2017

Design of Novel Algorithms for Comparative Analysis of Complex Gene Families and their Application to Nematode Detoxification Pathways

Curran, David Michael

UNIVERSITY OF CALGARY

Design of Novel Algorithms for Comparative Analysis of Complex Gene Families and their Application to Nematode Detoxification Pathways

by

David Michael Curran

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN VETERINARY MEDICAL SCIENCES

CALGARY, ALBERTA

JANUARY, 2017

Abstract

Parasitic nematodes present a current and devastating global problem, infecting billions of people, and causing huge production losses in both crops and livestock. There are a limited number of anthelmintic drugs available to treat these infections, and resistance has arisen quickly and spread across the globe. Xenobiotic metabolism is a well-known mechanism of drug resistance in insects, and evidence suggests it may also play a role in the development of drug resistance in parasitic nematodes. Identifying candidate enzymes in the free-living nematodes may help to understand or combat the rising resistance crisis in the parasites. However, identifying many of the protein sequences that may be involved in xenobiotic metabolism can prove challenging due to high sequence divergence and draft quality genome assemblies. This work focuses on novel software to detect hard-to-find genes, as well as methods of performing comparative phylogenetic analyses, both between species and within a population.

In the absence of specific selective pressures, the phylogeny of a multi-species gene family will tend to agree with the underlying species tree. Conversely, adaptive evolution tends to manifest as incongruence in the gene tree as well as lineage-specific expansions and contractions. These properties, collectively termed phylogenetic instability, have been found to be good predictors of proteins that directly interact with the environment. I have developed an algorithm to quantify phylogenetic instability in a gene tree, and show that it correlates exceptionally well with the known substrate specificity of human cytochrome P450 enzymes. I then apply this technique to five detoxification gene families (*cyp, fmo, sdr, ugt, gst*) from five free-living nematode species (*Caenorhabditis elegans*, *C. briggsae*, *C. brenneri*, *C. remanei*, and *Pristionchus pacificus*). These gene families are known to act on both endogenous and xenobiotic molecules, and these new methods allow me to predict which are likely involved in xenobiotic metabolism. This will aid in the study of these enzymes, including their orthologs in the parasitic species.

Acknowledgements

I am very appreciative of the support and guidance I received from my supervisors Drs. James Wasmuth and John Gilleard, without whom I would have never have learned to be the researcher I am today. They took a chance on a molecular biologist that badly wanted to become a bioinformatician, and for that I will always be grateful.

Thanks as well to my committee, Drs. Guido van Marle, Gordon Chua, and Alastair Cribb. The tough questions and helpful answers I've received over the past four years have been incredibly important to my graduate experience. I am lucky to have had Drs. Edwin Wang and Dannie Durand as my examiners, both bringing invaluable experience and expertise. In particular, I appreciate Dr. Durand's enthusiasm as I took some of her excellent work and modified it to my own purposes.

Last but certainly not least, I am immensely grateful to my family who have always stood by me during the highs and lows of this degree. I can't imagine how much harder this PhD would have been without my amazing wife and fellow scientist Jamie by my side. She supported me when things were tough, celebrated with me when things were good, talked science with me over beers, and provided an ear when I absolutely had to wax loquacious on obscure technical points nobody else wanted to hear.

Table of Contents

List of Tables

List of Figures

List of Symbols, Abbreviations, Nomenclature

Abbreviation	Definition
AR	Armin Rouhi; summer student.
CEGMA	Core Eukaryotic Genes Mapping Approach.
CNV	Copy-number variation.
Cyp	Cytochrome P450.
DMC	David Michael Curran; me.
Fmo	Flavin-containing monooxygenase
Gst	Glutathione S-transferase.
HGT	Horizontal gene transfer.
HSP	High-scoring segment pair.
HMM	Hidden Markov model.
ILS	Incomplete lineage sorting.
JDW	Dr. James D. Wasmuth; my supervisor.
JSG	Dr. John S. Gilleard; my co-supervisor.
Pgp	P-glycoprotein.
pHMM	Profile hidden Markov model.
Sdr	Short chain dehydrogenase.
Ugt	UDP-glycosyltransferases.
WHO	World Health Organization.

Epigraph

In short, if all the matter in the universe except the nematodes were swept away, our world would still be dimly recognizable, and if, as disembodied spirits, we could then investigate it, we should find its mountains, hills, vales, rivers, lakes, and oceans represented by a film of nematodes.

The location of towns would be decipherable, since for every massing of human beings there would be a corresponding massing of certain nematodes. Trees would still stand in ghostly rows representing our streets and hiways. The location of the various plants and animals would still be decipherable, and, had we sufficient knowledge, in many cases even their species could be determined by an examination of their erstwhile nematode parasites.

— Nathan Cobb, Nematodes and their relationships (1914)

1. Introduction

The aim of this thesis is to study the evolution of large gene families, how an organism's environment might shape this evolution, and apply this analysis to the detoxification genes in nematodes. Studying gene family evolution is not a new endeavour, but pursuing this analysis required the generation of several new software applications. These algorithms and programs represent the bulk of the novel contribution of this work. I also present my characterization of the xenobiotic potential (the likelihood that a protein acts on or responds to a molecule produced outside of the organism) of these detoxification gene families, as those with the strongest signals of adaptive evolution are most likely to be acting on molecules from the environment. Nematodes represent ideal organisms to study, as the tractability of the free-living species in laboratory settings has led to an extensive body of knowledge and genetic resources. This work is motivated by the parasitic species that present a current and devastating global problem, infecting billions of people, and causing huge production loss of crops and livestock. There are a limited number of anthelmintic drugs available to treat these infections, and resistance has arisen quickly and spread across the globe. I hypothesize that drug metabolism is playing a major role in the development of this resistance, and that it is due to detoxification enzymes with high xenobiotic potential.

1.1 Nematodes

Nematodes are unsegmented roundworms, and are one of the most abundant forms of multicellular life on the planet. They have been found in nearly every biosphere, including many traditionally considered to be the realm of single-cell organisms; even kilometers below the surface in a South African gold mine, where they thrive by feeding on bacteria (Borgonie *et al.*, 2011). These animals are very diverse, ranging from sub-millimetre to several meters in length, and are typically phylogenetically classified into five clades with approximately half designated as parasites. Though their environmental niches vary wildly, they are very constrained morphologically. *Caenorhabditis elegans* (*C. elegans*) and *C. briggsae* are two free-living nematodes that are difficult for trained

taxonomists to differentiate visually, yet they are estimated to have diverged as species 80-110 million years ago (Stein *et al.*, 2003). This number is likely to be an overestimate (Cutter, 2015), but if true it would be older even than the divergence times between mouse and human.

1.1.1 Free living nematodes

One of the most prevalent species in biology laboratories today is the nematode *C. elegans*. It was selected in the 1960s by Sydney Brenner as a model to study neurobiology, work for which he shared the 2002 Nobel Prize in Physiology or Medicine. In a proposal to the Medical Research Council in 1963 he wrote “Thus we want a multicellular organism which has a short life cycle, can be easily cultivated, and is small enough to be handled in large numbers, like a microorganism. It should have relatively few cells, so that exhaustive studies of lineage and patterns can be made, and should be amenable to genetic analysis” (Félix, 2008). All of these properties are still important today, and recent technological advances continue to favour the organism; that it is transparent for its entire life cycle means that fluorescent proteins can be put to good use, and it has turned out to be one of the most tractable organisms for genetic manipulation in the form of RNA interference. *C. elegans* can be easily maintained in the laboratory, storing individuals on agar plates at room temperature for several months. Once their food supply of *E. coli* has been exhausted, the larvae enter a dauer stage of arrested development where they are protected against desiccation and other environmental stresses. A piece of this media can be simply transferred to a new plate containing a fresh lawn of *E. coli*, at which point the nematodes will resume their development.

C. elegans was originally isolated by Sydney Brenner from a mushroom compost in Bristol, UK, but it can be found in rotting plant matter from temperate regions in every continent except Antarctica (Blaxter and Denver, 2012). Its genome was sequenced in 1998 (The *C. elegans* Sequencing Consortium, 1998), the first from a multicellular organism. Unlike most eukaryotes, genes in *C. elegans* and some related nematodes are occasionally organized in polycistronic transcriptional units, or operons (Spieth *et*

al., 1993; Guiliano and Blaxter, 2006). While many other species of *Caenorhabditis* are dioecious (reproduce with male-female pairs), populations of *C. elegans* are androdioecious (self-fertile hermaphrodites coexisting with rare males), an adaptation that appears to have arisen independently in *C. elegans* and *C. briggsae* (Kiontke *et al.*, 2004). It is thought that *C. elegans* associates with many different invertebrates and arthropods primarily as a means of long-distance dispersal (termed phoresy), as it has been shown to readily disembark in favourable conditions (Lee *et al.*, 2011). Microsatellite data have found identical genotypes distributed not only throughout a country, but even between continents (Haber *et al.*, 2005). This suggests that humans are likely aiding in the distribution, which may not be surprising as *C. elegans* is almost exclusively found in anthropogenic habitats (Barrière and Félix, 2005).

C. briggsae is related to *C. elegans* and shares many of the same characteristics that make it an excellent model organism. It also reproduces via self-fertilizing hermaphrodites, the genome was published 13 years ago (Stein *et al.*, 2003), and the current assembly is of extremely high quality. It can often be isolated in conjunction with *C. elegans*, even from the same piece of fruit (Félix and Duveau, 2012). These species appear to take a temporal approach to avoiding competition, with *C. briggsae* dominating in the summers and *C. elegans* taking over during the fall. These seasonal shifts also correlate with their temperature preferences *in vitro*. This species has been thought to live in necromenic association with snails, where dauer larvae colonize a host until it dies, at which point the nematodes resume their development and feed on the corpse (Kiontke and Sudhaus, 2006). However, a recent study suggests that at least some strains appear to be entomopathogenic, which is where the dauer larvae in conjunction with certain bacterial species hasten the demise of an insect host; development of the nematode then resumes and they feed on the corpse (Abebe *et al.*, 2010).

C. remanei (formerly *C. vulgaris*) has been isolated along with *C. briggsae* and *C. elegans*, though never both at the same time. It has primarily been found in North America and Germany in compost heaps, and so its natural habitat remains unknown (Baird *et al.*, 1994; Baird, 1999). It has been found associated with terrestrial molluscs

and isopods, and it has been suggested that this is a phoretic association. In contrast to *C. elegans* and *C. briggsae*, both *C. remanei* and *C. brenneri* are gonochoristic and reproduce via male-female couplings. The current version of the *C. remanei* genome assembly was completed 2007 and is of good quality.

C. brenneri (formerly *C. sp 4*, *C. sp CB5161,* or *C. sp PB2801*) is the most recently described species of those studied in this work (Sudhaus and Kiontke, 2007). It is virtually morphologically identical to *C. remanei*, which is why it was not formally described for 30 years. Extensive mating tests eventually showed that the two species are reproductively isolated. The most recent genomic assembly was completed in 2010 and is of good quality. Its natural distribution appears to mirror that of *C. remanei*, but south of the Tropic of Cancer instead of to the north (Sudhaus and Kiontke, 2007). Unexpectedly, this organism is reported to harbour the most intra-species molecular diversity of any eukaryote, at levels approaching those of hyperdiverse bacteria (Dey *et al.*, 2013).

Pristionchus pacificus is a related free-living nematode, though it is estimated to have diverged from the *Caenorhabditis* ancestor species between 280 and 430 million years ago (Dieterich *et al.*, 2008). Like *C. elegans* and *C. briggsae*, *P. pacificus* is hermaphroditic (Herrmann *et al.*, 2006). Its genome was published in 2008, and is currently of good, but slightly lower quality than *C. remanei* or *C. brenneri* (Dieterich *et al.*, 2008). It has twice the median number of introns per gene compared to *C. elegans*, and the median intron size is twice as long. Compared to *C. elegans*, it has major expansions in several classes of protein families involved in xenobiotic metabolism including the cytochrome P450s, UDP-glucoronosyl transferases, and ABC transporters. Originally it was thought to live exclusively in a necromenic association with the oriental beetle (Herrmann *et al.*, 2006; Cinkornpumin *et al.*, 2014), but recently it has been found to also be a facultative predator of other nematodes (Serobyan *et al.*, 2014).

1.1.2 Parasitic nematodes

The parasitic nematodes, along with the parasitic cestodes and trematodes, form the group of organisms termed the helminths. It is thought that the ancestor nematode species was likely free-living, and that parasitism evolved multiple independent times throughout the evolution of this family (Blaxter *et al.*, 1998). While the evolutionary details are still under discussion, parasitic nematodes are currently a serious global problem. Two billion people are estimated to be infected, and though deaths are generally low, the impact in terms of morbidity rivals that of malaria, HIV/AIDS, or tuberculosis (World Health Organization, 2005; Lustigman *et al.*, 2012). Of the neglected tropical diseases, the soil-transmitted helminths *Ascaris lumbricoides*, *Trichuris trichiura*, *Necator americanus*, and *Ancylostoma duodenale* are responsible for the highest burden in humans (Murray *et al.*, 2012). The effects of an infection can be insidious, as it has been suggested that the blood-feeding hookworms may be responsible for up to half of the literacy gap between Southern and Northern USA (Bundy *et al.*, 2013). Parasitic nematodes also cause billions of dollars of production loss in livestock (Stromberg and Gasbarre, 2006; Sutherland and Leathwick, 2011). The situation in small ruminants is particularly serious, primarily due to the prevalence and pathogenicity of the trichostrongylid *Haemonchus contortus*, which can severely impact income and food in resource-poor settings (Vattaa and Lindberg, 2006). The impact goes beyond animals, and nematodes that parasitize plants may be responsible for the destruction of up to 15% of the world's crops (Trudgill and Blok, 2001; Abad *et al.*, 2008; Atkinson *et al.*, 2012).

1.2 Anthelmintics

The WHO only recommends four drugs for the treatment of the soil-transmitted helminth infections in humans: albendazole and mebendazole (both benzimidazoles), levamisole, and pyrantel pamoate (World Health Organization, 2006). There are only three other drugs recommended for treatment of any helminth in humans: ivermectin, diethylcarbamazine, and praziquantel. These seven drugs belong to six classes of

anthelmintic, and are pictured in **Figure 1.1** along with two additional classes used in veterinary medicine.

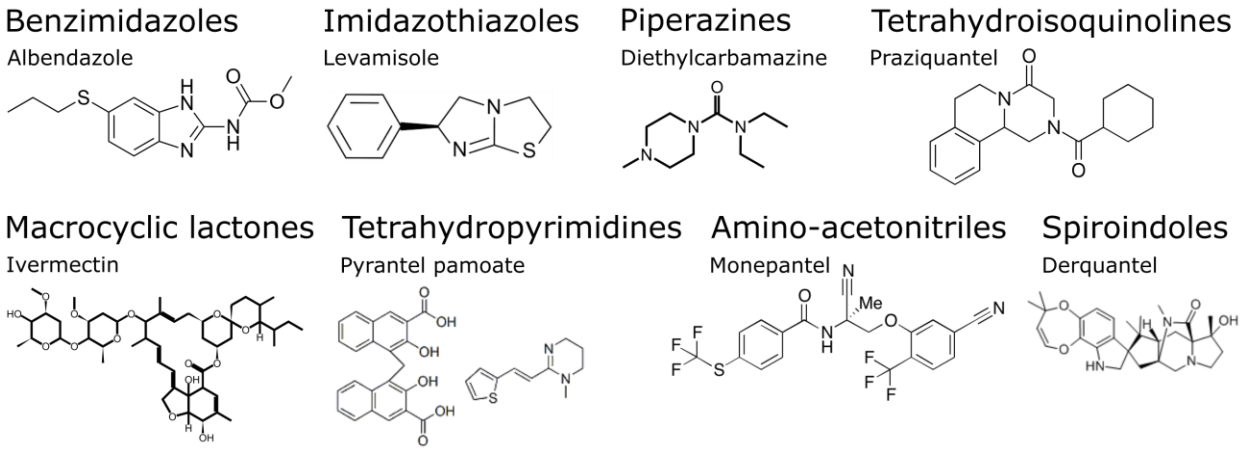


Figure 1.1 Chemical structures of current anthelmintics. Examples of the six classes of anthelmintics licensed for use in humans, and two used only in veterinary medicine (monepantel and derquantel).

The mechanism of action for most of these anthelmintics has been elucidated over the years. The benzimidazoles act on the helminth β-tubulin protein inhibiting its polymerization into microtubules, and leading to immobilization of the parasite (Driscoll *et al.*, 1989; Beech *et al.*, 2011). The macrocyclic lactones act on glutamate and GABA-gated chloride channels, thereby inhibiting pharyngeal pumping, egg laying, and motility of the parasite (Martin and Pennington, 1989; Forrester *et al.*, 2002). The imidazothiazoles, tetrahydropyrimidines, and amino-acetonitriles all act as agonists on members of the nicotinic acetylcholine-gated cation channels, leading to rigid paralysis of the parasite (Martin *et al.*, 2004). Conversely, the spiroindoles act as antagonists to these channels which also interferes with the parasite motility (Lee *et al.*, 2002; Johnson *et al.*, 2004). The tetrahydroisoquinolines increase calcium permeability through an unknown mechanism, leading to spastic paralysis; they also cause disintegration of the tegument in *Schistosoma* species (Pax *et al.*, 1978; Becker *et al.*, 1980). The mechanism of action of the piperazines is still unclear even after nearly six decades of use, but appears to alter host arachidonic acid and nitric oxide pathways, which somehow leads to immobilization of the parasite (Buxton *et al.*, 2014).

Mass drug administration programs have been launched in an attempt to eradicate certain human parasites in endemic regions. Ivermectin has been deployed against river blindness (*Onchocerca volvulus*) (Gustavsen *et al.*, 2011; Tekle *et al.*, 2012), praziquantel against schistosomiasis (Rollinson *et al.*, 2013), and albendazole and mebendazole against hookworms (both *Necator americanus* and *Ancylostoma duodenale*) and ascariasis (Keiser and Utzinger, 2010; Levecke *et al.*, 2014).

1.2.1 Anthelmintic resistance

Until the introduction of monepantel (Kaminsky *et al.*, 2008), there had not been a new class of anthelmintic since ivermectin 30 years ago. The advances since have occurred almost entirely in the field of veterinary medicine, as every licensed human anthelmintic with the exception of diethylcarbamazine was developed first for use in animals before being adopted for humans (Geary, 2012). This lack of new drugs poses a major problem, as resistance has arisen rapidly (**Table 1.1**) and spread around the globe (**Table 1.2**); it is no longer uncommon to find sheep or goat farms where the animals are infected with helminths exhibiting high levels of resistance to all available anthelmintic drugs, commonly *H. contortus* (Kaplan, 2004; Kaplan and Vidyashankar, 2012). To combat this, novel approaches are being used to identify potential new anthelmintics. One recent study has used large chemical libraries to identify compounds that impact *C. elegans* more strongly than mammalian cells, and has identified 14 novel compounds, many of which appear to act via mechanisms distinct from current anthelmintics (Mathew *et al.*, 2016).

Table 1.1 Interval between anthelmintic introduction and resistance. The report of resistance column indicates the first documented case of resistance, though in most cases there were earlier reports suspecting resistance. This table has been adapted from (Kaplan, 2004).

Anthelmintic	Host	Initial approval	Report of resistance
Thiabendazole	Sheep	1961	1964
	Horse	1962	1965
Levamisole	Sheep	1970	1979
Pyrantel	Horse	1974	1996
Ivermectin	Sheep	1981	1988
	Horse	1983	2002
Moxidectin	Sheep	1991	1995
	Horse	1995	2003

Table 1.2 Percent of sheep and goat farms reporting anthelmintic resistance by country. The 'combination' column indicates resistance was reported when administering a combination of all four drugs at once. N/A indicates that no data were available. These data were collected from many reports summarized in (Kaplan and Vidyashankar, 2012).

Country	Benzimidazole	Levamisole	Ivermectin	Moxidectin	Combination
USA	98%	54%	85%	37%	48%
Brazil	100%	100%	100%	100%	95%
Australia	100%	100%	80%	30%	N/A
New Zealand	41%	24%	25%	N/A	8%

The first report of resistance against the benzimidazoles came from sheep infected with *H. contortus* (Conway, 1964), and the mechanism of this resistance has been studied in detail ever since. It has been established that much of this effect is due to one of three single amino acid mutations in the β-tubulin gene (Kwa *et al.*, 1994, 1995; Ghisi

et al., 2007; Saunders *et al.*, 2013), though other studies have implicated additional mechanisms such as drug efflux (Blackhall *et al.*, 2008; Lespine *et al.*, 2012). One study induced resistance to ivermectin in *C. elegans* by gradually increasing the concentration over several months, and found the effect was due in part to increased expression of *pgp-1*, an efflux pump (James and Davey, 2009). Resistance to the other classes of anthelmintic drugs is not as well understood and many genome-wide association studies have failed to identify a conserved mechanism, suggesting that the mechanism may be more complex (Gilleard, 2013). In mosquitos and flies, wild isolates with resistance to the insecticide dichlorodiphenyltrichloroethane (DDT) do so by modification or overexpression of a cytochrome P450 (*cyp*) gene (Dunkov *et al.*, 1997; Amichot *et al.*, 2004; Chiu *et al.*, 2008). Evidence also suggests that drug metabolism plays an important mechanistic role in resistance in the trematode *Fasciola hepatica* (Devine *et al.*, 2009; Alvarez *et al.*, 2005), so I hypothesize that this may be an important mechanism in the rest of the helminths.

1.3 Xenobiotic metabolism

Xenobiotics are compounds that an organism encounters that do not serve as a source of energy or as precursors for biomolecules. As the free-living *Caenorhabditis* species feed on the bacteria present around rotting organic matter they are constantly exposed to an array of xenobiotics, including toxic molecules produced by bacteria, yeast, and other microbes (Alegado *et al.*, 2003). In order to survive this environment, they have developed an extensive array of detoxification enzymes. Xenobiotic metabolism proceeds in three phases (Omiecinski *et al.*, 2011). The first phase is functionalization, in which the compound is metabolized to expose some reactive group. This is carried out by protein families including the Cyps, the Flavin-containing monooxygenases (Fmos), and short-chain dehydrogenases (Sdrs). The second phase is conjugation, which involves adding an endogenous hydrophilic molecular moiety. This is performed by protein families such as the UDP-glycosyltransferases, glutathione-S-transferases, and sulfotransferases. The third phase is excretion of the substrate out of the cell by one of several ABC transporter proteins, often a member of the P-

glycoproteins. Phase I reactions may or may not be necessary prior to Phase II, and different compounds may be able to be excreted at any point in the detoxification process.

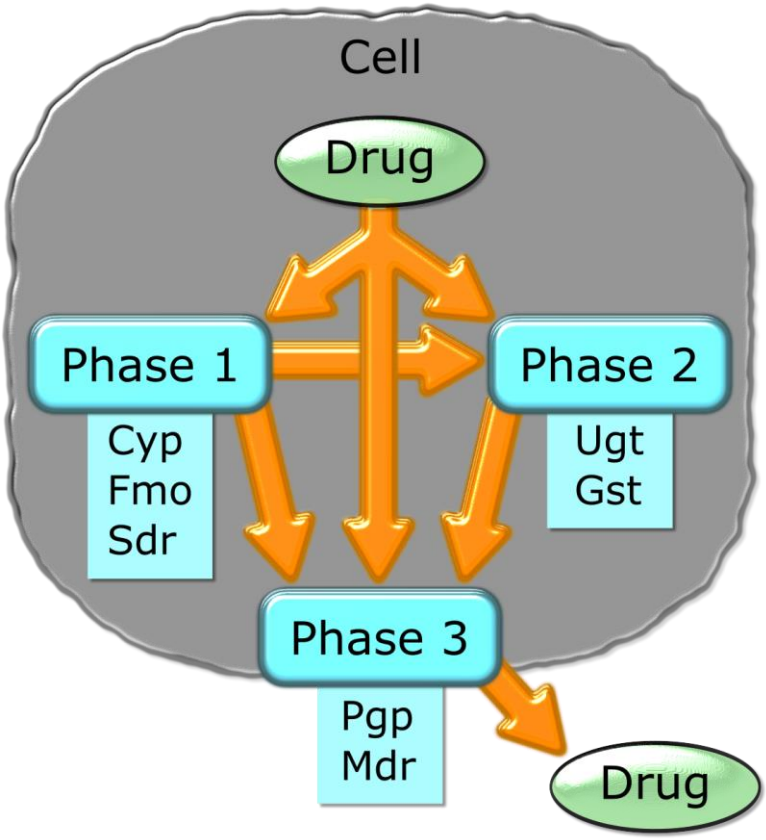


Figure 1.2 Schematic of xenobiotic metabolism. This figure shows how the detoxification gene families act in concert to protect the cell from harmful xenobiotics.

In contrast to free-living species, parasites typically experience a narrower array of xenobiotics. Most do not spend significant portions of their life cycles feeding in the environment, but instead devote their genetic repertoire to surviving within their host species. This concept appears to be supported by the general trend of free-living species having larger gene families involved with xenobiotic metabolism (Matoušková *et al.*, 2016). While parasites may have smaller gene families, there has been mounting evidence that this capacity has not been lost; *H. contortus* has the ability to metabolize benzimidazoles (Cvilink *et al.*, 2008 and Stasiuk, unpublished), as do the flukes *Dicrocoelium dendriticum* (V. Cvilink *et al.*, 2009) and *Fascioloides magna* (Prchal *et al.*,

2016), and the tapeworms *Moniezia expansa* (Prchal *et al.*, 2015) and *Mesocestoides vogae* (Munguía *et al.*, 2015). Resistant isolates of the liver fluke *Fasciola hepatica* have been shown to metabolize albendazole and triclabendazole 21% better than susceptible isolates, which suggests metabolism may be a major mechanism of resistance for this organism (Robinson *et al.*, 2004; Scarcella *et al.*, 2012, 2013). Studies have also suggested that *H. contortus* possesses the ability to metabolize moxidectin (Alvinerie *et al.*, 2001) and monepantel (Stuchlikova *et al.*, 2014).

1.2.1 Cytochrome P450s

This gene family has been well-studied since its identification in 1958 (Klingenberg, 1958; Omura, 1999; Nelson, 2009, 2011). These oxidase enzymes can be found in all kingdoms of life, even including a virus, and are found in every known species of eukaryote except for the protists (Nelson, 2011). It is a generally large gene family, with 59 members in humans and 75 in *C. elegans*. The general mode of action is to bind a redox partner, normally cytochrome P450 reductase, which donates two electrons. These are used to split an oxygen molecule at the Cyp's heme group, and the resulting hydroxyl group is attached to the substrate (Munro *et al.*, 2003).

While many of these enzymes do function as oxidases, other members of the family can act as peroxidases or reductases. The Cyps are known to catalyze a wide range of reactions, acting on endogenous molecules such as fatty acids, steroids, and prostaglandins, as well as xenobiotics including drugs, anaesthetics, ethanol, and many more (Bernhardt, 2006). The substrate specificity of a single enzyme can be staggeringly permissive; human Cyp-3A4 is estimated to be involved in the metabolism of 50%, and Cyp-2D6 of 30%, of all prescribed drugs (Guengerich, 2003; Krau, 2013).

For many years it was thought that helminths lacked the capacity for oxidative metabolism, in particular that they had no xenobiotic Cyp enzymes (Precious and Barrett, 1989; Pemberton and Barrett, 1989; Barrett, 1998, 2009). Recent work has called these conclusions into question as many parasites have been found with dozens of *cyp* genes in their genomes (Laing *et al.*, 2013, 2015). Further, it has been shown

that triclabendazole-resistant isolates of the liver fluke *Fasciola hepatica* can be made more susceptible by administration of either of two Cyp inhibitors (Devine *et al.*, 2010, 2012).

1.3.2 Flavin-containing monooxygenases

The Fmos are thought to be a very ancient gene family, as homologs were identified in such diverse organisms as bacteria, protozoa, mammals, and sharks (Petalcorin *et al.*, 2005). Subsequent genome sequencing has confirmed this hypothesis, with Fmo sequences being identified in essentially all bacteria, fungi, animals, and plants. It is a much smaller gene family than the Cyps, with humans and *C. elegans* possessing five proteins each. *C. remanei* appears to have five genes as well, though the other *Caenorhabditis* only have four; *P. pacificus* on the other hand appears to have 10 or 11 (see **Figure 5.5**).

Though they carry out similar reactions and require oxygen and NADPH as cofactors, the Fmos are a class of enzymes distinct from the Cyps (Ziegler and Mitchell, 1972; Krueger and Williams, 2005). Compared to Cyps, Fmos seem to prefer substrates with nucleophilic nitrogen or sulphur atoms. They have a somewhat unusual mechanism where the energy to carry out the oxidation reaction is present in the enzyme prior to the substrate binding, which means that any soft nucleophile that can physically fit into the active site is a candidate for metabolism (Ziegler, 2002). Like the Cyps, the Fmos have evolved incredibly broad specificity and are capable of metabolizing thousands of different substrates; the two families are able to act on many of the same substrates, though they often produce different metabolites (Krueger and Williams, 2005).

1.3.3 Short chain dehydrogenases

Alcohols, ketones, and aldehydes are typically metabolized by Sdrs, medium-chain dehydrogenases, and aldo-keto reductases, which are all distinct enzyme superfamilies (Jez and Penning, 2001; Persson *et al.*, 2009). Much like the Cyps the Sdrs appear to be very ancient, appearing in all kingdoms of life, with low pairwise sequence identities (Jornvall *et al.*, 1999; Jörnvall *et al.*, 2010). In contrast to the Cyps, these enzymes have

been considered important xenobiotic metabolizers in helminths for many years (Barrett, 1997). *H. contortus* has been shown to be able to reduce the anthelmintic flubendazole, which may be via a Sdr enzyme (Vokral *et al.*, 2010). The Sdrs are a massive superfamily, containing a quarter of all known dehydrogenases (Kallberg and Persson, 2006). Members are most easily identified by motifs and their conserved tertiary structures, in particular the presence of a Rossmann fold (Kavanagh *et al.*, 2008). The Sdrs are made of many protein families, including the *Drosophila* alcohol dehydrogenases (but not the human version) and the Flavin reductases (Kallberg *et al.*, 2010).

Sdrs have very diverse substrates, with functions ranging from the metabolism of sugars and steroids, to the synthesis of antibiotics, to the detoxification of xenobiotics like aromatic ketones and quinones (Hoffmann and Maser, 2007). Humans have between 73 and 82 Sdrs (Bray *et al.*, 2009; Kallberg *et al.*, 2010), and based on my analysis the free-living nematodes possess a similar range of between 78 and 119 (see **Table 5.2**). Though the function of half of the human Sdrs are unknown, many are predicted to have important endogenous roles as one third are implicated in diseases.

1.3.4 Uridine 5'-diphospho-glycosyltransferases

The Ugts are the most important phase II detoxification enzymes, catalyzing the addition of various UDP sugar moieties to their substrates (Meech *et al.*, 2012; Rowland *et al.*, 2013; Matoušková *et al.*, 2016). This protein family is also ancient, and can be found in bacteria, animals, and plants. One hypothesis to explain the wide array of substrates metabolized by these proteins is that like Cyps, they are the result of an evolutionary arms race between ancient plants and insects (Gonzalez and Nebert, 1990; Meech *et al.*, 2012). This idea is strengthened by the apparent convergent evolution of cyanogenic glycosides as a defense mechanism in many plants and the Burnet moth; two Cyps and one Ugt make up the entirety of the biosynthetic pathway in the moth (Jensen *et al.*, 2011). The majority of insect Ugts are involved in xenobiotic metabolism, primarily of plant defense molecules (Meech *et al.*, 2012). Plants also use

an array of Ugts for more mundane processes, including the production of signalling molecules and the alteration of those molecules to direct their sequestration.

Ugts may act on a wide array of nucleophilic atoms such as aliphatic alcohols, phenols, carboxylic acids, aromatic and aliphatic amines, thiols, and acidic carbons (Rowland *et al.*, 2013). Ugts are involved in many vital endogenous reactions, such as the production of bilirubin, bile acids, fatty acids, steroids, thyroid hormones, and fat-soluble vitamins (Tukey and Strassburg, 2000; Kiang *et al.*, 2005). In vertebrates, UDP glucuronic acid is most commonly attached to xenobiotics as they are metabolized, but *C. elegans* and *H. contortus* appear to use UDP glucose instead (Cvilink *et al.*, 2008; Laing *et al.*, 2010).

1.3.5 Glutathione S-transferases

The glutathione S-transferases (Gsts) have traditionally be known for their role detoxifying endogenous and xenobiotic compounds, but more recently some members have been shown to be involved in signalling pathways controlling cell proliferation and apoptosis (Enayati *et al.*, 2005; Laborde, 2010). Some of their endogenous detoxification roles involve protection against reactive oxygen species (Hayes and Pulford, 1995). The general activity of Gsts is due to their ability to lower the pK_a of the sulfhydryl group of reduced glutathione from 9 to 6.5 when bound to the enzyme's active site (Hayes and Pulford, 1995). The group is then bound to an electrophilic atom, often carbon, nitrogen, or sulfur; this makes their range of potential substrates very large.

There are at least 13 described classes of Gst, with some being specific to mammals, plants, bacteria, or insects (Viktor Cvilink *et al.*, 2009); 4 major classes are detectable by differing patterns of sequence motifs in the *Caenorhabditis* (**Figure 5.3**). There the standard Gsts, the Gstos (omega class; includes Gst-44), the Gstks (kappa class), and an unnamed class (likely the zeta class; includes Gst-42, Gst-43, and Y53G8B.1). There are at least 3 sub-patterns of motifs detected in the 'standard' Gsts, which may be evidence for some of these remaining classes.

These proteins have been found to be overexpressed in nematodes in response to anthelmintics, and so have been considered as potential drug targets (Gupta and Rathaur, 2005). The study that induced resistance to ivermectin in *C. elegans* also found that the mechanism appeared to reduce cellular levels of glutathione (James and Davey, 2009). It may be that efflux pumps co-transport this compound along with the drug, or it could be evidence that the Gsts are involved directly in metabolism of this anthelmintic.

1.4 Genomics

1.4.1 Genome assembly and quality

As the costs have decreased exponentially in the past two decades, genomic sequencing has become a staple technology in research around the world. This technology has been immensely important in applications from characterizing novel species in ecology to personalized cancer genomics in medicine. Unfortunately, the methods to process these data have not kept pace, leading to the notion of a $1,000 genome requiring $100,000 analysis (Mardis, 2010). Genomics research is in a new position where groups are generating sequencing data faster than they can assemble it. This is a problem that will only get worse, barring some ground-breaking advance in assembly algorithms, which seems unlikely, or until the error rates and costs of long-read sequencing technology such as that from Oxford Nanopore drop significantly.

The first step in a genomics analysis involves assembling the short sequence read data into contiguous regions (contigs), and assembling contigs into scaffolds using longer-distance mapping information. Repetitive genomic regions must be dealt with, a task that may be impossible with solely the information contained in the short reads, and haplotypic and allelic variants must be separated from sequencing or assembly errors. This requires specialized computational resources for most eukaryotic genomes, and is complicated by the glut of available software tools (Nagarajan and Pop, 2013). The genomics community routinely holds competitions to compare the current state of assembly methods, such as the Assemblathon (Earl *et al.*, 2011) or GAGE (Salzberg *et*

al., 2012), but choosing the best assembler for some data set still depends on the sequencing technology used, the properties of the genome of interest, and the availability of additional data.

The second step in the analysis involves annotating genes and other genetic elements on the genome. This is non-trivial for bacterial genomes, but is an order of magnitude more difficult when introns are involved. If a high quality reference genome for the species of interest is not available, the annotations from related species may be used to aid in *de novo* predictions, but additional sequencing data such as expressed sequence tags or RNA-seq are likely to be required to generate a set of protein sequences of reasonable quality (Yandell and Ence, 2012). Even genomes that are under continual improvement are very likely to contain many errors, as demonstrated by Wakaguri *et al.* (2009) who found that 41% of *Toxoplasma gondii* gene models contained at least one error, and 14% of the exons from well-studied apicomplexan parasites were not supported by any cDNA evidence (Wakaguri *et al.*, 2009).

The quality of a genome is measured in several ways, with the N50 being perhaps the most important. When the scaffolds of an assembly are arranged by decreasing size, the N50 is the length of the scaffold where 50% of the total genome is contained on scaffolds of at least that size. CEGMA is a tool used to measure the quality of eukaryotic genome annotations (Parra *et al.*, 2007, 2009). The authors collected a set of 248 conserved genes that are present in exactly one copy in virtually all eukaryotes, and the assembly of interest is searched for orthologs. As these genes are likely to be the easiest to annotate, this method yields an upper bound on the likely fraction of correctly annotated genes. A genome is considered to be of ‘draft’ quality if it meets minimum standards for submission to a database, and is considered ‘high-quality’ if it is at least 90% complete (Yandell and Ence, 2012).

1.4.2 Gene family evolution

Gene duplication is one of the major mechanisms behind the present-day eukaryotic biodiversity (Lynch and Conery, 2000; Ponting, 2008; Mendivil Ramos and Ferrier,

2012). These duplicated genes are referred to as paralogs, and for a short time afterwards both genes experience an elevated rate of evolution, as measured by the ratio of non-synonymous over synonymous mutations accumulated in the coding sequence (Panchin *et al.*, 2010; Rosello and Kondrashov, 2014). The original copy tends to return to its pre-duplication levels, while the novel copy maintains a higher level for an extended time, often as it acquires a novel specificity or function. It is not feasible in all cases to test the type of selective pressures that drive a duplication to fixation in a population, but Kondrashov (2012) describes several specific examples of adaptive gene duplications, including multi-drug resistance in *Plasmodium falciparum*.

There are several biological mechanisms that can lead to gene duplications, and they result in distinct genomic signatures. A tandem duplication leaves the two paralogs adjacent in the genome, an intrachromosomal duplication results in the new gene copy being located on the same chromosome, and an interchromosomal duplication results in the new copy being located on another chromosome (Mendivil Ramos and Ferrier, 2012). It is unclear which mechanism will be most prevalent in a given species; tandem duplications are the most common form in cows, rodents, and dogs, but intrachromosomal duplications make up the majority in humans.

A gene family is a set of paralogs that have arisen by duplication, and often carry out related functions (Gao *et al.*, 2014). It is accepted in the literature that the presence of a gene family confers some adaptive advantage, but there is no clear consensus on the factors that lead to its formation. Some studies suggest that the fixation of paralogs may easily arise from non-adaptive processes – genetic drift, mutation, and recombination (Lynch, 2007) – while others invoke positive selection to explain the retention of the new duplicates (Innan and Kondrashov, 2010).

1.5 Overview

This project began while our lab was working on publishing the genome of *Haemonchus contortus* (Laing *et al.*, 2013). I was attempting to characterize the Cyp family, and the apparent failure of the annotation pipeline on this gene family – though

the other detoxification families appeared unaffected – combined with inability of BLAST to yield acceptable results led me to develop Figmop as a more sensitive gene finder (Chapter 2). MIPhy and RDepth (Chapters 3 and 4) are software applications that allow comparative phylogenetic analyses on gene families between, and within species, respectively. MIPhy was developed as a way to quantify the lineage-specific variations in the size of the Cyp family in related organisms, and was especially motivated by the analysis of Thomas (2007). I originally intended to use existing algorithms to carry out my analysis, but after evaluation none proved suitable. NOTUNG (Stolzer *et al.*, 2012) was the closest, so I adapted their algorithm into a method that performs simultaneous gene tree reconciliation and clustering. RDepth began as a simple analysis of copy-number variation in the wild isolates of *C. elegans* sequenced by Thompson *et al.* (2013). Members of my lab kept asking me for data on their genes of interest, so I decided to preserve the methods as a web tool.

1.5.1 Contributions

Chapter 2 (Figmop) has been previously published in the journal *Bioinformatics* (Curran *et al.*, 2014) (see **Appendices A** for the reproduction license), though substantial material has been added to the introduction and discussion sections, and the supplemental material from the original manuscript has been incorporated into the main text. DMC was involved in all aspects of that work, including the conception, design, and implementation of the algorithm, and preparation of the manuscript. JDW and JSG assisted with the conception and testing of the software, and with editing the manuscript. Chapter 3 (MIPhy) is a manuscript I intend to publish once I settle on a suitable journal. DMC was involved in all aspects of the work, including the conception, design, and implementation of the algorithm, and preparation of the manuscript. JDW and JSG assisted with the direction and testing of the software, and editing the manuscript. Chapter 4 (RDepth) is a suite of web tools that were a collaboration between DMC and AR, a summer student, and which may be published in the future. DMC was involved in all aspects of the work, including training and managing AR, designing and implementing the framework, and implementing the first tool (GDepth;

except for the ‘Formatting options’ box and the ‘Save’ functionality). DMC guided the design of the second tool (CDepth), but it was implemented by AR. Chapter 5 is a comparative analysis of the five major detoxification gene families in five species of free-living nematode, making use of the software described in the previous three chapters. This work is described in a manuscript that is currently being prepared. DMC was involved in all aspects of this work, and JDW and JSG aided in the design of the study as well as the interpretation of the results. Chapter 6 is a discussion of the entirety of this work, as well as detailed future applications.

2. Figmop: a profile HMM to identify genes and bypass troublesome gene models in draft genomes

2.1 Abstract

Gene models from draft genome assemblies of metazoan species are often incorrect, missing exons or entire genes, particularly for large gene families. Consequently, labour intensive manual curation is often necessary. I present Figmop (Finding Genes using Motif Patterns) to help with the manual curation of gene families in draft genome assemblies. The program uses a pattern of short sequence motifs to identify putative genes directly from the genome sequence. Using a large gene family as a test case, Figmop was found to be more sensitive and specific than a BLAST-based approach. The visualization employed allows the validation of gene hits to be carried out very quickly and easily, saving hours if not days from an analysis.

Source code of Figmop is freely available for download at https://github.com/dave-the-scientist, implemented in C and Python and is supported on Linux, Unix, and MacOSX.

2.2 Introduction

The majority of published metazoan genome assemblies are in draft form, representing an *in silico* prediction of the *in vivo* genome. Errors include sequence reads that are not clustered into contigs, and contigs that are misplaced on scaffolds, either in the wrong location or the wrong orientation. This can have a significant impact on annotating the genes in the genome, an already complex process in eukaryotes where genes may have tiny exons and introns of variable length (Picardi and Pesole, 2010). The best algorithms get many of the gene models mostly correct, but this may not be enough to make and test evolutionary and functional hypotheses in many cases, particularly for large and complex gene families. The implications are that the majority of metazoan genomes contain many gene models that are incorrect, and that the absence of a sequence in the gene annotations should not be sufficient evidence to conclude that the organism is actually lacking that gene (Gilabert *et al.*, 2016). An older study also

echoes this warning, as they found that 41% of *Toxoplasma gondii* gene models contained at least one error, and 14% of the exons from well-studied apicomplexan parasites were not supported by any cDNA evidence (Wakaguri *et al.*, 2009).

Here I present Figmop (Finding Genes using Motif Patterns), which will guide the user to identify the correct gene models, or provide a measure of the accuracy of the current gene models of their gene family of interest. The software extends the use of MEME and MAST (version 4.5) (Bailey *et al.*, 2006) to identify regions of the genome containing genes of interest. It uses amino acid motifs to capture conservation within short stretches of sequence in a gene family, allowing the user to specify an architecture for that family as a specific pattern of those motifs. For highly variable regions, alternative or optional motifs can be included. Figmop implements a profile hidden Markov model (pHMM) that conducts a fuzzy match of this motif architecture against the given genome sequence, accounting for variation and introns as random or unmatched motifs. One of the most important aspects of this software is that it takes as input a pattern of protein sequence motifs, and applies this to a DNA sequence that may contain extensive introns, without the user having to specify any information about those introns.

Figmop has proven invaluable in my efforts to manually curate cytochrome P450 (Cyp) gene family members, which have high sequence variability and variable intron/exon structure, within a draft genome assembly of the parasitic nematode *Haemonchus contortus* (project ID PRJEB506, version CAVP000000000.1) (Laing *et al.*, 2013). Figmop is a general tool, and besides Cyps it has proven useful in identifying other diverse gene families, including the glutathione-S-transferases, UDP-glucuronosyl transferases, and ABC-transporters; it has been run on many different genomes as well, including human, fruit fly, and several nematodes and protists.

2.3 Materials and methods

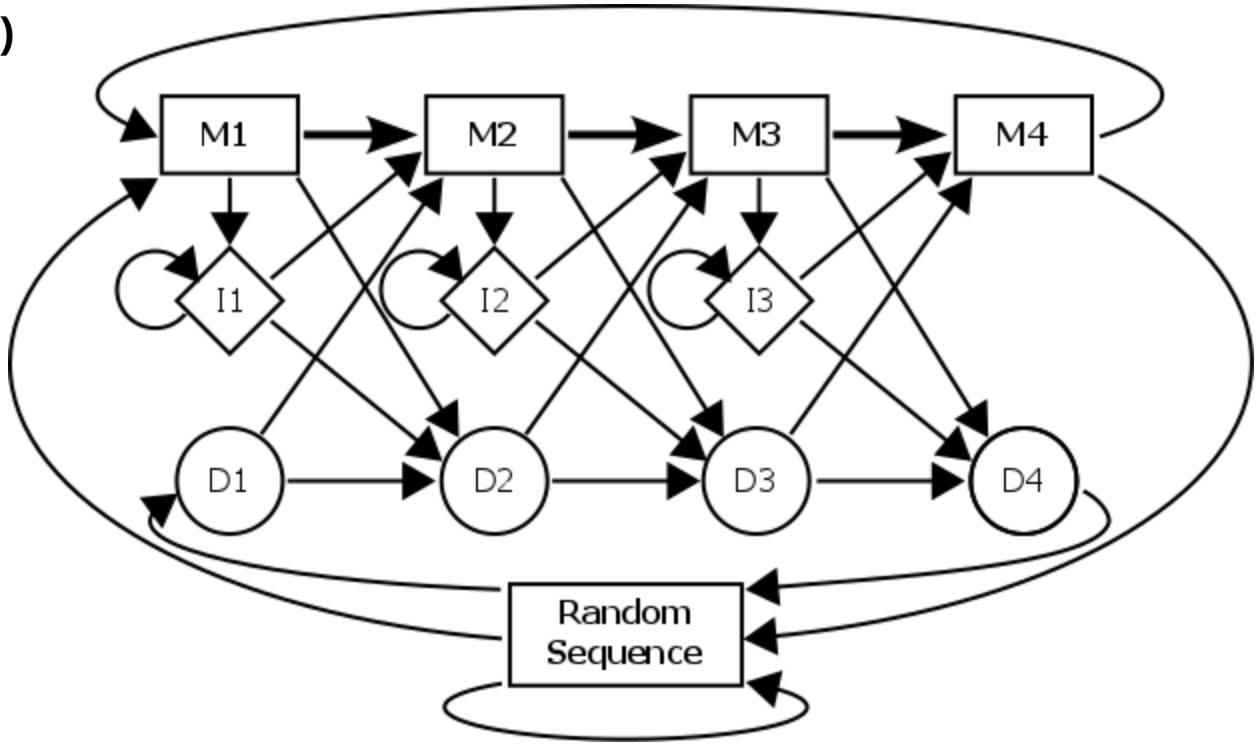
2.3.1 Motif generation and detection

Figmop takes as input a set of protein sequences representative of the user's gene family of interest. These sequences should be from related species and/or confirmed full-length proteins from the test species. The user runs the MEME software to generate a set of motifs, and specifies a pattern from these. Figmop is then run, where it first uses the program MAST to detect these motifs across the test genome sequence (Bailey *et al.*, 2006). This is the most computationally intensive step of Figmop, taking approximately eight minutes to process a 370 Mb genome on a 2.6 GHz computer, but must only be run once per genome per full set of motifs.

2.3.2 Figmop pHMM

The Figmop configuration file is automatically generated and contains the motif pattern and a set of configurable probabilities that constitute a pHMM (shown in **Figure 2.1A**). The 'Random Sequence' state emits any symbol and is used to identify background sequence that is not part of the pattern, while each match state emits one of the symbols in the pattern of interest. The insert states emit any symbol and allow the pHMM to identify a pattern containing some spurious symbols, and the delete states do not emit anything and allow some of the symbols in the pattern to be skipped. As an example, consider identifying the pattern 'ABCD' in the sequence of symbols shown in the first line of **Figure 2.1B**. The first four symbols, 'fedd', are classified as 'Random Sequence', before the path through the pHMM passes through states M1, M2, M3, and M4, which emit 'A', 'B', 'C', and 'D', respectively. The path then returns to 'Random Sequence' to emit the next eight symbols. The second instance of the pattern contains spurious symbols, and is emitted by states M1, M2, M3, I3, I3, M4, respectively. The third instance of the pattern is missing 'C', and is emitted by the path through states M1, M2, D3, M4. The fourth instance is missing 'B' and contains a spurious symbol, and is emitted by the path through states M1, I1, D2, M3, M4.

A)



B) feddABCDbtklavsfABCxhDzrcajcABDkaqwtiAqCDxb
rrrr**MMMM**rrrrrrrr**MMMIIM**rrrrrr**MMM**rrrrrr**MIMM**rr

Figure 2.1 The Figmop profile HMM. A) shows a schematic representation of the profile hidden Markov model used by Figmop to detect patterns, in this case a pattern with four symbols. M1 – M4 are the match states, I1 – I3 are the insert states, and D1 – D4 are the delete states. The first line of **B)** is an example sequence of symbols containing four instances of the pattern 'ABCD' (indicated with capital letters). The second line indicates the sequence of states that produce the sequence of symbols, where r indicates 'Random Sequence', M indicates one of the match states, and I is one of the insert states.

The software scans the motif complement of the user's genome and uses the Viterbi algorithm to detect significant architectures. This step is coded in C, and takes approximately 9 seconds to run on a 370 Mb genome. Different architectures or probabilities using the same set of motifs can be searched in the same amount of time without the need to run MAST on the genome again.

2.3.3 Output

The genomic regions containing the detected architectures are collected, and the sequences are extracted and saved in FASTA format. MAST is then run on these sequences in order to produce an HTML output with which the significant architectures can be visualized.

2.3.4 Reciprocal BLAST search

When validating Figmop, I compared it to a BLAST-based approach previously used to detect homologous genes in a genome, called reciprocal BLAST. This method takes as input a file of query protein sequences, the subject genome scaffolds file, a set of representative protein sequences, and a threshold E-value. The first step in this procedure was to search the query set against the subject genome using TBLASTN (version 2.2.27) with relaxed constraints using the command:

tbastn -db genome_scaffolds -query query_sequences.fa -num_threads 10 -evalue 10 -threshold 11 -window_size 0 -outfmt 6 -out initial_blast.txt

This file was then filtered, and those high-scoring segment pairs (HSPs) with an E-value lower than the given threshold were saved to a new file. Any overlaps in these remaining HSPs were merged, so for each scaffold in the genome this yielded a list of regions that were matched by at least one BLAST HSP from any of the query sequences. The nucleotide sequence for each of the above hits were saved to a new file. Representative Cyp sequences were chosen as five of the most phylogenetically diverse Cyps from *C. elegans*, which were then searched against the *H. contortus* genomic regions with another TBLASTN:

tblastn -subject genomic_regions.fa -query representative_sequences.fa -evalue 1e-2 -outfmt 6 -out final_blast.txt

These HSPs were then filtered by the given E-value threshold and any overlaps merged, yielding a final list of reciprocal BLAST hits.

In order to produce **Figure 2.3** in such a way that introns were not depicted, I required a mapping from the genomic coordinates to a consistent amino acid numbering scheme across a Cyp protein sequence. The five representative *C. elegans* Cyps were aligned using M-COFFEE (Wallace *et al.*, 2006), and this alignment was used as the consistent Cyp amino acid coordinates. Comparing each of the reciprocal BLAST hits to this alignment allowed each to be mapped to the representative alignment, and so mapped to consistent Cyp coordinates.

In order to quantify false positives and negatives, the Cyp pattern described in the main text was assumed to be correct, and a minimum of 5 of the 13 motifs conforming to the pattern were required for a region to be said to contain a Cyp gene. An HSP was considered a true positive if it was found between the first and last motif of the region, or within 1 kb of that region. While admittedly somewhat circular, there is currently no better method of positively identifying a Cyp that I am aware of. If BLAST predicted that a region of the genome contained a Cyp, 30 kb of sequence around that location was analyzed using MAST to manually search for evidence of the Cyp pattern. None of the regions detected by BLAST, but missed by Figmop, contained any evidence of this pattern, and so were all considered false positives.

2.4 Results

2.4.1 Validation

The original motivation for developing Figmop was to find the Cyp genes in nematode genome assemblies. A total of 530 nematode Cyp genes from *Caenorhabditis* species and *Pristionchus pacificus* were collected and an architecture of 13 MEME-derived motifs was generated (**Figure 2.2A**). This pattern of motifs was then used by Figmop to search the genome assembly of *Drosophila melanogaster*, where it returned all of the 85 defined Cyp genes, as well as two probable pseudogenes (results not shown).

I then used Figmop to search the current genome assembly of the parasitic nematode *H. contortus* (Laing *et al.*, 2013), and found 28 regions (*e.g.* **Figure 2.2C**).

However, all of the published gene model predictions for these likely Cyp genes were truncated or fragmented (*e.g.* **Figure 2.2B**). Further inspection revealed that all of the expected sequence motifs for this Cyp were present in the genome (**Figure 2.2C**), but were not identified as coding sequence by the gene prediction software. This was also found to be the case for the rest of the 28 putative Cyp genes.

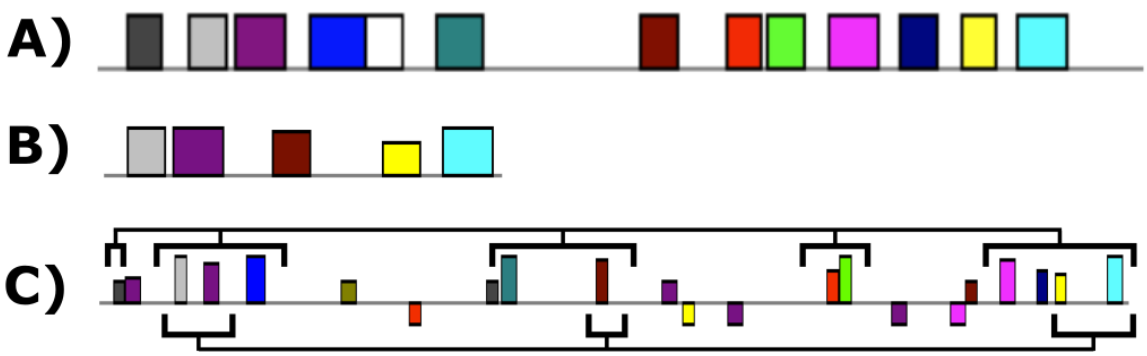


Figure 2.2 Improving a Cyp gene model using Figmop. The coloured bars represent sequence motifs identified by MAST, where the height is proportional to the strength of the match to the query sequence. **A)** shows the pattern of motifs common to all Cyp genes, while **B)** shows one of the fragmented gene models annotated in the published genome assembly (Laing *et al.*, 2013). **C)** shows the corresponding genomic region for **B),** where the bottom indicators map the currently annotated coding regions and the top indicators map the Cyp pattern from **A)** as found by Figmop.

2.4.2 Comparison to BLAST

A common approach to find a gene of interest in a genome in the absence of gene models is to compile a set of known homologous proteins, and run tBLASTn (version 2.2.27) against the genomic sequence, accepting those high-scoring segment pairs (HSPs) that satisfy an E-value threshold (Camacho *et al.*, 2009). When searching a whole genome, a small E-value should be used to avoid false positives; 1E-40 might be suitable in this case. At this threshold only four of the 28 regions found by Figmop were matched by one or more HSPs, but there were an additional seven HSPs that, following manual inspection, were most likely false positives, as the genomic regions contained no evidence of the Cyp pattern described previously. The E-value threshold had to be relaxed to 1E-10 before all 28 regions were found by at least one HSP, at which point

there were an additional 432 HSPs that were likely false positives. A reciprocal BLAST approach was used to improve the specificity. Using this approach tBLASTn still required an E-value threshold of 1E-10 before it could detect all 28 putative Cyp regions, though it still returned eight false positives. Even at this very relaxed threshold, many of the Cyp regions were only matched by a single HSP (**Figure 2.3B**). A threshold of 1E-25 was required to exclude all false positives, though at this point only 11 of the 28 putative Cyp regions were returned.

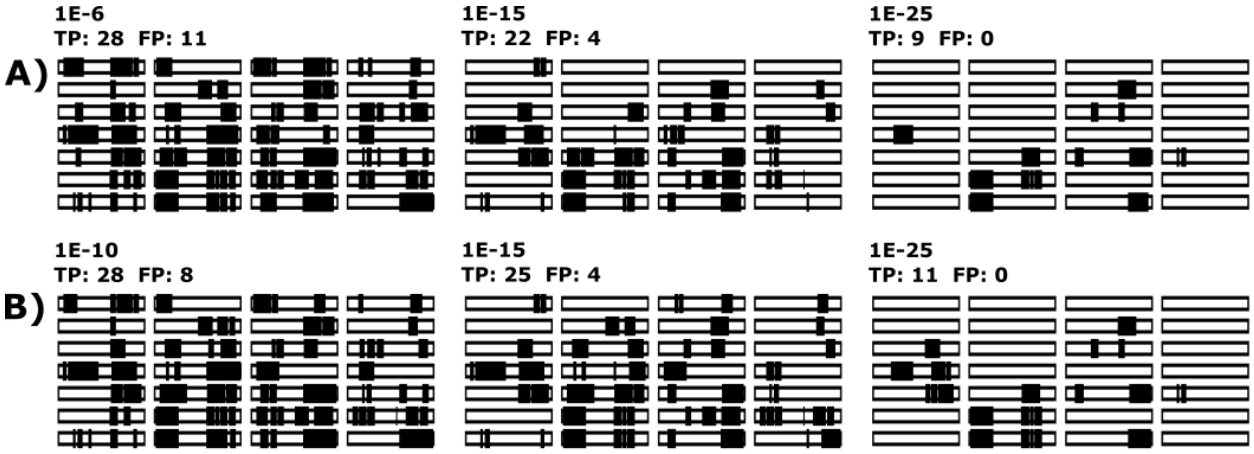


Figure 2.3 Mapping the BLAST hits onto the 28 Cyp regions found by Figmop. A) shows the results using the set of 81 unique *C. elegans* Cyp protein sequences as the query, while **B)** used the set of 530 putative nematode Cyps. For each plot the E-value threshold is shown on top, followed by the number of true positives (TP) and false positives (FP). For both sets, the leftmost plot represents the smallest E-value threshold that could still detect all 28 Cyp regions, the rightmost plot is the largest threshold that returned no false positives, and the middle plot is set at a reasonable compromise.

Figure 2.3 shows the results of the reciprocal BLAST procedure using various E-value thresholds, and using either the 81 unique Cyp genes from *C. elegans* or a set of 530 putative Cyp genes from four *Caenorhabditis* species and *Pristionchus pacificus*. For the two sets of query proteins, thresholds of 1E-6 and 1E-10, respectively, were required before they could detect all 28 putative Cyp regions found by Figmop. These thresholds are both very high for searching an entire genome, hence why the searches returned 11 and eight false positives, respectively. For both sets, a threshold of 1E-25 was required before all false positives were excluded, though this lowered the sensitivity

so that nine and 11 of the Cyp regions were found, respectively. 1E-15 was found to be a reasonably good threshold in terms of true and false positives for both sets, but is quite high for the purpose of searching a genome.

2.4.3 Tandem repeats test case

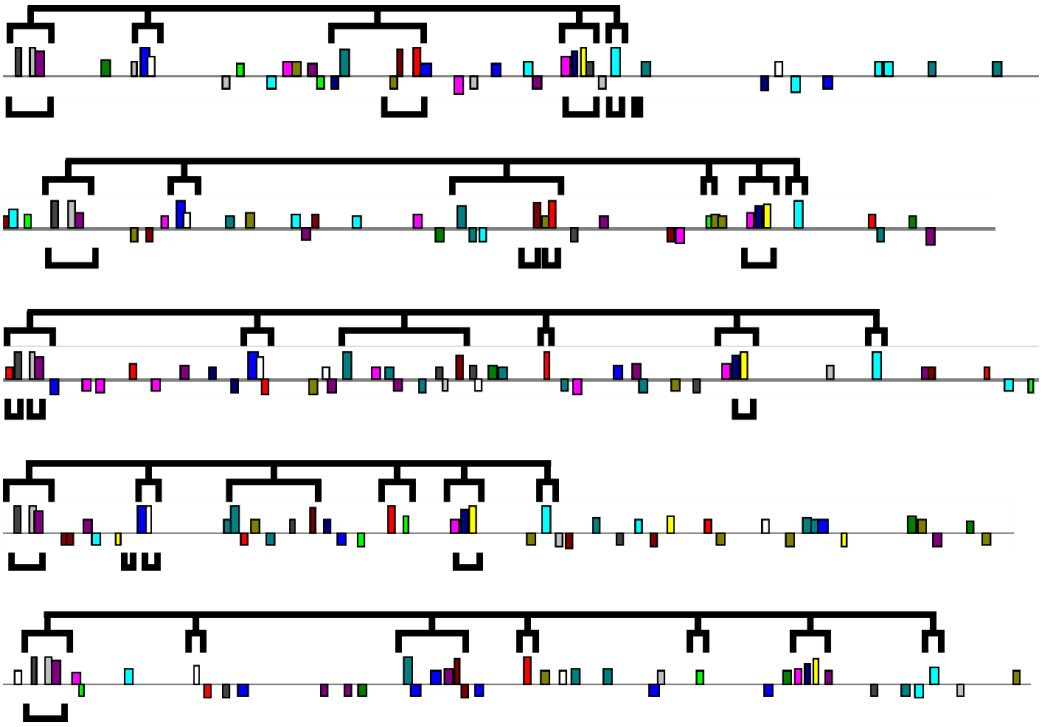


Figure 2.4 Identification of a tandem array of genes with Figmop. The five lines above comprise a continuous 51 kb region of one scaffold from the genome assembly of *H. contortus* that was found to contain 5 putative Cyp genes. The coloured bars are the sequence motifs found in this region by MAST searching a 6-frame translation of the nucleotide sequence. The top connected indicators show the five regions as found by Figmop, and the bottom indicators show the reciprocal BLAST HSPs (E-value threshold of 1E-10).

Figure 2.4 shows a 51 kb region on one of the *H. contortus* scaffolds that was found to contain five Cyp genes in a tandem array by Figmop, and illustrates some of the differences between Figmop and a more standard BLAST-based approach to detecting genes. Originally the reciprocal BLAST approach was used, but in order to maximize the sensitivity I didn't filter the results by a second round of TBLASTN, and instead kept

all of the HSPs from the first TBLASTN that were below my threshold. After merging all overlapping HSPs, I ended up with 17 distinct hits in the 51 kb region.

While this BLAST approach was able to detect at least one HSP in each of the five putative Cyp regions, Figmop was able to identify several more motifs in each, and these motifs covered the entire length of each gene. Perhaps more useful, all Figmop designated matches are automatically joined together into a single cohesive region. Further, because of the visual output from MAST it is generally obvious to the user which parts of their gene pattern have been detected. This allows easy identification of duplications, truncations, and other modifications that might be difficult or time-consuming to identify using a local method like BLAST.

2.5 Discussion

The disparity between the Cyps identified in the genome compared to the gene models in *H. contortus* seems to indicate that the gene annotation software performed particularly badly for this gene family. I cannot completely discount the possibility that it is due to Figmop detecting pseudogenes, but this seems unlikely to be the case for all 28 regions. Further, I have observed a similar disparity between Cyp protein sequences and Cyp regions in the genomes of other parasitic nematodes, in particular *Strongyloides ratti* (**Table 2.1**). This suggests that the phenomenon is not limited to *H. contortus*, and that care should be taken when analyzing the annotated Cyp protein sequences from parasitic nematodes.

Though the examples discussed here involve the Cyp gene family in different species, the software is for general use. I have used it for the manual curation of several other large divergent gene families such as the Fmos, Ugts, and Gsts (all detoxification families, described in Chapter 5), as well as identifying 47 genes from small or single-member gene families in 23 nematode genomes (Gilabert *et al.*, 2016). For these uses I found it to be preferable to BLAST in sensitivity and specificity, but mostly it was beneficial in terms of ease of use and time. The motif patterns are not restricted only to

genes, and so Figmop could also prove valuable in detecting genetic elements such as transcription factor binding sites or endogenous viruses (Cotton *et al.*, 2016).

Table 2.1 Identifying Cyps in parasitic nematodes. A comparison between Cyps identified in the gene annotations and those identified in the genome using Figmop. All genome assemblies are taken from WormBase release 239, except for *Haemonchus contortus* which is from Laing *et al.* (2013). There were two independent assemblies for *Ascaris suum*, the first is assembly PRJNA80881, while *Ascaris suum 2* is assembly PRJNA62057. For all columns, a sequence was considered a Cyp if it had at least 5 of the 13 Cyp motifs in the appropriate order. 'From genome' indicates the number of Cyps Figmop detected in the genome and 'From gene models' indicates the number detected in the annotated gene models; this number is presented in the 'Models' column, and is further broken down in the following four columns. '> 62%' indicates the number of mostly complete protein sequences containing more than 8 of the Cyp motifs, 'Front' indicates those with 8 or fewer of the first motifs, 'Back' indicates those with 8 or fewer of the last motifs, and 'Fragmented' indicates those that appear to be missing many motifs throughout the sequence.

Species	**From genome**	**From gene models**				
		Models	> 62%	Front	Back	Fragmented
Ascaris suum	15	16	8	3	5	0
Ascaris suum 2	12	13	10	2	1	0
Brugia malayi	7	7	6	0	1	0
Haemonchus contortus	28	17	2	6	6	3
Strongyloides ratti	19	6	3	1	2	0
Trichinella spiralis	4	2	2	0	0	0

I have created Figmop, software that uses a pHMM to compare the motif patterns of a set of similar sequences against a test genome. There are other sequence searching tools that use pHMMs due to their sensitivity, such as HMMER, which performs protein/protein searches (Eddy, 2008), or nHMMER, which performs DNA/DNA searches (Wheeler and Eddy, 2013), but I am not aware of any that are trained on

protein sequences and suitable to search DNA sequences that include introns. Further, those programs take an alignment as input, while Figmop allows the user to inject additional knowledge and expertise when specifying their pattern of motifs.

2.6 Acknowledgements

I thank Dr. Aude Gilabert for breaking earlier versions of the program. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) CREATE programme in Host–Parasite Interactions (#413888-2012).

3. MIPhy: A new approach to cluster and score inter-species phylogenetic instability in large gene families

3.1 Abstract

Under neutral evolutionary conditions, the phylogeny of a multi-species gene family will tend to agree with the underlying species tree. Adaptive evolutionary pressures may manifest as incongruence between these trees, or as inter-species copy number variation. Taken together these characteristics, termed phylogenetic instability, have been linked to adaptations in sampling and responding to an organism's environment, though there has never been a method to quantify it.

Here, I present novel algorithms to score this incongruence and cluster a phylogenetic tree by minimizing the overall phylogenetic instability. The software package, named MIPhy, also includes an interactive HTML-based tree viewer. I demonstrate the usefulness of MIPhy by predicting which members of the cytochrome P450 gene superfamily metabolize xenobiotics and which metabolize endogenous compounds. My predictions correlate very well with known substrate specificities of the human enzymes. The software is available for download or as an online web tool at http://www.miphy.wasmuthlab.org

3.2 Introduction

In the absence of specific selective pressures, the phylogeny of a multi-species gene family will tend to agree with the underlying species tree. However, gene events such as gene duplication/loss, horizontal gene transfer (HGT), and incomplete lineage sorting (ILS) – where a polymorphic locus in an ancestral species results in incongruence with the species tree – may become fixed in a species due to adaptive or non-adaptive evolutionary processes. These events can result in lineage-specific variations in gene family size or incongruence between the gene family phylogeny and the species tree, properties that have collectively been referred to as 'phylogenetic instability' (Thomas, 2007) (see **Figure 3.1** for an example). Attempting to work backwards and elucidate the

sequence of events that led from the species tree to the observed gene family is a process called event-inference reconciliation.

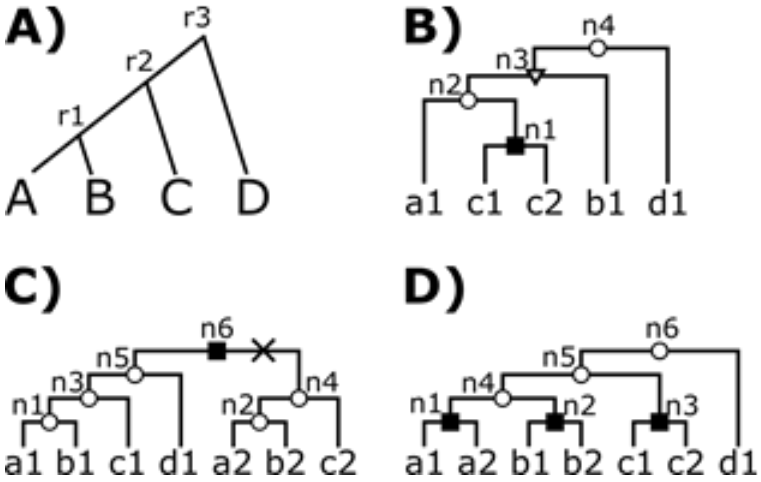


Figure 3.1 Example of phylogenetic instability. A) shows a species tree, with its three ancestor species nodes labeled (r1-r3). **B-D)** show example gene trees, where a1 indicates a gene originating from species A. These trees have their gene events labeled with filled squares, open circles, open triangles, and Xs, representing duplication, speciation, incongruence, and loss events, respectively.

It has been hypothesized that the change in environment during a speciation event may lead to higher levels of phylogenetic instability (Lynch and Conery, 2000; Zhang, 2003; Hurley *et al.*, 2005), especially in genes involved in responding to molecules from the environment (xenobiotics). This has been observed in gene families involved in the immune response (de Bono *et al.*, 2004; Nei *et al.*, 1997; Su *et al.*, 1999), chemosensory receptors (Niimura and Nei, 2005; Thomas *et al.*, 2005), detoxification (Thomas, 2007), and host-pathogen interactions (Wasmuth *et al.*, 2012).

A detailed analysis was conducted on the vertebrate cytochrome P450 gene family (Thomas, 2007), which found the enzymes with known xenobiotic substrates (about half of the gene family) exhibited high phylogenetic instability, while those with known endogenous substrates were strikingly phylogenetically stable. That work relied upon the author's detailed knowledge of the gene family under study, and was not quantified. I am not aware of an algorithm that assigns a score of phylogenetic instability that can be compared between clusters in a given family's phylogeny. As the genomes of an

increasing number of species are deposited, manual analysis of large gene families from hundreds of species will become intractable. Further, it is desirable to use an algorithm that is consistent and deterministic instead of relying on human instinct.

Here, I propose using phylogenetic instability to predict the functional roles of the members of a gene family. Specifically, to identify which family members are under pressure to duplicate and contributing to altered or new functions, with the possibility of new phenotypes. Such a tool can be used to prioritize genes for further study: identifying the mechanism of some species-specific function, or identifying new therapeutic targets in pathogens. The process to detect phylogenetically unstable genes is two-fold. First, a tree of a large multi-member gene family is split into meaningful clusters (termed phylogenetic groups), by incorporating an event-inference model of gene evolution. Second, each phylogenetic group is independently scored for phylogenetic instability.

3.2.1 Related work

There are several existing algorithms for species/gene tree reconciliation, but none are designed to quantify the stability of gene clusters from large gene families. Some quantify the gene events along every branch of a large gene family phylogeny, but they do not segregate that tree into meaningful clusters, nor do they score each gene in order to compare and rank the individual family members.

CAFE 3 uses a stochastic birth-death model of gene family evolution, inferring the size of ancestral families (De Bie *et al.*, 2006; Han *et al.*, 2013). It implements a sampling procedure to determine the statistical significance of those gene families that differ from their expected values, and models the effects of genome assembly and gene annotation errors to provide a more accurate estimate of its evolutionary rates. CAFE only uses the gene family sizes without considering the phylogenetic relationships within them, and so would be unable to distinguish between the cases in **Figure 3.1C** and **Figure 3.1D**, though they are the product of different evolutionary histories. The algorithm only calculates whether an entire gene family is under adaptive evolution; I

am interested in the specific clusters of genes within that family. Because of this, it is more suited for large-scale analyses of many gene families at once.

BadiRate is similar to CAFE, implementing several additional stochastic models of evolution, and providing three statistical frameworks to calculate significance (Librado et al., 2012). While it allows for more detailed analyses of species traits, it still relies on gene count data and so is unsuitable here for the same reasons as CAFE.

NOTUNG (Chen *et al.*, 2000; Vernot *et al.*, 2008; Stolzer *et al.*, 2012) implements a parsimony-based reconciliation algorithm. It finds the sequence of gene events (gene duplication, gene loss, HGT, and ILS) explaining the differences between the observed gene tree and the underlying species relationships that minimizes a weighted sum. Uniquely amongst other reconciliation methods, it allows for the species or gene tree to be non-binary; as the true history for many species is unclear, polytomies can be useful to describe the current state of knowledge. Important in the consideration is that NOTUNG explicitly models HGT, and it assumes that ILS is a very rare event, only considering it at polytomies in the gene tree. Identifying HGT is a computationally intensive process and is unlikely to play an important role in gene families from multicellular organisms, and I assume that incongruence (as produced by ILS or any other mechanism) is a common enough event to include (Carstens and Knowles, 2007; Mirarab *et al.*, 2016; Scally *et al.*, 2012). RANGER-DTL is another reconciliation method, and has been reported to be 1,000-1,000,000x faster than similar software (Bansal *et al.*, 2012). Unfortunately, this model proved unsuitable as it too does not allow for ILS events.

There are also several probabilistic reconciliation methods available (Rasmussen and Kellis, 2007, 2011; Ma *et al.*, 2008; Doyon *et al.*, 2010, 2012). While these models make use of more sophisticated models of evolution, they are far more computationally intensive and are only applicable to species for which speciation times and/or ancestral population size estimates are available, which is not the case for most species. PHYLDOG overcomes some of these limitations as it is able to estimate the most likely gene trees, species tree, and evolutionary history of a large number of gene families at

once (Boussau *et al.*, 2013). Though it does not explicitly model ILS, the software can accommodate it as long as the signal is not too strong. Unfortunately, I expect gene families involved in direct environmental interactions to have a strong ILS signal. Further, this software is designed to combine the information from many gene families at once, and requires extremely significant computational resources (Chaudhary *et al.*, 2015).

While an individual is able to manually cluster a small tree without much trouble, the large size of some gene families and the ever-expanding availability of sequence data mean that this will quickly become intractable. There are several software packages used to automatically cluster a phylogenetic tree, but because of the ill-defined nature of clustering problems in general, the methods often come to different conclusions on the same data sets. I am aware of no method that is targeted towards multi-species gene families, which means that none make use of the additional information such as an event-inference model of gene evolution. The clustering algorithm described here combines the similarity between each gene with the most parsimonious explanation of gene events, to predict the ancestry of each observed member of the gene family.

3.2.2 My contributions

The algorithm described in this work is derived from the core reconciliation methods of NOTUNG (Chen *et al.*, 2000; Vernot *et al.*, 2008; Stolzer *et al.*, 2012), with key modifications. I have called the software package MIPhy, which implements my clustering and instability scoring algorithms and provides an HTML-based phylogenetic tree viewer. The first step of the process involves classifying internal nodes of the gene tree as representing a duplication, speciation, or incongruence event (used hereafter to refer to any incongruence with the species tree, no matter the origin) (**Figure 3.1**). The NOTUNG software applies a cost for each duplication event, but not for speciation or incongruence. Incongruence may appear due to minor errors in sequencing/gene-finding or to an incompletely resolved branch in tree-building software, or it may be due to novel environmental selective pressures acting on one or more species. As such I include these events in my reconciliation. As an example, node n3 in **Figure 3.1B** is an

incongruence event because gene a1 is closer to genes c1 and c2 than it is to gene b1, which would be expected from the species tree (**Figure 3.1A**). If I did not allow incongruence events, n3 would instead be classified as a duplication, and would require three additional loss events to reconcile.

I am interested in identifying members of a gene family under adaptive evolution, and so I must also cluster the given gene tree into phylogenetic groups. Each phylogenetic group is defined to include only those gene events at or below the recent common ancestor of that group. This algorithm extends reconciliation methods into a cost function, and then finds the clustering pattern that minimizes this weighted sum. Every node in the gene tree is evaluated in a depth-first post-order traversal; if the node is a leaf a new phylogenetic group is defined as containing only that node. At each non-leaf node, the cost function is used to compare merging all of that node's descendants into a single phylogenetic group, versus leaving them with their existing clustering pattern. This is the first phase, and it generates a preliminary clustering pattern. By way of example: **Figure 3.1C** and **Figure 3.1D** have the same number of genes from each species, but exhibit different phylogenies. MIPhy clusters **Figure 3.1C** into two groups; the first with zero gene events and containing the leaves a1, b1, c1, and d1; and the second with one loss event and containing the leaves a2, b2, and c2. MIPhy clusters **Figure 3.1D** as one group with three duplication events.

NOTUNG and related parsimony-based methods do not consider the relative branch lengths when predicting a gene event; if a gene tree leaf is separated by an uncommonly large phylogenetic distance from its closest group, there should be a cost associated with the decision to include it in that cluster. This is accomplished by transforming the pairwise phylogenetic distances onto a coordinate system using multidimensional scaling (Torgerson, 1952). With each sequence represented as a point in coordinate space, the second phase of the algorithm refines the preliminary clustering pattern by extending the cost function to include a clustering metric. MIPhy uses a simple standard deviation for each cluster, but more sophisticated methods like the Davies-Bouldin index (Davies and Bouldin, 1979) or silhouette (Rousseeuw, 1987) may be easily substituted.

3.3 Materials and methods

3.3.1 Algorithm notation

Given a gene tree T_G, g represents some node, and l and r are its children. If g is a terminal leaf, its originating species $ori(g)$ is defined to be the species node in T_S from which the gene g was collected. If g is an internal node, $ori(g)$ is defined to be the most recent common ancestor in T_S of $ori(l)$ and $ori(r)$. The lineage of a node $lin(g)$ is the set of species nodes (including ancestral species) tracing $ori(g)$ back to the root of T_S. The set of all terminal leaves in the subtree of T_G rooted by g is given by $lvs(g)$. The species represented in the subtree of T_G rooted at g, $spc(g)$, is the set obtained by applying $ori(c)$ to every leaf c in $lvs(g)$; S_G is the set of species with at least one gene in T_G. The represented species of a node g not present in the represented species of a node h is given by $miss(g,h) = spc(g) - spc(h)$. One of three mutually exclusive gene events must take place at each internal node in T_G: duplication, speciation, or incongruence. These are denoted by the binary variables $E_D(g)$, $E_S(g)$, and $E_I(g)$, respectively, constrained such that $E_D(g) + E_S(g) + E_I(g) = 1$. The contribution of each node in some phylogenetic group to the overall score of that group is $E_D(g) \cdot \theta_D + E_I(g) \cdot \theta_I + L(g) \cdot \theta_L$, where θ_D, θ_I, and θ_L are the given weights for duplication, incongruence, and loss gene events, respectively.

In the initial clustering (Section 3.3.3) the overall clustering pattern is described by $clusters(root)$, which is a set of sets, where each inner set describes one phylogenetic group in T_G. All sequences from T_G are contained in $clusters(root)$, while $clusters(g)$ contains the most parsimonious clustering pattern for those sequences in $lvs(g)$. These patterns are built iteratively by comparing $sep(g)$ (which is the score if the existing clustering patterns $clusters(l)$ and $clusters(r)$ are kept intact) with $cmb(g)$ (which is the score if all descendants are combined into a single cluster); the minimum of these values is stored as $best(g)$. Examples of these variables can be found in **Figure 3.2** and **Table 3.1**.

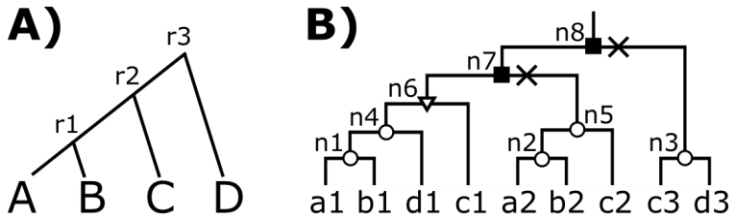


Figure 3.2 Example trees for the worked MIPhy example. A) is a species tree and **B)** is a gene tree, where the gene events are indicated with filled squares, open circles, open triangles, and Xs, representing duplication, speciation, incongruence, and loss events, respectively.

Table 3.1 Worked MIPhy example. This table provides the variables used in the Event inference and Initial clustering phases of the MIPhy algorithm applied to **Figure 3.2**. Only one set of terminal leaves is included for brevity. The $best(g)$ value for each node isn't explicitly stated, but is indicated by the bolded score value; the values θ_D, θ_I, and θ_L, indicate the weight of one duplication, incongruence, or loss event, respectively, and define $score(g)$. The final clustering pattern is found at the root, $clusters(n8)$, and predicts three clusters. The clustering pattern for this tree is invariant for all parameter weights such that $\theta_I < 3\theta_L$.

	Event inference values					**Initial clustering values**		
g	$ori(g)$	$lin(g)$	$lvs(g)$	$spc(g)$	Event	$cmb(g)$	$sep(g)$	$clusters(g)$
a1	A	$\{A, r1, r2, r3\}$	$\{a1\}$	$\{A\}$	-	-	$\boldsymbol{3\theta_L}$	$\{\{a1\}\}$
b1	B	$\{B, r1, r2, r3\}$	$\{b1\}$	$\{B\}$	-	-	$\boldsymbol{3\theta_L}$	$\{\{b1\}\}$
c1	C	$\{C, r2, r3\}$	$\{c1\}$	$\{C\}$	-	-	$\boldsymbol{2\theta_L}$	$\{\{c1\}\}$
d1	D	$\{D, r3\}$	$\{d1\}$	$\{D\}$	-	-	$\boldsymbol{\theta_L}$	$\{\{d1\}\}$
n1	r1	$\{r1, r2, r3\}$	$\{a1, b1\}$	$\{A, B\}$	E_S	$\boldsymbol{\theta_L}$	$6\theta_L$	$\{\{a1, b1\}\}$
n2	r1	$\{r1, r2, r3\}$	$\{a2, b2\}$	$\{A, B\}$	E_S	$\boldsymbol{2\theta_L}$	$6\theta_L$	$\{\{a2, b2\}\}$
n3	r3	$\{r3\}$	$\{c3, d3\}$	$\{C, D\}$	E_S	$\boldsymbol{\theta_L}$	$3\theta_L$	$\{\{c3, d3\}\}$
n4	r3	$\{r3\}$	$\{a1, b1, d1\}$	$\{A, B, D\}$	E_S	$\boldsymbol{\theta_L}$	$3\theta_L$	$\{\{a1, b1, d1\}\}$
n5	r2	$\{r2, r3\}$	$\{a2, b2, c2\}$	$\{A, B, C\}$	E_S	$\boldsymbol{\theta_L}$	$4\theta_L$	$\{\{a2, b2, c2\}\}$
n6	r3	$\{r3\}$	$\{a1, b1, c1, d1\}$	$\{A, B, C, D\}$	E_I	$\boldsymbol{\theta_I}$	$3\theta_L$	$\{\{a1, b1, c1, d1\}\}$
n7	r3	$\{r3\}$	$\{a1, b1, c1, d1, a2, b2, c2\}$	$\{A, B, C, D\}$	E_D	$\theta_D + \theta_I + \theta_L$	$\boldsymbol{\theta_I + \theta_L}$	$\{\{a1, b1, c1, d1\}, \{a2, b2, c2\}\}$
n8	r3	$\{r3\}$	$\{a1, b1, c1, d1, a2, b2, c2, c3, d3\}$	$\{A, B, C, D\}$	E_D	$2\theta_D + \theta_I + 2\theta_L$	$\boldsymbol{\theta_I + 2\theta_L}$	$\{\{a1, b1, c1, d1\}, \{a2, b2, c2\}, \{c3, d3\}\}$

3.3.2 Event inference

Unlike NOTUNG, my algorithm does not attempt to model HGT explicitly. Instead, these events will be classified as incongruence or duplications, which this algorithm quantifies in the cost function. I also allow incongruence events to take place at any node in a gene tree, instead of restricting them to polytomies. Moreover, I assume that incongruence is more likely than a duplication event followed by several independent loss events, so the latter case is not considered as possible history. It should be noted that this algorithm defines a duplication to be the presence of at least one gene from the same species in both children of some node. This is not sufficient to rigorously prove that a duplication has taken place in some ancestral species, but this definition has been found to perform well in practice. Taking **Figure 3.2B** as an example, if $E_I(n6) = 1$ then $clusters(n6)$ is scored with 1 incongruence event. If I did not allow incongruence events (so $E_D(n6) = 1$), $clusters(n6)$ would be scored with 1 duplication and 3 losses. This algorithm identifies gene events using purely local information in a single pass through T_G, decreasing the time complexity by orders of magnitude compared to similar software that models HGT.

Here I present my reconciliation and clustering algorithm, adapted from those used by NOTUNG. It proceeds in two phases, the first inferring the gene events and using them to generate an initial clustering, and the second phase incorporating traditional data clustering techniques to refine the clusters. If T_S is given by **Figure 3.2A** and T_G by **Figure 3.2B**, the gene events taking place at every internal node are inferred as follows:

- $E_D(g) = 1\ if\ spc(l) \cap spc(r) \neq \emptyset$: Node g is a duplication event if its children share any represented species. As an example, $E_D(n7) = 1$ because $spc(\mathrm{n6}) = \{A, B, C, D\}$ and $spc(n5) = \{A, B, C\}$, so $spc(n6) \cap spc(n5) = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$.
- $E_S(g) = 1\ if\ E_D(g) = 0\ and\ ori(l) \notin lin(r) \wedge ori(r) \notin lin(l)$: Node g is a speciation event if it is not a duplication event, and the originating species of neither child is contained in the lineage of the other. As an example, $E_S(n4) = 1$ because $E_D(n4) = 0$ and $ori(n1) = r1 \notin lin(d1) = \{D, r3\}$ and $ori(d1) = D \notin lin(n1) = \{r1, r2, r3\}$.

- $E_I(g) = 1\ if\ E_S(g) + E_I(g) = 0$: More explicitly, node g is an incongruence event if it is not a duplication event, and the originating species of one child is contained in the lineage of the other. As an example, $E_I(n6) = 1$ because $E_D(n6) = 0$ and $ori(n4) = r3 \in lin(c1) = \{C, r2, r3\}$.

Phylogenetic groups are defined by the most recent common ancestor of the leaves in that group, and the numbers of duplication and incongruence events counted in a group defined by node g are found by the recursive equations:

$$D(g) = E_D(g) + D(l) + D(r), \quad (1)$$

and

$$I(g) = E_I(g) + I(l) + I(r). \quad (2)$$

A speciation event indicates that genes from one species (or ancestral species) will be found exclusively in the descendants of one child and not the other. Conversely, for both children of a duplication event node there should be one gene from every species that has not yet been excluded by a previous speciation or incongruence event. Loss events are therefore counted at duplication nodes, as the number of total species of each child not present in the other:

$$L'(g) = E_D(g) \cdot loss(g) + L'(l) + L'(r),$$

where $loss(g) = |miss(l, r)| + |miss(r, l)|$

This would only be accurate if every species is represented by at least one gene in the total species of each cluster. To complete this concept, I introduce a new quantity $M(g)$ that compares the represented species at g with the represented species in T_G. Thus, the total loss events counted in the descendants of some node g would be given by:

$$L(g) = L'(g) + M(g),$$

where $M(g) = |S_G - spc(g)|$.

If T_G is **Figure 3.2B**, $L(n5) = L'(n5) + M(n5) = 0 + 1 = 1$ because no genes from species D are present in $leaves(n5)$, while $L(n7) = L'(n7) + M(n7) = 1 + 0 = 1$. As

demonstrated here, the M term does not propagate up the tree, and tends to disappear as the algorithm progresses further from the leaves.

The above equations are somewhat naïve, as they do not account for loss in ancestral species. If T_S is given by **Figure 3.2A** and I consider the group defined by n3, the above equations would calculate that 2 loss events have occurred, once each for species A and B. However, a more parsimonious explanation is that the homolog was only lost once, in species r1, the ancestor of A and B. The following equations account for these processes and are those used by MIPhy:

$$L(g) = L'(g) + M(g), \tag{3}$$

$$\text{where } L'(g) = E_D(g) \cdot loss(g) + L'(l) + L'(r),$$

$$loss(g) = \max_{s \in spc(l)} |\{ori(s,t) | t \in miss(r,l)\}| + \max_{s \in spc(r)} |\{ori(s,t) | t \in miss(l,r)\}|,$$

$$\text{and } M(g) = \max_{s \in spc(g)} \left|\{ori(s,t) | t \in \{S_G - spc(g)\}\}\right|.$$

An example from **Figure 3.2B**:

$$M(n3) = \max(|\{ori(C,A), ori(C,B)\}|, |\{ori(D,A), ori(D,B)\}|)$$

$$M(n3) = \max(|\{r2\}|, |\{r3\}|) = \max(1,1) = 1$$

$$\therefore L(n3) = L'(n3) + 1 = 0 + 1 = 1$$

3.3.3 Initial clustering

This section of the algorithm combines equations (1), (2), and (3) into the following score function:

$$score(g) = \theta_D \cdot D(g) + \theta_I \cdot I(g) + \theta_L \cdot L(g) + \theta_P \cdot P(g), \tag{4}$$

where the θ values are the weights applied to each event, and $P(g)$ is a clustering metric defined as equation (5) in Section 3.3.4; for this initial phase of the algorithm it is set to 0. Each node g in T_G is visited in a post-order depth-first traversal. The algorithm is described by the following pseudocode:

If g is a terminal node:

$$best(g) = M(g)$$

$$clusters(g) = \{\{g\}\}$$

Otherwise:

$$cmb(g) = score(g)$$

$$sep(g) = best(l) + best(r)$$

If $cmb(g) \leq sep(g)$, all descendants of g are merged into one cluster:

$$best(g) = score(g)$$

$$clusters(g) = \{lvs(g)\}$$

Otherwise the existing cluster patterns of nodes l and r are kept intact:

$$best(g) = sep(g)$$

$$clusters(g) = clusters(l) \cup clusters(r)$$

3.3.4 Cluster refinement

This initial clustering pattern arises from the most parsimonious history of gene events required to reconcile T_G with T_S. It indicates which groups of genes are most likely, for the given weights and assuming all branch lengths in T_G were equal, to have evolved from a single homologue in the ancestral species. This second phase of the algorithm refines this prediction by incorporating branch length information, specifically the pairwise distance information between the genes.

Many metrics exist to measure and compare the spread between points in a cluster. However, many of them require that these points be embedded into a coordinate system, where the concept of a mean exists, such as Euclidian space. This is accomplished here using multi-dimensional scaling. First, the full pairwise distance matrix from T_G is extracted into the matrix D, such that D_{ij} is the phylogenetic distance (measured as the sum of the branch lengths) between the leaves i and j. A Gram matrix M can then be generated by:

$$M_{ij} = \frac{D_{i1}^2 + D_{1j}^2 - D_{ij}^2}{2}$$

where ‘sequence 1’ is an arbitrary choice held constant throughout the calculation of the matrix (this sequence will be represented by the origin). I can then find the coordinate points by eigenvalue decomposition. If $M = USU^T$ is solved, each row of the matrix $X = U\sqrt{S}$ contains the coordinates for a point that represents one leaf from T_G.

Using this set of coordinate points, MIPhy calculates the standard deviation of every phylogenetic group, and defines the spread function for the group rooted at g:

$$P(g) = \frac{\sigma(g)}{\bar{\sigma}} - 1, \tag{5}$$

$$\text{where } \bar{\sigma} = \frac{1}{|clusters(root)|} \cdot \sum_{g \in clusters(root)} \sigma(g),$$

and where $\sigma(g)$ is the standard deviation of the points representing the sequences in the phylogenetic group rooted by g, and $\bar{\sigma}$ is the mean standard deviation of all phylogenetic groups. The quantity $P(g)$ is normalized around 0, so $P(g) = 1.0$ indicates that the spread of group g is 100% larger than the average spread, while $P(h) = -0.3$ indicates that the spread of cluster h is 30% smaller than $\bar{\sigma}$.

As in Section 3.3.3, each node g in T_G is again visited in turn, and the clustering procedure is repeated using the full score function in equation (4). This has the effect of separating phylogenetic groups with very long branch lengths.

3.3.5 Running MIPhy on a large phylogeny of 5,498 animal Cyp proteins

The NCBI genome database (https://www.ncbi.nlm.nih.gov/assembly/organism/) was filtered for all animal genomes that were at a ‘Chromosome’ or ‘Complete’ level of assembly on July 26, 2016, yielding 98 hits. When there were multiple genome assemblies for a single species, only that with the highest number of annotated proteins was kept. Finally, *Bos indicus, Capra aegagrus, Mus spretus, Nasalis larvatus,* and *Nomascus leucogenys* were discarded as they were judged to contain too few protein sequences (these had less than 1,500) to have reliable annotations. The full set of protein sequences for the remaining 58 species was concatenated into one file, which was queried with the 628 vertebrate cyp proteins from Thomas (2007) using BlastP

(Camacho *et al.*, 2009), and resulting in 5,498 hits with an E-value $< 10^{-10}$. These sequences were aligned using Clustal Omega (Sievers *et al.*, 2011) with the command:

clustalo -i INPUT_FILE.fa --threads 10 --log INPUT_FILE-clustalO.log -v --force --use-kimura --iter 10 -o OUT_FILE

The columns of this alignment with <75% gaps were used to build a phylogenetic tree using RAxML (Stamatakis, 2014) with the command:

raxml -s INPUT_FILE.phylip -T 10 -# 5 -m PROTGAMMAWAG -j -p 12345 -n OUT_FILE

3.4 Results

3.4.1 Program input and workflow

This software requires two input files: the gene tree in Newick format, and an information file that contains the species tree (topology only) as well as the assignment of each sequence to one species. The specifics of the tree are not important, and MIPhy can be used to analyze nucleotide or amino acid trees, generated using any kind of algorithm. The cluster analysis is performed in Python, a small local daemon server is started, and an HTML document is launched to display the results. This page has interactive controls and communicates directly with the Python server, allowing the user to see the effects of modifying any of the parameter weights in real time.

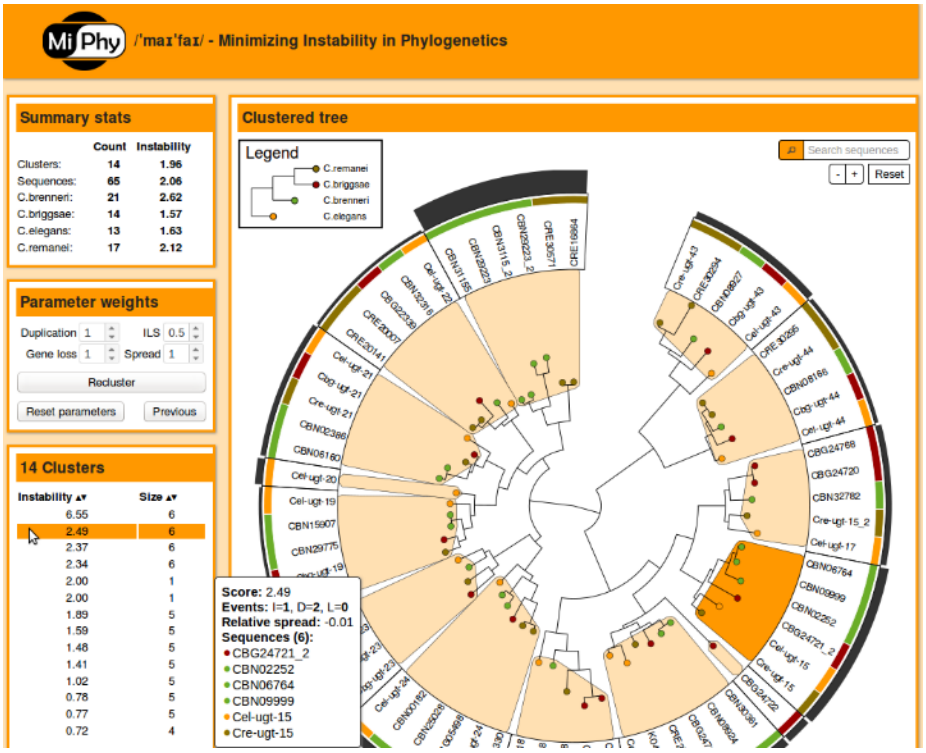


Figure 3.3 MIPhy results interface. Visualization for a subset of the *ugt* gene family sequences from four species of *Caenorhabditis* nematode. The clusters are listed in the table on the left as well as indicated by the light orange shapes on the interior of the tree. The instability of each cluster is visualized by the bar charts on the outside of the tree. The colours of the band just inside of the circle match the colours of the tree nodes, and represent the originating species of each sequence.

3.4.2 The MIPhy interface and run times

The visualization page displays the given gene and species trees, summary statistics for the species and clusters, the current parameter values, and a sortable list of the current clusters (**Figure 3.3**). Clicking on a sequence name will provide additional details, as will clicking on a cluster on the tree or in the list. The page also contains a usage description, and provides options to modify font sizes, the tree width, and the colour of each element (not shown). The tree and legend can be exported and saved as an SVG image file, and the sequences and instability scores from one or more species can be exported and saved as a CSV file.

MIPhy was used to analyze a dataset of annotated vertebrate *cyp* genes, which consists of 628 sequences from 10 species (Thomas, 2007). The algorithm calculated the optimal clustering pattern for the given T_G and T_S in less than 0.2 seconds on a 2.7 GHz laptop. Viewing the clustering information in a web-browser loaded in ~5 seconds. Modifying parameter weights causes the clustering analysis to be rerun, and redrawing the new results is sped up as only a subset of the page elements need to be modified or recreated (<1 second). To determine how MIPhy will scale to cope with forthcoming exponential increases in genome sequences, I analyzed a tree of 5,498 Cyp protein sequences from 58 animal species. MIPhy completed the initial clustering phase in 30 seconds, the optional cluster refinement phase in 7 minutes, and loaded the results in a web browser in 1.5 minutes.

3.4.3 Phylogenetic instability of human Cyp proteins

MIPhy was run with default parameters on the phylogenetic tree of 628 Cyp protein sequences from 10 vertebrate species from Thomas (2007), and the human scores were extracted and graphed (**Figure 3.4**). These scores fell into two broad categories: 31 were unstable with scores in the interval [18.2, 97.5], and 28 were stable with scores in [0.1, 10.9]; of these, 23 had scores in [0.1, 5.7], and the remaining 5 had intermediate scores in [7.8, 10.9].

The highest intermediate scores, Cyp-11B1 (steroid 11β-hydroxylase) and Cyp-11B2 (aldosterone synthase) appear to have been recently duplicated in the terrestrial vertebrates, and likely played a role in the ancient transition from sea to land (Colombo *et al.*, 2006). Their instability score is elevated because rats appear to have two additional genes in that cluster, and no proteins were present in chicken or frog. It is unclear whether they are truly missing or simply absent from the assemblies. The next highest intermediate sequences, Cyp-8A1 and Cyp-8B1, have a likely over-estimated score of 9.5. They are in a phylogenetic group that should likely be split into two, but was not due to the complementary pattern of species with missing genes in each sub-cluster. The fifth intermediate protein, Cyp-27A1, has a high instability score because zebrafish possess a species-specific, paralogous array of three sequences, and frog

possesses three sequences that are fairly distinct from each other. There are also two pufferfish and one additional zebrafish genes nearby in the tree, and it is unclear if they should be included in the Cyp-27A1 cluster.

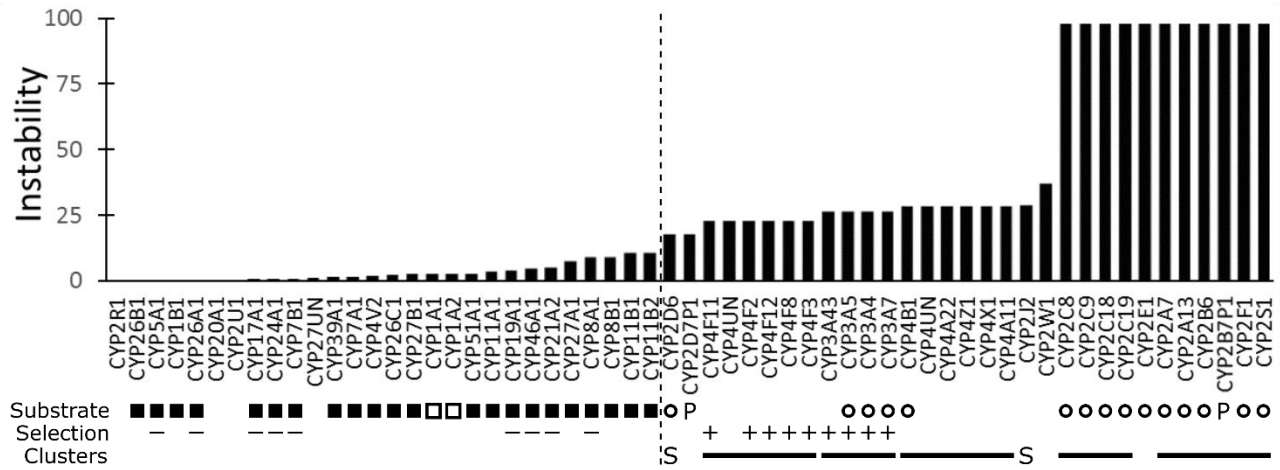


Figure 3.4 The phylogenetic instability of the 59 human Cyp proteins. The vertical dashed line separates the stable from the unstable sequences. 'Substrate' indicates those proteins with primarily endogenous roles (filled squares), primarily xenobiotic roles (empty circles), xenobiotic and endogenous roles (empty squares), and pseudogenes (P). 'Selection' indicates those of the 18 sequences tested that showed evidence of positive selection (+), or no positive selection (-). In the 'Clusters' row, the solid lines indicate those genes that are located in tandem arrays in the genome, or are syntenic with a tandem array in the mouse genome (S). All substrate, selection, and clustering information was taken from Thomas (2007).

The default weight values are set at 1.0, 1.0, 0.5, and 1.0, for duplications, loss, incongruence, and spread, respectively. These values are somewhat arbitrary, though in practice they tend to be robust. The effects of modifying these values are discussed in terms of the clustering pattern, which indicates which sequences are clustered together, and of the cluster rankings, which indicates the instability score of each group relative to the others. Increasing the weight for gene loss had very little effect; at even triple its default value it only caused four small groups out of the 47 from the vertebrate Cyp gene tree to be merged with their sister groups. Decreasing the gene duplication weight was much the same, causing five groups to be merged when it was set to 1/3 its default value. Increasing the weights for duplication and loss together had no effect on

the clustering pattern, and very minimal effect on the cluster rankings. Decreasing both weights together had the same effect as increasing the spread weight, which tended to break up groups. Decreasing the spread weight to zero had very little impact, only merging two singleton groups with their neighbors. Decreasing the incongruence weight had no effect, and increasing it had little impact until it became very high, at which point it tended to break up groups.

3.5 Discussion

Positive selection, pseudogenization, and the presence of tandem arrays of genes are characteristic of rapidly evolving genes, such as those involved in xenobiotic interactions (Thomas, 2007). Though MIPhy did not incorporate any of this information, every human Cyp sequence with these characteristics received a high instability score (**Figure 3.4**). These predictions appear to extend to the functional role of the enzymes as well, as MIPhy performed very well at classifying the human Cyp proteins into those primarily acting on xenobiotic or endogenous substrates. All enzymes with known endogenous functions had low scores, while all but two with primarily xenobiotic substrates had high instability scores; these exceptions were Cyp-1A1 and Cyp-1A2. While the latter is one of the most important human enzymes involved in xenobiotic metabolism, it has been suggested that it also has important endogenous roles (Zhou *et al.*, 2009; Kapitulnik and Gonzalez, 1993), which may have shaped its evolutionary history in the vertebrate species studied here.

The predictions can be extended to species for which there exists less detailed substrate specificity information. The sequences from terrestrial species in the group containing human Cyp-27A1 appear stable, but those of the aquatic or amphibious species do not. This observation suggests that these paralogs may play some role specific to aquatic environments. A similar observation can be made about the cluster containing human Cyp-2W1. It has the second-highest instability score, and of the 43 total sequences there is only one each from human, macaque, mouse, and cow. There are 16 from frog, 10 from zebrafish, and 4 from pufferfish, which would suggest that

these paralogs may also have evolved to metabolize substrates specific to an aquatic environment.

The predictions from this analysis would be complimentary to a between-genes positive selection analysis, which is the most commonly used measure of adaptive evolution. While a codon-based positive selection test measures the patterns of sequence variation, my phylogenetic instability combines the relative sequence variation between species (from the cluster spread and incongruence events) with the history of duplications and losses. However, MIPhy does have its limitations. It is very sensitive to the given gene tree, and does not currently incorporate any measures of uncertainty such as bootstrapping. There are also exceptions to the assumption that phylogenetic instability is a hallmark of adaptive evolution; the most well-known is likely the beta-globin genes that form part of hemoglobin. These genes exhibit sequence polymorphism within and between human populations, lineage-specific expansions and contractions in gene cluster size, and yet continue to play a very vital endogenous role (Hill and Wainscoat, 1986; Opazo *et al.*, 2008).

An unanticipated use of MIPhy is in the naming of genes, specifically towards generating hierarchical naming conventions. Because a sequence identity threshold was used when naming Cyp proteins, one may reasonably assume that Cyp-3A4 and Cyp-3A5 have related functions, as they are likely closely related. Conversely, no such assumptions may be made about many other gene families, whose members appear to have been named in order of discovery. This can pose a problem with the discovery of novel genes. If two species possess the example genes Pqr-21 and Pqr-22, and one of them additionally possesses a paralog to Pqr-21, this paralog will be named with the next available number; perhaps Pqr-42. This single tiered naming system does not accommodate any way to suggest that these proteins are related to each other. I propose that a phylogenetic analysis like MIPhy could be used to cluster such a gene family into sub-families, and that these clusters could be used to inform a multi-tiered naming system that is better able to accommodate newly discovered gene members. This is an issue that is going to arise more often as increasing numbers of species are being sequenced.

This work presents the first algorithm for simultaneous reconciliation and clustering of large gene families that I am aware of. MIPhy has proven to be a valuable tool in identifying members of gene families that exhibit characteristics of adaptive evolution, and agrees very well with the known substrate specificity of human Cyp enzymes. It is a useful tool to gain an understanding of the evolution of large gene families, and to generate hypotheses about the potential functional roles of both the stable and unstable sequences.

3.6 Acknowledgements

I wish to thank Dr. James Thomas for providing me with his past data so that MIPhy could be validated on his published work. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant (#06239-2015) to JDW, a Collaborative Research and Training Experience Program (CREATE) program in Host-Parasite Interactions (#413888-2012) to JDW and JSG. DMC received a training scholarship from Alberta Innovates Technology Futures.

4. Visualizing *C. elegans* copy-number variation using RDepth

4.1 Introduction

The contribution of structural variation to the heterogeneity within a species is considerable; in humans these are responsible for more differences than single base-pair mutations (Conrad *et al.*, 2009; Pang *et al.*, 2010). Copy-number variation (CNV) is a major part of this; one study examined 80 *Arabidopsis* genomes and found 9% of genes were absent in at least one of the genomes (Tan *et al.*, 2012). A duplication or loss of even a single gene can have significant phenotypic impact – exemplified by several diseases (Fellermann *et al.*, 2006; Aitman *et al.*, 2006; Sebat *et al.*, 2007) – and may present a major mechanism of adaptive evolution (Feuk *et al.*, 2006).

Detecting such structural variations used to be done using microarrays (Iafrate *et al.*, 2004), but in recent years next-generation sequencing has surpassed it as the method of choice due to cost and sensitivity (Magi *et al.*, 2012). The process is still difficult using short-read sequencing technology, as a duplicated sequence will generate the same reads as the original, an inversion can only be detected by high quality reads exactly on the boundaries, and a deletion can simply appear to be a region of low coverage. Despite these difficulties, many methods have been developed to use this technology (Campbell *et al.*, 2008; Chiang *et al.*, 2009; Yoon *et al.*, 2009; Magi *et al.*, 2011).

The genome of the N2 strain of the nematode *Caenorhabditis elegans* was sequenced in 1998, and subsequent years of concerted effort have developed this genome assembly into arguably the highest quality of any eukaryotic organism (The C. elegans Sequencing Consortium, 1998). A few years ago, the ambitious Million Mutation Project (Thompson *et al.*, 2013) was launched to provide a resource of the tolerated mutations in *C. elegans* genes. They generated 2,007 mutagenized strains of the nematode, and sequenced the whole genome of each. They also fully sequenced 40 wild isolates from around the globe, producing a collection containing 1.5 million SNPs and deletions spanning 5 million bases. This has become an extremely valuable resource for large-scale phenotypic screens (Mathew *et al.*, 2016; Timbers *et al.*, 2016),

as well as population genetics studies on the wild isolates (Subirana *et al.*, 2015; Cook *et al.*, 2016; Lee *et al.*, 2016).

I have developed RDepth, a web-based suite of tools for visualizing CNV in the genomes of those 40 wild isolates of *C. elegans*. This method uses read depth as a predictor of CNV, in an effort to identify signatures of adaptive evolution without having to directly measure positive selection. The whole genome sequences were taken from the Million Mutation Project data (Thompson *et al.*, 2013), and mapped onto the N2 *C. elegans* reference genome (The C. elegans Sequencing Consortium, 1998). RDepth consists of two complementary tools; GDepth (Gene read Depth) is a tool to analyze a gene family or a list of genes, and displays the average read depth for each gene, from each isolate, to the user. It is available at http://rdepth.wasmuthlab.org/gdepth. CDepth (Coordinate read Depth) is a tool to analyze segments of genome from each isolate directly, and that collects the results from existing CNV detection methods. It is still under development, but is available at http://rdepth.wasmuthlab.org/cdepth.

4.2 Methods

The raw reads for all 40 wild isolates from the Million Mutation Project (Thompson et al., 2013) were downloaded from the NCBI Short Read Archive (accession SRP018046) in 2014. The paired reads for each strain were aligned to the N2 reference assembly using BowTie v2 (Langmead and Salzberg, 2012). The portion of correctly mapped reads was the same as described by the original authors (Thompson et al., 2013), and these data were converted to orar files (a custom file format described in Section 4.2.1).

4.2.1 The orar (ordered array) file format

The alignment files for the *C. elegans* isolates were large files (~320 GB in total), so a more compact binary file format was designed to include only the information required. Functions to read and write were written in Python. Each strain was represented by a dictionary, mapping the name of each scaffold in the reference genome to an array of zeros, one for each nucleotide in that scaffold. For each mapped read in the alignment, the appropriate array was incremented by 1 along the

coordinates of that mapping. This results in a collection of arrays, where the value at each position in the array indicates the read depth at that nucleotide. This is the data structure used in memory.

To save one of these structures to disk, the following values are collected: the orar protocol version (currently 2), the number of scaffolds, and the length of each of those scaffolds. These descriptor values are stored in a Python array of signed ints. For each scaffold (in the same order as the lengths), the following values are appended: the length of that scaffold name, the scaffold name in characters, and the read depths for that scaffold as an array of unsigned shorts.

To load one of these structures from a file, the reverse procedure is followed. The first integer worth of data in the file is read, which indicates the orar protocol version that should be used. The next block is also read as an integer, which indicates the number of scaffolds. This instructs the software to read that many integers worth of data from the current file position, which are the lengths of those scaffolds. The next integer indicates the length of the first scaffold name, and then that many bytes of data are read interpreted as characters. The first scaffold length indicates how much of the file to read as an array of unsigned shorts, and this array is then associated with the name of that scaffold. One integer worth of data is then read, which indicates the length of the second scaffold's name, and this process is repeated until all of the data has been parsed. This file format allowed the read depth data for every nucleotide to be stored using 200 MB for each *C. elegans* isolate, reducing the footprint of all 40 isolates from 320 GB to 8 GB. It is suitable to encode read depths up to 65,534.

4.3 RDepth specifications

RDepth is designed to be a collection of connected tools used to explore intra-species CNV in some population. GDepth can be used to visualize summary statistics for some genes of interest, and analyze them as a whole. CDepth is designed to explore a single region of the genome in greater detail, from either a single isolate or the entire sequenced population. The two tools will be linked, so that it is simple for a user

to move between the higher-level information from GDepth and the greater detail of CDepth. This can be useful in determining if some gene with apparently high read depth is part of a segmental duplication, or if the depth is simple due to a localized sequencing or mapping bias.

4.3.1 GDepth structure

The genomic coordinates and name of every gene and pseudogene were extracted from the N2 annotation file by a small Python daemon running on our lab server. It also loads the read depth data, and provides the interface for the client HTTP requests to access. The RDepth web page is written using HTML and CSS, with Javascript providing the functionality. When the user indicates their selection of genes to visualize, the list of names is sent to the lab's server, which returns the read depth information for those genes. The graph is drawn using the D3 Javascript library (https://d3js.org).

4.3.2 GDepth interface

Upon reaching the GDepth page, the user is presented with three text boxes (**Figure 4.1A**). The first is labelled 'Gene families to include', and will include all genes with names starting with the user's query. If the user enters 'gst', the site will return data for the *gst* genes, but also the *gstk* and *gsto* genes. This can be used to analyze subsets of gene families; for example, entering 'gst-2' will display results for *gst-2* but also the genes *gst-20* – *gst-29*. The middle text box is labelled 'Gene families to exclude', and 'gstk, gsto' could be entered here to limit the results to only the *gst* genes. The final text box is labelled 'Genes to match exactly', and this will display data for only those genes with names entered in a comma-, space-, tab-, or newline-separated list. WormBase sequence identifiers can also be used here instead of gene names (ex: K08F4.11 instead of gst-3).

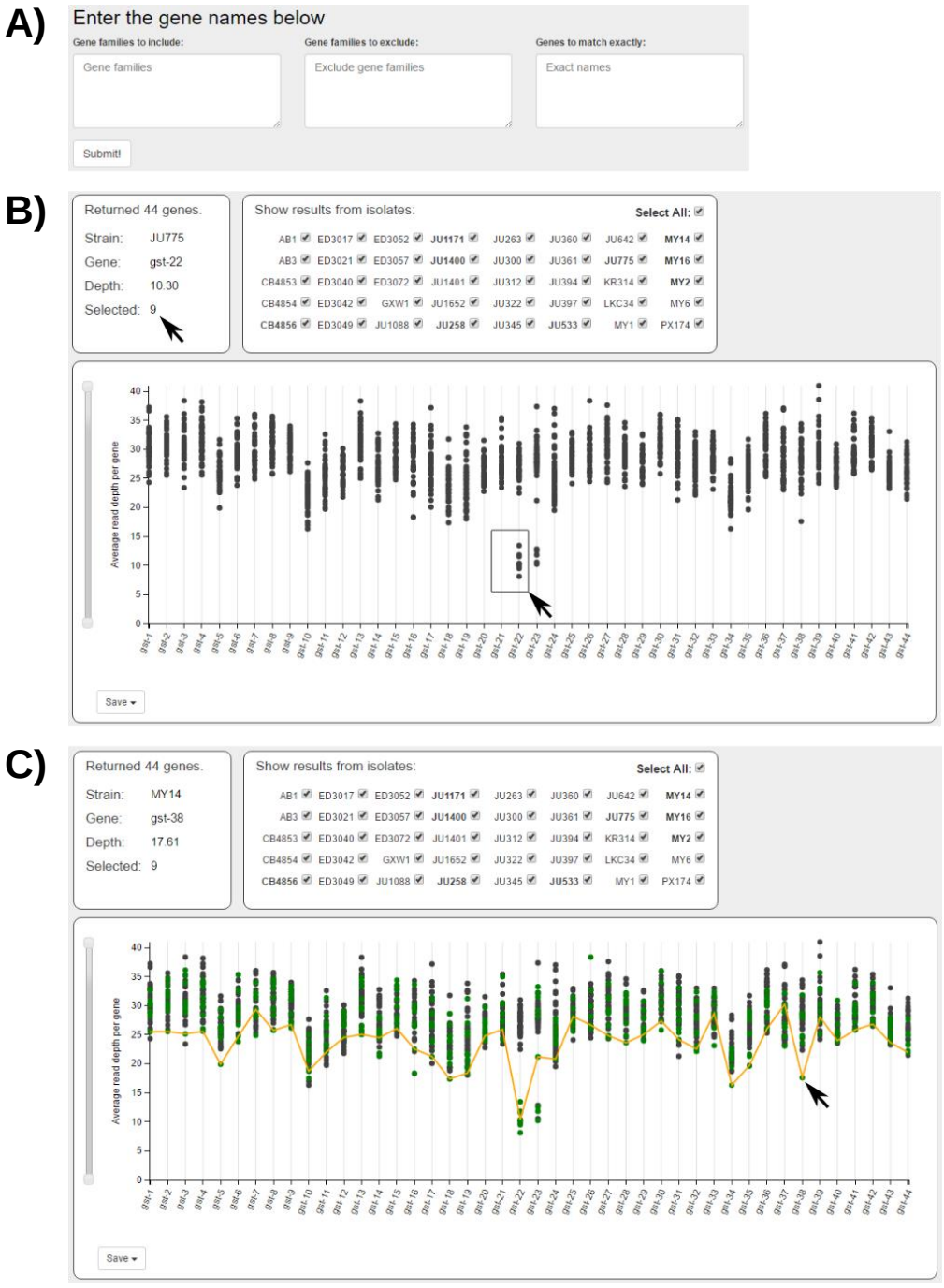


Figure 4.1 The GDepth interface. A) shows the search input boxes. **B)** shows the read depth for the *gst* gene family, and the click-and-drag functionality; the top arrow indicates there are data points from nine isolates inside of the box. **C)** shows the effects of colouring those nine isolates green, and hovering the mouse over a single point; the orange line indicates the read depth of each other gene from that isolate.

The results interface is displayed in **Figure 4.1B**. The top-left box displays information, including the number of genes returned by the user's search, and the isolate, gene, and read depth of the data point under the user's mouse. The 40 isolates are listed in the top-middle box, and the checkboxes by each name allow the user to display only a subset of the isolates if desired. It can be difficult to discern how many data points are overlapping at one location, so the user can drag a box with the mouse to display this. Isolates can be recoloured individually or as a group (**Figure 4.1C**), which can be useful to check for read depth biases. The orange line in **Figure 4.1C** indicates the read depth of the *gst* genes from isolate MY14, and it is apparent that it is consistently lower than in the other isloates.

The dark grey vertical bar beside the graph is a zooming interface, which allows the user to expand the Y-axis resolution of the graph. The 'Save' button in the bottom-left corner of the graph allows the user to save the current view of the graph as an SVG image. Finally, the 'Formatting options' box allows the user to specify the width and height of the graph, and is used to colour the data from isolates.

4.4 GDepth analysis of the cyp gene family

Figure 4.2 shows several potential cases of CNV in the *C. elegans cyp* genes from these 40 isolates. Both *cyp-31A2* and *cyp-35A4* appear to be duplicated in two or more isolates, as the average read depth of these genes is over double what is seen in the other isolates. The most prevalent CNV is found in *cyp-33D3*, which appears to be completely lost in 25 of the 40 isolates. There are three genes that appear to have a read depth of approximately half the background level, *cyp-13A5*, *cyp-14A4*, and *cyp-35B2*, which could indicate a heterozygous deletion in five isolates. The read depth depends on mapping short reads to the N2 reference genome, so it is possible that these genes in these isolates have high sequence divergence, causing many of the reads to simply fail to be mapped. If true, this would still be an alternate indication of adaptive evolution. Or, as there is only a single isolate with low read depth for those genes, these could simply be false positives.

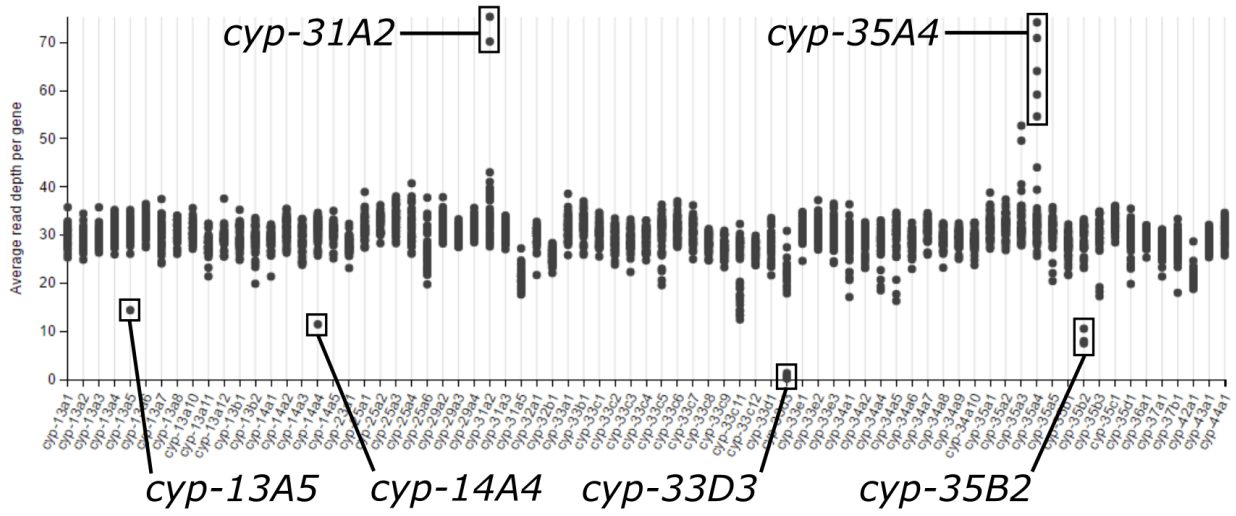


Figure 4.2 GDepth results for the *cyp* gene family. This is the view after entering 'cyp' into the 'Gene families to include' search box. The top-left box displays several descriptive statistics, and the *C. elegans* isolates are indicated in the top-middle box.

4.5 Discussion

This tool was developed to detect potential signatures of adaptive evolution in wild isolates of *C. elegans*. Genes that exhibit positive selection may also experience intra-species CNV, but it has been suggested that the converse may not hold in general; that intra-species CNV is not sufficient evidence to predict positive selection (Bush *et al.*, 2014). Instead, it may identify genes under relaxed purifying selective pressure. This could still be a useful indication of xenobiotic functionality, as proteins with vital functions are less likely to be able to tolerate duplications or losses (Kondrashov, 2012). However, based on the correlation between phylogenetic instability and adaptive evolution seen in Chapter 3, I hypothesize that there also exists a correlation between intra-species CNV and adaptive evolution, at least for the *C. elegans* Cyps. This could be tested by measuring the ratio of non-synonymous to synonymous mutations in the *cyp* genes within this population, but that is beyond the scope of this thesis. If the correlation does hold, it would be very useful in an analysis of Cyp enzymes, which contain some members that metabolize endogenous compounds, and others that

detoxify xenobiotics (Thomas, 2007); these two classes of functions experience dramatically different selective pressures.

4.5.1 Future directions

Another useful extension would be to extend GDepth to other species. Full genome sequencing data for 37 wild isolates of *C. briggsae* has been generated (Thomas *et al.*, 2015), and it should be simple to incorporate that information into the existing GDepth framework. The first steps towards this goal have already been taken, as the sequence reads from these samples have been mapped to the reference *C. briggsae* genome assembly, and those results are contained in orar files. Another research group under Erik Andersen has recently sequenced the genomes of 500 isolates of *C. elegans*, and they are interested in visualization similar to RDepth as a comparative genomics tool. While this software would certainly require substantial modification to accommodate such a data set, I feel that the general design of RDepth would be appropriate and valuable. A related extension that would significantly broaden the use of the software would be to enable users to run GDepth on their own sequencing data from their population of interest. These data sets tend to be very large, so this application would be unlikely to remain as a web tool. Instead, I could modify the existing code so that a user could run it locally and still use the same HTML interface, much like with MIPhy (Chapter 3). The user would require a genome assembly, a mapping of their isolates to that assembly, and an annotation file for that assembly. I would develop a script to extract the necessary information from those files into a format usable by the RDepth tools.

5. A comparative analysis of the detoxification gene families of free-living nematodes

5.1 Introduction

Parasitic nematodes are a global problem, responsible for an estimated two billion current human infections and a greater burden of disease than malaria or tuberculosis (World Health Organization, 2005; Hotez *et al.*, 2009). Parasites of livestock arguably cause a greater burden to humans by reducing animal production; the costs to the North American cattle industry are estimated at $2 billion annually (Stromberg and Gasbarre, 2006). The situation in small ruminants is even worse, primarily due to the prevalence and pathogenicity of the trichostrongylid *Haemonchus contortus*, which can severely impact income and food availability in resource-poor settings (Vattaa and Lindberg, 2006). These parasites have traditionally been controlled using anthelmintic drugs, but resistance has arisen rapidly and spread around the globe; it is no longer uncommon to find sheep or goat farms with high levels of resistance to all available anthelmintic drugs (Kaplan, 2004; Kaplan and Vidyashankar, 2012).

5.1.1 Species of interest

Studying these parasitic species can be challenging, as most cannot be grown *in vitro*. The quality of their genomic resources is also quite variable, so many studies first analyze related free-living organisms like *C. elegans* (Burns *et al.*, 2015). Besides being easily manipulatable in a laboratory setting, *C. elegans* may have the highest-quality genome assembly of any organism. As this work is a comparative analysis I included three other free-living *Caenorhabditis* species with high quality genomes. I used *Pristionchus pacificus* as an outgroup, which is another free-living Clade V nematode, though it is estimated to have diverged from the *Caenorhabditis* ancestor species between 280 and 430 million years ago (Dieterich *et al.*, 2008). The quality of these genomes is summarized in **Table 5.1**.

Table 5.1 Genome statistics for the nematodes used in this work. These values describe the genome assemblies used here. The quality is described by the final four columns. The ideal number of scaffolds should equal the number of chromosomes, and the scaffold N50 should be as large as possible. The 'CEGMA found' column provides an upper bound on the percentage of gene models likely to have been found, and the 'CEGMA ratio' provides a measure of homozygosity (ideally at 1.0).

Nematode	Size (Mb)	Number genes	Number scaffolds	Scaffold N50 (kb)	CEGMA found (%)	CEGMA ratio
C. elegans	100.2	20,056	7	17,493	97 / 99	1.1 / 1.2
C. briggsae	108.5	22,425	367	14,512	98 / 99	1.1 / 1.2
C. brenneri	190.4	30,670	3,305	382	98 / 100	1.7 / 1.8
C. remanei	145.4	31,476	3,670	436	93 / 97	1.3 / 1.5
P. pacificus	172.5	23,500	18,083	1,244	85 / 94	1.2 / 1.3

5.1.2 Anthelmintic resistance

Drug resistance often arises by mutation of the drug target, but may also be via the metabolism and excretion of the compound by the pathogen. In parasitic nematodes, the former appears to be the case with benzimidazole resistance (Kwa *et al.*, 1994, 1995; Ghisi *et al.*, 2007; Saunders *et al.*, 2013), though other studies have also implicated additional mechanisms such as drug efflux (Blackhall *et al.*, 2008; Lespine *et al.*, 2012). The drug targets of the other classes of anthelmintic drugs are not as well understood and many genome-wide association studies have failed to identify a conserved mechanism of resistance, which suggests that the mechanism may be more complex (Gilleard, 2013). In mosquitos and flies, wild isolates with resistance to the insecticide dichlorodiphenyltrichloroethane (DDT) do so by modification or overexpression of a cytochrome P450 (*cyp*) gene (Dunkov *et al.*, 1997; Amichot *et al.*, 2004; Chiu *et al.*, 2008). Evidence also suggests that drug metabolism plays an important mechanistic role in resistance in the trematode *Fasciola hepatica* (Devine *et*

al., 2009; Alvarez *et al.*, 2005), so I hypothesize that this may be an important mechanism in helminths in general.

5.1.3 Xenobiotic metabolism

Xenobiotics are compounds that an organism encounters that do not serve as a source of energy or as precursors for biomolecules, such as secondary metabolites from plants, fungi, or bacteria, environmental pollutants, or drugs. Because many of these compounds may be toxic, all organisms have evolved various means of protecting themselves via detoxification, which proceeds in three phases (Omiecinski *et al.*, 2011). The first phase is called functionalization, in which the compound is metabolized to expose some reactive group. This is carried out by gene families including the Cyps, the Flavin-containing monooxygenases (Fmos), and short-chain dehydrogenases (Sdrs). The second phase is conjugation, which involves adding an endogenous hydrophilic molecular moiety to the compound. This is performed by gene families such as the UDP-glycosyltransferases (Ugts), glutathione-S-transferases (Gsts), and sulfotransferases. The third phase is excretion of the substrate out of the cell by one of several ABC transporter proteins, often a member of the P-glycoproteins. Phase I reactions may or may not be necessary prior to Phase II, and different compounds may be able to be excreted at any point in the detoxification process. There is very little known about potential sulfotransferases in *C. elegans*, so for this work I will focus on the other five Phase I & II gene families.

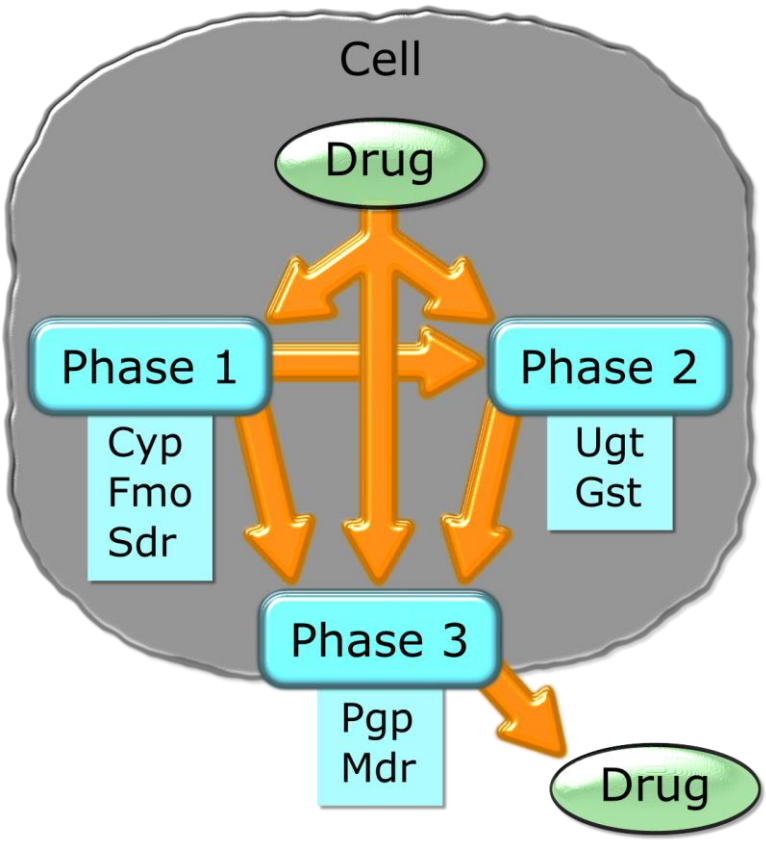


Figure 5.1 Schematic of xenobiotic metabolism. This figure shows how the detoxification gene families act in concert to protect the cell from harmful xenobiotics. This is **Figure 1.2**, which has been reproduced here for ease of reading.

5.1.4 Metabolism and drug resistance

Though mechanisms have not yet been elucidated, several parasitic species have shown evidence of drug metabolism, including the nematode *H. contortus* (Vokřál *et al.*, 2012), the trematodes *F. hepatica* (Scarcella *et al.*, 2012, 2013) and *Dicrocoelium dendriticum* (Skálová *et al.*, 2010), and the tapeworms *Hymenolepis diminuta* (Bártíková *et al.*, 2012) and *Moniezia expansa* (Prchal *et al.*, 2015). If the mechanisms of anthelmintic metabolism can be identified, novel drugs targeting these enzymes may restore efficacy to the existing anthelmintic drugs (Kerboeuf *et al.*, 2003). In previous studies, up to 50% of the resistance to benzimidazoles has been shown to be reverted by using P-glycoprotein (Pgp) inhibitors in *H. contortus* (Beugnet *et al.*, 1997; Kerboeuf

et al., 1999). Significant reductions in ivermectin and moxidectin resistance were also seen when co-administered with a Pgp inhibitor (Molento and Prichard, 1999).

5.1.5 Objectives

Here, I present a comparative genomics analysis of five detoxification gene families (Cyps, Fmos, Sdrs, Ugts, and Gsts) in five species of free-living nematode: *C. elegans*, *C. brenneri*, *C. briggsae*, *C. remanei*, and *P. pacificus*. Hundreds of these genes are currently unannotated, so this work will help to describe these important gene families. The comparative analysis will also use MIPhy (Chapter 3) to classify these genes based on their likelihood of interacting with environmental substrates. Potential xenobiotic metabolizers may provide a mechanism for the reintroduction of existing anthelmintics, while those proteins predicted to act on endogenous compounds may prove to be effective therapeutic targets for future drug discovery endeavours.

5.2 Materials and methods

5.2.1 Nematode genomes and their quality

The genomic resources for the nematode species were downloaded from WormBase (www.wormbase.org) release 255. These include the genome, the protein sequences, and the gff3 annotations file from these assemblies: *C. brenneri* PRJNA20035, *C. briggsae* PRJNA10731, *C. elegans* PRJNA13758, *C. remanei* PRJNA53967, and *P. pacificus* PRJNA12644.

CEGMA v2.5 (Parra *et al.*, 2007, 2009) was used as an indication of genome quality. It contains a core set of 248 genes that all eukaryotes are expected to have exactly one copy of, and uses BLAST to search the genome of interest for evidence of those genes. These are likely to be the easiest genes to find in any eukaryote genome, so it provides an upper bound on the proportion of other genes that are likely to be present in the assembly. It was run on using the command:

cegma -T 7 -g GENOME_FILE

5.2.2 Pipeline summary and notation

In this section, I developed a multi-step pipeline and processed five gene families from five nematode species. This involves many intermediate files, so I refer to them with the following naming scheme: TYPE_SPECIES_GENEFAMILY_subtype. TYPE refers to the type of file: INIT means the initial set of sequences, FIG means sequences found by Figmop, EXO is the output of the program Exonerate, SEQ is a final set of sequences, ALN is an alignment, PROTTEST is the output of the program ProtTest, and TREE is a phylogenetic tree. SPECIES indicates the originating species of the data in the file: 4CP indicates the four species of *Caenorhabditis* plus *P. pacificus*, CEL indicates *C. elegans*, NO-CEL indicates all species except *C. elegans*, etc., and GENEFAMILY denotes the gene family (SET is used is a general placeholder in examples). Subtype is not used for some files, but further classifies others.

The main purpose of this pipeline is to extract high-quality coding sequences so that the subsequent phylogenetic analysis will be accurate. I first gathered all named protein sequences for each gene family of interest, and then used Figmop (described in Chapter 2) to expand this set with existing unnamed protein sequences. This set was then matched back to the individual genomes using a splice-aware aligner, to expand it to include unannotated sequences (those for which there are no current gene models). For each species this was done using the total set from Figmop, and using all sequences except those from that species. This second search was done to generate a set of sequences that were unbiased by any errors currently in each species' gene models. Both sets of protein sequences were manually inspected to ensure they contained the expected patterns of sequence motifs before I aligned them and built phylogenetic trees. These steps are described in greater detail in the following sections, and illustrated in **Figure 5.2**.

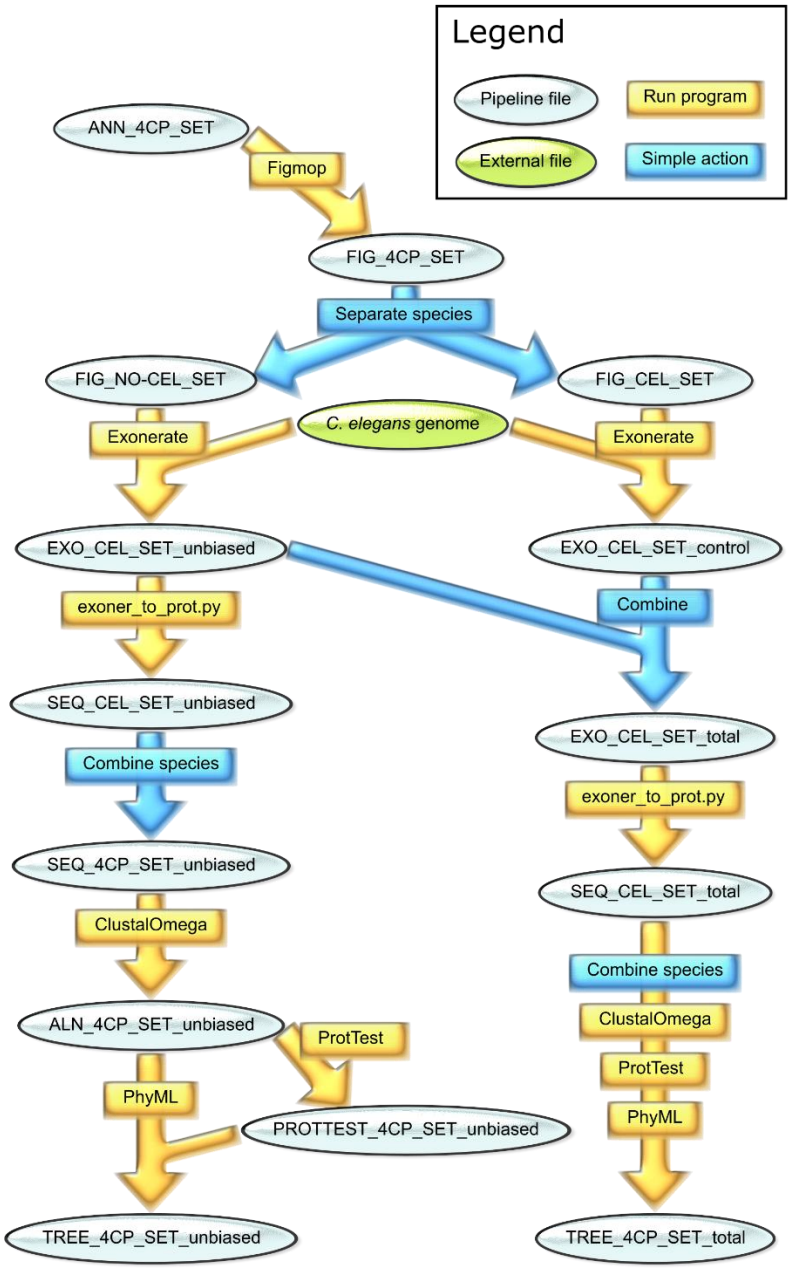


Figure 5.2 Flowchart of the gene family analysis pipeline. This pipeline begins with a positive set of protein sequences and yields two phylogenetic trees. It was run for each of the five detoxification gene families. The last arrow on the right, containing four steps together, indicates that SEQ_CEL_SET_total undergoes the same steps as SEQ_CEL_SET_unbiased; the details have been hidden for the sake of brevity.

5.2.3 Collecting the protein sequences

I have previously found that many gene models in the genome assemblies appear to be wrong, most commonly because a gene model is fused with its closest neighbour on the genome, or a gene model is split into two. To overcome this, I gathered the protein sequences for each gene family using a multi-step pipeline. First, I gathered all annotated sequences from WormBase, yielding a high-quality set of sequences (termed ANN_4CP_SET), and ran MEME v4.9.1 (Bailey and Elkan, 1994; Bailey *et al.*, 2006) to find a robust pattern of motifs. The protein sequences for each species were then searched with this pattern using Figmop, yielding all gene models that fit the pattern, annotated or not (termed FIG_4CP_SET). The exception is the short-chain dehydrogenases, as they appear to be annotated as several different, but related, gene families. For these, I collected all of the protein sequences from the five nematodes that were a matched by the Pfam motif PF00106 (accessed at http://pfam.xfam.org/family/PF00106 on November 1, 2016). Many of the hits in *C. elegans* are from the Dhs gene family, but four are from the Stdh family, seven are named as Let-767, Drd-5, Sdz-8, Ard-1, Dhrs-4, Decr-1.2, and Decr-1.3, and the rest are currently unnamed. For clarity, I will collectively refer to these hits as the Sdr gene family.

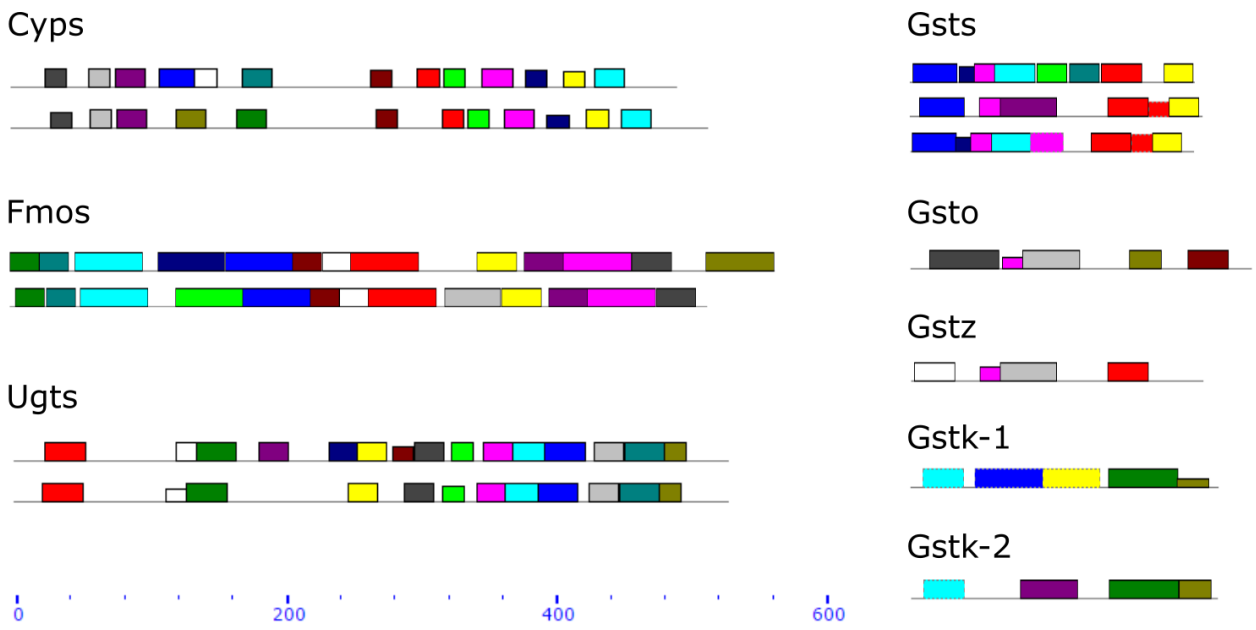


Figure 5.3 Patterns of sequence motifs used by Figmop. The motif pattern diversity for each gene family is shown, where each coloured bar represents one sequence motif, and the height of the bar indicates how well that motif matches the sequence; the scale on the bottom indicates the length of each sequence in amino acids. Each set of motifs was generated separately, so while the cyan motifs in the two Cyp patterns are the same, that motif is in no way related to the cyan motifs in the other gene families. The Gst subtypes are displayed on the right, and these do share one common set of motifs.

5.2.4 Extracting coding sequences

The FIG_4CP_SET files contain many incorrect gene models (see **Figure 5.4**), primarily because they were predicted using general-purpose gene finding software such as AUGUSTUS (Stanke and Waack, 2003; Stanke *et al.*, 2008). These programs work very well for most genes, but one set of parameters is not expected to perform perfectly for every gene.

Figure 5.4 Examples of malformed gene models from *P. pacificus*. These patterns should be interpreted by comparing them to the Cyp motif pattern in **Figure 5.3**. **A)** shows a gene model with an apparent 400 amino acid insertion in the middle, **B)** shows two Cyps that have apparently been fused into a single gene model, **C)** shows two gene models that appear to be the front half of Cyps, and **D)** shows two gene models that appear to be the back half of Cyps.

To overcome these errors with targeted information, I used the splice-aware aligner Exonerate v2.4.0 (Slater and Birney, 2005) to align these protein sequences back to the raw genomes. As the sequences from one species will align best to that genome and I wanted to produce gene models unbiased by errors already present in the annotated sequences, I aligned the sequences from all species except that one to the genome of that species. Exonerate searches for a set of exons that, when translated, most closely align to the query protein sequences. It chooses sets of splice donors and acceptor sites to optimize this matching, which occasionally results in stop codons being included in the coding sequence. I therefore also included a control procedure to test the accuracy of this process, by aligning the protein sequences from one species to its own genome. To perform these searches, FIG_4CP _SET was split into ten subsets. Those sequences from *C. elegans* were extracted as FIG_CEL_SET, and those sequences from all species except *C. elegans* were extracted as FIG_NO-CEL_SET; this was repeated for all five species. FIG_CEL_SET was aligned against the *C. elegans* genome assembly as a control to yield EXO_CEL_SET_control, FIG_NO-CEL_SET was aligned against the *C. elegans* genome to yield EXO_CEL_SET_unbiased, and these two files were merged to yield EXO_CEL_SET_total; this was repeated for each species resulting in 15 sets of Exonerate hits for each gene family. An example of the commands used:

exonerate --model protein2genome --dpmemory 2000 –refine region --showalignment no FIG_CEL_SET ELEGANS_GENOME > EXO_CEL_SET_control

Each query sequence may match to dozens or hundreds of locations in the subject genome, and so each region of the genome with a potential hit tends to have many overlapping but distinct predicted gene models. This required a method to choose a single predicted gene model for each genomic region, so they were ranked using the quality of the match with the script exoner_to_prot.py. This originally proceeded by dividing the bit score of the alignment by the length of the generated coding sequence. Small spurious matches between the genome and some query sequence can subtly influence that hit, resulting in shifted splice donor and acceptor sites. This can in turn influence the predicted coding sequence, by including sequence that should be excluded, or missing true coding sequence. This often manifests as stop codons just before or after a true exon, so as a penalty the match quality is divided by the number of stop codons. Many of the existing gene models are only the front or rear half of the true sequence, and if it is a good match, these fragments may have the best match quality. So, this score is further modified by penalizing coding sequences that deviate from the expected coding sequence length. Finally, if there are tandem arrays of genes in some region, it is also common for some Exonerate hits to include the front half of one true gene, followed by a large intron that covers the rest of that true gene and the front half of the next true gene, and then include the back half of that second true gene. To avoid this, predicted gene models that deviate from the expected gene model length are also penalized. After some experimentation, I found an effective ranking function to be:

$$score = \frac{b - \theta_p \cdot |E[p] - p|}{\theta_g + \frac{|E[g] - g|}{E[g]}} \cdot \frac{1}{s + 1}$$

where b is the match bit score; $p, E[p], \theta_p$ are the predicted protein length, the expected protein length, and the weight of this penalty; $g, E[g], \theta_g$ are the predicted gene model length, the expected gene model length, and the weight of that penalty; and s is the number of stop codons in the predicted coding sequence. In this equation $\theta_p =$

1 places more weight on deviations from the expected protein length while $\theta_p = 0.1$ places less, and $\theta_g = 0.1$ places more weight on deviations from the expected gene model length while $\theta_g = 2.0$ places less. Various parameter values were evaluated for each gene family, attempting to optimize the number of hits returned and the correctness of those hits (as measured by the number of stop codons and how well the hits match the Figmop motifs). The best θ_p values were found between 0.1 and 0.5, and the best θ_g values were found to be 1 or 2. Though this ranking procedure does not have a theoretical basis, the quality is demonstrated by the control hits in **Table 5.2**.

The protein files were examined again using Figmop and any sequences with fewer than 5 or 6 of the 15 – 20 expected sequence motifs (depending on the gene family) were discarded as being false positives. The remaining gene models were then annotated using the gff3 file for each species, where the existing gene model name was kept that covered the highest percent of each predicted gene model. If there was no existing gene model that covered at least 35% of the predicted model, a new name was assigned as 'Spc-unk-1', where 'Spc' is the 3-letter abbreviation for each species (Cel for *C. elegans*, Cbn for *C. brenneri*, Cbr for *C. briggsae*, Cre for *C. remanei*, and Ppa for *P. pacificus*). The unbiased hits were collected together (for all species except *P. pacificus*, for which I used the total hits) into SEQ_4CP_SET_unbiased, and the total hits for all species were collected as SEQ_4CP_SET_total (see **Table 5.2**).

5.2.5 Phylogenetic analyses

Both of the final protein sequence files (SEQ_4CP_SET_unbiased and SEQ_4CP_SET_total) from each gene family were aligned using ClustalOmega v1.2.3 (Sievers *et al.*, 2011) with the command:

clustalo -i SEQ_4CP_SET_unbiased -o ALN_4CP_SET_unbiased --threads 10 --iterations 100 -v

Columns that contained >90% gaps were removed from the alignment, and the model of evolution that best fit these filtered alignments was determined with ProtTest v3.4.2 (Darriba *et al.*, 2011) using the command:

prottest3 -i ALN_4CP_SET_unbiased -o PROTTEST_4CP_SET_unbiased -threads 15 -all-distributions -AICC -BIC -verbose

The LG+G model was found to be most appropriate for all alignments, and so the phylogenetic trees TREE_4CP_SET_unbiased and TREE_4CP_SET_total were built using PhyML v20120412 (Guindon and Gascuel, 2003) with the command:

phyml -i ALN_4CP_SET_unbiased -d aa -m LG -f e -v 0 -a e -s BEST -b -4 -o tlr --no_memory_check

5.3 Results

5.3.1 Sequence collection and alignment

The three protein files generated for each species using Exonerate (**Table 5.2**) have different properties. For all species, the control set was not missing any genes, though in nearly all cases a few stop codons were incorporated into some sequences. Manual inspection of these sequences shows that these are due to exon boundaries that had been incorrectly extended into the adjacent intron. This indicates how well Exonerate can take the protein sequences of one gene family and align them back to their originating genome. The unbiased hits are derived using non-native query sequences, so any hits returned should be free of the fragmentation and fusion errors that led to existing malformed gene models, but they are expected to incorporate extended exon errors at least as often as the control set. The total set is expected to be like the control set, but any malformed gene models may be corrected by the sequences from the other species. There are a few cases where the total set has more hits than the control set, which may be evidence of pseudogenes, or due to the inclusion of true genes that were missed in the annotation of that genome assembly, or simply false positives. In most

cases the unbiased hits were of similar quality to the control set, but in every case the unbiased hits for *P. pacificus* were of very poor quality, returning low numbers of hits and with coding sequences that were often only half of their expected length. For this reason I used the *P. pacificus* total hits moving forward, even in SEQ_4CP_SET_unbiased.

Table 5.2 Results of extracting coding sequences using Exonerate. The three values under each gene family are the number of control, unbiased, and total hits, respectively, where the number in parentheses indicates the number of stop codons in the sequences. The two values in each column of the 'Totals' row are the lengths of SEQ_4CP_SET_unbiased and SEQ_4CP_SET_total, respectively.

	Cyp			**Ugt**			**Gst**			**Sdr**			**Fmo**		
C. brenneri	95 (9)	90 (1)	91 (1)	112 (5)	110 (7)	110 (4)	65 (0)	65 (0)	66 (0)	114 (2)	118 (3)	119 (1)	7 (0)	7 (0)	7 (0)
C. briggsae	72 (2)	74 (4)	74 (2)	80 (1)	82 (10)	83 (6)	39 (0)	40 (0)	40 (0)	76 (1)	78 (0)	78 (0)	4 (0)	4 (0)	4 (0)
C. elegans	80 (2)	76 (4)	79 (2)	74 (3)	72 (2)	74 (1)	54 (2)	54 (1)	54 (2)	80 (2)	79 (1)	80 (1)	5 (0)	5 (0)	5 (0)
C. remanei	81 (0)	78 (0)	82 (1)	84 (2)	81 (4)	81 (1)	35 (0)	37 (2)	37 (0)	99 (0)	95 (0)	99 (0)	7 (0)	7 (0)	7 (0)
P. pacificus	171 (5)	65 (13)	168 (3)	148 (7)	78 (8)	148 (7)	73 (3)	36 (2)	76 (4)	115 (5)	91 (3)	113 (0)	11 (2)	9 (1)	11 (2)
Totals		486 (12)	494 (9)		493 (30)	496 (19)		272 (7)	273 (6)		483 (4)	489 (2)		34 (2)	34 (2)

The alignment of each set of sequences is described in **Table 5.3**. The total alignment length is a function of the sequence length, the number of sequences, and the diversity present in each set. The difference between the total alignment length and the filtered length represents the sequence variation discarded because it was common to fewer than 10% of sequences, a property most evident in the Cyps. The difference between the filtered alignment length and the average protein length provides an indication of the number of informative positions of each sequence, with more

conserved sequences requiring fewer additional columns (which is why the Fmo alignments are so similar). However, a small difference could also indicate that the sequences have both few informative positions and little in common with the rest of the set, so the number of residues remaining in the five least informative sequences were also measured; the difference between these values and the average protein length indicates how much of each sequence was discarded as being uninformative. The very low informative sequence lengths from the ALN_4CP_SDRS_total indicates that it is a very poor alignment, and so it was abandoned at this step. This was unexpected, as all other measures seemed to indicate it was very similar to the unbiased alignment.

Table 5.3 Alignment qualities of the five gene families. This table shows the quality of the alignments generated for the detoxification gene families, as described in Section 5.2.5. The two columns under each gene family show the stats for ALN_4CP_SET_unbiased and ALN_4CP_SET_total, respectively. 'Least informative sequences' indicates the five shortest lengths of non-gap columns in the alignment.

	Cyp		**Ugt**		**Gst**		**Sdr**		**Fmo**	
Total alignment length	2798	3095	1861	2074	876	905	1151	1349	625	658
Filtered alignment length	652	758	573	611	252	266	506	506	592	592
Least informative sequences	199 - 201	193 - 203	174 - 205	185 - 205	111 - 133	96 - 136	71 - 134	11 - 21	347 - 378	347 - 378
Average protein length	462	476	460	473	213	215	282	286	495	497

5.3.2 MIPhy analyses of the detoxification gene families

The nine phylogenetic trees of the nematode detoxification gene families were analyzed using MIPhy (described in Chapter 3). One property common to every tree was that most of the *P. pacificus* sequences were well separated from the *Caenorhabditis* sequences, and tended to be found in many adjacent small clusters (see **Figure 5.5** for a representative example). It is also important to note that the assemblies of *C. remanei* and *C. brenneri* are known to contain many misassembled alleles (Erik C. Andersen, personal communication, 2016). This occurs when one allele that is nearly identical to its partner is placed onto a small contig instead of being merged with the original. The effect is that these species may appear to have a paralog to some protein, but the higher the sequence identity the higher the likelihood that is an artifact of the assembly instead of a true protein sequence.

5.3.3 MIPhy analysis of Fmos

The Fmos trees were identical except for one short branch, and so only TREE_4CP_FMOS_total is shown in **Figure 5.5**. From this figure, it appears that the most stable cluster is Fmo-4, as the sequences are very similar and each species has one copy. The apparently duplicated *C. brenneri* Fmo-4 genes are identical, except that one is missing 89 amino acids of the C-terminus, and so this seems likely to be an artifact of the genome assembly instead of a true paralog. The tree also suggests that the lone *P. pacificus* sequence near that cluster is likely to be an ortholog of Fmo-4, even though the current version of the annotation contains no gene model at this genomic region.

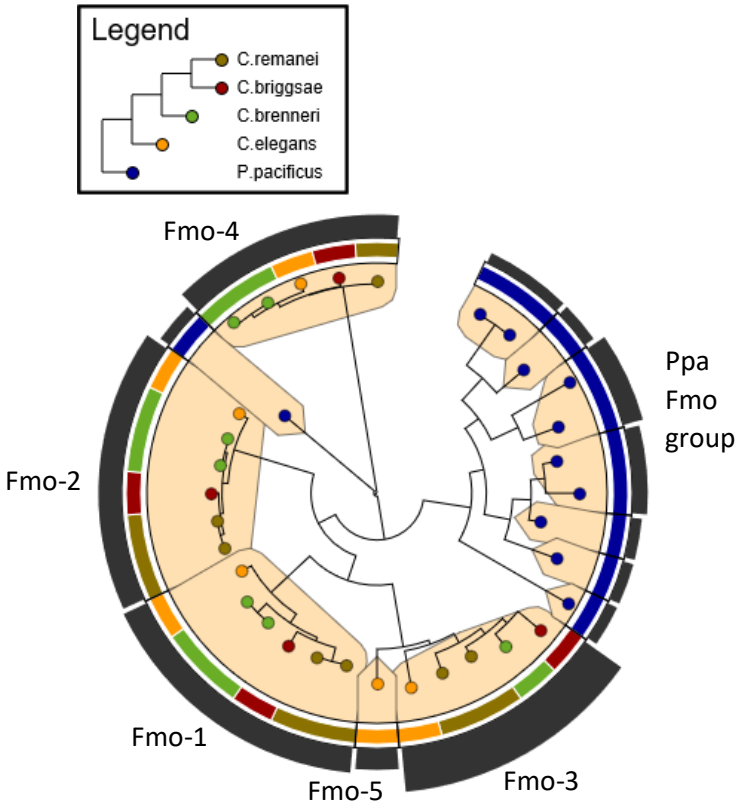


Figure 5.5 MIPhy tree of the Fmo gene family. The clusters were predicted using the default MIPhy parameter weight values, and they are named following the annotations of the *C. elegans* sequences.

The remaining sequences all appear to have arisen from a sequence distinct from the ancestor of Fmo-4. The ancestor of the *Caenorhabditis* species appears to have had three paralogs of this gene that have all survived into the present-day species. In the case of Fmo-1 and Fmo-2, the gene relationships mirror the species relationships, and are all very similar. There appear to be duplications in *C. brenneri* and *C. remanei*, but as the paralogs share 98-100% sequence identity they too are likely artifacts of the genome assemblies rather than separate genes.

The Fmo-3 cluster appears to be more unstable, with duplications in *C. remanei* and *C. elegans* and an incongruent phylogeny. Finally, the *P. pacificus* Fmo cluster is very unstable, as it contains 10 sequences well-separated from any other genes. Based on these characteristics I would predict that Fmo-1, Fmo-2, and Fmo-4 likely act on

endogenous substrates, while Fmo-3, Fmo-5, and the *P. pacificus* Fmo group may act on xenobiotic molecules. There is evidence that Fmo-5 is upregulated in *C. elegans* in response to ethanol (Kwon *et al.*, 2004).

5.3.4 MIPhy analysis of Cyps

The other gene families are much larger and more difficult to visualize, and so their clustered trees are not presented in a figure. Instead, MIPhy was run on the full tree for each gene family, and the instability scores for the *C. elegans* genes were extracted and graphed (**Figure 5.6**).

For the Cyps, it is not immediately clear where to place the boundary to separate stable and unstable sequences in **Figure 5.6**. While a hard boundary is not strictly necessary, Daf-9 is an important regulatory component in the dauer pathway, and this function is conserved throughout Clades III, IV, and V of the nematodes (Gilabert *et al.*, 2016). It is unlikely that this enzyme would also perform a detoxification role, so the boundary will be placed just after this sequence. Barring the exceptions discussed below, all of the *C. elegans* sequences with higher instability scores are predicted to be xenobiotically active and are indicated with bold names in **Table 5.4**. This table also correlates those predictions with the results of ten studies that have measured gene expression in *C. elegans* in response to various xenobiotics.

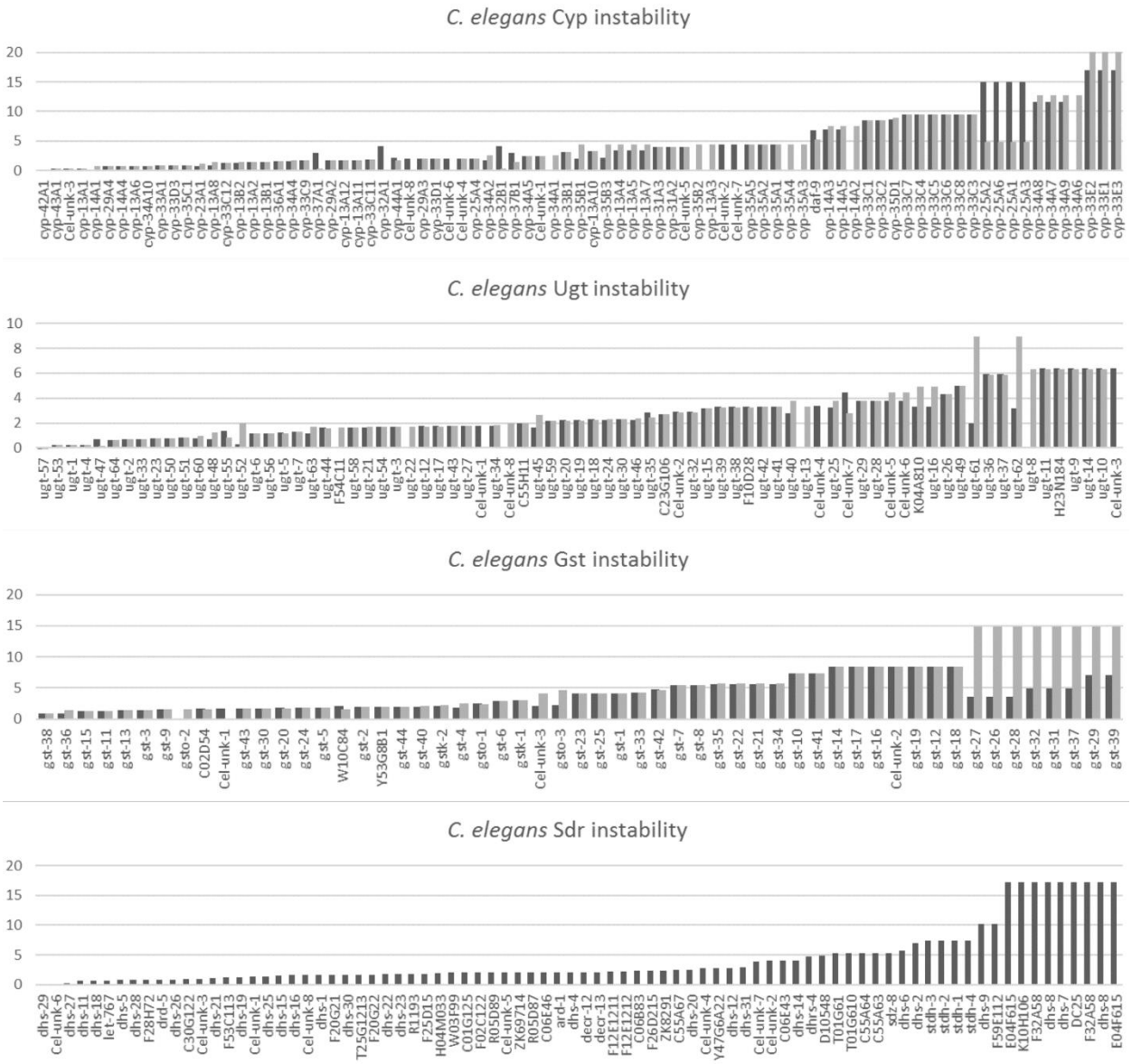


Figure 5.6 The instability of *C. elegans* detoxification genes. These instability values for Cyp, Ugt, Gst, and Sdr sequences were extracted from the full phylogenetic trees of all five nematodes. The black and grey values are the instability scores from TREE_4CP_SET_unbiased and TREE_4CP_SET_total, respectively. The 4CP_SDR_TREE-totals data were not included due to the poor alignment quality as described in **Table 5.3**. All sequences named 'unk' are not members of a gene family, but from genomic regions with no existing gene model. No relatedness should be inferred between any of these sequences, within or between species.

Table 5.4 Cyp proteins involved in the xenobiotic response. Studies that have found evidence of upregulation of these *C. elegans* genes in response to various xenobiotics: A (Menzel *et al.*, 2001), B (Menzel *et al.*, 2005), C (Reichert and Menzel, 2005), D (Chakrapani *et al.*, 2008), E (Jones *et al.*, 2013), and five benzimidazoles in F (S Stasiuk, 2016, unpublished). Other studies have found overexpression of genes in response to specific compounds: ethanol in G (Kwon *et al.*, 2004), 2,2′,5,5′-tetrachlorobiphenyl in H (Menzel *et al.*, 2007), cadmium sensitivity in I (Cui *et al.*, 2007), albendazole in J (Laing *et al.*, 2010), ivermectin and moxidectin in K (Menez *et al.*, 2016). The sequences with bold names are those classified as unstable by MIPhy, and *cyp-31A1* is italicized as it is a known pseudogene.

Cyp	Evidence	Cyp	Evidence	Cyp	Evidence	Cyp	Evidence
13A4	I	**25A3**		**33C5**		**35A1**	A,B,C,G,K
13A5	I	25A4	H	**33C6**		**35A2**	A,B,C,D,J
13A6	F	**25A6**		**33C7**	I	**35A3**	A,E,F
13A7	D,I	29A2	A,E,H,J	**33C8**	G	**35A4**	A,E,F
13A8	G	*31A1*	A	**33E1**	A,D	**35A5**	A,B,C,E,F,G,J
13A10	E	31A3	A	**33E2**		**35B1**	A,C,E
14A2	K	33A1	G	**33E3**		**35B2**	A,C,E,F
14A3	D,H	33B1	H	**34A6**		**35B3**	E
14A4	I	**33C1**	E	**34A7**	E	35C1	A,B,C,E,F,J
14A5	A,C,G,K	**33C2**	D	**34A8**		**35D1**	E
25A1		**33C3**		**34A9**	A,E	37B1	C,H,K
25A2		**33C4**		34A10	E,H		

Both trees agree that the cluster containing the *C. elegans* sequences Cyp-33E1, Cyp-33E2, and Cyp-33E3 is the most unstable. This cluster also contains 8 sequences from *C. remanei*, and 2 each from *C. brenneri* and *C. briggsae*. The next group the trees agree as scoring very unstable is the cluster containing Cyp-34A6 - Cyp-34A9, though the unbiased set did not detect *cyp-34A6* in the *C. elegans* genome.

The trees disagree on the instability of Cyp-25A1, Cyp-25A2, Cyp-25A3, and Cyp-25A6, with the unbiased tree scoring them as the second most unstable cluster, and the total tree spliting that cluster into two. This difference is primarily because Cbr-unk-7 in the totals tree was missing 114 amino acids that were present in the unbiased version

of that sequence. These missing amino acids contain the appropriate Cyp sequence motifs (see **Figure 5.3**), so I predict that this is truly an unstable cluster of sequences. The next unstable cluster, containing Cyp-33C3 – Cyp-33C8, is consistent between the two trees. It contains 6 sequences from *C. elegans*, 3 from *C. remanei*, 2 from *C. brenneri*, and 1 from *C. briggsae*, and the phylogeny of the cluster mirrors the species relationships.

The Cyp-35D1 cluster contains a single *C. elegans* sequence, but 3 each from *C. brenneri* and *C. remanei* and 2 from *C. briggsae*. The next unstable cluster contains Cyp-33C1 and Cyp-33C2 from *C. elegans*, 5 each from *C. brenneri* and *C. briggsae*, and 2 from *C. remanei*. It interesting that even with 8 apparent duplication events, the phylogeny of this cluster exactly matches the species relationships with no incongruence. One of these duplications appears to have occurred in the ancestral species after *C. elegans* split off, so that the other 3 *Caenorhabditis* species have two subclusters of sequences, while *C. elegans* only has one. The next likely unstable cluster contains Cyp-14A2, Cyp-14A3, and Cyp-14A5 from *C. elegans* (though the unbiased data set missed *cyp-14A2* in the genome), 3 sequences from *C. briggsae*, 2 from *C. remanei*, and 1 from *C. brenneri*.

The cluster containing the *C. elegans* sequences Cyp-35A1 – Cyp-35A5 also appears to be very unstable. Besides these 5 sequences it contains 1 from each other *Caenorhabditis* species with no incongruent events, which is why it has an intermediate instability score. However, immediately adjacent to it are 4 small clusters containing between 1 and 3 sequences from the other *Caenorhabditis* species, and the phylogeny within and between these clusters is highly incongruent. These 5 clusters are well-separated from any other sequences, and I believe they should be merged into one that would contain 5 sequences from *C. elegans*, 4 from *C. brenneri*, 4 from *C. briggsae*, and 3 from *C. remanei*; this cluster would then have one of the higher instability scores in the tree. The two trees disagree about the cluster containing Cyp-35B3, primarily because the unbiased tree is missing Cyp-35B2 from *C. elegans*. In both trees this cluster has a low or moderate score, but is adjacent to a cluster containing several sequences from the other *Caenorhabditis*; these two clusters are well separated from

any other cluster, and should be merged. This new cluster would contain 3 sequences from *C. elegans*, 2 or 3 from *C. remanei* (the totals tree is missing one), and 2 each from *C. brenneri* and *C. briggsae*.

5.3.5 MIPhy analysis of Ugts

The Ugt instability scores appear much like the Cyps, in that it is not immediately obvious where to place the classifying boundary in **Figure 5.6**, but here there is no protein with a known developmental function to suggest it. Instead, the clusters are analyzed in descending instability scores (summarized in **Table 5.5**).

The most unstable cluster in the total tree, and the biggest disagreement between the two trees is about the cluster(s) containing Ugt-61 and Ugt-62 from *C. elegans*. This difference appears to be due to several sequences, notably CBG17955 from *C. briggsae* and CRE06643 from *C. remanei*; in TREE_4CP_UGT_unbiased these sequences are more similar to their respective ugt-62 sequences. It is unclear which tree is correct, but the presence of multiple duplications and high incongruence with the species tree leads me to believe these sequences should be considered as unstable.

The most unstable cluster in the unbiased tree, and second-most in the total tree contains 10 sequences from the non-elegans *Caenorhabditis* species. It is adjacent to a moderately unstable cluster containing *C. elegans* Ugt-32 and a pseudogene. Based on the long branches within the clusters, the relatively short branches separating them, and the long branches to the nearest neighbour, they should be merged. This would then contain 2 sequences from *C. elegans* (one a pseudogene), 6 each from *C. brenneri* and *C. briggsae*, and 2 from *C. remanei*.

Table 5.5 Other detoxification proteins involved in the xenobiotic response. Studies have found evidence of upregulation of these *C. elegans* genes in response to various xenobiotics. The identity of the evidence studies is the same as for **Table 5.4**, and the bold sequence names are those predicted as unstable using MIPhy.

Sdr	Evidence	Ugt	Evidence	Gst	Evidence
Dhs-2	H	1	C,G,I,J	**1**	G
Dhs-7		5	J	4	C,E,K
Dhs-8		**8**	C,E,F	5	E,J,K
DC2.5		**9**	G	**7**	
E04F6.15		**10**		**8**	
F32A5.8		**11**		**10**	
K10H10.6	H	13	E	**12**	E
Dhs-14		**14**	F	13	E
C01G12.5		**H23N18.4**		**14**	E
C06B8.3		**16**	C,J	**16**	E
F12E12.11		**K04A8.10**	G	**17**	
F12E12.12		22	J	**18**	
F26D21.5		**25**	E,G,J	**19**	I
R05D8.9	I	**C55H1.1**		20	C
Dhs-23	G,H	**26**		**21**	E
F25D1.5	H	**28**	G	**22**	
Dhs-31		**29**		**23**	
Stdh-1		**30**		**25**	E
Stdh-2		**32**		**26**	
Stdh-3		33	E,F	**27**	
Stdh-4		**36**	E	**28**	
R05D8.7		**37**	E	**29**	
Sdz-8		41	J	30	E
C55A6.3		46	E	**31**	E
C55A6.4		**49**		**32**	
T01G6.1		**61**		**33**	I
T01G6.10		**62**		**34**	
		63	G,J	**35**	
				37	
				38	C,I
				39	E
				41	
				42	G

Both trees agree that the next most unstable cluster contains Ugt-8 – Ugt-11, Ugt-14, and H23N18.4 (an unnamed Ugt) from *C. elegans*. Cel-unk-3 is also part of this cluster in the total tree, but this is actually Ugt-8. It was not named correctly due to the predicted gene model for Cel-unk-3 missing the first 700 bp, which caused it to be less than half of the length of the existing gene model and so receive the unknown 'unk' designation. This cluster currently consists of only those 6 *C. elegans* sequences and is flanked by two stable clusters that contain Ugt-12 and Ugt-13. It is unlikely that it should be merged with either of those but it may be related to the next adjacent cluster, which contains 14 sequences from *C. brenneri, C. briggsae,* and *C. remanei*. These clusters would have been merged in TREE_4CP_UGT_unbiased if a single node had been rearranged, and so this seems to be the most evolutionarily parsimonious explanation. If merged it would contain 20 sequences, and would easily be the most unstable cluster in the gene family.

The next most unstable cluster contains Ugt-36 and Ugt-37 from *C. elegans*, 2 sequences from *C. brenneri*, 3 from *C. remanei*, and 1 from *C. briggsae*. After that there is the cluster with Ugt-49 from *C. elegans*, 3 sequences from *C. brenneri*, 2 from *C. remanei*, 2 from *C. briggsae*, and this is one of the few clusters that also contains a sequence from *P. pacificus*. Next is the cluster with Ugt-26 from *C. elegans*, also containing 3 sequences from *C. briggsae*, 2 from *C. brenneri*, and 1 from *C. remanei*.

The cluster containing Ugt-25 from *C. elegans* is moderately unstable, but is adjacent to the singleton cluster of the unnamed sequence C55H1.1 and an unstable cluster containing four sequences from *C. brenneri*. I predict these clusters should be merged as they are not well separated from each other and are highly incongruent. This cluster would then contain 2 sequences from *C. elegans*, 5 from *C. brenneri* (one misassembled allele), 3 from *C. remanei* (one misassembled allele), and 2 from *C. briggsae*. The clusters containing Ugt-28 – Ugt-30 are also moderately unstable, and also adjacent to other poorly-separated groups. Three of them should be merged, which would result in a cluster containing 3 sequences each from *C. elegans* and *C. remanei*, 5 from *C. brenneri*, and 2 from *C. briggsae*. This cluster would then be one of the most unstable in the tree.

5.3.6 MIPhy analysis of Gsts

There is one large disagreement between the two Gst trees, involving the *C. elegans* sequences Gst-26 – Gst-29, Gst-31, Gst-32, Gst-37, and Gst-39. In the unbiased tree these 8 sequences are found in two separate clusters, while in the total tree they are merged together. This disagreement is caused by the placement of six of the sequences themselves, though their positions in the two trees are separated by only a few very short branches. This may have occurred due to many slight differences in the 6 sequences between the two data sets, or it could simply be an artifact of the tree-building heuristic. Regardless, I believe the total tree is more correct, where all mentioned sequences are merged into one very unstable cluster instead of several of moderate instability. This cluster also contains 1 sequence each from *C. briggsae* and *C. remanei*, and 6 from *C. brenneri*, though I believe that three of the *C. brenneri* sequences are actually misassembled alleles. The first pair differ by 1 out of 209 amino acids, the second by 4 amino acids, and the third pair by 11 amino acids.

The second most unstable cluster contains 7 *C. elegans* sequences, Gst-12, Gst-14, Gst-16 – Gst-19, and Cel-unk-2 (which is a genomic region that currently contains no gene models). This is an expansion specific to this species, as the cluster has no incongruence and contains one sequence each from the other *Caenorhabditis* species (*C. brenneri* has an apparent paralog, but it only differs at 2 positions so it is almost certainly a misassembled allele).

The third most unstable cluster contains Gst-10 and Gst-41 from *C. elegans*. It also contains 3 sequences each from *C. brenneri* and *C. remanei*, and zero from *C. briggsae*. The 3 adjacent clusters are small and contain sequences from only *C. briggsae* and *C. brenneri*, and so I believe they should be merged into one. This would then be one of the most unstable clusters, with 2 sequences from *C. elegans*, 3 from *C. remanei*, 3 from *C. briggsae*, and 8 from *C. brenneri*.

The next unstable cluster contains Gst-21, Gst-22, Gst-34, and Gst-35 from *C. elegans*, 2 sequences from *C. brenneri*, and 1 each from *C. briggsae* and *C. remanei*. The next cluster contains Gst-7 and Gst-8 from *C. elegans*, 2 sequences each from *C.*

briggsae and *C. brenneri* (there appear to be 3 but CBN32903 is 100% identical to Cbn-gst-8), and 1 from *C. remanei*.

The cluster containing Gst-42 from *C. elegans* has a high instability score because it contains 4 sequences from *P. pacificus*; the four *Caenorhabditis* sequences are very similar and are separated by very short branches. The *P. pacificus* sequences are very well-separated from any other sequences, which indicates they likely have arisen from the same Gst-42 ancestor sequence. Because of this reason the whole cluster is classified as unstable. The cluster containing Gst-1, Gst-23, and Gst-25 is one of the few containing a sequence from *P. pacificus*; the duplications also appear to be specific to *C. elegans*.

5.3.7 MIPhy analysis of Sdrs

As described in Section 5.3.1, ALN_4CP_SDRS_total was abandoned due to poor quality so **Figure 5.6** only contains instability scores from TREE_4CP_SDRS_unbiased. The single most unstable cluster on this tree does not contain any *C. elegans* sequences, and Cre-srt-41 is the only named sequence out of the 17. It contains 2 sequences from *C. remanei*, 3 from *C. briggsae*, and 12 from *C. brenneri*. Of the seven adjacent clusters, five contain sequences from only one species, and most are not well separated. Because of this these eight clusters should be merged into one, which would contain 7 sequences from *C. elegans*, 19 from *C. brenneri*, 6 from *C. briggsae*, and 4 from *C. remanei*.

The Dhs-7 & Dhs-8 cluster is also very unstable, with a high degree of internal structure and many independent species-specific expansions. It contains 6 sequences each from *C. elegans* and *C. briggsae*, 5 each from *C. brenneri* and *C. remanei*, and 1 from *P. pacificus*. The next most unstable cluster contains the Stdh sequences, with 4 from *C. elegans*, 1 each from *C. remanei* and *C. briggsae*, and 2 or 3 sequences from *C. brenneri* (there appear to be 4 but one is very likely a misassembled allele, and a second may be as well).

The next most unstable cluster contains only 3 sequences from *C. briggsae*, but falls in a section of the tree with 3 stable sister clusters and 8 small unstable adjacent clusters (7 contain sequences from only one species). Most of these small clusters are separated by very short branch lengths, which means that they would have been clustered together by MIPhy if the phylogenetic tree was output with only a few rearrangements. Because of these short branches and the uncertainty in the collected protein sequences, these nine clusters should be merged into one unstable group, which would contain 1 sequence from *C. elegans*, 5 each from *C. brenneri* and *C. briggsae*, and 4 from *C. remanei*. The cluster containing CBN12803 from *C. brenneri* is in a similar situation, as it has one stable sister cluster and five adjacent clusters that contain one or two sequences. Those six unstable clusters should be merged, which would result in 2 sequences from each of the *Caenorhabditis* (there appear to be 3 from both *C. brenneri* and *C. remanei*, but 1 from each appears to be a misassembled allele that is identical to the paralog but missing one or two exons), where Dhs-31 and Cel-unk-5 (a known pseudogene) are from *C. elegans*.

The next unstable cluster contains Sdz-8 and two unnamed sequences from *C. elegans*, 2 from *C. remanei*, and 1 each from *C. briggsae* and *C. brenneri* (there appear to be 2 sequences, but they are 100% identical). After that is the group containing T01G61 and T01G610 from *C. elegans* and 2 each from the other *Caenorhabditis* (though the 2 *C. remanei* sequences may be misassembled alleles as they share 97.5% sequence identity).

The cluster containing the Dhs-2 sequences has 1 from each *Caenorhabditis* species (*C. brenneri* appears to have 2, but they differ at 3 of 342 positions and so are almost certainly misassembled alleles) in agreement with the species phylogeny. It has a high instability score because it also contains 6 sequences from *P. pacificus*. These are very well separated from any other sequences from this nematode, which implies they are descended from the ancient Dhs-2 sequence. Because of this it is classified as unstable.

The Dhs-9 cluster is the 3rd highest scoring, but I believe it should be much lower in the instability rankings. It appears to be two distinct clusters that have been merged only because of the placement of Ppa-dhs-26 in the tree; in fact, the cluster is split up if the spread weight is increased from 1.0 to 1.1. The next unstable cluster contains the Dhs-6 sequences, and is only scored as unstable because of one sequence, Cre-unk-7. This sequence is at the end of a very long branch, so may be an old pseudogene or simply a poorly assembled gene model, but I do not believe it is sufficient evidence to classify any of the Dhs-6 sequences as unstable. The instability score for the Dhrs-4 cluster is also likely overestimated, due entirely to the placement of CBN01504, which is well separated from the other sequences by a long branch. If it was absent or had been placed closer to the *C. brenneri* Dhrs-4 sequence (achievable by rearranging two nodes with very short branches) this cluster would have a very low instability score. This likely indicates that this sequence is a pseudogene, is incorrect, or perhaps that it is an unstable paralog specific to *C. brenneri*; but in any case, is not sufficient to classify the Dhrs-4 sequences as unstable. The cluster containing the *C. elegans* sequence D1054.8 appears to be moderately unstable, but this is entirely due to two unannotated sequences from *C. brenneri*. They are both at the end of long branches, well separated from the other sequences in that cluster, and as there are no gene models predicted in these regions of the genome, they may be misassembled sequence, or perhaps pseudogenes. Based on the stability of the other sequences I do not classify them as unstable.

5.4 Discussion

In the MIPhy analysis of each gene family, each cluster was manually examined and in many cases conclusions were reached by subjective override of the MIPhy output. I do not believe that this indicates a limitation with MIPhy as this was not required with the vertebrate dataset in Chapter 3, but rather was made necessary by the quality of the gene models and the unbalanced divergence in the species used. In fact, I feel that this is an example of the usefulness of MIPhy as the same analysis would have been much more difficult without this software.

A common property of these gene family trees is that the *P. pacificus* sequences were most often found sequestered with other *P. pacificus* sequences, instead of being found in clusters with sequences from the *Caenorhabditis*. This is most likely due to the unbalanced level of sequence divergence within the set of nematodes used here; the divergence within the *Caenorhabditis* is much less than the divergence between *P. pacificus* and any of the *Caenorhabditis*. Unfortunately, this means that many of the *P. pacificus* sequences do not impart very much information to the MIPhy analysis, as it expects clusters to have sequences from each species at an equal frequency. Future analyses would likely benefit from including additional *Caenorhabditis* genomes, or by choosing a more closely related outgroup species.

There are two *cyp* genes inducible by many different xenobiotics that are not classified as unstable by MIPhy: *cyp-35C1* and *cyp-29A2* (**Table 5.4**). Based on the sequence conservation of the orthologs in the *Caenorhabditis* species, I do not expect that MIPhy would ever classify these sequences as xenobiotic. It is possible that their instability scores would increase with the inclusion of additional species data, or that these enzymes also function in some conserved endogenous role that has shaped their evolutionary history, as has been hypothesized about Cyp-1A2 in humans (Zhou *et al.*, 2009; Kapitulnik and Gonzalez, 1993). It could also be that these enzymes were already able to metabolize a wide variety of xenobiotics, and so did not experience adaptive evolution when the ancestral *Caenorhabditis* nematode population split into the different niches that became the present-day species. There are six other genes with some evidence of xenobiotic induction that were classified as stable by MIPhy: *cyp-13A6*, *cyp-13A8, cyp-14A4, cyp-33A1, cyp-34A10*, and *cyp-37B1*. The explanations above may also apply to these sequences, but as there is less evidence to support their expression patterns they may simply be false positives like *cyp-31A1*, which is a known pseudogene in *C. elegans*.

There are two sequences from *C. briggsae* in the Daf-9 cluster, and the position of the paralog has strongly impacted the instability score of the whole cluster. As this enzyme has a known endogenous role, I used it as the boundary between stable and unstable sequences. The unexpectedly high instability score indicates it may not have

been an ideal choice, and if the boundary had been placed lower MIPhy would have classified eight more xenobiotically inducible genes as being unstable: *cyp-13A4 – cyp-13A7, cyp-13A10, cyp-25A4, cyp-31A3,* and *cyp-33B1*.

There are 13 Cyp sequences that MIPhy predicts as being unstable that have not been substantiated with published biochemical studies. However, every one of these is found in a cluster with at least one inducible gene except for the cluster containing Cyp-25A1, Cyp-25A2, Cyp-25A3, and Cyp-25A6. It may be that these sequences are false positives and do not act on xenobiotics, but the sparsity of the evidence in **Table 5.5** indicates that many genes are only induced by a small number of compounds; it may simply be that the appropriate molecule has not yet been tested.

There is less known about the substrate specificity of the Ugt proteins, which may explain why there are four clusters (Ugt-26, Ugt-32, Ugt-49, and Ugt-61/Ugt-62) classified as unstable with no evidence of xenobiotic upregulation. Further studies using appropriate inducers may identify these as responding to xenobiotics. There are also four *C. elegans* sequences (Ugt-1, Ugt-5, Ugt-33, and Ugt-63) that do respond to xenobiotics that MIPhy could not classify as unstable. Finally, there are four additional sequences (Ugt-13, Ugt-22, Ugt-41, and Ugt-46) with evidence of xenobiotic induction that received intermediate scores.

There are equal numbers of *ugt* and *gst* genes with evidence of upregulation, but MIPhy classifies more Gst proteins as unstable. Even though more predictions were made, there are only two clusters classified as unstable that have no evidence of xenobiotic induction: the first containing Gst-7 & Gst-8, and the second containing Gst-10 & Gst-41 from *C. elegans*. The Gst-7/8 cluster contains duplications in all of the *Caenorhabditis* except *C. remanei*, and the Gst-10/41 cluster is one of the most unstable in the tree. I hypothesize that both of these clusters are very likely to act on xenobiotic substrates, and that the correct induction compound has simply not yet been tested. However, this gene family also has more examples of genes that have been shown to be xenobiotically induced that appear to be phylogenetically stable: Gst-4, Gst-5, Gst-13, Gst-20, Gst-30, and Gst-38. Each of these is supported by only one or

two studies so it is possible that some are false positives, but it also seems likely that others of these enzymes may act on many substrates, and thus play both an endogenous and xenobiotic role.

The short-chain dehydrogenases have the fewest studies with evidence of xenobiotic induction. Perhaps because of this there are five clusters (Dhs-31, Stdh-1 – Stdh-4, R05D8.7, Sdz-8/C55A6.3/C55A6.4, and T01G6.1/T01G6.10) classified as unstable with no supporting gene expression studies, while only a single cluster (Dhs-23 & F25D1.5) has evidence of upregulation but was classified as stable.

With the construction of this pipeline, this analysis is almost ready to be extended to the parasitic nematodes. It would be necessary to collect an unbiased set of protein sequences from these species (as described in Section 5.2.4), but the poor results obtained on *P. pacificus* lead me to believe the Exonerate-based method as described above would perform even worse on more divergent species. Further, I acknowledge that this method is not the ideal way to extract a protein sequence from a genome. It chooses the best whole predicted gene model for some genomic location, but with the huge complexity inherent in each gene model prediction it is most likely that the best sequence for most regions will require a mix of exons from several of the top predicted gene models. It is for exactly such a situation that an extension of Figmop to a gene-finding algorithm would be a valuable tool (described in Section 6.1). A profile hidden Markov model implicitly contains all combinations of potential gene model features, and outputs their optimal combination. Barring this extension, a simpler improvement could be made by constructing a hidden Markov model to piece together the Exonerate results to generate an optimal coding sequence. For a genomic region, the likelihood of some nucleotide being part of an exon could be calculated based on how often it is found in an exon in the set of hits, modified by the quality of each hit (potentially evaluated in a similar manner as in exoner_to_prot.py). The design of the HMM itself would ensure that the reading frame was respected, as well as the order of features like splice donors and acceptors (as depicted in **Figure 6.1**).

The output of this work could aid in novel anthelmintic development. If a specific metabolic enzyme can be found that is responsible for resistance to some anthelmintic, a second drug targeting that metabolic enzyme could potentially allow the existing anthelmintic to become useful once again. Elucidating the whole complement of enzymes that are used to break down drugs in general would also be valuable, as it would allow future drug candidates to be evaluated for their susceptibility to metabolism. Identifying the potential for cross resistance between novel and existing drugs is important as this could lead to the rapid development of resistance. Pursuing this aspect of resistance mechanisms has wide implications, particularly as all anthelmintics are potential candidates for metabolism and efflux. Gaining a better understanding of this process is therefore beneficial for both current and future anthelmintics.

I hypothesize that drug metabolism plays an important role in the development of drug resistance in the parasitic nematodes, and so this sort of comparative analysis would be vital to future studies. It could also be extended to any other group of organisms, as nothing in this pipeline is nematode-specific. As DDT resistance in mosquitos and flies appears to arise from overexpression or modification of a single *cyp* gene, a similar analysis on insect detoxification gene families could prove valuable. Metabolism may play a role in the mechanism of resistance to other insecticides, and so could be of global importance to elucidate.

6. Discussion

The work presented here describes several algorithms designed for an in-depth comparative analysis of a gene family across several species. First I described Figmop (Chapter 2), a program that can identify related protein sequences, or can bypass the gene model predictions and run directly on draft-quality genome assemblies. This was originally motivated by the high sequence variability of the cytochrome P450 proteins, and the apparent failure of traditional gene predictors to identify these genes in the genome of the hookworm *Haemonchus contortus*. Other tools exist that also use profile hidden Markov models (HMM) to identify matches, such as Nhmmer (Wheeler and Eddy, 2013). However, I am not aware of any that can be trained on protein sequences but run on a genome with long introns, that provide a high level of control over the underlying HMM, or that output graphical results allowing for easy interpretation of complex combinations of gene fragments, fusions, and arrays.

MIPhy (Chapter 3) is an algorithm designed to analyze the phylogenetic tree of a multi-species gene family, cluster that tree into parsimonious groups, and score the phylogenetic instability of each group. Similar to a positive selection analysis, MIPhy is a method to identify signatures of adaptive evolution. Other reconciliation tools are available, and in fact MIPhy is adapted from NOTUNG (Stolzer *et al.*, 2012). However, none were found that allowed for widespread incomplete lineage sorting (ILS), nor were any designed to cluster a phylogenetic tree and so assign a score to each sequence.

Finally I described RDepth (Chapter 4), a web tool to visualize copy-number variation in populations of *C. elegans*. This allows identification of those genes that can tolerate duplications or loss, which are unlikely to play important endogenous roles. Instead, they may act on substrates or signals from the nematode’s environment. This tool is designed such that in the future a user could run it on their own species of interest.

These tools allowed an analysis of the evolution of large and divergent gene families, which often contain paralogs that can confound existing methods (Casola and Hahn,

2009; Katju and Bergthorsson, 2010). They were all used in a study of five detoxification gene families in five species of nematode (Chapter 5).

6.1 Future extensions to Figmop to predict coding sequences

Figmop (Chapter 2) was developed to identify hard-to-find genes, and has proven to be highly sensitive. One of the primary reasons is that it identifies short regions of sequence similarity and looks for a specified pattern of those regions. It is therefore insensitive to the intervening sequence, both in terms of identity and length, which allows it to identify genes with almost no detriment from large and diverse introns. These same properties mean it also has great potential to be a targeted gene finder.

Because of the non-specific definition of splice-donor and splice-acceptor sites, there are typically an exponentially large number of possible combinations of these sites that could produce a coding sequence at one genomic location. Any evidence identifying which nucleotides form part of the mRNA can greatly improve the accuracy of gene-finding algorithms like AUGUSTUS (Stanke *et al.*, 2006, 2008). RNAseq datasets can be used for this purpose, but they may suffer from mapping errors and intron contamination. There is a program, AUGUSTUS-PPX, that combines *ab initio* predictions with a profile of conserved sequence blocks that describes one protein family (Keller *et al.*, 2011). This is similar to the pattern of motifs used by Figmop, and so the simplest extension to this software would be to implement an option to collect the regions matched by the user's motif pattern, and combine them into a set of sequence blocks in a format that could be used by AUGUSTUS-PPX.

However, there are a few compelling reasons to implement my own algorithm. Firstly, AUGUSTUS is designed to identify coding regions from an entire genome, much of which is non-coding. Figmop can extract only the region containing the gene of interest, so the gene-finding problem becomes simpler. The extracted sequence can be safely assumed to begin prior to the start codon, and the likely distance can even be supplied; this also means the predictive model does not have to consider overlapping gene models. Secondly, Figmop could automatically compare the observed pattern

between the protein sequences and the genomic hits, and infer the probable intron placements and lengths. Other gene finding software allows the user to specify this sort of information, but in my case this would be estimated separately for each genomic region. This could be important, as I commonly observe differing intron numbers and lengths among members of one gene family in the genome of a single species. Finally, AUGUSTUS has always been designed for high-throughput workflows, where manual inspection of predictions is intractable. This Figmop extension would be designed to identify members of a single gene family at a time, typically no more than a few dozen genes, and so for every genomic region it could present the user with several potential gene models.

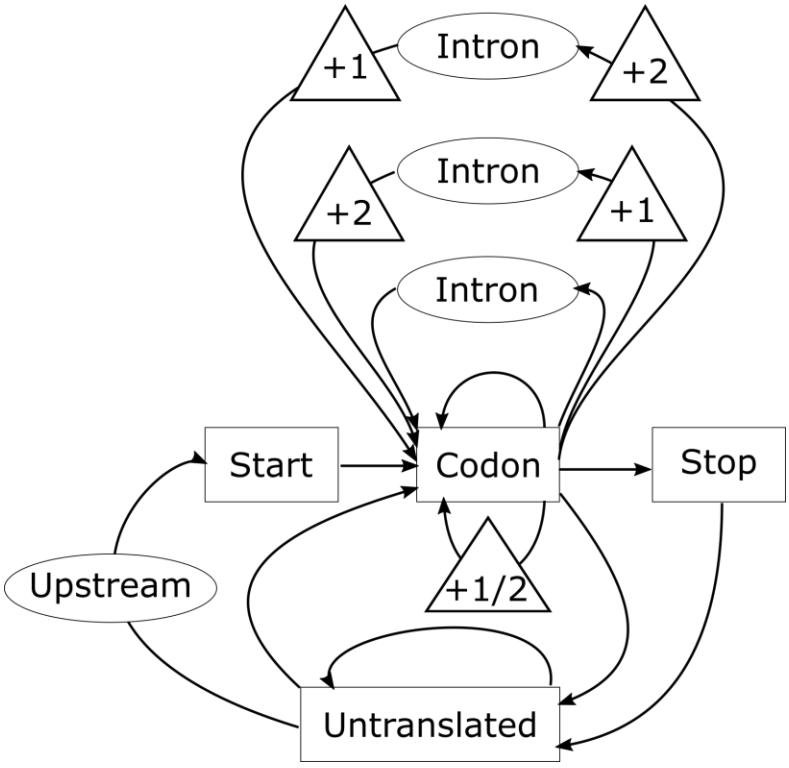


Figure 6.1 pHMM to predict the coding sequence of Figmop hits. The boxes represent a simple state in the model, the circles indicate more complex states (sub-models definable by the user), and the triangles are not true states but represent instructions for the Viterbi decoder (used to preserve the reading frame, and allow for frameshifts). The ‘Start’ state emits the initial ‘ATG’, ‘Codon’ is a triplet that forms part of the coding sequence, ‘Stop’ emits one of the stop codons, and ‘Untranslated’ emits the sequence on either side of the gene.

Figure 6.1 shows a profile hidden Markov model (pHMM) that has been designed to predict the coding sequence of some gene when given the raw genome and the Figmop results. The oval shapes in that figure are themselves small state-models, which would be modifiable by the user. The 'Upstream' model would search for patterns of information such as that depicted in **Figure 6.2**, and the 'Intron' model could be automatically generated for the user given the genome annotations (such as a gff3 file). One advantage of partitioning the complexity of these two searches into their own sub-models is that parallel computing can be leveraged to dramatically speed up the implementation.

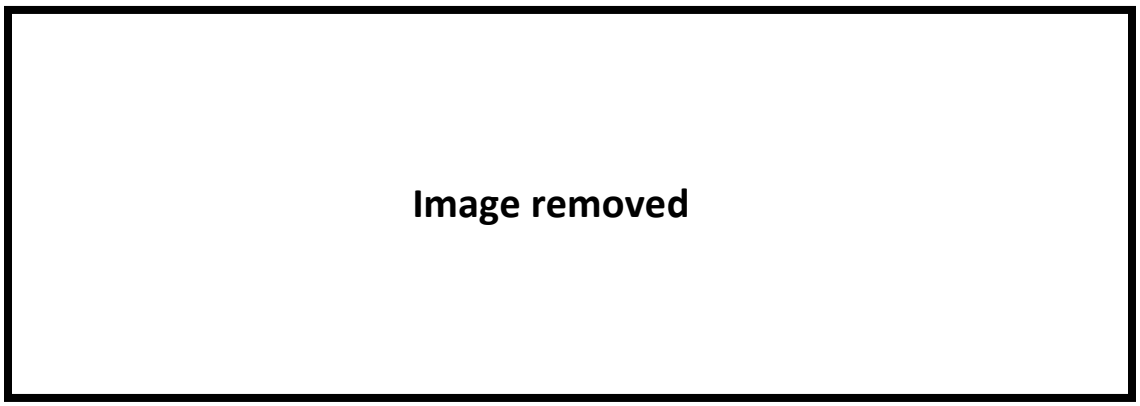


Figure 6.2 Conserved features upstream of *Caenorhabditis* genes. Figure 3 from the WormBook chapter on transcriptional regulation of gene expression in *C. elegans* (https://www.ncbi.nlm.nih.gov/books/NBK19715/) exemplifies the sequence patterns that could be identified by the 'Upstream' state in **Figure 6.1**. **A)** shows the sequence logos of the various features and their conservation in five species of *Caenorhabditis*, and **B)** indicates how often and how far upstream each feature is found in the respective genomes. This image was removed as copyright permissions could not be obtained.

When searching for motifs in some segment of DNA, there will be many spurious matches with poor E-values. When identifying the user's pattern with Figmop, these motifs will be skipped by the model, and so the E-values of the individual motif matches are not accounted for by the algorithm. However, predicting gene structures requires a much more complex model, and so I would incorporate the strength of the motif matches into this new pHMM. The most obvious way to do this is with the emission

probabilities of the ‘Codon’ state in **Figure 6.1**; this is the likelihood that the current nucleotide under consideration is part of the coding sequence of the gene.

This probability e would normally be set to 1.0 for any codon that was part of a motif, and some minimum value (perhaps $e = 0.5$) for those that formed the coding sequence between motifs. This can be modified using a scaling calculation on the E-value of a motif match:

$$e = 1 - \frac{a}{\log_{10}(E\ value)},$$

where a is a parameter that dictates the minimum level of significance for an E-value to matter. If the codon is not part of a motif match, or if $E\ value > 1 \cdot 10^{-a}$, $e = m$, where m is the minimum probability of a codon. This constrains the range of the emission probability to $m \leq e < 1$, where it approaches 1 as the strength of the motif match increases. A permutation that more strongly rewards motif matches could retain the minimum value of e as m, and modify the equation to:

$$e = 1 - \frac{a}{\log_{10}(E\ value)} + m.$$

This now means that the range of the emission probability is $m \leq e < m + 1$ (it can be greater than 1), which is typically not done in a pHMM. While this does not invalidate the general decoding using the Viterbi algorithm, it does have implications described below.

As mentioned previously, one of the advantages to extending Figmop to be its own gene finder, is that the user could be presented with several potential gene models for each genomic region. To my knowledge, existing software only outputs the single most likely gene model as predicted by its model, but when any HMM is unrolled to a state graph, finding the k-likely paths is analogous to finding the k-shortest paths through a directed graph. There has been substantial work done on this general problem, and so if I used the first definition of e where it cannot exceed 1, I could use Yen’s algorithm (Yen, 1971) to present the k-likely gene structure predictions to the user. If I used the second definition of e that allows it to exceed 1, this problem becomes analogous to finding the k-shortest paths through a directed graph that allows negative path lengths. I

may still be able to use Yen's algorithm, or I may have to use a modification such as Jonhson's algorithm (Kamburowski, 1997) or Eppstein's algorithm (Eppstein, 1994).

It is important to note that the extensions to Figmop described here are not intended to produce competitive software to current gene finders like AUGUSTUS, but rather complimentary software. This new program would allow researchers to reanalyze some genome for their gene family of interest in a low-throughput but high-impact manner. It would only be useful for those genes that can be well described by a pattern of motifs, but has the potential to produce more accurate sequences and so aid in the continual and incremental improvements of genome assemblies.

6.2 Future extensions to MIPhy

MIPhy (Chapter 3) has proven to be valuable software in an analysis of the detoxification gene families of free-living *Caenorhabditis*. While the software is publicly available, there remain several improvements or modifications that could make it more useful for users.

6.2.1 Alternatives to relative spread

MIPhy currently uses a standard deviation calculation as its metric for relative spread, but other measures could be easily implemented and provided as options to the user. A standard deviation calculation is very sensitive to outliers from the mean, so it would yield a higher score for a cluster that contained four similar sequences and one at the end of a long branch. The mean absolute deviation is a similar calculation that is less sensitive to outliers, and the median absolute deviation is less sensitive still. There is no reason to believe one of these is always a better choice than the others, but the mean absolute deviation may be more appropriate in cases where one species is expected to be much more divergent than the others (such as *P. pacificus* in Chapter 5). A common general-use metric of cluster goodness is the silhouette (Rousseeuw, 1987), which is the ratio of how well a point fits its current cluster over how well it fits the nearest neighbouring cluster. It is essentially the mean absolute deviation modified by both the absolute mean deviation of the nearest cluster and the distance to that cluster.

It was deemed unsuitable here as it is not a measure of the divergence within a cluster, but could be easily implemented in MIPhy.

6.2.2 Cluster robustness

The default MIPhy parameters are somewhat arbitrary, and different parameter combinations may lead to different clustering patterns, which may suggest different conclusions for the same data set. It would therefore be useful to explore the conclusions of the entire parameter space. This is not possible with clustering or optimization problems in general due to the exponentially large number of possible outputs, and so such surveys are typically performed with heuristics or some sampling procedure. However, the fact that hierarchical clustering is being performed on a gene tree means that there are relatively few valid clustering patterns.

For a tree with $n + 1$ sequences there are n internal tree nodes, and at each the clustering algorithm decides whether to merge all descendants into one group or leave them as they are. This means there are at most n decision variables, and as each choice is binary there are at most n^2 total configurations of these variables. However, the real number is far lower as a single choice can strongly constrain the decision space. Consider the root node. If the decision is made to merge its child clusters, the entire tree becomes one cluster. Therefore, there is only this single valid clustering pattern that includes the ‘merge’ decision at the root node. This one observation reduces the number of configurations to $n^2/2 + 1$, and this number will be similarly reduced further by the rest of the tree structure. I leave the proof of this limit as beyond the scope of this thesis, but from empirical testing it appears that a tree with $n + 1$ sequences can have between $n + 1$ and $n \cdot (n - 1)/2$ valid clustering patterns, depending on the specifics of the tree structure.

There are additional factors that reduce the size of the decision space even further. While not an exhaustive list, the following properties are common in real data. The effect of removing a single decision node can be substantial, as it directly decrements n in the above limits, and so can remove up to $n - 1$ possibilities from the total decision

space. Each clustering decision is made based on the score function, which is primarily a function of the chosen weights and the identity of that node's terminal children (the relative spread typically has little impact). If I set the relative spread weight to 0, then all nodes with the same subtrees become equivalent (from **Figure 3.1C** node $n1 \equiv n2$ and $n3 \equiv n4$); that is, no matter the chosen weights the same decision will always be made at these nodes, so only one needs to be considered. It is also common for a decision node to be invariant, especially at speciation event nodes. Consider node n5 from **Figure 3.2B** and **Table 3.1**; the decision here compares keeping the child clusters separate ($\{\{a2, b2\}, c2\}$, with score $sep(n5) = 4\theta_L$) to merging them into one cluster ($\{\{a2, b2, c2\}\}$, with score $cmb(n5) = \theta_L$). As the weights are strictly non-negative and the lowest scoring option is chosen, the merge decision here can never score worse than the separate decision. Besides removing this node from the decision space, any invariant 'merge' node will also remove all descendant nodes from consideration (so node n5 from **Figure 3.2B** removes node n2). The nodes near the root are also often invariant, though they instead choose to separate (their decision tends to reduce to $sep(g) = score(l) + score(r)$ vs $cmb(g) = score(l) + score(r) + \theta_D$).

All of these reductions in the decision space should make it easily searchable with a simple constraint propagation strategy. Further, as exemplified above, for each decision node the inequality it uses to make the clustering decision is known (node n5 from **Figure 3.2B** chooses 'separate' when $4\theta_L < \theta_L$; that this is impossible is why it is invariant). The inequalities from all decision nodes for each valid clustering pattern could then be combined, which would yield the range or ratio of parameter values that lead to that particular pattern. For each internal node, the inequalities from all valid clustering patterns could be combined, which would define the ratios of parameter values that lead to the formation of each phylogenetic group. This could be provided to the user as raw information, or could be used to calculate a measure of robustness for each cluster, somewhat akin to a bootstrap analysis.

6.2.3 Predicting the stable / unstable classification boundary

The ultimate purpose of MIPhy is to classify enzymes as acting on either endogenous or xenobiotic substrates, which involves classifying their sequences as either stable or unstable. The placement of the boundary line in **Figure 3.4** correlates well with the known substrate specificity information, positive selection analyses, and genomic tandem arrays, which are markers of adaptive evolution independent from phylogenetic instability. It was placed where the instability score jumped by 67%, but there is another jump even larger, and there is no definitive reason to place it where I did. In this way MIPhy presents different hypotheses, and allows the user to interpret them in the context of their own knowledge.

However, while testing permutations of the parameters (relative weights for duplications, loss and incongruence), I found the genes occasionally changed ranking within the groups classified as stable or unstable, but almost never crossed from one to the other. I recognize that this property may be useful in predicting the stable/unstable boundary, but did not pursue it in the current version of the software. In order to explore how the gene rankings change as the parameters change, one approach would be to evaluate all possible orderings. The theoretical number of total orderings to explore would be $s!$, where s is the maximum number of genes found in any one species. This term grows exceptionally quickly ($10! = 3.6 \cdot 10^6$, while $20! = 2.3 \cdot 10^{18}$) and so a brute-force approach is clearly intractable.

Instead, each sequence can be analyzed individually along with all possible scores it can take. The instability score MIPhy calculates for some sequence is a function of the weight parameters and the gene events predicted for its group. This means that all sequences in a group will have the same score, as will all sequences in all clusters with the same gene events (common with stable clusters). As discussed in Section 6.2.2, there is a manageable number of valid clustering patterns for a gene tree. An additional consequence of this is that a given sequence can be found in only a few different cluster configurations. As the instability score is a function of the gene events within a group, this means that all possible instability scores for a gene can be expressed by only a few combinations of the weight parameters. While this does not completely describe the

ranking of the sequences without actually setting the parameters, it allows many observations to be made about their relative ordering.

Consider node n6 from **Figure 3.2B** and **Table 3.1**. This node chooses ‘merge’ for all parameter values where $\theta_I \leq 3\theta_L$, which would mean that the instability scores for sequences c1 and d1 are equal as they would be in the same group. However, if the parameter weights are chosen such that $\theta_I > 3\theta_L$, node n6 will choose ‘separate’ instead. This means that sequences c1 and d1 will be in different clusters with different scores: $score(c1) = 2\theta_L$, while $score(d1) = score(n4) = \theta_L$. Therefore, no matter what the parameter weights are set at, the instability score of sequence c1 will be greater than or equal to that of sequence d1. Similar relative ranking statements will be possible throughout the tree, and each such constraint will strongly reduce the number of total possible orderings. It may be the case that these constraints reduce the number of possible orderings to a manageable level so that each one can be explored. This would allow boundaries in the rankings to be identified that sequences never or rarely cross, one of which is likely to be a useful stable / unstable classifier.

One final approach could also take advantage of the small number of possible instability score equations for each sequence. With these, MIPhy could very quickly explore the parameter space around the values chosen by the user in a coarse manner (for example all combinations of the θ_D, θ_I, and θ_L weights $\pm$ 1.0, with a step of 0.1; this would be only 8,000 parameter combinations). At each setting the rank of every sequence within its own species would be recorded. At the end of these combinations, the average rank of each sequence could be calculated, weighted by how different that parameter combination was to that chosen by the user. If it is the case that sequences change ranking within their group but rarely cross into another group, this process would accentuate the differences between the groups and therefore make the classification boundary between them clear (see **Figure 6.3** for a simulated example). A real implementation would likely be approximately this dramatic, as I saw genes crossing groups far less than 10% of the time, especially near the default values.

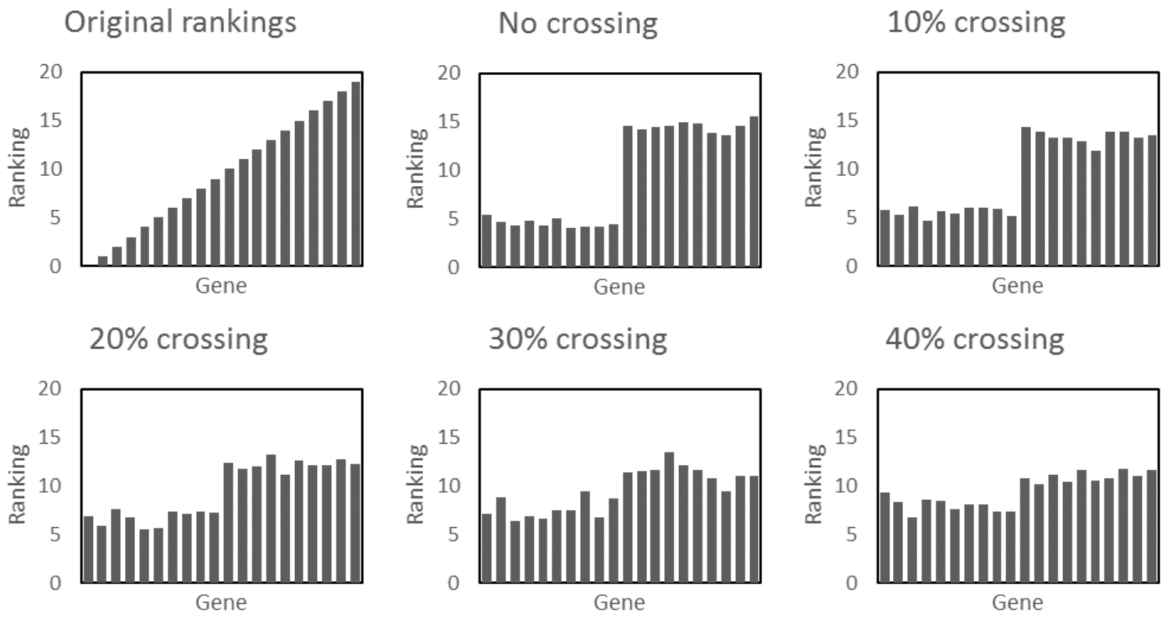


Figure 6.3 Simulation of averaging gene rankings over changing parameters. Beginning with one group of 10 stable sequences (ranked 1-10) and one group of 10 unstable sequences (ranked 11-20), each sequence was given a chance (indicated on each graph) to cross to the other group. The order of the sequences in each group was then randomized, simulating sequences changing rankings within their own group because of changing parameter values. This procedure was repeated 50 times, and the average rank of each sequence is graphed here.

6.2.4 Other algorithm modifications

Moving even a single sequence a short distance in the gene tree can dramatically impact that cluster and those around it. MIPhy assumes that the given tree is correct, but it would be useful to explore the uncertainty in the topology. One method could involve iterating over each internal tree node, collapsing that node into a polytomy, and re-running the MIPhy algorithm on every topology suggested by the polytomy. Any improvement in the overall instability score could be weighted by the uncertainty at that node (bootstrap, SH test (Shimodaira, 2002), or similar), and identified for the user. This would identify single nodes of the tree that may be erroneous. A related approach could be applied to the sequences, by considering the impact each would have if it were removed from the tree. This could not only help identify sequences with errors, but also

provide some intuition about why certain clusters appear as they do. If very many of the sequences were from one species, it might also be an indication that the given species relationships do not match well with the observed sequence relationships. This could simply mean that the species relationships were input incorrectly, or it could be an observation on the selective pressure experienced by the sequences from that species.

Bibliography

Abad,P. *et al.* (2008) Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat. Biotechnol.*, **26**, 909–915.

Abebe,E. *et al.* (2010) An entomopathogenic *Caenorhabditis briggsae*. *J. Exp. Biol.*, **213**, 3223–3229.

Aitman,T.J. *et al.* (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, **439**, 851–855.

Alegado,R.A. *et al.* (2003) Characterization of mediators of microbial virulence and innate immunity using the *Caenorhabditis elegans* host-pathogen model. *Cell. Microbiol.*, **5**, 435–444.

Alvarez,L.I. *et al.* (2005) Altered drug influx/efflux and enhanced metabolic activity in triclabendazole-resistant liver flukes. *Parasitology*, **131**, 501–510.

Alvinerie,M. *et al.* (2001) In vitro metabolism of moxidectin in *Haemonchus contortus* adult stages. *Parasitol. Res.*, **87**, 702–704.

Amichot,M. *et al.* (2004) Point mutations associated with insecticide resistance in the *Drosophila* cytochrome P450 Cyp6a2 enable DDT metabolism. *Eur. J. Biochem.*, **271**, 1250–1257.

Atkinson,H.J. *et al.* (2012) Strategies for transgenic nematode control in developed and developing world crops. *Curr. Opin. Biotechnol.*, **23**, 251–256.

Bailey,T.L. *et al.* (2006) MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369-73.

Bailey,T.L. and Elkan,C. (1994) Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. *Proc. Second Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Baird,S.E. *et al.* (1994) *Caenorhabditis vulgaris* sp.n. (Nematoda: Rhabditidae): A Necromenic associate of pill bugs and snails. *Nematologica*, **40**, 1–11.

Baird,S.E. (1999) Natural and experimental associations of *Caenorhabditis remanei* with Trachelipus rathkii and other terrestrial isopods. *Nematology*, **1**, 471–475.

Bansal,M.S. *et al.* (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**, i283–i291.

Barrett,J. (1998) Cytochrome P450 in parasitic protozoa and helminths. *Comp. Biochem. Physiol. - C Pharmacol. Toxicol. Endocrinol.*, **121**, 181–183.

Barrett,J. (2009) Forty years of helminth biochemistry. *Parasitology*, **136**, 1633–1642.

Barrett,J. (1997) Helminth detoxification mechanisms. *J. Helminthol.*, **71**, 85–89.

Barrière,A. and Félix,M.A. (2005) High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. *Curr. Biol.*, **15**, 1176–1184.

Bártíková,H. *et al.* (2012) The activity of drug-metabolizing enzymes and the biotransformation of selected anthelmintics in the model tapeworm *Hymenolepis diminuta*. *Parasitology*, **139**, 809–18.

Becker,B. *et al.* (1980) Light and electron microscopic studies on the effect of praziquantel on *Schistosoma mansoni*, *Dicrocoelium dendriticum*, and *Fasciola hepatica* (trematoda) *in vitro*. *Zeitschrift für Parasitenkd. Parasitol. Res.*, **63**, 113–128.

Beech,R.N. *et al.* (2011) Anthelmintic resistance: markers for resistance, or susceptibility? *Parasitology*, **138**, 160–174.

Bernhardt,R. (2006) Cytochromes P450 as versatile biocatalysts. *J. Biotechnol.*, **124**, 128–145.

Beugnet,F. *et al.* (1997) Partial in vitro reversal of benzimidazole resistance by the free-living stages of *Haemonchus contortus* with verapamil. *Vet. Rec.*, **141**, 575–576.

De Bie,T. *et al.* (2006) CAFE: A computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.

Blackhall,W.J. *et al.* (2008) P-glycoprotein selection in strains of *Haemonchus contortus*

resistant to benzimidazoles. *Vet. Parasitol.*, **152**, 101–107.

Blaxter,M. and Denver,D.R. (2012) The worm in the world and the world in the worm. *BMC Biol.*, **10**, 57.

Blaxter,M.L. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.

de Bono,B. *et al.* (2004) VH gene segments in the mouse and human genomes. *J. Mol. Biol.*, **342**, 131–143.

Borgonie,G. *et al.* (2011) Nematoda from the terrestrial deep subsurface of South Africa. *Nature*, **474**, 79–82.

Boussau,B. *et al.* (2013) Genome-scale coestimation of species and gene trees. *Genome Res.*, **23**, 323–330.

Bray,J.E. *et al.* (2009) The human short-chain dehydrogenase/reductase (SDR) superfamily: A bioinformatics summary. *Chem. Biol. Interact.*, **178**, 99–109.

Bundy,D.A.P. *et al.* (2013) Worms, wisdom, and wealth: Why deworming can make economic sense. *Trends Parasitol.*, **29**, 142–148.

Burns,A.R. *et al.* (2015) *Caenorhabditis elegans* is a useful model for anthelmintic discovery. *Nat. Commun.*, **6**, 7485.

Bush,S.J. *et al.* (2014) Presence-absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints. *Mol. Biol. Evol.*, **31**, 59–69.

Buxton,S.K. *et al.* (2014) Diethylcarbamazine Increases Activation of Voltage-Activated Potassium (SLO-1) Currents in *Ascaris suum* and Potentiates Effects of Emodepside. *PLoS Negl. Trop. Dis.*, **8**.

Camacho,C. *et al.* (2009) BLAST plus: architecture and applications. *BMC Bioinformatics*, **10**, 1.

Campbell,P.J. *et al.* (2008) Identification of somatically acquired rearrangements in

cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Carstens,B.C. and Knowles,L.L. (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Syst. Biol.*, **56**, 400–411.

Casola,C. and Hahn,M.W. (2009) Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *J. Mol. Evol.*, **68**, 679–687.

Chakrapani,B.P.S. *et al.* (2008) Development and evaluation of an in vivo assay in *Caenorhabditis elegans* for screening of compounds for their effect on cytochrome P450 expression. *J. Biosci.*, **33**, 269–277.

Chaudhary,R. *et al.* (2015) Assessing approaches for inferring species trees from multi-copy genes. *Syst. Biol.*, **64**, 325–339.

Chen,K. *et al.* (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comp Biol*, **7**, 429–447.

Chiang,D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.

Chiu,T.-L. *et al.* (2008) Comparative molecular modeling of *Anopheles gambiae* CYP6Z1, a mosquito P450 capable of metabolizing DDT. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 8855–8860.

Cinkornpumin,J.K. *et al.* (2014) A host beetle pheromone regulates development and behavior in the nematode *Pristionchus pacificus*. *Elife*, **3**, 1–21.

Colombo,L. *et al.* (2006) Aldosterone and the conquest of land. *J. Endocrinol. Invest.*, **29**, 373–379.

Conrad,D.F. *et al.* (2009) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Conway,D.P. (1964) VARIANCE IN THE EFFECTIVENESS OF THIABENDAZOLE AGAINST *HAEMONCHUS*. *Am. J. Vet. Res.*, **25**, 844–846.

Cook,D.C. *et al.* (2016) The Genetic Basis of Natural Variation in *Caenorhabditis elegans* Telomere Length. *Genetics*, **204**, 371–83.

Cotton,J.A. *et al.* (2016) An expressed, endogenous Nodavirus-like element captured by a retrotransposon in the genome of the plant parasitic nematode *Bursaphelenchus xylophilus*. *Sci. Rep.*, **6**, 39749.

Cui,Y. *et al.* (2007) Toxicogenomic analysis of *Caenorhabditis elegans* reveals novel genes and pathways involved in the resistance to cadmium toxicity. *Genome Biol.*, **8**, R122.

Curran,D.M. *et al.* (2014) Figmop: A profile HMM to identify genes and bypass troublesome gene models in draft genomes. *Bioinformatics*, **30**, 3266–3267.

Cutter,A.D. (2015) *Caenorhabditis* evolution in the wild. *BioEssays*, **37**, 983–995.

Cvilink,V. *et al.* (2008) LC-MS-MS identification of albendazole and flubendazole metabolites formed ex vivo by *Haemonchus contortus*. *Anal. Bioanal. Chem.*, **391**, 337–343.

Cvilink,V. *et al.* (2009) Phase I biotransformation of albendazole in lancet fluke (*Dicrocoelium dendriticum*). *Res. Vet. Sci.*, **86**, 49–55.

Cvilink,V. *et al.* (2009) Xenobiotic metabolizing enzymes and metabolism of anthelminthics in helminths. *Drug Metab. Rev.*, **41**, 8–26.

Darriba,D. *et al.* (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinforma.* , **27**, 1164–1165.

Davies,D.L. and Bouldin,D.W. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1**, 224–227.

Devine,C. *et al.* (2009) Effect of the metabolic inhibitor, methimazole on the drug susceptibility of a triclabendazole-resistant isolate of *Fasciola hepatica*.

Parasitology, **136**, 183–92.

Devine,C. *et al.* (2010) Inhibition of cytochrome P450-mediated metabolism enhances ex vivo susceptibility of *Fasciola hepatica* to triclabendazole. *Parasitology*, **137**, 871–80.

Devine,C. *et al.* (2012) Potentiation of triclabendazole action in vivo against a triclabendazole-resistant isolate of *Fasciola hepatica* following its co-administration with the metabolic inhibitor, ketoconazole. *Vet. Parasitol.*, **184**, 37–47.

Dey,A. *et al.* (2013) Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 11056–60.

Dieterich,C. *et al.* (2008) The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat. Genet.*, **40**, 1193–1198.

Doyon,J.P. *et al.* (2010) An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*., pp. 93–108.

Doyon,J.P. *et al.* (2012) An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, **9**, 26–39.

Driscoll,M. *et al.* (1989) Genetic and molecular analysis of a *Caenorhabditis elegans* beta-tubulin that conveys benzimidazole sensitivity. *J. Cell Biol.*, **109**, 2993–3003.

Dunkov,B.C. *et al.* (1997) The *Drosophila* Cytochrome P450 Gene Cyp6a2 : Localization, Heterologous Expression, and Induction by Phénobarbital. *DNA Cell Biol.*, **16**, 1345–1356.

Earl,D. *et al.* (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res.*, **21**, 2224–2241.

Eddy,S.R. (2008) A probabilistic model of local sequence alignment that simplifies

statistical significance estimation. *PLoS Comput. Biol.*, **4**.

Enayati,A.A. *et al.* (2005) Insect glutathione transferases and insecticide resistance. *Insect Mol. Biol.*, **14**, 3–8.

Eppstein,D. (1994) Finding the k shortest paths. *35th Annu. Symp. Found. Comput. Sci.*, 154–165.

Félix,M.-A. (2008) RNA interference in nematodes and the chance that favored Sydney Brenner. *J. Biol.*, **7**, 34.

Félix,M.-A. and Duveau,F. (2012) Population dynamics and habitat sharing of natural populations of *Caenorhabditis elegans* and *C. briggsae*. *BMC Biol.*, **10**, 59.

Fellermann,K. *et al.* (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.*, **79**, 439–448.

Feuk,L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Forrester,S.G. *et al.* (2002) A glutamate-gated chloride channel subunit from *Haemonchus contortus*: Expression in a mammalian cell line, ligand binding, and modulation of anthelmintic binding by glutamate. *Biochem. Pharmacol.*, **63**, 1061–1068.

Gao,F. *et al.* (2014) Genome structure drives patterns of gene family evolution in ciliates, a case study using chilodonella uncinata (protista, ciliophora, phyllopharyngea). *Evolution (N. Y).*, **68**, 2287–2295.

Geary,T.G. (2012) Are new anthelmintics needed to eliminate human helminthiases? *Curr. Opin. Infect. Dis.*, **25**, 709–17.

Ghisi,M. *et al.* (2007) Phenotyping and genotyping of *Haemonchus contortus* isolates reveals a new putative candidate mutation for benzimidazole resistance in nematodes. *Vet. Parasitol.*, **144**, 313–320.

Gilabert,A. *et al.* (2016) Expanding the view on the evolution of the nematode dauer signalling pathways: refinement through gene gain and pathway co-option. *BMC Genomics*, **17**, 476.

Gilleard,J.S.J. (2013) *Haemonchus contortus* as a paradigm and model to study anthelmintic drug resistance. *Parasitology*, **140**, 1506–22.

Gonzalez,F.J. and Nebert,D.W. (1990) Evolution of the P450 gene superfamily:. animal-plant 'warfare', molecular drive and human genetic differences in drug oxidation. *Trends Genet.*, **6**, 182–186.

Guengerich,F.P. (2003) Cytochromes P450, drugs, and diseases. *Mol. Interv.*, **3**, 194–204.

Guiliano,D.B. and Blaxter,M.L. (2006) Operon conservation and the evolution of trans-splicing in the phylum nematoda. *PLoS Genet.*, **2**.

Guindon,S. and Gascuel,O. (2003) A Simple, Fast, and Accurate Method to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.*, **52**, 696–704.

Gupta,S. and Rathaur,S. (2005) Filarial glutathione S-transferase: Its induction by xenobiotics and potential as drug target. *Acta Biochim. Pol.*, **52**, 493–500.

Gustavsen,K. *et al.* (2011) Onchocerciasis in the Americas: from arrival to (near) elimination. *Parasit. Vectors*, **4**, 205.

Haber,M. *et al.* (2005) Evolutionary history of *Caenorhabditis elegans* inferred from microsatellites: Evidence for spatial and temporal genetic differentiation and the occurrence of outbreeding. *Mol. Biol. Evol.*, **22**, 160–173.

Han,M. V. *et al.* (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.*, **30**, 1987–1997.

Hayes,J.D. and Pulford,D.J. (1995) The glutathione S-transferase supergene family: regulation of GST and the contribution of the isoenzymes to cancer chemoprotection and drug resistance. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 445–600.

Herrmann,M. *et al.* (2006) Nematodes of the genus *Pristionchus* are closely associated with scarab beetles and the Colorado potato beetle in Western Europe. *Zoology*, **109**, 96–108.

Hill,A. V and Wainscoat,J.S. (1986) The evolution of the alpha- and beta-globin gene clusters in human populations. *Hum. Genet.*, **74**, 16–23.

Hoffmann,F. and Maser,E. (2007) Carbonyl reductases and pluripotent hydroxysteroid dehydrogenases of the short-chain dehydrogenase/reductase superfamily. *Drug Metab. Rev.*, **39**, 87–144.

Hotez,P.J. *et al.* (2009) Rescuing the bottom billion through control of neglected tropical diseases. *Lancet*, **373**, 1570–1575.

Hurley,I. *et al.* (2005) Duplication events and the evolution of segmental identity. In, *Evolution and Development*., pp. 556–567.

Iafrate,A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Innan,H. and Kondrashov,F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.*, **11**, 97–108.

James,C.E. and Davey,M.W. (2009) Increased expression of ABC transport proteins is associated with ivermectin resistance in the model nematode *Caenorhabditis elegans*. *Int. J. Parasitol.*, **39**, 213–220.

Jensen,N.B. *et al.* (2011) Convergent evolution in biosynthesis of cyanogenic defence compounds in plants and insects. *Nat. Commun.*, **2**, 273.

Jez,J.M. and Penning,T.M. (2001) The aldo-keto reductase (AKR) superfamily: An update. In, *Chemico-Biological Interactions*., pp. 499–525.

Johnson,S.S. *et al.* (2004) Interrelationships among physicochemical properties, absorption and anthelmintic activities of 2-desoxoparaherquamide and selected analogs. *J. Vet. Pharmacol. Ther.*, **27**, 169–181.

Jones,L.M. *et al.* (2013) Adaptive and Specialised Transcriptional Responses to Xenobiotic Stress in *Caenorhabditis elegans* Are Regulated by Nuclear Hormone Receptors. *PLoS One*, **8**, e69956.

Jornvall,H. *et al.* (1999) SDR and MDR: completed genome sequences show these protein families to be large, of old origin, and of complex nature. *FEBS Lett.*, **445**, 261–264.

Jörnvall,H. *et al.* (2010) Superfamilies SDR and MDR: From early ancestry to present forms. Emergence of three lines, a Zn-metalloenzyme, and distinct variabilities. *Biochem. Biophys. Res. Commun.*, **396**, 125–130.

Kallberg,Y. *et al.* (2010) Classification of the short-chain dehydrogenase/reductase superfamily using hidden Markov models. *FEBS J.*, **277**, 2375–2386.

Kallberg,Y. and Persson,B. (2006) Prediction of coenzyme specificity in dehydrogenases/reductases: A hidden Markov model-based method and its application on complete genomes. *FEBS J.*, **273**, 1177–1184.

Kamburowski,J. (1997) The nature of simplicity of Johnson's algorithm. *Omega*, **25**, 581–584.

Kaminsky,R. *et al.* (2008) Identification of the amino-acetonitrile derivative monepantel (AAD 1566) as a new anthelmintic drug development candidate. *Parasitol. Res.*, **103**, 931–939.

Kapitulnik,J. and Gonzalez,F.J. (1993) Marked endogenous activation of the CYP1A1 and CYP1A2 genes in the congenitally jaundiced Gunn rat. *Mol.Pharmacol.*, **43**, 722–725.

Kaplan,R.M. (2004) Drug resistance in nematodes of veterinary importance: A status report. *Trends Parasitol.*, **20**, 477–481.

Kaplan,R.M. and Vidyashankar,A.N. (2012) An inconvenient truth: Global worming and anthelmintic resistance. *Vet. Parasitol.*, **186**, 70–78.

Katju,V. and Bergthorsson,U. (2010) Genomic and population-level effects of gene conversion in *Caenorhabditis* paralogs. *Genes (Basel).*, **1**, 452–468.

Kavanagh,K.L. *et al.* (2008) Medium- and short-chain dehydrogenase/reductase gene and protein families: The SDR superfamily: Functional and structural diversity within a family of metabolic and regulatory enzymes. *Cell. Mol. Life Sci.*, **65**, 3895–3906.

Keiser,J. and Utzinger,J. (2010) The Drugs We Have and the Drugs We Need Against Major Helminth Infections. *Adv. Parasitol.*, **73**, 197–230.

Keller,O. *et al.* (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, **27**, 757–763.

Kerboeuf,D. *et al.* (1999) Flow cytometry analysis of drug transport mechanisms in *Haemonchus contortus* susceptible or resistant to anthelmintics. *Parasitol. Res.*, **85**, 118–23.

Kerboeuf,D. *et al.* (2003) P-glycoprotein in helminths: Function and perspectives for anthelmintic treatment and reversal of resistance. *Int. J. Antimicrob. Agents*, **22**, 332–346.

Kiang,T.K.L. *et al.* (2005) UDP-glucuronosyltransferases and clinical drug-drug interactions. *Pharmacol. Ther.*, **106**, 97–132.

Kiontke,K. *et al.* (2004) *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 9003–8.

Kiontke,K. and Sudhaus,W. (2006) Ecology of *Caenorhabditis* species. *WormBook*, 1–14.

Klingenberg,M. (1958) Pigments of rat liver microsomes. *Arch. Biochem. Biophys.*, **75**, 376–386.

Kondrashov,F.A. (2012) Gene duplication as a mechanism of genomic adaptation to a

changing environment. *Proceedings. Biol. Sci.*, **279**, 5048–5057.

Krau,S.D. (2013) Cytochrome p450, Part 1. What Nurses Really Need to Know. *Nurs. Clin. North Am.*, **48**, 671–680.

Krueger,S.K. and Williams,D.E. (2005) Mammalian flavin containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism. *Pharmacol. Ther.*, **106**, 357–387.

Kwa,M.S. *et al.* (1995) Beta-tubulin genes from the parasitic nematode *Haemonchus contortus* modulate drug resistance in *Caenorhabditis elegans*. *J. Mol. Biol.*, **246**, 500–510.

Kwa,M.S.G. *et al.* (1994) Benzimidazole resistance in *Haemonchus contortus* is correlated with a conserved mutation at amino acid 200 in β-tubulin isotype 1. *Mol. Biochem. Parasitol.*, **63**, 299–303.

Kwon,J.Y. *et al.* (2004) Ethanol-response genes and their regulation analyzed by a microarray and comparative genomic approach in the nematode *Caenorhabditis elegans*. *Genomics*, **83**, 600–614.

Laborde,E. (2010) Glutathione transferases as mediators of signaling pathways involved in cell proliferation and cell death. *Cell Death Differ.*, **17**, 1373–1380.

Laing,R. *et al.* (2015) The cytochrome P450 family in the parasitic nematode *Haemonchus contortus*. *Int. J. Parasitol.*, **45**, 243–251.

Laing,R. *et al.* (2013) The genome and transcriptome of *Haemonchus contortus*, a key model parasite for drug and vaccine discovery. *Genome Biol.*, **14**, R88.

Laing,S.T. *et al.* (2010) Characterization of the xenobiotic response of *Caenorhabditis elegans* to the anthelmintic drug albendazole and the identification of novel drug glucoside metabolites. *Biochem. J.*, **432**, 505–514.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357–359.

Lee,B.H. *et al.* (2002) Marcfortine and paraherquamide class of anthelmintics: discovery of PNU-141962. *Curr Top Med Chem*, **2**, 779–793.

Lee,H. *et al.* (2011) Nictation, a dispersal behavior of the nematode *Caenorhabditis elegans*, is regulated by IL2 neurons. *Nat. Neurosci.*, **15**, 107–112.

Lee,Y. *et al.* (2016) Inverse correlation between longevity and developmental rate among wild *C. elegans* strains. *Aging (Albany. NY).*, **8**, 986–999.

Lespine,A. *et al.* (2012) P-glycoproteins and other multidrug resistance transporters in the pharmacology of anthelmintics: Prospects for reversing transport-dependent anthelmintic resistance. *Int. J. Parasitol. Drugs Drug Resist.*, **2**, 58–75.

Levecke,B. *et al.* (2014) Assessment of anthelmintic efficacy of mebendazole in school children in six countries where soil-transmitted helminths are endemic. *PLoS Negl. Trop. Dis.*, **8**, e3204.

Librado,P. *et al.* (2012) BadiRate: Estimating family turnover rates by likelihood-based methods. *Bioinformatics*, **28**, 279–281.

Lustigman,S. *et al.* (2012) A research agenda for helminth diseases of humans: The problem of helminthiases. *PLoS Negl. Trop. Dis.*, **6**, e1582.

Lynch,M. (2007) The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.*, **8**, 803–13.

Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–5.

Ma,J. *et al.* (2008) DUPCAR: reconstructing contiguous ancestral regions with duplications. *J. Comput. Biol.*, **15**, 1007–27.

Magi,A. *et al.* (2011) Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res.*, **39**.

Magi,A. *et al.* (2012) Read count approach for DNA copy number variants detection. *Bioinformatics*, **28**, 470–478.

Mardis,E.R. (2010) The $1,000 genome, the $100,000 analysis? *Genome Med.*, **2**, 84.

Martin,R.J. *et al.* (2004) Oxantel is an N-type (methyridine and nicotine) agonist not an L-type (levamisole and pyrantel) agonist: Classification of cholinergic anthelmintics in Ascaris. *Int. J. Parasitol.*, **34**, 1083–1090.

Martin,R.J. and Pennington, a J. (1989) A patch-clamp study of effects of dihydroavermectin on *Ascaris* muscle. *Br. J. Pharmacol.*, **98**, 747–756.

Mathew,M.D. *et al.* (2016) Using *C. elegans* Forward and Reverse Genetics to Identify New Compounds with Anthelmintic Activity. *PLoS Negl. Trop. Dis.*, **10**, e0005058.

Matoušková,P. *et al.* (2016) The Role of Xenobiotic-Metabolizing Enzymes in Anthelmintic Deactivation and Resistance in Helminths. *Trends Parasitol.*, **32**, 481–491.

Meech,R. *et al.* (2012) The glycosidation of xenobiotics and endogenous compounds: Versatility and redundancy in the UDP glycosyltransferase superfamily. *Pharmacol. Ther.*, **134**, 200–218.

Mendivil Ramos,O. and Ferrier,D.E.K. (2012) Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution. *Int J Evol Biol*, **2012**, 846421.

Menez,C. *et al.* (2016) Acquired tolerance to ivermectin and moxidectin after drug selection pressure in the nematode *Caenorhabditis elegans*. *Antimicrob. Agents Chemother.*, **60**, 4809–4819.

Menzel,R. *et al.* (2001) A systematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 genes reveals CYP35 as strongly xenobiotic inducible. *Arch. Biochem. Biophys.*, **395**, 158–168.

Menzel,R. *et al.* (2005) CYP35: xenobiotically induced gene expression in the nematode *Caenorhabditis elegans*. *Arch Biochem Biophys*, **438**, 93–102.

Menzel,R. *et al.* (2007) Cytochrome P450s and Short-chain Dehydrogenases Mediate

the Toxicogenomic Response of PCB52 in the Nematode *Caenorhabditis elegans*. *J. Mol. Biol.*, **370**, 1–13.

Mirarab,S. *et al.* (2016) Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst. Biol.*, **65**, 366–380.

Molento,M.B. and Prichard,R.K. (1999) Effects of the multidrug-resistance-reversing agents verapamil and CL 347,099 on the efficacy of ivermectin or moxidectin against unselected and drug-selected strains of *Haemonchus contortus* in jirds (*Meriones unguiculatus*). *Parasitol. Res.*, **85**, 1007–1011.

Munguía,B. *et al.* (2015) Development of novel valerolactam-benzimidazole hybrids anthelmintic derivatives: Diffusion and biotransformation studies in helminth parasites. *Exp. Parasitol.*, **153**, 75–80.

Munro,A.W. *et al.* (2003) Cytochromes P450: novel drug targets in the war against multidrug-resistant *Mycobacterium tuberculosis*. *Biochem. Soc. Trans.*, **31**, 625–630.

Murray,C.J.L. *et al.* (2012) Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, **380**, 2197–2223.

Nagarajan,N. and Pop,M. (2013) Sequence assembly demystified. *Nat. Rev. Genet.*, **14**, 157–67.

Nei,M. *et al.* (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 7799–7806.

Nelson,D.R. (2011) Progress in tracing the evolutionary paths of cytochrome P450. *Biochim. Biophys. Acta - Proteins Proteomics*, **1814**, 14–18.

Nelson,D.R. (2009) The cytochrome p450 homepage. *Hum. Genomics*, **4**, 59–65.

Niimura,Y. and Nei,M. (2005) Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene*, **346**, 23–28.

Omiecinski,C.J. *et al.* (2011) Xenobiotic metabolism, disposition, and regulation by receptors: From biochemical phenomenon to predictors of major toxicities. *Toxicol. Sci.*, **120**, S49–S75.

Omura,T. (1999) Forty years of cytochrome P450. *Biochem. Biophys. Res. Commun.*, **266**, 690–8.

Opazo,J.C. *et al.* (2008) Genomic evidence for independent origins of beta-like globin genes in monotremes and therian mammals. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 1590–1595.

Panchin,A.Y. *et al.* (2010) Asymmetric and non-uniform evolution of recently duplicated human genes. *Biol. Direct*, **5**, 54.

Pang,A.W. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.*, **11**, R52.

Parra,G. *et al.* (2009) Assessing the gene space in draft genomes. *Nucleic Acids Res.*, **37**, 289–297.

Parra,G. *et al.* (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Pax,R. *et al.* (1978) A Benzodiazepine Derivative and Praziquantel: Effects on Musculature of *Schistosoma mansoni* and *Schistosoma japonicum*. *Naunyn. Schmiedebergs. Arch. Pharmacol.*, **304**, 309–315.

Pemberton,K.D. and Barrett,J. (1989) The detoxification of xenobiotic compounds by *Onchocerca gutturosa* (Nematoda: Filarioidea). *Int. J. Parasitol.*, **19**, 875–878.

Persson,B. *et al.* (2009) The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Biol. Interact.*, **178**, 94–98.

Petalcorin,M.I.R. *et al.* (2005) The fmo genes of *Caenorhabditis elegans* and *C. briggsae*: Characterisation, gene expression and comparative genomic analysis. *Gene*, **346**, 83–96.

Picardi,E. and Pesole,G. (2010) Computational methods for ab initio and comparative gene finding. *Methods Mol. Biol.*, **609**, 269–284.

Ponting,C.P. (2008) The functional repertoires of metazoan genomes. *Nat. Rev. Genet.*, **9**, 689–698.

Prchal,L. *et al.* (2015) Biotransformation of anthelmintics and the activity of drug-metabolizing enzymes in the tapeworm *Moniezia expansa*. *Parasitology*, **142**, 648–659.

Prchal,L. *et al.* (2016) Metabolism of drugs and other xenobiotics in giant liver fluke (*Fascioloides magna*). *Xenobiotica.*, **46**, 132–140.

Precious,W.Y. and Barrett,J. (1989) The possible absence of cytochrome P-450 linked xenobiotic metabolism in helminths. *BBA - Gen. Subj.*, **992**, 215–222.

Rasmussen,M.D. and Kellis,M. (2011) A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol.*, **28**, 273–290.

Rasmussen,M.D. and Kellis,M. (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.*, **17**, 1932–1942.

Reichert,K. and Menzel,R. (2005) Expression profiling of five different xenobiotics using a *Caenorhabditis elegans* whole genome microarray. *Chemosphere*, **61**, 229–237.

Robinson,M.W. *et al.* (2004) The comparative metabolism of triclabendazole sulphoxide by triclabendazole-susceptible and triclabendazole-resistant *Fasciola hepatica*. *Parasitol. Res.*, **92**, 205–210.

Rollinson,D. *et al.* (2013) Time to set the agenda for schistosomiasis elimination. *Acta Trop.*, **128**, 423–440.

Rosello,O.P.I. and Kondrashov,F.A. (2014) Long-Term asymmetrical acceleration of protein evolution after gene duplication. *Genome Biol. Evol.*, **6**, 1949–1955.

Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation

of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Rowland,A. *et al.* (2013) The UDP-glucuronosyltransferases: Their role in drug metabolism and detoxification. *Int. J. Biochem. Cell Biol.*, **45**, 1121–1132.

Salzberg,S.L. *et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, **22**, 557–567.

Saunders,G.I. *et al.* (2013) Characterization and comparative analysis of the complete *Haemonchus contortus* B-tubulin gene family and implications for benzimidazole resistance in strongylid nematodes. *Int. J. Parasitol.*, **43**, 465–475.

Scally,A. *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**, 169–75.

Scarcella,S. *et al.* (2012) Increase of carboxylesterase activity in *Fasciola hepatica* recovered from triclabendazole treated sheep. *Mol. Biochem. Parasitol.*, **185**, 151–153.

Scarcella,S. *et al.* (2013) Increase of gluthatione S-transferase, carboxyl esterase and carbonyl reductase in *Fasciola hepatica* recovered from triclabendazole treated sheep. *Mol. Biochem. Parasitol.*, **191**, 63–65.

Sebat,J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316**, 445–9.

Serobyan,V. *et al.* (2014) Adaptive value of a predatory mouth-form in a dimorphic nematode. *Proc. Biol. Sci.*, **281**, 20141334.

Shimodaira,H. (2002) An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst. Biol.*, **51**, 492–508.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

Skálová,L. *et al.* (2010) Biotransformation of Xenobiotics in Lancet Fluke (*Dicrocoelium dendriticum*). In, *Veterinary Parasitology*., pp. 251–271.

Slater,G.S.C. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Spieth,J. *et al.* (1993) Operons in *C. elegans*: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell*, **73**, 521–532.

Stamatakis,A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stanke,M. *et al.* (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Stanke,M. *et al.* (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.

Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, ii215-ii225.

Stein,L.D. *et al.* (2003) The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.*, **1**.

Stolzer,M. *et al.* (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, **28**, 409–415.

Stromberg,B.E. and Gasbarre,L.C. (2006) Gastrointestinal Nematode Control Programs with an Emphasis on Cattle. *Vet. Clin. North Am. - Food Anim. Pract.*, **22**, 543–565.

Stuchlikova,L. *et al.* (2014) Metabolic pathways of anthelmintic drug monepantel in sheep and in its parasite (*Haemonchus contortus*). *Drug Test. Anal.*, **6**, 1055–1062.

Su,C. *et al.* (1999) Diversity and evolution of T-cell receptor variable region genes in mammals and birds. *Immunogenetics*, **50**, 301–308.

Subirana,J.A. *et al.* (2015) High evolutionary turnover of satellite families in *Caenorhabditis*. *BMC Evol. Biol.*, **15**, 218.

Sudhaus,W. and Kiontke,K. (2007) Comparison of the cryptic nematode species

Caenorhabditis brennen sp. n. and *C. remanei* (Nematoda: Rhabditidae) with the stem species pattern of the Caenorhabditis Elegans group. *Zootaxa*, 45–62.

Sutherland,I.A. and Leathwick,D.M. (2011) Anthelmintic resistance in nematode parasites of cattle: A global issue? *Trends Parasitol.*, **27**, 176–181.

Tan,S. *et al.* (2012) Variation of presence/absence genes among *Arabidopsis* populations. *BMC Evol. Biol.*, **12**, 1.

Tekle,A.H. *et al.* (2012) Impact of long-term treatment of onchocerciasis with ivermectin in Kaduna State, Nigeria: first evidence of the potential for elimination in the operational area of the African Programme for Onchocerciasis Control. *Parasit. Vectors*, **5**, 28.

The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

Thomas,C.G. *et al.* (2015) Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. *Genome Res.*, **125**, 667–678.

Thomas,J.H. *et al.* (2005) Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 4476–4481.

Thomas,J.H. (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. *PLoS Genet.*, **3**, 720–728.

Thompson,O. *et al.* (2013) The million mutation project: A new approach to genetics in *Caenorhabditis elegans*. *Genome Res.*, **23**, 1749–1762.

Timbers,T.A. *et al.* (2016) Accelerating Gene Discovery by Phenotyping Whole-Genome Sequenced Multi-mutation Strains and Using the Sequence Kernel Association Test (SKAT). *PLoS Genet.*, **12**, e1006235.

Torgerson,W.S. (1952) Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**, 401–419.

Trudgill,D.L. and Blok,V.C. (2001) APOMICTIC, POLYPHAGOUS ROOT-KNOT NEMATODES: Exceptionally Successful and Damaging Biotrophic Root Pathogens. *Annu. Rev. Phytopathol*, **39**, 53–77.

Tukey,R.H. and Strassburg,C.P. (2000) Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annu. Rev. Pharmacol. Toxicol.*, **40**, 581–616.

Vattaa, a F. and Lindberg, a L.E. (2006) Managing anthelmintic resistance in small ruminant livestock of resource-poor farmers in South Africa. *J. S. Afr. Vet. Assoc.*, **77**, 2–8.

Vernot,B. *et al.* (2008) Reconciliation with non-binary species trees. *J. Comput. Biol.*, **15**, 981–1006.

Vokral,I. *et al.* (2010) Effect of Flubendazole on Biotransformation Enzymes Activities in *Haemonchus contortus*. *Open Parasitol. J.*, **4**, 24–28.

Vokřál,I. *et al.* (2012) The metabolism of flubendazole and the activities of selected biotransformation enzymes in *Haemonchus contortus* strains susceptible and resistant to anthelmintics. *Parasitology*, **139**, 1309–16.

Wakaguri,H. *et al.* (2009) Inconsistencies of genome annotations in apicomplexan parasites revealed by 5'-end-one-pass and full-length sequences of oligo-capped cDNAs. *BMC Genomics*, **10**, 312.

Wallace,I.M. *et al.* (2006) M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.

Wasmuth,J.D. *et al.* (2012) Integrated bioinformatic and targeted deletion analyses of the SRS gene superfamily identify SRS29C as a negative regulator of toxoplasma virulence. *MBio*, **3**, e00321-12.

Wheeler,T.J. and Eddy,S.R. (2013) Nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, **29**, 2487–2489.

World Health Organization (2005) Deworming for health and development: report of the third global meeting of the partners for parasite control.

World Health Organization (2006) Preventive chemotherapy in human helminthiasis. *… Use Anthelminthic Drugs Control …*, 62.

Yandell,M. and Ence,D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, **13**, 329–42.

Yen,J.Y. (1971) Finding the K Shortest Loopless Paths in a Network. *Manage. Sci.*, **17**, 712–716.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

Zhang,J. (2003) Evolution by gene duplication: An update. *Trends Ecol. Evol.*, **18**, 292–298.

Zhou,S.F. *et al.* (2009) Insights into the substrate specificity, inhibitors, regulation, and polymorphisms and the clinical impact of human cytochrome P450 1A2. *Aaps J*, **11**, 481–494.

Ziegler,D.M. (2002) An overview of the mechanism, substrate specificities, and structure of FMOs. *Drug Metab. Rev.*, **34**, 503–511.

Ziegler,D.M. and Mitchell,C.H. (1972) Microsomal oxidase IV: Properties of a mixed-function amine oxidase isolated from pig liver microsomes. *Arch. Biochem. Biophys.*, **150**, 116–125.

Appendices A: Reproduction license for Figmop manuscript

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Oct 12, 2016

This Agreement between David M Curran ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	3966720992351
License date	Oct 12, 2016
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	Figmop: a profile HMM to identify genes and bypass troublesome gene models in draft genomes:
Licensed content author	David M. Curran, John S. Gilleard, James D. Wasmuth
Licensed content date	11/15/2014
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Comparative Genomics Analysis of the Detoxification Gene Families in Nematodes
Publisher of your work	n/a
Expected publication date	Dec 2016
Permissions cost	0.00 CAD
Value added tax	0.00 CAD
Total	0.00 CAD
Requestor Location	David M Curran Room G351F, 3330 Hospital Dr N.W. Calgary, AB T2N1N4 Canada Attn: David M Curran
Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	David M Curran Room G351F, 3330 Hospital Dr N.W. Calgary, AB T2N1N4 Canada Attn: David M Curran
Total	0.00 CAD
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.
5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.
6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com
7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.
8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.
9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.
10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
12. Other Terms and Conditions:
v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.
