *11*, 399-402.

Moreno, S. N., Ip, H. S., and Cross, G. A. (1991). An mRNA-dependent in vitro translation system from Trypanosoma brucei, Mol Biochem Parasitol *46*, 265-74.

Moyle, M., Foster, D. L., McGrath, D. E., Brown, S. M., Laroche, Y., De Meutter, J., Stanssens, P., Bogowitz, C. A., Fried, V. A., Ely, J. A., and *et. al.* (1994). A hookworm glycoprotein that inhibits neutrophil function is a ligand of the integrin CD11b/CD18, J Biol Chem *269*, 10008-15.

Muller, F., Bernard, V., and Tobler, H. (1996). Chromatin diminution in nematodes, Bioessays *18*, 133-8.

Murphy, W. J., Watkins, K. P., and Agabian, N. (1986). Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing, Cell *47*, 517-25.

Murray, J., Gregory, W. F., Gomez-Escobar, N., Atmadja, A. K., and Maizels, R. M. (2001). Expression and immune recognition of Brugia malayi VAL-1, a homologue of vespid venom allergens and Ancylostoma secreted proteins, Mol Biochem Parasitol *118*, 89-96.

Mutafova, T. (1976). Comparative cytological studies of mitotic and male meiotic karyotype of Ascaridia dissimilis (Vigueras, 1931) and Ascaridia galli (Schrank, 1788), Z Parasitenkd *48*, 239-45.

Mutafova, T., Dimitrova, Y., and Komandarev, S. (1982). The karyotype of four Trichinella species, Z Parasitenkd *67*, 115-20.

Myler, P. J., Audleman, L., deVos, T., Hixson, G., Kiser, P., Lemley, C., Magness, C., Rickel, E., Sisk, E., Sunkin, S., *et. al.* (1999). Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes, Proc Natl Acad Sci U S A *96*, 2902-6.

Myler, P. J., Beverley, S. M., Cruz, A. K., Dobson, D. E., Ivens, A. C., McDonagh, P. D., Madhubala, R., Martinez-Calvillo, S., Ruiz, J. C., Saxena, A., *et. al.* (2001). The Leishmania genome project: new insights into gene organization and function, Med Microbiol Immunol (Berl) *190*, 9-12.

Myler, P. J., Sisk, E., McDonagh, P. D., Martinez-Calvillo, S., Schnaufer, A., Sunkin, S. M., Yan, S., Madhubala, R., Ivens, A., and Stuart, K. (2000). Genomic organization and gene function in Leishmania, Biochem Soc Trans *28*, 527-31.

Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, Trends Biochem Sci *24*, 34-6.

Nakai, K., and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells, Genomics *14*, 897-911.

Nicolas, R. H., and Goodwin, G. H. (1996). Molecular cloning of polybromo, a nuclear protein containing multiple domains including five bromodomains, a truncated HMG-box, and two repeats of a novel domain, Gene *175*, 233-40.

Nishihira, J. (2000). Macrophage migration inhibitory factor (MIF): its essential role in the immune system and cell growth, J Interferon Cytokine Res *20*, 751-62.

Nomura, T., Carlton, J. M., Baird, J. K., del Portillo, H. A., Fryauff, D. J., Rathore, D., Fidock, D. A., Su, X., Collins, W. E., McCutchan, T. F., *et. al.* (2001). Evidence for different mechanisms of chloroquine resistance in 2 Plasmodium species that cause human malaria, J Infect Dis *183*, 1653-61.

O'Brien, S. J., Menotti-Raymond, M., Murphy, W. J., Nash, W. G., Wienberg, J., Stanyon, R., Copeland, N. G., Jenkins, N. A., Womack, J. E., and Marshall Graves, J. A. (1999). The promise of comparative genomics in mammals, Science *286*, 458-62, 479-81.

Oberlander, U., Adam, R., Berg, K., Seeber, F., and Lucius, R. (1995). Molecular cloning and characterization of the filarial LIM domain proteins AvL3-1 and OvL3-1, Exp Parasitol *81*, 592-9.

Ohta, Y., Okamura, K., McKinney, E. C., Bartl, S., Hashimoto, K., and Flajnik, M. F. (2000). Primitive synteny of vertebrate major histocompatibility complex class I and class II genes, Proc Natl Acad Sci U S A *97*, 4712-7.

Okimoto, R., Macfarlane, J. L., Clary, D. O., and Wolstenholme, D. R. (1992). The mitochondrial genomes of two nematodes, Caenorhabditis elegans and Ascaris suum, Genetics *130*, 471-98.

Onodera, S., Kaneda, K., Mizue, Y., Koyama, Y., Fujinaga, M., and Nishihira, J. (2000). Macrophage migration inhibitory factor up-regulates expression of matrix metalloproteinases in synovial fibroblasts of rheumatoid arthritis, J Biol Chem *275*, 444-50.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures, Structure *5*, 1093-108.

Orita, M., Yamamoto, S., Katayama, N., Aoki, M., Takayama, K., Yamagiwa, Y., Seki, N., Suzuki, H., Kurihara, H., Sakashita, H., *et. al.* (2001). Coumarin and chromen-4-one analogues as tautomerase inhibitors of macrophage migration inhibitory factor: discovery and X-ray crystallography, J Med Chem *44*, 540-7.

Ottesen, E. A. (2000). The global programme to eliminate lymphatic filariasis, Trop Med Int Health *5*, 591-4.

Ottesen, E. A., and Ramachandran, C. P. (1995). Lymphatic filiariasis infection and disease: control strategies, Parasitology Today *11*, 129-131.

Page, A. P. (1997). Cyclophilin and protein disulfide isomerase genes are co-transcribed in a functionally related manner in Caenorhabditis elegans, DNA Cell Biol *16*, 1335-43.

Palfi, Z., Xu, G. L., and Bindereif, A. (1994). Spliced leader-associated RNA of trypanosomes. Sequence conservation and association with protein components common to trans-spliceosomal ribonucleoproteins, J Biol Chem *269*, 30620-5.

Parkinson, J., Whitton, C., Guiliano, D., Daub, J., and Blaxter, M. (2001). 200000 nematode expressed sequence tags on the Net, Trends Parasitol *17*, 394-396.

Parsons, J. D. (1995). Improved tools for DNA comparison and clustering, Comput Appl Biosci *11*, 603-13.

Parsons, J. D., and Rodriguez-Tome, P. (2000). JESAM: CORBA software components to create and publish EST alignments and clusters, Bioinformatics *16*, 313-25.

Pastrana, D. V., Raghavan, N., FitzGerald, P., Eisinger, S. W., Metz, C., Bucala, R., Schleimer, R. P., Bickel, C., and Scott, A. L. (1998). Filarial nematode parasites secrete a homologue of the human cytokine macrophage migration inhibitory factor, Infect Immun *66*, 5955-63.

Paterson, A. H., Bowers, J. E., Burow, M. D., Draye, X., Elsik, C. G., Jiang, C. X., Katsar, C. S., Lan, T. H., Lin, Y. R., Ming, R., and Wright, R. J. (2000). Comparative genomics of plant chromosomes, Plant Cell *12*, 1523-40.

Pennock, J. L., Behnke, J. M., Bickle, Q. D., Devaney, E., Grencis, R. K., Isaac, R. E., Joshua, G. W., Selkirk, M. E., Zhang, Y., and Meyer, D. J. (1998a). Rapid purification and characterization of L-dopachrome-methyl ester tautomerase (macrophage-migration-inhibitory factor) from Trichinella spiralis, Trichuris muris and Brugia pahangi, Biochem J *335*, 495-8.

Pennock, J. L., Wipasa, J., Gordge, M. P., and Meyer, D. J. (1998b). Interaction of macrophage-migration-inhibitory factor with haematin, Biochem J *331*, 905-8.

Perler, F. B., and Karam, M. (1986). Cloning and characterization of two Onchocerca volvulus repeated DNA sequences, Mol Biochem Parasitol *21*, 171-8.

Perry, K., and Agabian, N. (1991). mRNA processing in the Trypanosomatidae, Experientia *47*, 118-28.

Pogonka, T., Oberlander, U., Marti, T., and Lucius, R. (1999). Acanthocheilonema viteae: characterization of a molt-associated excretory/secretory 18-kDa protein, Exp Parasitol *93*, 73-81.

Post, R. J., McCall, P. J., Trees, A. J., Delves, C. J., and Kouyate, B. (1989). Chromosomes of six species of Onchocerca (Nematoda: Filarioidea), Trop Med Parasitol *40*, 292-4.

Postlethwait, J. H., Woods, I. G., Ngo-Hazelett, P., Yan, Y. L., Kelly, P. D., Chu, F., Huang, H., Hill-Force, A., and Talbot, W. S. (2000). Zebrafish comparative genomics and the origins of vertebrate chromosomes, Genome Res *10*, 1890-902.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences, Nucleic Acids Res *28*, 141-5.

Rand, J. B. (1989). Genetic analysis of the cha-1-unc-17 gene complex in Caenorhabditis, Genetics *122*, 73-80.

Ranz, J. M., Casals, F., and Ruiz, A. (2001). How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*, Genome Res *11*, 230-9.

Ravel, C., Dubessay, P., Britto, C., Blaineau, C., Bastien, P., and Pages, M. (1999). High conservation of the fine-scale organisation of chromosome 5 between two pathogenic *Leishmania* species, Nucleic Acids Res *27*, 2473-7.

Reboul, J., Vaglio, P., Tzellas, N., Thierry-Mieg, N., Moore, T., Jackson, C., Shin-i, T., Kohara, Y., Thierry-Mieg, D., Thierry-Mieg, J., *et. al.* (2001). Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in C. elegans, Nat Genet *27*, 332-6.

Redmond, D. L., and Knox, D. P. (2001). Haemonchus contortus SL2 trans-spliced RNA leader sequence, Mol Biochem Parasitol *117*, 107-10.

Richards, F. O., Boatin, B., Sauerbrey, M., and Seketeli, A. (2001). Control of onchocerciasis today: status and challenges, Trends Parasitol *17*, 558-63.

Richer, J. K., Hunt, W. G., Sakanari, J. A., and Grieve, R. B. (1993). Dirofilaria immitis: effect of fluoromethyl ketone cysteine protease inhibitors on the third- to fourth-stage molt, Exp Parasitol *76*, 221-31.

Roger, T., David, J., Glauser, M. P., and Calandra, T. (2001). MIF regulates innate immune responses through modulation of Toll-like receptor 4, Nature *414*, 920-4.

Rosengren, E., Aman, P., Thelin, S., Hansson, C., Ahlfors, S., Bjork, P., Jacobsson, L., and Rorsman, H. (1997). The macrophage migration inhibitory factor MIF is a phenylpyruvate tautomerase, FEBS Lett *417*, 85-8.

Rothstein, N., Stoller, T. J., and Rajan, T. V. (1988). DNA base composition of filarial nematodes, Parasitology *97*, 75-9.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000). Artemis: sequence visualization and annotation, Bioinformatics *16*, 944-5.

Sakaguchi, Y., Tada, I., Ash, L. R., and Aoki, Y. (1983). Karyotypes of Brugia pahangi and Brugia malayi (Nematoda: Filarioidea), J Parasitol *69*, 1090-3.

Satoskar, A. R., Bozza, M., Rodriguez Sosa, M., Lin, G., and David, J. R. (2001). Migration-inhibitory factor gene-deficient mice are susceptible to cutaneous Leishmania major infection, Infect Immun *69*, 906-11.

Schacher, J. F. (1962a). Developmental stages of *Brugia pahangi* in the final host, Journal of Parasitology *48*, 693-706.

Schacher, J. F. (1962b). Morphology of the microfilaria of *B. pahangi* and the larval stages in the mosquito, Journal of Parasitology *48*, 679-692.

Scott, A. L., Dinman, J., Sussman, D. J., Yenbutr, P., and Ward, S. (1989). Major sperm protein genes from Onchocerca volvulus, Mol Biochem Parasitol *36*, 119-26.

Selkirk, M. E., Yazdanbakhsh, M., Freedman, D., Blaxter, M. L., Cookson, E., Jenkins, R. E., and Williams, S. A. (1991). A proline-rich structural protein of the surface sheath of larval Brugia filarial nematode parasites, J Biol Chem *266*, 11002-8.

Shah, J. S., Karam, M., Piessens, W. F., and Wirth, D. F. (1987). Characterization of an Onchocerca-specific DNA clone from Onchocerca volvulus, Am J Trop Med Hyg *37*, 376-84.

Shimizu, T., Abe, R., Nakamura, H., Ohkawara, A., Suzuki, M., and Nishihira, J. (1999). High expression of macrophage migration inhibitory factor in human melanoma cells and its role in tumor cell growth and angiogenesis, Biochem Biophys Res Commun *264*, 751-8.

Siridewa, K., Karunanayake, E. H., and Chandrasekharan, N. V. (1996). Polymerase chain reaction-based technique for the detection of Wuchereria bancrofti in human blood samples, hydrocele fluid, and mosquito vectors, Am J Trop Med Hyg *54*, 72-6.

Siridewa, K., Karunanayake, E. H., Chandrasekharan, N. V., Abeyewickreme, W., Franzen, L., Aslund, L., and Pettersson, U. (1994). Cloning and characterization of a repetitive DNA sequence specific for Wuchereria bancrofti, Am J Trop Med Hyg *51*, 495-500.

Smith, H. L., and Rajan, T. V. (2000). Tetracycline inhibits development of the infective-stage larvae of filarial nematodes in vitro, Exp Parasitol *95*, 265-70.

Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. (1993). Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions, Cell *73*, 521-32.

Stamps, S. L., Taylor, A. B., Wang, S. C., Hackert, M. L., and Whitman, C. P. (2000). Mechanism of the phenylpyruvate tautomerase activity of macrophage migration inhibitory factor: properties of the P1G, P1A, Y95F, and N97A mutants, Biochemistry *39*, 9671-8.

Stec, I., Nagl, S. B., van Ommen, G. J., and den Dunnen, J. T. (2000). The PWWP domain: a potential protein-protein interaction domain in nuclear proteins influencing differentiation?, FEBS Lett *473*, 1-5.

Stec, I., Wright, T. J., van Ommen, G. J., de Boer, P. A., van Haeringen, A., Moorman, A. F., Altherr, M. R., and den Dunnen, J. T. (1998). WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a Drosophila dysmorphy gene maps in the Wolf-Hirschhorn syndrome critical region and is fused to IgH in t(4;14) multiple myeloma, Hum Mol Genet *7*, 1071-82.

Stover, N. A., and Steele, R. E. (2001). Trans-spliced leader addition to mRNAs in a cnidarian, Proc Natl Acad Sci U S A *98*, 5693-8.

Sturm, N. R., and Campbell, D. A. (1999). The role of intron structures in trans-splicing and cap 4 formation for the Leishmania spliced leader RNA, J Biol Chem *274*, 19361-7.

Sturm, N. R., Yu, M. C., and Campbell, D. A. (1999). Transcription termination and 3'-End processing of the spliced leader RNA in kinetoplastids, Mol Cell Biol *19*, 1595-604.

Sugimoto, H., Suzuki, M., Nakagawa, A., Tanaka, I., and Nishihira, J. (1996). Crystal structure of macrophage migration inhibitory factor from human lymphocyte at 2.1 A resolution, FEBS Lett *389*, 145-8.

Sugimoto, H., Taniguchi, M., Nakagawa, A., Tanaka, I., Suzuki, M., and Nishihira, J. (1999). Crystal structure of human D-dopachrome tautomerase, a homologue of macrophage migration inhibitory factor, at 1.54 A resolution, Biochemistry *38*, 3268-79.

Sulston, J. E., and Horvitz, H. R. (1977). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans, Dev Biol *56*, 110-56.

Sun, H. W., Bernhagen, J., Bucala, R., and Lolis, E. (1996). Crystal structure at 2.6-A resolution of human macrophage migration inhibitory factor, Proc Natl Acad Sci U S A *93*, 5191-6.

Sun, L. V., Foster, J. M., Tzertzinis, G., Ono, M., Bandi, C., Slatko, B. E., and O'Neill, S. L. (2001). Determination of Wolbachia genome size by pulsed-field gel electrophoresis, J Bacteriol *183*, 2219-25.

Sutton, R. E., and Boothroyd, J. C. (1986). Evidence for trans splicing in trypanosomes, Cell *47*, 527-35.

Suyama, M., and Bork, P. (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species, Trends Genet *17*, 10-3.

Suzuki, M., Sugimoto, H., Nakagawa, A., Tanaka, I., Nishihira, J., and Sakai, M. (1996). Crystal structure of the macrophage migration inhibitory factor from rat liver, Nat Struct Biol *3*, 259-66.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic Inference. In Molecular Sytematics, D. M. Hillis, C. Moritz, and B. K. Mable, eds. (Sunderland, Ma, Sinauer), pp. 407-514.

Swope, M., Sun, H. W., Blake, P. R., and Lolis, E. (1998). Direct link between cytokine activity and a catalytic site for macrophage migration inhibitory factor, Embo J *17*, 3534-41.

Szabo, G., Katarova, Z., and Greenspan, R. (1994). Distinct protein forms are produced from alternatively spliced bicistronic glutamic acid decarboxylase mRNAs during development, Mol Cell Biol *14*, 7535-45.

Tan, T. H., Edgerton, S. A., Kumari, R., McAlister, M. S., Rowe, S. M., Nagl, S., Pearl, L. H., Selkirk, M. E., Bianco, A. E., Totty, N. F., *et. al.* (2001). Macrophage migration inhibitory factor of the parasitic nematode Trichinella spiralis, Biochem J *357*, 373-83.

Tawe, W., Pearlman, E., Unnasch, T. R., and Lustigman, S. (2000). Angiogenic activity of Onchocerca volvulus recombinant proteins similar to vespid venom antigen 5, Mol Biochem Parasitol *109*, 91-9.

Taylor, A. B., Johnson, W. H., Jr., Czerwinski, R. M., Li, H. S., Hackert, M. L., and Whitman, C. P. (1999). Crystal structure of macrophage migration inhibitory factor complexed with (E)-2-fluoro-p-hydroxycinnamate at 1.8 A resolution: implications for enzymatic catalysis and inhibition, Biochemistry *38*, 7444-52.

Taylor, M. J., Cross, H. F., and Bilo, K. (2000). Inflammatory responses induced by the filarial nematode Brugia malayi are mediated by lipopolysaccharide-like activity from endosymbiotic Wolbachia bacteria, J Exp Med *191*, 1429-36.

Tchavtchitch, M., Fischer, K., Huestis, R., and Saul, A. (2001). The sequence of a 200 kb portion of a Plasmodium vivax chromosome reveals a high degree of conservation with Plasmodium falciparum chromosome 3, Mol Biochem Parasitol *118*, 211-22.

Thacker, C., Marra, M. A., Jones, A., Baillie, D. L., and Rose, A. M. (1999). Functional genomics in *Caenorhabditis elegans*: An approach involving comparisons of sequences from related nematodes, Genome Res *9*, 348-59.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology., Science *282*, 2012-8.

Thomas, W. K., and Wilson, A. C. (1991). Mode and tempo of molecular evolution in the nematode caenorhabditis: cytochrome oxidase II and calmodulin sequences, Genetics *128*, 269-79.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res *25*, 4876-82.

Townson, S., Hutton, D., Siemienska, J., Hollick, L., Scanlon, T., Tagboto, S. K., and Taylor, M. J. (2000). Antibiotics and Wolbachia in filarial nematodes: antifilarial activity of rifampicin, oxytetracycline and chloramphenicol against Onchocerca gutturosa, Onchocerca lienalis and Brugia pahangi, Ann Trop Med Parasitol *94*, 801-16.

Treinin, M., Gillo, B., Liebman, L., and Chalfie, M. (1998). Two functionally dependent acetylcholine subunits are encoded in a single Caenorhabditis elegans operon, Proc Natl Acad Sci U S A *95*, 15492-5.

Triantaphyllou, A. C., and Moncol, D. J. (1977). Cytology, reproduction, and sex determination of Strongyloides ransomi and S. papillosus, J Parasitol *63*, 961-73.

Underwood, A. P., and Bianco, A. E. (1999). Identification of a molecular marker for the Y chromosome of Brugia malayi, Mol Biochem Parasitol *99*, 1-10.

Unnasch, T. R., and Williams, S. A. (2000). The genomes of Onchocerca volvulus, Int J Parasitol *30*, 543-52.

Vandenberghe, A. E., Meedel, T. H., and Hastings, K. E. (2001). mRNA 5'-leader trans-splicing in the chordates, Genes Dev *15*, 294-303.

Vanfleteren, J. R., Van de Peer, Y., Blaxter, M. L., Tweedie, S. A., Trotman, C., Lu, L., Van Hauwaert, M. L., and Moens, L. (1994). Molecular genealogy of some nematode taxa as based on cytochrome c and globin amino acid sequences, Mol Phylogenet Evol *3*, 92-101.

Vassilev, I., and Mutafova, T. (1974). Comparative studies on the karyotype of Ascaris suum and "Ascaris ovis", Z Parasitenkd *43*, 115-21.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et. al.* (2001). The sequence of the human genome, Science *291*, 1304-51.

Walker, D. R., and Koonin, E. V. (1997). SEALS: a system for easy analysis of lots of sequences, Proc Int Conf Intell Syst Mol Biol *5*, 333-9.

Walker, W. H., Girardet, C., and Habener, J. F. (1996). Alternative exon splicing controls a translational switch from activator to repressor isoforms of transcription factor CREB during spermatogenesis, J Biol Chem *271*, 20145-1050.

Ware, J., Moran, L., Foster, J., Posfai, J., Vincze, T., Guiliano, D., Blaxter, M., Eisen, J., and Slatko, B. (2002). Sequencing and analysis of a 63 kb bacterial artificial chromosome insert from the Wolbachia endosymbiont of the human filarial parasite Brugia malayi, Int J Parasitol *32*, 159-66.

Weiser, W. Y., Temple, P. A., Witek-Giannotti, J. S., Remold, H. G., Clark, S. C., and David, J. R. (1989). Molecular cloning of a cDNA encoding a human macrophage migration inhibitory factor, Proc Natl Acad Sci U S A *86*, 7522-6.

Williams, C., Xu, L., and Blumenthal, T. (1999). SL1 trans splicing and 3'-end formation in a novel class of Caenorhabditis elegans operon, Mol Cell Biol *19*, 376-83.

Williams, S. A. (1999). Deep within the filarial genome: progress of the filarial genome project, Parasitol Today *15*, 219-24.

Williams, S. A., Lizotte-Waniewski, M. R., Foster, J., Guiliano, D., Daub, J., Scott, A. L., Slatko, B., and Blaxter, M. L. (2000). The filarial genome project: analysis of the nuclear, mitochondrial and endosymbiont genomes of Brugia malayi, Int J Parasitol *30*, 411-9.

Witt, C., and Ottesen, E. A. (2001). Lymphatic filariasis: an infection of childhood, Trop Med Int Health *6*, 582-606.

Woods, I. G., Kelly, P. D., Chu, F., Ngo-Hazelett, P., Yan, Y. L., Huang, H., Postlethwait, J. H., and Talbot, W. S. (2000). A comparative map of the zebrafish genome, Genome Res *10*, 1903-14.

Xie, H., Bain, O., and Williams, S. A. (1994). Molecular phylogenetic studies on filarial parasites based on 5S ribosomal spacer sequences, Parasite *1*, 141-51.

Xu, Y., Liu, L., and Michaeli, S. (2000). Functional analyses of positions across the 5' splice site of the trypanosomatid spliced leader RNA. Implications for base-pair interaction with U5 and U6 snRNAs, J Biol Chem *275*, 27883-92.

Xue, Y., Canman, J. C., Lee, C. S., Nie, Z., Yang, D., Moreno, G. T., Young, M. K., Salmon, E. D., and Wang, W. (2000). The human SWI/SNF-B chromatin-remodeling complex is related to yeast rsc and localizes at kinetochores of mitotic chromosomes, Proc Natl Acad Sci U S A *97*, 13015-20.

Yenbutr, P., and Scott, A. L. (1995). Molecular cloning of a serine proteinase inhibitor from Brugia malayi, Infect Immun *63*, 1745-53.

Zang, X., Yazdanbakhsh, M., Jiang, H., Kanost, M. R., and Maizels, R. M. (1999). A novel serpin expressed by blood-borne microfilariae of the parasitic nematode Brugia malayi inhibits human neutrophil serine proteinases, Blood *94*, 1418-28.

Zeng, F. Y., Weiser, W. Y., Kratzin, H., Stahl, B., Karas, M., and Gabius, H. J. (1993). The major binding protein of the interferon antagonist sarcolectin in human placenta is a macrophage migration inhibitory factor, Arch Biochem Biophys *303*, 74-80.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences, J Comput Biol *7*, 203-14.

Zilka, A., Garlapati, S., Dahan, E., Yaolsky, V., and Shapira, M. (2001). Developmental regulation of heat shock protein 83 in Leishmania. 3' processing and mRNA stability control transcript abundance, and translation id directed by a determinant in the 3'-untranslated region, J Biol Chem *276*, 47922-9.

Zorio, D. A., Cheng, N. N., Blumenthal, T., and Spieth, J. (1994). Operons as a common form of chromosomal organization in C. elegans, Nature *372*, 270-2.