THE CYTOCHROME P450 SUPERFAMILY COMPLEMENT IN *CAPITELLA TELETA*

THE CYTOCHROME P450 SUPERFAMILY COMPLEMENT (CYPome) IN THE ANNELID *CAPITELLA TELETA*

By:

CHRISTOPHER DEJONG, H.B.SC.

A thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2013) McMaster University

(Biology) Hamilton, Ontario

TITLE: The Cytochrome P450 superfamily Complement (CYPome) in the annelid *Capitella teleta*

AUTHOR: Chris A. Dejong H.B.Sc (McMaster University)

SUPERVISOR: Dr. Joanna Y. Wilson

NUMBER OF PAGES: iv, 116

Abstract

CYPs are a large and diverse protein superfamily found in all domains of life and are able to metabolize a wide array of both exogenous and endogenous molecules. The CYPome of the polychaete annelid *Capitella teleta* has been robustly identified and annotated with the genome assembly available (version 1). Annotation of 84 full length and 12 partial CYP sequences predicted a total of 96 functional CYPs in *C. teleta.* A further 13 CYP fragments were found but these may be pseudogenes. The *C. teleta* CYPome contained 24 novel CYP families and seven novel CYP subfamilies within existing families. A phylogenetic analysis was completed, primarily with vertebrate sequences, and identified that the *C teleta* sequences were found in 9 of the 11 metazoan CYP clans. Clan 2 was expanded in this species with 51 CYPs in 14 novel CYP families containing 20 subfamilies. There were five clan 3, four clan 4, and six mitochondrial clan full length CYPs. Two CYPs, CYP3071A1 and CYP3072A1, did not cluster with any metazoan CYP clan. *C. teleta* had a CYP51A1 gene with ~65% identity to vertebrate CYP51A1 sequences and was predicted to have lanosterol 14 α-demethylase activity. Several CYPs (CYP376A1, CYP3068A1, CYP3069A1, and CYP3070A1) are discussed as candidate genes for steroidogenesis. There are two CYP1-like CYPs and a total of four CYP331s found in *C. teleta*, which may play a role in PAH metabolism and warrant further analysis. The sand goby, *Pomatoschistus minutus*, had its CYPome annotated but a majority of the CYPs could only be partially annotated because the current sand goby genome assembly was still in the contig stage. Based on the annotations, the sand goby

CYPome is predicted to contain approximately 50 CYPs with members from all expected families and subfamilies.

Acknowledgements

I always said I'd name as few names as possible here, and I think I'll succeed in doing that. Thank you to everyone in the Wilson lab, and everyone that dealt with my loud and often crass nature around LSB, you've been great to talk to and relax with, each of you know who you are. Thank you Derek for the consistent counsel over the years I've been here. Thank you to Andrew McArthur for edits on chapter 2 and all other help along the way. Thank you to all of my housemates over the years here in Hamilton, you girls kept me alive for that last writing stretch... seriously. To my family (Mama, daddy, Yiayia and Pa, my sibs and the list keeps going) for their endless support and understanding. To Noah, my brother , thanks for keeping me sane and always being there no matter what, and listening to my endless science.

To my committee - Jon Stone and Ben Evans. Thanks for the input for the project and council.

And of course to Joanna Wilson, you're honestly the best supervisor a guy could ask for. You have dealt with all my crap from the very lows to the crunch-time highs. Thank you for the countless, and scarily quick and thorough edits. Thank you for being my science and life council.

This work represents two years of my life. I hope you get something out of it!

Table of Contents

List of Tables and Figures

Chapter 1:

Chapter 2:

Appendix I

Abbreviations

AA: Amino Acid

AHR: Aryl Hydrocarbon Receptor

BAP: Benzo(a)pyrene

BLAST: Basic Local Alignment Sequence Tool

bp: Base pair

CYP: Cytochrome P450

CYPome: Cytochrome P450 genome complement

DNA: Deoxyribonucleic acid

ER: Estrogen Receptor

EST: Expressed Sequence Tag

FASTA: Fast-All

gb: Gigabase

JGI: Joint Genome Institute

MEME: Multiple Em for Motif Elicitation

MUSCLE: Multiple Sequence Comparison by Log-Expectation

NCBI: National Center for Biotechnology Information

PAH: Polycyclic Aromatic Hydrocarbon

PASA: Program to Assemble Spliced Alignments

RaxML: Randomized axelerated Maximum Likelihood

RNA: Ribonucleic Acid

RNAseq: RNA-Sequencing

XRE: Xenobiotic Response Element

Chapter 1: General Introduction

Cytochrome P450

Superfamily Overview

The cytochrome P450 (CYP) superfamily of proteins is found in all domains of life (Nelson, 2011; Nelson, 1999). The main function the CYP enzyme is the catalysis of a monooxygenase reaction (Nebert and Gonzalez, 1987). These reactions are important in metabolism of both exogenous compounds, such as drugs and pollutants (e.g. polycyclic aromatic hydrocarbons (PAHs)), and the anabolism and catabolism of endogenous compounds, such as lipids and steroid hormones (Nelson, et al., 2013). Metazoans typically have between 40 and 120 CYPs in their genome (Nelson, 2011). The high CYP gene copy number in metazoans is due to the wide array of molecules CYPs act on, a need for specificity towards target molecules, and to provide differential gene expression (Goldstone et al., 2006). There are over 12,000 CYP genes in 1000 named CYP families identified across all organisms (Nelson, 2011); 196 named CYP families are found in metazoa (Nelson et al., 2013). Additional CYP families are identified on a regular basis, with almost all newly annotated CYP genome complements (CYPomes) from non-vertebrates adding novel families to this gene superfamily (Nelson, 2011).

CYPs are best known for their role in the metabolism of drugs and contaminants. CYPs metabolize a majority of known drugs (Nebert and Russell, 2002; Nebert, et al., 2013). In humans, CYP3A4 and genes from the CYP2C and 2D subfamilies are

responsible for the majority of the reactions, with some activity provided by other CYP2s, CYP1A1 and CYP1A2 (Nebert and Russell, 2002; Nebert et al., 2013). In general, the CYPs responsible for metabolism of drugs and contaminants are found in CYP families 1-4.

All newly identified CYPs are named by the Cytochrome P450 nomenclature committee, using standard conventions for this gene superfamily. CYPs are named by amino acid sequence identity; genes with 40% and 55% identify are placed in the same family and subfamily, respectively. CYPs are named by family using a numeral and named by subfamily using a letter. The specific gene is given a number, by order of discovery. For example CYP19A1 is in family 19, subfamily A and has a gene number of 1 (Nelson et al., 1996). Phylogenetics and synteny can play a role in naming genes that have evident orthology and conserved synteny yet do not satisfy the arbitrary identity cut-off (Nelson et al., 1996). Synteny is playing a smaller role in nomenclature assignments because there is a high rate of synteny loss outside vertebrates (Nelson, 2011).

Beyond the traditional naming of CYPs, David Nelson (1998) coined the term 'clan' (similarity to 'clade' is intentional) as a broad grouping of CYP families that consistently cluster together on phylogenetic trees. The clans are meant to capture the some of the evolutionary history between and across CYP families (Nelson, 1999). Though there are no strict parameters for what grouping of sequences make up a clan, or how much diversity there is between the clans, the designation makes naming and describing CYPs less cumbersome (Nelson, 1999). There are 11 clans in metazoa, (2, 3,

4, 7, 19, 20, 26, 46, 51, 74 and mitochondrial) but only amphioxus, of all organisms with a defined CYPome, has all metazoan clans (Figure 1; Nelson et al., 2013). Jawed vertebrates (gnathostomes) have all clans except clan 74 (Figure 1; Nelson et al., 2013). Clans are traditionally named after the lowest number CYP family in the clan, with the exception of clan 2, which also includes the CYP1 family (Nelson et al., 2013). Some clans are very large and have more diverse functions than the smaller clans. For example, clan 2 contains CYP families with roles in steroidogenesis (CYP17A and CYP21A) and xenobiotic metabolism (CYP1 and CYP2). There are very small clans that contain only a single CYP family, for example the 19 and 51 clans. Typically, the large CYP families (families 1-4) are responsible for metabolism of exogenous substrates. CYP families with single genes or small copy number typically occur in a one-to-one manner across vertebrate species and these genes are generally responsible for metabolism of endogenous substrates such as CYP51 (cholesterol) and CYP19 (androgens; Lewis, 2004). CYP families primarily responsible for exogenous metabolism generally have a higher copy number. These CYPs have a more flexible active site than CYPs involved in metabolism of endogenous substrates, this is likely to allow the enzymes to metabolize the wide range of xenobiotics that may be encountered in the environment (Lewis, 2004).

The ancestral eukaryotic CYP is believed to be CYP51; it is the only CYP found across plants (*Arabidopsis* has two), metazoa, and fungi (Nelson, 1999). The ancestor of ascomycete fungi was thought to have only contained CYP51 (lanosterol 14a-demethylase) and CYP61 (22-sterol desaturase). CYP51 is involved in the early steps of the ergosterol pathway in fungi, converting the first sterol, lanosterol, in the pathway;

CYP61 is further in the ergosterol pathway in fungi and so likely evolved after CYP51 (Nelson, 1999). Because of its deduced evolutionary history, CYP51 is typically used to root phylogenetic trees of metazoan CYPomes (Nelson, 1999; Nelson, 2011; Yoshida, et al., 2000).

Metazoan CYP Diversity

CYPs have been extensively studied on an evolutionary level in vertebrates, and increasingly in invertebrates (as reviewed by Sezutsu, et al., 2013). Vertebrate CYPomes such as human (Lewis, 2004), zebrafish (*Danio rerio*; Goldstone et al., 2010) and pufferfish (*Fugu rubripes;* Nelson, 2003) have helped to characterize the expected complement of CYPs in vertebrates. Vertebrates typically have CYPs from families 1-5, 7, 8, 11, 17, 19, 20, 21, 24, 26, 27, 39, 46 and 51. Due to a whole genome duplication in the teleost line, teleost fish have an increased number of CYPs and duplicated CYPs that are typically found in only a single copy in the other vertebrate groups (Goldstone et al., 2010).

Invertebrate CYPomes such as *Drosophila melanogaster* (Tijet, et al., 2001), *Daphnia pulex* (Baldwin, Marko, and Nelson, 2009), *Strongylocentrotus purpuratus* (Goldstone et al., 2006), and *Nematostella vectensis* (Goldstone, 2008) have begun to fill in our knowledge of CYP evolution outside vertebrates. There are many other CYPomes that have been catalogued without comparative analyses or peer reviewed publication, these CYPomes are located on the Cytochrome P450 webpage (Nelson, 2009). Non-chordate metazoans typically have fewer CYP clans. Many of the clans responsible for

production and metabolism of endogenous molecules important in vertebrate species are missing in invertebrate species (Figure 2; Nelson et al., 2013), which sometimes use different signalling molecules (Baker, 2011). The CYPome in invertebrates contain novel families not found in vertebrates. For example, clan 3 includes the CYP6 family from *D. melanogaster* and the mitochondrial clan includes CYP12, 301, 302, and 315 (Tijet et al., 2001); families that are not found in vertebrate species. Enzymes from the CYP12 family are capable of xenobiotic metabolism; a function which is unusual for a mitochondrial CYP (Guzov et al., 1998). Because of the incredible diversity of invertebrates there are many more families expected in invertebrates than vertebrates; vertebrates have 19 (Nelson, 2011) of the 196 CYP families identified across animals (Nelson et al., 2013). CYP family saturation has yet to happen in invertebrates and each new invertebrate CYPome has added novel families to the CYP superfamily of proteins (Baldwin et al., 2009; Nelson, 2009; Tijet et al., 2001).

General Sequence Analysis and Protein Behaviour

CYPs average 500 amino acids in length and contain 1 to 13 exons (Nelson, 2009). There is an active site with a heme group, which is responsible for binding oxygen for the monooxygenase reaction (Nebert and Gonzalez, 1987). CYPs are membrane bound proteins, typically bound to the endoplasimic reticulum or in the mitochondria (Williams, et al., 2000). When bound to carbon monoxide, CYPs have a spectrometry peak at 450nm, giving rise to the 'P450' portion of the name (Garfinkel, 1958).

Cytochrome P450 is a very diverse superfamily, with some P450s having less than 10% identity. Yet, there have been many CYPs with experimentally derived 3D protein structures through X-ray crystallography, and those CYPs are structurally similar after folding despite the sequence divergence (Williams et al., 2000). The most divergent portions of CYPs are in the C- and N- terminal regions. The N-terminal region is important for determining the localization of the CYP (Pernecky, et al., 1993). Neither terminal end are considered important in the catalytic activity of the protein so the termini sequences are more flexible than the core (Williams et al., 2000).

The heme binding region starts at approximately amino acid 430 and has a well conserved motif of PFXXGXXRXCXG (the 'X' represents a non conserved amino acid); the cysteine, until recently, was considered the only absolutely conserved amino acid in all known cytochrome P450s; Sezutsu et al., 2013). *Ciona intestinalis* CYP20 has a histidine substituted for cysteine, suggesting that this amino acid is not absolutely required (Hideki et al., 2013). There are three other well conserved motifs: portions of the I-helix, [A/G]GX[E/D]T[T/S], located around amino acid 300; K-helix, EXXR, located around amino acid 360; and an area known as the 'meander coil', FDPER, located around amino acid 410. The K-helix motif is incredibly well conserved in CYPs, with only a handful of known exceptions to the two conserved amino acids (Rupasinghe et al., 2006). The K-helix motif is thought to have a stabilising function of the core structure (Werck-Reichhart and Feyereisen, 2000). The I-helix motif makes up the proton transfer groove distal to the heme (Werck-Reichhart and Feyereisen, 2000). These motifs are important when analysing potential CYPs, if one or more of these regions are missing, or out of

place, it is likely that the gene was constructed incorrectly, a pseudogene or not a CYP at all.

Steroidogenesis of Vertebrate-like Signalling Steroids

Vertebrate steroidogenesis is well understood; the specific genes and proteins and the substrates and intermediates involved have been identified (Figure 2). The first step in the steroid pathway is the long-chain cleavage of cholesterol to pregnenolone via CYP11A (Baker, 2011). CYPs and the hydroxysteroid dehydrogenases (HSDs) are the primary enzymes responsible for vertebrate steroidogenesis. The production of estradiol (18 carbon) from lanosterol (30 carbon) is a six to eight enzymatic step process (Figure 2) and involves CYPs from families 11, 17, 19, 21 (Baker, 2011).

All animals create various types of steroids, though the steroids that are produced vary across phyla. For example, CYP21A has 21-hydroxylase activity, catalyzing the production of cortisol and aldosterone. CYP21A was not found in amphioxus (Nelson, 2009), and is thought to be exclusive to vertebrates. Ecdysteroids are major steroids in arthropods, having roles in molting and reproduction (Lafont and Mathieu, 2007). The presence and production of ecdysteroids in other protostomes is still inconclusive.

The sex steroids are one of the end products of vertebrate steroidogensis. CYP19 has the aromatase function, which is responsible for estrogen production from androgen precursors (Nebert et al., 1991). Vertebrates have two estrogen receptors (ERs) and one androgen receptor (AR; Baker, 2011). The effects of estrogens and androgens are mediated through their respective receptors. The ER is postulated to be the ancestral

steroid receptor: ancestral-ER (AncSR1) was deduced using available sequences, expressed *in vitro* and had function like vertebrate ERs (Thornton, et al., 2003). AncSR1 bound E2 and increased transcription through an estrogen response element (Thornton et al., 2003).

The ER in the mollusc *Aplysia californica* was found to be insensitive to estrogens but was constitutively expressed (Thornton et al., 2003). *Capitella teleta* and *Platynereis dumerilii*, found in the sister phylum to molluscs, have ERs that respond to exogenous estrogen, the first species with this function identified outside the vertebrates (Keay and Thornton, 2009). Thus, estrogen activated ERs appear to be lost in molluscs (Keay and Thornton, 2009). Endogenous estrogen production has been identified in the annelid *Nereis virens*, likely having a role during early stages of oogenesis through activating expression of vitellogenin, a protein important as a source of nutrients during development (Garciaa-Alonso and Rebscher, 2005). Steroid synthesis in *N. virens* is thought to occur in the gut epithelium (Garciaa-Alonso and Rebscher, 2005) but the enzymes involved have not been identified.

CYP19 has recently been identified in the amphioxus genome and through *in silico* methods, it was postulated to have aromatase function (Callard et al., 2011). CYP19 is interesting because it has no known closely related CYPs and is in a clan all on its own (Nelson et al., 2013). It would be expected that there would be a *CYP19-like* gene in hemi-chordates or invertebrates, but no sequence with orthology to vertebrate CYP19 has yet to be found in any of the sequenced genomes. There are two hypotheses that have

been proposed to explain the lack of a *CYP19* ortholog in current invertebrate genomes: (1) That *CYP19* or a *CYP19-like* gene has been lost in all non-chordate animals that have been analysed so far and is eluding discovery, or (2) there was such a high evolutionary drive for the creation of the aromatase gene in chordates that it evolved incredibly quickly (Callard et al., 2011; Nelson et al., 2013). In the future more cephalochordate genomes will help shed light on this issue. With estrogen activated ERs in annelids documented and quantifiable amounts of endogenous estrogen, a gene with aromatase activity must be present in annelids. If there is no *CYP19* gene found in *C. teleta,* aromatase function could be mediated by another CYP with convergent functional evolution to CYP19 or be provided by another protein family all together.

Capitella teleta

General Species Overview

Capitella teleta is a marine polychaete annelid (a lophotrochozoan) found all along the shores of North America, Japan and the Mediterranean. The incredibly opportunistic nature of this species is a probable explanation for its wide distribution (Blake, et al, 2009). High population density of *C. teleta* is an indication of disturbed marine environments and the species is considered an important bio-indicator by the U.S Environmental Protection agency. *C. teleta* was found to be the most opportunistic invertebrate species after the Massachusetts oil spill in 1969 (Sanders et al., 1980) and is considered to be the most opportunistic of the *Capitella* genus (Grassle, 1980) with the fastest population increase and largest maximum population in disturbed habitats.

C. teleta has been long known to significantly degrade PAHs in sediment, and is suggested that CYPs are responsible (Gardner, et al., 1979; Lee, 1998). Benzo[α]pyrene (BaP) has been shown to be metabolized by CYPs in another annelid *Nereis virens* (Lee, 1998). There have been mixed results in the literature whether there is an increase in CYP concentration due to PAH exposure in *N. virens* (primarily tested with BaP), with some studies reporting 2-fold increase in CYPs and others reporting no change (reviewed by Lee, 1998). But in *C. teleta* two identified CYPs (CYP4AT1 and CYP331A1) had increased expression with exposure to some PAHs (Li, et al., 2004). CYP331A1 and had 1.9-2.6 fold increase gene expression after exposure from BaP (Li et al., 2004). CYP4AT1 had 1.25 to 1.9 fold increase in gene expression after exposure to PAH contaminated sediments (Li et al., 2004). CYP331A1 is located in clan 3 with the CYP3s, which are traditionally known for their importance in exogenous metabolism (McArthur et al., 2003).

Beyond the interest of exogenous metabolism, *C. teleta* is particularly interesting in the search for steroidogenesis genes because this species has an ER that binds estrogen and activates transcription, the first time this has been observed outside vertebrates (Keay and Thornton, 2009). It has been postulated that there is a protein with aromatase function in this species.

Capitella **Taxonomy**

Capitella telata has recently undergone taxonomic review as it has been misidentified as *Capitella capitata* in many scientific publications. Two sister papers

designated *C. teleta* (Blake, 2009) and *C. capitata* (Blake, 2009) as different species. Many previous studies completed on these animals were using older species designations. For example, the studies of Keay and Thornton (2009) and Li and colleagues (Li et al., 2004), were completed on *C. teleta* although the species was referred as *C. capitata* in the publications. It is important to check the taxonomy of these species in all studies done prior to 2010 to determine which species the study is referring to. If the study refers to the sequenced genome of the species on JGI, or call it *Capitella sp. I* they are referring to *C. teleta*. Blake (2009) has identified over 200 papers that focused on *C. teleta* but with incorrect species identification.

The *Capitella* Genome

The *C. teleta* genome was released in its current state as version 1 in 2007. The genome is 333.7mb large and is still in the scaffold stage of assembly with 21,042 scaffolds (JGI, 2007). Approximately half the genome is on 454 scaffolds each larger than 188Kb (JGI, 2007). JGI's automated pipeline predicted 32,415 gene models. Of the gene models predicted on JGI, 114 are labeled as CYPs with varying confidence. There is a small EST database of 138,404 reads with a median length of 742bp. The EST database has a slight 5' bias.

Research Goals

The main objective of this research has been to identify and annotate the complete CYP complement (CYPome) of *C. teleta.* Through evolutionary and sequence analyses, I raise functional hypotheses for some of the proposed proteins. Though there is no

functional data on any CYPs in *C. teleta,* studies demonstrating PAH biotransformation in *C. telata* (Linke-Gamenick, et al., 2000), an increased abundance in sites with high PAH exposure and an increase in expression of two CYPs during PAH exposure (Bach, et al., 2005; Li et al., 2004; Ramskov and Forbes, 2008) makes the *C. teleta* CYPome an important protein family to annotate from a toxicological perspective. Furthermore, the *C. teleta* CYPome is important for an exploration of steroidogenic CYPs because they produce estradiol and are known to be responsive to estrogens. Chapter 2 contains the *C. teleta* CYPome with phylogenetic analyses, motif comparison, and functional predictions of the identified CYPs in a scientific manuscript format. Chapter 2 provides names of the genes in the CYPome, as approved by the nomenclature committee, a phylogenetic analysis of the entire CYPome with vertebrate CYPs, and more detailed phylogenies of clans 2, 3 and 4, and mitochondrial. *In silico* protein analyses were completed for two CYPs (CYP51A1 and CYP376A1) to examine if they have a potential role in early steroidogensis metabolism stages and these results are discussed in Chapter 3.

References

Bach, L., Palmqvist, A., Rasmussen, L. J., and Forbes, V. E. (2005). Differences in PAH tolerance between capitella species: Underlying biochemical mechanisms. *Aquatic Toxicology, 74*(4), 307-319.

Baker, M. E. (2011). Origin and diversification of steroids: Co-evolution of enzymes and nuclear receptors. *Molecular and Cellular Endocrinology, 334*(1–2), 14-20. doi: 10.1016/j.mce.2010.07.013

Baldwin, W., Marko, P., and Nelson, D. R. (2009). The cytochrome P450 (CYP) gene superfamily in daphnia pulex. *BMC Genomics, 10*(1), 169. doi: 10.1186/1471-2164-10-169

Blake, J. A. (2009). Redescription of capitella capitata (fabricius) from west greenland and designation of a neotype (polychaeta, capitellidae). *Zoosymposia, 2*, 55-80.

Blake, J. A., Grassle, J. P., and Eckelbarger, K. J. (2009). Capitella teleta, a new species designation for the opportunistic and experimental capitella sp. I, with a review of the literature for confirmed records.

Callard, G. V., Tarrant, A. M., Novillo, A., Yacci, P., Ciaccia, L., Vajda, S., . . . Cotter, K. A. (2011). Evolutionary origins of the estrogen signaling system: Insights from

amphioxus. *The Journal of Steroid Biochemistry and Molecular Biology*, *127*(3–5), 176-188. doi: 10.1016/j.jsbmb.2011.03.022

Garciaa-Alonso, J., and Rebscher, N. (2005). Estradiol signalling in nereis virens reproduction. *Invertebrate Reproduction and Development*, *48*(1-3), 95-100. doi: 10.1080/07924259.2005.9652175

Gardner, W. S., Lee, R. F., Tenore, K. R., and Smith, L. W. (1979). Degradation of selected polycyclic aromatic hydrocarbons in coastal sediments: Importance of microbes and polychaete worms. *Water, Air, and Soil Pollution*, *11*(3), 339-347.

Garfinkel, D. (1958). Studies on pig liver microsomes. I. enzymic and pigment composition of different microsomal fractions. *Archives of Biochemistry and Biophysics*, *77*(2), 493. doi: http://dx.doi.org/10.1016/0003-9861(58)90095-X"

Goldstone, J. V. (2008). Environmental sensing and response genes in cnidaria: The chemical defensome in the sea anemone nematostella vectensis. *Cell Biology and Toxicology*, *24*(6), 483-502.

Goldstone, J. V., McArthur, A., Kubota, A., Zanette, J., Parente, T., Jonsson, M., . . . Stegeman, J. (2010). Identification and developmental expression of the full complement of cytochrome P450 genes in zebrafish. *BMC Genomics*, *11*(1), 643. doi: 10.1186/1471-2164-11-643

Goldstone, J. V., Hamdoun, A., Cole, B., Howard-Ashby, M., Nebert, D., Scally, M., . . . Stegeman, J. (2006). The chemical defensome: Environmental sensing and response genes in the strongylocentrotus purpuratus genome. *Dev Biol, 300,* 366-384. doi: 10.1016/j.ydbio.2006.08.066

Grassle, J. (1980). Polychaete sibling species. *Aquatic oligochaete biology* (pp. 25-32) Springer.

Keay, J., and Thornton, J. W. (2009). Hormone-activated estrogen receptors in annelid invertebrates: Implications for evolution and endocrine disruption. *Endocrinology, 150*(4), 1731-1738. doi: 10.1210/en.2008-1338

Lafont, R., and Mathieu, M. (2007). Steroids in aquatic invertebrates. *Ecotoxicology, 16*(1), 109-130. doi: 10.1007/s10646-006-0113-1

Lee, R. F. (1998). Annelid cytochrome P-450. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology, 121*(1), 173-179.

Lewis, D. F. V. (2004). 57 varieties: The human cytochromes P450. *Pharmacogenomics, 5*(3), 305-318. doi: 10.1517/phgs.5.3.305.29827

Li, B., Bisgaard, H. C., and Forbes, V. E. (2004). Identification and expression of two novel cytochrome P450 genes, belonging to CYP4 and a new CYP331 family, in the polychaete capitella capitata sp.I. *Biochemical and Biophysical Research Communications, 325*(2), 510. doi: http://dx.doi.org/10.1016/j.bbrc.2004.10.066"

Linke-Gamenick, I., Forbes, V. E., and M'endez, N. (2000). Effects of chronic fluoranthene exposure on sibling species of capitella with different development modes.

McArthur, A., Hegelund, T., Cox, R., Stegeman, J., Liljenberg, M., Olsson, U., . . . Celander, M. (2003). *Phylogenetic analysis of the cytochrome P450 3 (CYP3) gene family* Springer New York. doi: 10.1007/s00239-003-2466-x

Nebert, D. W., and Gonzalez, F. J. (1987). P450 genes: Structure, evolution, and regulation. *Annual Review of Biochemistry, 56,* 945-993. doi: 10.1146/annurev.bi.56.070187.004501

Nebert, D. W., Nelson, D. R., Coon, M. J., Estabrook, R. W., Feyereisen, R., Fujii-Kuriyama, Y., . . . Johnson, E. F. (1991). The P450 superfamily: Update on new sequences, gene mapping, and recommended nomenclature. *DNA and Cell Biology, 10*(1), 1-14.

Nebert, D. W., and Russell, D. W. (2002). Clinical importance of the cytochromes P450. *The Lancet, 360*(9340), 1155-1162. doi: 10.1016/S0140-6736(02)11203-7

Nebert, D. W., Wikvall, K., and Miller, W. L. (2013). Human cytochromes P450 in health and disease. *Philosophical Transactions of the Royal Society B: Biological Sciences}, 368*(1612) doi: 10.1098/rstb.2012.0431

Nelson, D. R. (2009). The cytochrome P450 homepage. *Hum Genomics, 4*(1), 59-65.

Nelson, D. R. (2011). Progress in tracing the evolutionary paths of cytochrome P450. *Biochimica Et Biophysica Acta.Proteins and Proteomics, 1814*(1), 14-18.

Nelson, D. R. (1999). Cytochrome \P450\ and the individuality of species. *Archives of Biochemistry and Biophysics, 369*(1), 1. doi: http://dx.doi.org/10.1006/abbi.1999.1352"

Nelson, D. R. (2003). Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. *Archives of Biochemistry and Biophysics, 409*(1), 18-24. doi: 10.1016/S0003-9861(02)00553-2

Nelson, D. R., Goldstone, J. V., and Stegeman, J. J. (2013). The cytochrome P450 genesis locus: The origin and evolution of animal cytochrome P450s. *Philosophical Transactions of the Royal Society B: Biological Sciences}, 368*(1612) doi: 10.1098/rstb.2012.0474

Nelson, D. R., Koymans, L., Kamataki, T., Stegeman, J. J., Feyereisen, R., Waxman, D. J., . . . Nebert, D. W. (1996). P450 superfamily: Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics and Genomics, 6*(1) Retrieved from http://journals.lww.com/jpharmacogenetics/Fulltext/1996/02000/P450_superfamily__update_on_new_sequences,_gene.2.aspx

Pernecky, S. J., Larson, J. R., Philpot, R. M., and Coon, M. J. (1993). Expression of truncated forms of liver microsomal P450 cytochromes 2B4 and 2E1 in escherichia

coli: Influence of NH2-terminal region on localization in cytosol and membranes. *Proceedings of the National Academy of Sciences, 90*(7), 2651-2655. Retrieved from http://www.pnas.org/content/90/7/2651.abstract

Ramskov, T., and Forbes, V. E. (2008). Life history and population dynamics of the opportunistic polychaete capitella sp. I in relation to sediment organic matter. *Marine Ecology Progress Series, 369*, 181. doi: doi: 10.3354/meps07584

Rupasinghe, S., Schuler, M. A., Kagawa, N., Yuan, H., Lei, L., Zhao, B., . . . Lamb, D. C. (2006). The cytochrome P450 gene family CYP157 does not contain EXXR in the K-helix reducing the absolute conserved P450 residues to a single cysteine. *FEBS Letters, 580*(27), 6338. doi: http://dx.doi.org/10.1016/j.febslet.2006.10.043"

Sanders, H. L., Grassle, J. F., Hampson, G. R., Morse, L. S., Garner-Price, S., and Jones, C. C. (1980). Anatomy of an oil spill: Long-term effects from the grounding of the barge florida off west falmouth, massachusetts.

Sezutsu, H., Le Goff, G., and Feyereisen, R. (2013). Origins of P450 diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences, 368*(1612)

Thornton, J. W., Need, E., and Crews, D. (2003). Resurrecting the ancestral steroid receptor: Ancient origin of estrogen signaling. *Science}, 301*(5640), 1714-1717. doi: 10.1126/science.1086185

Tijet, N., Helvig, C., and Feyereisen, R. (2001). The cytochrome \P450\ gene superfamily in drosophila melanogaster: Annotation, intron-exon organization and phylogeny. *Gene, 262*(12), 189. doi: http://dx.doi.org/10.1016/S0378-1119(00)00533-3"

Werck-Reichhart, D., and Feyereisen, R. (2000). Cytochromes P450: A success story. *Genome Biology, 1*(6), reews3003.1-reews3003.9. doi: 10.1186/gb-2000-1-6-reviews3003

Williams, P. A., Cosme, J., Sridhar, V., Johnson, E. F., and McRee, D. E. (2000). Mammalian microsomal cytochrome P450 monooxygenase: Structural adaptations for membrane binding and functional diversity. *Molecular Cell, 5*(1), 121. doi: http://dx.doi.org/10.1016/S1097-2765(00)80408-6"

Yoshida, Y., Aoyama, Y., Noshiro, M., and Gotoh, O. (2000). Sterol 14-demethylase P450 (CYP51) provides a breakthrough for the discussion on the evolution of cytochrome P450 gene superfamily. *Biochemical and Biophysical Research Communications, 273*(3), 799-804.

Figures

Figure 1.1. CYP clan distribution across metazoa and fungi. Black circles indicate presence of the clan and white circle indicate absence of the clan in a given taxonomic group. The asterisk indicates a likely gene transfer from metazoan. Figure from Nelson et al, 2013(Nelson et al., 2013).

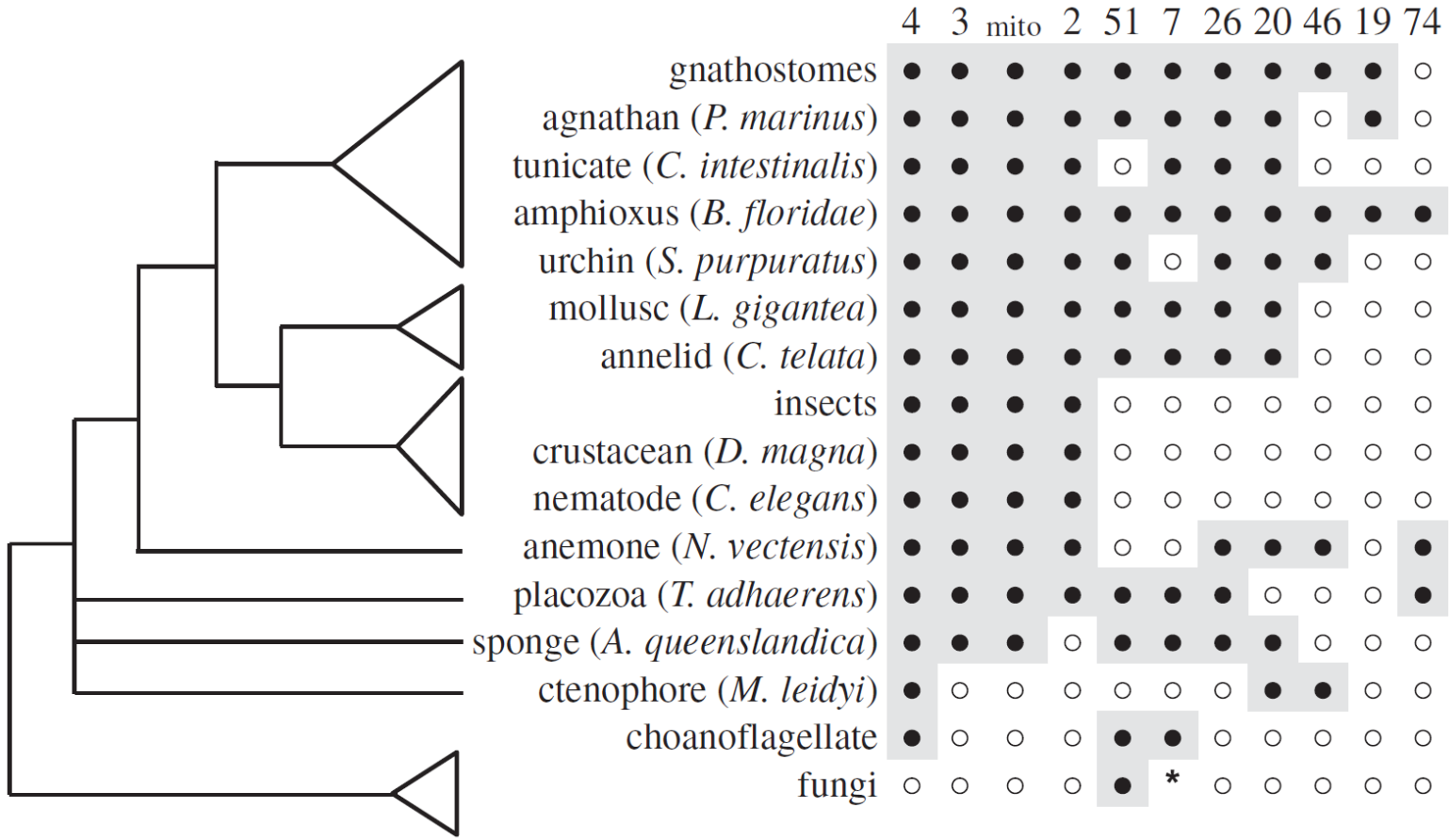
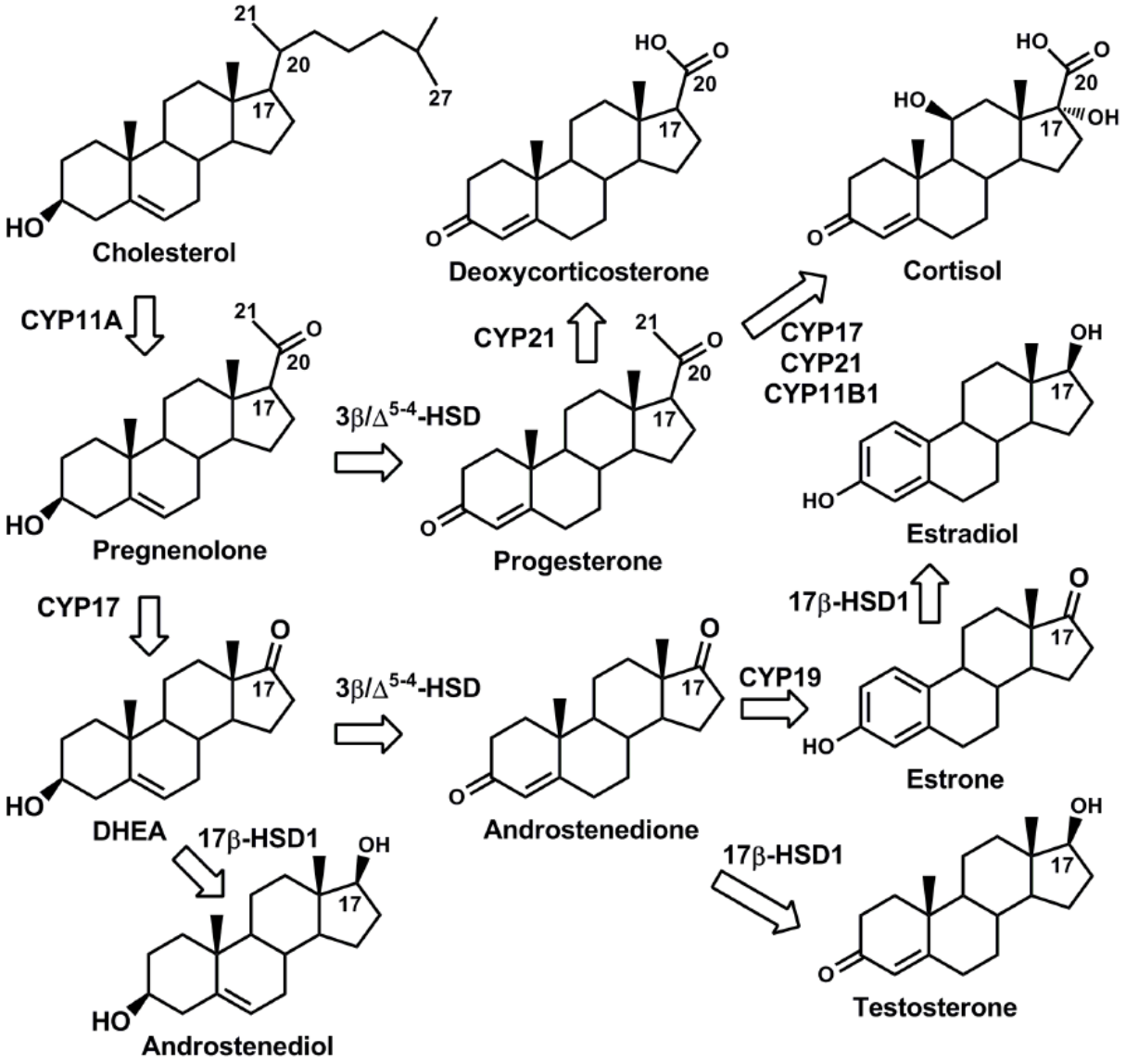

4 3 mito 2 51 7 26 20 46 19 74
gnathostomes ● ● ● ● ● ● ● ● ● ● ○
agnathan (P. marinus) ● ● ● ● ● ● ● ● ○ ● ○
tunicate (C. intestinalis) ● ● ● ● ○ ● ● ● ○ ○ ○
amphioxus (B. floridae) ● ● ● ● ● ● ● ● ● ● ●
urchin (S. purpuratus) ● ● ● ● ● ○ ● ● ● ○ ○
mollusc (L. gigantea) ● ● ● ● ● ● ● ● ○ ○ ○
annelid (C. telata) ● ● ● ● ● ● ● ● ○ ○ ○
insects ● ● ● ● ○ ○ ○ ○ ○ ○ ○
crustacean (D. magna) ● ● ● ● ○ ○ ○ ○ ○ ○ ○
nematode (C. elegans) ● ● ● ● ○ ○ ○ ○ ○ ○ ○
anemone (N. vectensis) ● ● ● ● ○ ○ ● ● ● ○ ●
placozoa (T. adhaerens) ● ● ● ● ● ● ● ○ ○ ○ ●
sponge (A. queenslandica) ● ● ● ○ ● ● ● ● ○ ○ ○
ctenophore (M. leidyi) ● ○ ○ ○ ○ ○ ○ ● ● ○ ○
choanoflagellate ● ○ ○ ○ ● ● ○ ○ ○ ○ ○
fungi ○ ○ ○ ○ ● * ○ ○ ○ ○ ○

Figure 1.2. The vertebrate steroidogenesis pathway. The first step of the pathway is side chain
cleavage of cholesterol by CYP11A. Hydroxysteroid dehydrogenases (HSDs) and CYPs catalyze
the steps in the pathway to create several major steroids used as signalling molecules including cortisol, estradiol, testosterone, aldosterone, and progesterone. Figure from Baker 2011 (Baker, 2011).


Cholesterol
Deoxycorticosterone
Cortisol
CYP11A
CYP21
CYP17
CYP21
CYP11B1
Pregnenolone
3β/Δ5-4-HSD
Progesterone
Estradiol
CYP17
17β-HSD1
3β/Δ5-4-HSD
CYP19
DHEA
Androstenedione
Estrone
17β-HSD1
17β-HSD1
Androstenediol
Testosterone

Chapter 2: The Cytochrome P450 Superfamily Complement (CYPome) in the annelid Capitella teleta.

Chris A. Dejong and Joanna Y. Wilson

Department of Biology, McMaster University

Author Contributions:

CAD: Planned the experiment, data analysis, writing the manuscript

JYW: Planned the experiment, data analysis, writing the manuscript, funded the research

Abstract

The Cytochrome P450 super family (CYP) is responsible for a wide range of functions in metazoans, having roles in both exogenous and endogenous metabolism. In this study the CYP complement (CYPome) of the annelid *Capitella teleta* has been robustly identified and annotated with the genome assembly available. Annotation of 84 full length and 12 partial CYP sequences predicted a total of 96 functional CYPs in *C. teleta.* A further 13 CYP fragments were found but these may be pseudogenes. The *C. teleta* CYPome contained 24 novel CYP families and seven novel CYP subfamilies within existing families. A phylogenetic analysis was completed with primarily vertebrate sequences and identified that the *C teleta* sequences were found in 9 of the 11 metazoan CYP clans. Clan 2 was expanded in this species with 51 CYPs in 14 novel CYP families containing 20 subfamilies. There were five clan 3, four clan 4, and six mitochondrial clan full length CYPs. Two CYPs, CYP3071A1 and CYP3072A1, did not cluster with any metazoan CYP clans. We found xenobiotic response elements upstream of several *C. teleta* CYPs: in genes related to vertebrate CYP1 (CYP3060A1, CYP3061A1) and from families with reported transcriptional upregulation in response to PAH exposure (CYP4, CYP331). This may indicate a functional aryl hydrocarbon receptor in *C. teleta* and candidate CYPs for studies of PAH metabolism. *C. teleta* had a CYP51A1 with ~65% identity to vertebrate CYP51A1 sequences and has been predicted to have similar lanosterol 14 α-demethylase activity. Several CYPs (CYP376A1, CYP3068A1, CYP3069A1, and CYP3070A1) are discussed as candidate genes for steroidogenesis.

Introduction

The Cytochrome P450 (CYP) superfamily of protein enzymes are found in all domains of life (Nelson, 2011; Nelson, 1999). CYPs catalyze a monooxygenase reaction (Nebert and Gonzalez, 1987) of compounds that fall into two general categories: exogenous (i.e. xenobiotics) and endogenous (e.g. steroids and lipids) substrates. CYPs are involved in both the synthesis and catabolism of important biological signaling molecules. CYPs involved in metabolism of endogenous substrates typically act on a small number of very similar, structurally related molecules. CYPs responsible for metabolism of xenobiotics generally have more flexible active sites to allow them to act on a wider array of substrates.

All newly identified CYPs are named by the Cytochrome P450 nomenclature committee, using standard conventions for this gene superfamily. CYPs are named by amino acid sequence identity; genes with 40% and 55% identify are placed in the same family and subfamily, respectively (Nelson et al., 1996). CYPs are named by family and subfamily using a numeral and letter, respectively. The specific gene is given a number, by order of discovery (Nelson et al., 1996). For example CYP19A1 is in family 19, subfamily A and has a gene number of 1.

Since the early 2000's there have been several studies focused on the CYP genome complements (CYPomes) in metazoans, with studies completed on vertebrates (Goldstone et al., 2010; Lewis, 2004; Nelson, 2003), hemichordates (Goldstone et al., 2006), insects (Tijet, et al., 2001), crustaceans (Baldwin, et al., 2009), and Cnidaria (Goldstone, 2008).

Many more CYPomes have been partially completed and unpublished CYPomes have been made available on the Cytochrome P450 webpage (Nelson, 2009). The smallest number of genes in a metazoan CYPome was found in the sponge *Amphimedon queenslandica* (35 CYP genes) and the largest metazoan CYPome identified so far included ~235 genes in the lancelet, *Branchiostoma floridae* (Nelson, et al., 2013). Vertebrate genomes typically contain 57-102 CYP genes (Nelson, 2011).

Vertebrate steroidogenesis is well understood; the specific genes and proteins and the substrates and intermediates involved have been identified. CYPs and the hydroxysteroid dehydrogenases (HSDs) are the primary enzymes responsible for vertebrate steroidogenesis. The first step in the steroid pathway is the long-chain cleavage of cholesterol to pregnenolone via CYP11A (Baker, 2011b). The production of estradiol (18 carbon) from lanosterol (30 carbon) is a six to eight enzymatic step process and involves CYPs from families 11, 17, 19, 21 (Baker, 2011b). The sex steroids are one of the end products of steroidogensis. CYP19A has the aromatase function, which is responsible for estrogen production from androgen precursors. The CYP19A gene has only been found in chordates, though is predicted to have more ancestral origins (Callard et al., 2011).

Capitella teleta is a polychaete annelid found in marine environments along the Pacific and Atlantic shores around the continental United States, Japan and the Mediterranean (Blake, et al., 2009). There has been an interest in determining the identify and function of CYPs in *C. teleta,* primarily focused on deciphering their ability to metabolize xenobiotics and polycyclic aromatic hydrocarbons (PAHs), in particular ((Li,

et al., 2004; Selck, et al., 2003). This stems from research that found *C. teleta* to be the most opportunistic invertebrate after a 1969 oil spill in Massachusetts (Sanders et al., 1980) and a concentration dependent increase in CYP-dependent activity with exposure to PAHs (Lee and Singer, 1980) in *Capitella* spp,. More recently, differences in tolerance to PAHs and capacity for PAH metabolism amongst *Capitella* species have been investigated (Bach,et al., 2005; Linke-Gamenick, et al., 2000). Two CYPs in *C. teleta,* CYP331A1 (a novel family) and CYP4AT1, have been identified and their expression was increased in response to various PAHs, suggesting a possible role of these CYPs in PAH metabolism (Li et al., 2004).

Invertebrate endocrine systems have been much less studied that their vertebrate counterparts. Yet, data show that multiple endocrine active agents, sometimes including steroids typical in vertebrates, are present in invertebrate lineages (Janer and Porte, 2007). Annelids are one group of invertebrates thought to produce and utilize the vertebrate sex steroid estradiol. *C. teleta* and *Platynereis dumerilii*, another marine annelid, had estrogen receptors (ERs) that responded to exogenous estrogen and regulated downstream gene expression, the first species with this function identified outside the vertebrates (Keay and Thornton, 2009). The annelid *Nereis virens* had detectable aromatase activity, likely occurring in the gut epithelium (Garciaa-Alonso and Rebscher, 2005). Despite having detectable aromatase activity, the protein responsible for this function remains unknown. CYPome studies in annelid species may provide clues to the evolution of the steroidogenesis pathway in metazoa and whether annelid invertebrates utilize the same enzymes for *de novo* sex steroid production.

The objective of this study was to annotate the *C. teleta* CYPome. The *C. teleta* CYPome is the first detailed analysis of a lophotrochozoan CYPome, providing important information on CYP content and evolution in an understudied metazoan superphyla. This study examines the potential role of the various CYPs in exogenous metabolism, particularly PAHs, and hypothesizes which CYPs may have a role in *C. teleta* steroidogenesis.

Methods

The *C. teleta* genome used for this study was version 1 of the assembly (Joint Genome Institute, University of California; JGI); the genome assembly had approximately 7.9x coverage with 21,042 scaffolds with a total size of 333.7 Mb. The EST database (National Center for Biotechnology Information, July 2012) had approximately 130,000 reads. The other sequences used in phylogenetic analyses were retrieved from the Cytochrome P450 web-page (Nelson, 2009). Many vertebrate sequences were used in the analyses, although there was a focus on *Danio rerio, Mus musculus,* and *Homo sapiens*; species which have had rigorous annotation of their CYPome (Goldstone et al., 2010; Nebert, et al., 2013; Nelson, 2009). A select number of CYP sequences from invertebrates were included: *Haliotis diversicolor* (Nelson, 2009), *Crassostrea gigas* (NCBI), *Daphnia pulex* (Baldwin et al., 2009) and *Helobdella robusta* (JGI). For some phylogenies, *Caenorhabditis spp.* sequences were added (Nelson, 2009).

Gene annotation

The *Capitella teleta* EST database was assembled with PASA (r2012-06-25, (Haas et al., 2003), to align and extend the ESTs to each other and to align them to the *C. teleta* genome. Homology searching of ESTs was performed using all CYPs from human and zebrafish using tBLASTn (v2.2.27; Altschul, et al., 1990). Hits were compiled and the regions hit were autonomously counted via a custom Perl script. This approach allowed for many CYPs to be used as inputs for homology searches. Since CYPs can have <15% sequence identity from each other in their amino acid sequences, the use of a wide variety of CYPs during homology searching maximizes the number of unique hits. The hit regions were checked against the PASA outputs, overlaps collected, and approximate gene regions predicted. The putative gene regions were compared to previously annotated CYPs and gene boundaries were adjusted using Artemis (v14.0.0, Rutherford et al., 2000) according to homology. Since there are no closely related species with their CYPome analyzed or a large EST/refseq database, exact exon boundaries were difficult to annotate with very high certainty. The exon boundaries were examined for appropriate splice signals (Mount, 1982) and to ensure that the boundaries were located appropriately to the reading frame. When there were large gaps in a gene, FASTA (36.3.5e; Pearson and Lipman, 1988) searches against genome scaffolds were completed to find these missing regions, rather than BLAST, because FASTA has increased sensitivity.

Once annotated, CYPs were compared to the automated gene calls in September 2012 on JGI. There were no CYPs on JGI that were not found by the above method. The

manual annotation made for more appropriate splice sites, with the JGI annotations at times leaving out segments or entire exons.

Phylogenetic Analyses of CYP sequences

Alignments were created in MUSCLE (v3.8.31; Edgar, 2004) and manually refined in Mesquite (v2.75; Maddison and Maddison D., 2011) at the amino acid level. The N- and C-termini of CYPs are more divergent and were hard masked from further analysis. ZORRO (r2011-12-01; Wu, et al., 2012), a soft masking tool, was used on the remainder of the alignment. The phylogenetic analysis was conducted using a total of 220 sequences on RAxML (v7.4.2; Stamatakis, 2006) with 100 bootstraps using the slower algorithm (-b) with a gamma distribution. The clan 2 phylogeny had additional *C. elegans* sequences, the clan 3 and 4 phylogeny included from *Daphnia pulex, Daphnia magna* and *C. elegans* sequences, and the mitochondrial phylogeny incorporated sequences from *D. pulex, Caenorhabditis spp., and H. robusta*. The maximum likelihood analyses were based on the VT substitution model with fixed base frequencies (phylogeny of all clans), MTMAM substitution model with fixed base frequencies (clan 2 phylogeny), LG substitution model with empirical base frequencies (clan 3 and 4 phylogeny), or JTT substitution model with empirical base frequencies (mitochondrial clan phylogeny). The appropriate models were determined by ProtTest (v3.2; Abascal, et al., 2005). To root the phylogenetic trees, CYPs outside the clans were chosen; the clan 2 phylogeny used CYP family 504 genes from fungus (*Magnaporthe grisea* and *Nassarius fischeri*), the clan 3 and 4 phylogeny used CYP72s from *Arabidopsis thaliana*, and the mitochondrial clan phylogeny used CYP86s from *A. thaliana* and vertebrate CYP19s. These roots were

selected based on closely related out-groups from David Nelson's "singlefam.tree" on the Cytochrome P450 webpage (Nelson, 2009).

All predicted *C. teleta* CYP genes were named by the cytochrome P450 nomenclature committee using the sequences provided, synteny data available and the phylogenetic trees generated in this study. Strap (r2013-02-26; Gille and Frommel, 2001) was used for the motif work and Figtree (v1.4.0) was used to generate the figures of the phylogenetic trees.

Searches for the xenobiotic response element (XRE, TNGCGTG; Sun, et al., 2004) in the 10kb and 20kb upstream region of the predicted start site in each gene of families *CYP331, CYP4* and *CYP3061* used the MEME suite (v4.9.1; Bailey and Elkan, 1994).

Results

Eighty-four full length CYPs were identified and annotated from the *C. teleta* assembly (v1); the entire list of CYPs, their genomic location, size, and nomenclature are provided in Supplementary Table 1. There were twelve partial CYP sequences identified that aligned well with existing ESTs but could not be completed based on the current assembly (Supplementary Table 2). There were thirteen partial CYP sequences identified that lacked any EST support (Supplementary Table 3); whether these were genes or pseudogenes remains unclear. Based on the names assigned by the cytochrome P450 nomenclature committee, the predicted *C. teleta* CYPs were found in 9 of the 11 known

metazoan CYP clans (Nelson et al., 2013) and predicted 24 novel CYP families and 7 novel CYP subfamilies.

All of the full length CYPs contained at least some signature CYP motifs (Table 1). The I-helix motif ([A/G]GX[D/E]T[T/S]; Werck-Reichhart and Feyereisen, 2000) had conservation of at least three of the six amino acids in all but thirteen *C. telata* CYPs (CYP3065A1-4, CYP3065B1, CYP3066A1-3, CYP3066C1, CYP3067A1, CYP372B1, and CYP39B1). The remaining CYPs had obvious sequence homology, with a majority of the conservation at the ends in the I-helix motif, even though this is the most poorly conserved motif of the four examined. The K-helix motif was fully conserved across all of the *C. teleta* sequences with no exceptions to the E-X-X-R consensus sequence(Werck-Reichhart and Feyereisen, 2000) (Table 1). The meander coil was conserved across all of the annotated sequences, although CYP372B1 and CYP4ED1 has substitutions for the first two amino acids in the motif (Table 1). Lastly, the cysteine residue in the heme binding loop is highly conserved, with very few exceptions (Sezutsu, et al., 2013), and this residue was present in all of the *C. teleta* sequences (Table 1). There was clear homology in the heme loop motif across all of the *C. teleta* sequences except for a gap in the motif in CYP3067A1. Interestingly, CYP3067A1 had a gap in both the heme loop and I helix motifs (Table1).

The phylogenetic relationships among the genes of the *C. teleta* CYPome is shown in Figure 1 and the distribution of these genes in the major clans (clans 2, 3, 4, and mitochondrial) shown in Figure 2. A majority of the *C. telata* CYPs were in clan 2, accounting for ~60% of the CYPome (Figure 2); Six of these genes were the most basal

sequences within this clan (Figure 1). 33 genes were clustered as a distinct sister group to the CYP1 and CYP2 genes, without a single vertebrate sequence (Figure 2). Five sequences clustered with the vertebrate CYP1s and eight sequences were clearly clustered within the CYP2s. As expected, CYP2U1 was the most basal of the CYP2 genes (Figure 2). There were no *C. telata* sequences that clustered with the CYP17 or CYP21 sequences.

Clan 3 and 4 contained nine and four CYPs, respectively, while six genes were from the mitochondrial clan (Figure 2). A single *C. telata* sequence was found to cluster with CYP4V (CYP4V25), CYP7s/CYP306s (CYP3067A1), CYP11A (CYP376A1), CYP20 (CYP20A1), CYP27s (CYP371B1), CYP39 (CYP39B1), CYP46 (CYP3070A1), CYP51 (CYP51A1). Interestingly, there were a small numbers of genes (4 sequences) that clustered with CYP26s (Figure 1).

Figure 3 shows phylogenies for clans 2 (A), 3 and 4 (B), and the mitochondrial clan (C). Invertebrate sequences were added to those sequences included in Figure 1 to help resolve and increase bootstrap support for internal branching arrangements within each clan (Figure 3). The addition of invertebrate sequences to the larger phylogeny interfered negatively with tree construction, producing a phylogeny with less robust bootstrap values. In the clan 2 phylogeny (Figure 3A), sequences from *C. elegans* and *D. pulex* were added to the analysis; the *C. elegans* sequences clustered with the CYP2s. The CYP3058 family, clustered closest with the *C. elegans* sequences. The large cluster of clan 2 *C. teleta* CYPs remained on their own, as in the large phylogeny (Figure 1), basal to the rest of the clan 2 sequences. Clan 3 and 4 were sister clans (Figure 1) and were

included together on the same clan phylogeny (Figure 3B) with added sequences from *C. elegans, H. robusta* and *D. pulex*. The additional *C. elegans* sequences clustered closest with the CYP331 family, which were basal in clan 3 in the large phylogeny (Figure 1). The *D. pulex* sequences clearly clustered with the CYP4Vs, including the *C. telata* CYP4V25 (Figure 3B). In the mitochondrial clan phylogeny (Figure 3C), CYP10B1, CYP362B1, CYP44C1, CYP372A1, and CYP372B1, clustered with CYP36 from *D. pulex* and CYP44 from *C. elegans.* CYP371B1 clustered with *H. robusta* CYP371A1.

Table 2 provides the upstream XREs of *C. teleta* genes from CYP families CYP331 and CYP4. The CYP1-like genes, CYP3060A1 and CYP3061A1, were also examined. These CYP genes were either closely related to vertebrate CYP1s (CYP3060A1, and 3061A1) or genes that were upregulated in response to PAH exposure (CYP331 and CYP4 (Li et al., 2004). CYP331A1 had three XREs within 10kb of the start site, the remaining CYP331A genes had no XREs. CYP331B1, CYP3060A1, CYP3061A1, and CYP4AT1 each had one XRE 10kb upstream. Multiple XREs were found upstream of CYP4V25 (two) and CYP4BK4 (four). Only CYP4ED1, of the *C. teleta* CYP4s, had zero XREs upstream of the start site.

Discussion

CYPome Annotation

Annotation of CYPomes can be challenging when working in species that are distantly related to those with a defined CYPome, because the searches are based on homology to known, yet distant sequences: *C. teleta* is the first lophotrochozoan to have

its CYPome annotated and vertebrate sequences were primarily used in our initial searches. These reference vertebrate sequences were well curated, with very high confidence in their annotation, including exon boundaries, making any manual corrections from the PASA output for *C. teleta* more reliable. Annotations of *C. teleta* were additionally verified using *C. elegans* and *D. pulex* sequences for unique hits and no significant regions were found that the vertebrate sequences missed. Overall, our analysis predicted eighty-four full length CYPs, and identified twelve partial CYP sequences that aligned well with existing ESTs, and thirteen partial CYP sequences that lacked any EST support. Our analysis of the *C. teleta* CYPome has identified 24 novel families and 7 novel subfamilies. CYP26 contained two new subfamilies and CYP4, CYP10, CYP39, CYP44, CYP352 each had one new subfamily in *C. teleta*. The CYPomes of non-chordate phyla often contain novel CYP families (Nelson, 2009; Nelson, 2011): *C. elegans* contained 14 unique families (Nelson, 2009) and the *D. melanogaster* CYPome contained 24 families with most families unique to arthropods (Tijet et al., 2001).

During manual annotation it was important to ensure that genes had a length of ~1500 bp, the average length for a CYP. Start (ATG) and stop (TAA/TAG/TGA) codons were noted and always present, as well as appropriate splice signals (GT/AG; Mount, 1982) at intron/exon boundaries. The numbers of exons were not well conserved between related CYPs in other species. Exon number was taken into consideration between related sequences within *C. teleta* during searches for missing exons where EST data was lacking for annotation.

All of the fully annotated *C. teleta* CYPs had EST support covering all or almost all of the gene. A notable exception was CYP3052C1, which was missing EST data for exons one and two. Homology searches in the expected upstream and downstream regions were able to identify the missing exons. There were 12 incompletely annotated CYPs with EST support and these were presumed to be functional, full length genes though they could not be resolved with the existing genome assembly. Thus, the total number of CYPs identified in *C. telata* was 96, which fall into the range predicted by Nelson and colleagues (Nelson et al., 2013) and is comparable to the 50-100 genes found in vertebrate CYPomes (Goldstone et al., 2010; Lewis, 2004; Nelson, 2003), 120 genes in the sea urchin *S. purpuratus* (Goldstone et al., 2006), 75 genes in the crustacean *Daphnia pulex* (Baldwin et al., 2009), 83 genes in the insect *D. melanogaster* (Tijet et al., 2001) and 82 genes in the sea anemone *N. vectensis* (Goldstone, 2008). *C. teleta* had an average number of CYPs for a metazoan CYPome.

There were 13 gene fragments that may be pseudogenes. These fragments lacked EST support or had identifiable early stop codons. The number of possible pseudogenes per functional gene (0.14) is higher than noted in other species: *Daphnia pulex* had .04 (Baldwin et al., 2009), *C. elegans* had 0.1 (Nelson, 2009) and *D. melanogaster* had 0.08 (Tijet et al., 2001) pseudogenes per functional gene. It is possible that a small number of these fragments were functional genes. One CYP on scaffold 342 had EST support but had an in frame stop codon in the first exon.

To provide support for the annotation process, the identified CYPs were examined for conserved CYP motifs. The heme binding region starts around amino acid 430 and

has a well conserved motif of PFXXGXXRXCXG (the 'X' represents a non conserved amino acid); the cysteine, until recently, was considered the only absolutely conserved amino acid in all known CYPs, although exceptions have been documented (Sezutsu, et al., 2013). There are three other well conserved motifs: portions of the I-helix, [A/G]GX[E/D]T[T/S], located around amino acid 300; K-helix, EXXR, located around amino acid 360; and an area known as the 'meander coil', FDPER, located around amino acid 410 (Sezutsu, et al., 2013). The K-helix motif is incredibly well conserved in CYPs, with only a handful of known exceptions to the two conserved amino acids (Rupasinghe et al., 2006). These motifs are important when analyzing potential CYPs, if one or more of these regions are missing, or out of place, it is likely that the gene was constructed incorrectly, a pseudogene or not a CYP at all.

The high conservation in the motifs (Table 1) was expected and supports our annotation of these genes as CYPs. The least conserved domain is the I-helix, and our findings in *C. teleta* support this; CYP3065A1-4, CYP3065B1, CYP3066A1-3, CYP3066C1, CYP3067A1, and CYP372B1, CYP39B1 all have lower conservation in the I-helix. The gaps in CYP3067A1 and the insertion in CYP372B1 are peculiar. Whether these genes are fully functional may be questioned, yet, there is EST support to show that they are expressed.

The *C. teleta* CYPome phylogenetic analysis (Figure 1) contains almost exclusively vertebrate sequences, along with the *C. telata* sequences we identified. The arrangement of the clans was consistent with previous work, down to the family level of the known sequences(Goldstone et al., 2010; Nelson, 2003). It was difficult to add any

sequences outside of vertebrates because of their divergence from vertebrate and *C. teleta* sequences and the lack of sequences that would help provide definitive phylogenetic relationships. When *Drosophila melanogaster* sequences were added to the phylogeny the bootstrap support was very weak, especially in clan 2-4 where most *D. melanogaster* sequences were added, and this is likely due to the evolutionary distance between vertebrates, insects and annelids. The *D. melanogaster* CYPome paper (Tijet et al., 2001) provides a prime example of the difficulty in creating phylogenies between vertebrates and invertebrate CYP sequences. There were major branches (i.e. those that separate clans) with less than 10% bootstrap support (Tijet et al., 2001). The more recent *D. pulex* CYPome (Baldwin et al., 2009) had much better support at the clan level (support beyond the clan level was not provided), which was due to increased saturation in arthropod CYPs from available insect CYP sequences and basal chordate CYPs. As genome sequences become available from a wider array of species across the major metazoan phyla, the evolutionary distance between CYPomes will be reduced and help improve the phylogenetic analyses.

The clan phylogenies (Figure 3) have additional invertebrate sequences to help resolve nodes within the major clans found in the *C. teleta* CYPome. The phylogenies are rooted using plant and fungal CYPs, these sequences chosen based on previous CYP clan phylogenies with diverse eukaryotic CYPs (Nelson, 2009). Using genes from closely related clans as an outgroup, such as clan 46 for the clan 3 and 4 phylogeny, resulted in trees with long branches for the root and poorer bootstrap support. The clan 2 phylogeny (Figure 4A) included sequences from *C. elegans*, the clan 3 and 4 phylogeny included *D.*

pulex, D. magna, and *C. elegans.* sequences, and the mitochondrial phylogeny incorporated sequences from *D. pulex, Caenorhabditis spp.,* and *H. robusta*. The addition of these sequences in the clan phylogenies increased the support for the internal nodes (data not shown). The sequence similarity data and phylogenetic analyses have provided information to infer the placement of the *C. teleta* CYPs into clans and assign nomenclature.

Clan Distribution

C. teleta possesses CYPs from all the metazoan clans except for 19 and 74. Clan 19 has not been found outside chordates (Reitzel and Tarrant, 2010) and clan 74 had not been found outside anemone and placozoa (Nelson et al., 2013), although it has been recently found in amphioxus (Callard et al., 2011). *C. teleta* is the first protostome analyzed to have representation in clan 46. CYP46 is the only clan 46 CYP gene in vertebrates and functions as a cholesterol 24-hydroxylase in the brain (Nebert et al., 2013). Since *C. teleta* CYP3070A1 had only 35% identity with human CYP46A1, it is difficult to predict whether the function is conserved in the *C. telata* ortholog. *In silico* molecular docking or 3D modeling of the protein may help support or refute the possibility that cholesterol is a substrate of CYP3070A1.

There are 53 full length clan 2 CYPs in *C. teleta*, representing ~60% of the total CYPome (Figure 2). This is the second largest in relative size for known CYPomes and is smaller only to *S. purpuratus* (~70%, Figure 2). Insects generally have only 5.5-10% of their CYPs in clan 2 (Baldwin et al., 2009). The function of many insect clan 2 CYPs

are unknown but some are known for their role in ecdysone synthesis (Baldwin et al., 2009). Clan 2 CYPs are much more important for metabolism of exogenous compounds in mammals (Nebert and Russell, 2002).

All of the *C. telata* clan 2 CYPs were located in novel CYP families; indeed the *C. telata* clan 2 sequences had 14 novel CYP families made up from 20 subfamilies. It has been postulated that a large number of CYPs related to families involved in exogenous metabolism (i.e. families 1-4) may suggest evolutionary pressure towards diverse function (Goldstone, 2008). The largest family was CYP3052 with 24 sequences; these sequences made up the majority of the large standalone cluster of 33 *C. teleta* CYPs on the phylogenetic tree (Figure 1). If this family of *C. teleta* CYPs follows the trend of other large CYP families, namely families CYP1-4, then these proteins may be involved in xenobiotic metabolism.

There were five novel families with a single sequence each (CYP3060-3063 and CYP3065) that were CYP1-like. There were fewer CYP1-like genes in *C. telata* than were found in *S. purpuratus* (11) but similar to what is typical (3-4 CYP1 genes) in vertebrates (Goldstone et al., 2006; Nelson, 2009). Of the two families that grouped with CYP2s, CYP3058 clustered more closely with *C. elegans* sequences than vertebrate CYP2 sequences in the clan 2 phylogeny (Figure 3A). The other family, CYP3059, clustered with vertebrate CYP2R, although the bootstrap support in the clan phylogeny was quite low (Figure 3A) suggesting the placement of this family is uncertain with respect to the vertebrate CYP2 families. The function of the *C. elegans* CYPs are

unknown but vertebrate CYP2s are well known for their role in xenobiotic metabolism (Nebert and Russell, 2002). CYP3057A1 and CYP3064A1 were basal in this clan and had high divergence from the remaining sequences..

The clan 3 phylogeny had sequences from across all metazoan phyla. Clan 3 contains families CYP3 and CYP5 in vertebrates, but is represented by different families in invertebrates such as families CYP6 and CYP9 (Verslycke, et al., 2006). Mammalian CYP3s are known to have very flexible active sites that can accommodate structurally diverse substrates. CYP3A4 is the most important enzyme involved in drug metabolism in humans but other CYP3s are also important in metabolism of endogenous and exogenous compounds (Nebert et al., 2013). Clan 3 CYPs are involved in both endogenous and exogenous metabolism in arthropods (Baldwin et al., 2009). *C. teleta* had two clan 3 families with a total of nine CYPs; both families were novel. *N. vectensis* had 20 clan 3 CYPs (Goldstone, 2008), *S. purpuratus* had 10 (Goldstone et al., 2006), and *D. melanogaster* has an expanded clan 3 with 36 CYPs (Tijet et al., 2001). Mammals appear to have a much smaller number of clan 3 genes than many invertebrate species; humans have just five clan 3 sequences from a single subfamily (Lewis, 2004).

The *C. teleta* clan 3 sequences included CYP331A1, which had been previously described (Li et al., 2004). CYP331A1 had increased expression from exposure to benzo[α]pyrene (BaP) and fluoranthene, two PAHs (Li et al., 2004). The CYP331 family has been expanded in this annotation with two more *CYP331A* genes and the *CYP331B1* gene.

The CYP4 family was expanded in *D. melanogaster* (32) and other insects (Tijet et al., 2001), but was relatively limited in *N. vectensis* (3) (Goldstone, 2008). There were five clan 4 CYPs in *C. teleta* and all were from the CYP4 family. CYP4V25 was an ortholog to CYP4Vs yet was below the 55% sequence identity threshold used during standard nomenclature. All of the top BLAST hits for CYP4V25 were CYP4Vs from various species (data not shown). Furthermore, CYP4Vs have been found in molluscs and crustaceans (D. Nelson, personal communication). Collectively, this information supports the placement of this sequence into the CYP4V family despite the low sequence identity to other gene members. Little is known of CYP4 function outside vertebrates. CYP4C has a role in juvenile hormone synthesis in the cockroach *Blaberus discoidalis* (Bradfield, et al., 1991). In vertebrates, CYP4s primarily metabolize endogenous compounds, specifically fatty acids, although they do metabolize some exogenous pharmaceuticals (e.g. erythromycin; Kalsotra, et al., 2004). Yet, even in mammals the function of CYP4V is unknown (Nebert et al., 2013).

Like the CYP4Vs, the function of CYP20A1 remains unclear in vertebrates. The *C. teleta* CYP20A1 is ~40% identical to other CYP20A1s but is a clear ortholog (Figure 1) with no other closely related sequences. CYP20A1 has been documented in invertebrates such as *S. purpuratus* and *H. robusta* (Nelson et al., 2013). It is interesting that CYP20A1 has unknown function yet has such clear homology between annelids and vertebrates.

CYP10 has been identified in molluscs and has been suggested as the only family in the mitochondrial clan in molluscs (Nelson, 1998). Since orthologs have now been

identified in two major phyla, the *C. telata* CYP10 may suggest that CYP10 is present in all lophotrochozoans. Interestingly, CYP10 was not the only mitochondrial CYP in *C. telata*. CYP44 was placed in the mitochondrial clan; a CYP44 homolog has also been found in *C. elegans* (Nelson, 2009), roundworms and molluscs (Nelson et al., 2013). Thus, CYP10 and CYP44 may be expected mitochondrial CYPs in lophotrochozoans.

PAH and xenobiotic metabolism

C. teleta has been long known to metabolize PAHs in sediment, and it has been suggested that CYPs were responsible (Gardner,et al., 1979; Lee, 1998). Benzo[α]pyrene was metabolized, likely by CYPs, in another annelid, *Nereis virens* (Lee, 1998). There is conflicting data on whether CYPs are transcriptionally upregulated after PAH exposure in *N. virens* (primarily tested with benzo[α]pyrene exposure), with some studies reporting a 2-fold increase in CYPs and others reporting no change (reviewed by Lee, 1998). In *C. teleta*, two CYPs (CYP4AT1 and CYP331A1) had a 1.9-2.6 fold increased expression with exposure to some PAHs, including benzo[a]pyrene (Li et al., 2004). CYP4AT1 had 1.25 to 1.9 fold increase in gene expression after exposure to PAH contaminated sediments (Li et al., 2004). Interestingly, there were 3 genes found for *C. teleta* in the CYP331 family, two of which were in the same (CYP331A) subfamiliy.

In vertebrate species, CYP1 gene expression is increased with exposure to PAHs (Review Oost, et al., 2003), through transcriptional activation via the aryl hydrocarbon receptor (AHR) pathway (Hahn, et al., 1998). The AHR is activated by planar PAHs and halogenated aromatic hydrocarbons; TCDD is the ligand with the highest affinity for this

receptor in many species (Hahn, 2002). AHRs transcriptionally regulate a battery of genes through interaction with a specific sequence, the xenobiotic or dioxin response element (XRE or DRE; Denison, et al., 1988). Many AHR ligands, including PAHs, are also substrates for CYP1 enzymes (Hahn, 2002). AHRs are present in invertebrates and the amino acid sequence of the DNA binding domain is similar to that found in vertebrates. Indeed, AHRs from *Drosophila* (Kozu et al., 2006), *C. elegans* (Powell-Coffman, et al., 1998), and *Mya arenaria* (Butler, et al., 2001) are capable of binding with the mammalian XRE sequence. Therefore, we examined the upstream region of the CYP1-like (CYP3060A1 CYP3061A1), CYP331 and CYP4AT genes in *C. telata* to determine if XREs were present (Table 2). CYP331A1 had three XREs within 10kb of the start site, but CYP4AT1 had only one. The difference in the number of XREs between these two CYPs may explain the difference in expression during BaP exposure, since there is a relationship between the number of XREs and the relative upregulation of the gene (Rushmore and Pickett, 1990). The CYP1-like *C. teleta* sequences had one XRE and CYP4BK4 had four XREs in the 10kb upstream region. Should the AHR have a role in regulating gene transcription in *C. teleta* after exposure to PAHs, we would predict that CYP4BK4 would have the greatest transcriptional response. Considering the structural link between AHR ligands and CYP1 substrates in vertebrates, we might speculate that CYP4BK4 be a primary candidate gene for studies of PAH metabolism in this species. Future PAH exposure studies in *C. teleta* will shed light on to role of the AHR and XREs in *C. teleta* and the potential role these CYPs may play in PAH metabolism.

Steroidogenic CYPs

CYP51A1 enzymes are responsible for lanosterol-14-alpha-demethylation; the conversion of lanosterol into cholesterol (Lamb, et al., 1998). Cholesterol is the precursor to steroids and this function is expected in all species with endogenous steroid production. The next step in vertebrate steroidogenesis is cholesterol-side-chain-cleavage, which is completed by CYP11A1 in vertebrates and converts cholesterol to pregnenolone (Baker, 2011b). There was one *C. teleta* CYP (CYP376A1) that clustered with the CYP11 family in the phylogenetic tree (Figure 1) and is the best candidate for cholesterol side-chain-cleavage function in *C. teleta*. CYP11B functions in the synthesis of cortisol and coticosterone (Baker, 2011a; Baker, 2011b), which are not expected in annelids since these molecules have not been found in amphioxus, *Ciona intestinalis,* or sea urchins (Holland et al., 2008).

CYP17A1 functions as a 17-alpha-hydroxylase, which is responsible for converting pregnenolone into DHEA. The production of DHEA is the next step in steroidogenesis after side-chain-cleavage and before the production of androgens (Baker, 2011b). Since there is no *C. telata* CYP that clusters with the CYP17 genes from vertebrates, it is difficult to predict which CYP is likely to complete this function at this time. There were many clan 2 CYPs identified but whether 17-alpha-hydroxylase activity is mediated by one of them is unclear. Analyzing the single copy clan 2 CYPs (e.g. CYP3057A1 and CYP3064A1) would be an appropriate place to begin the search for a 17-alpha-hydroxylase enzyme.

Detectable estrogen production has been documented in annelids (Garciaa-Alonso and Rebscher, 2005), yet there was no CYP19 identified in the *C. teleta* CYPome. This is not surprising, as a CYP19 has not been identified outside of chordates and sea anemone had no CYP19 (Goldstone, 2008), in spite of endogenous estrogen production (Twan, et al., 2003). It has been postulated that another CYP has the aromatase function outside of chordates (Goldstone, 2008). There are many CYPs identified in *C. teleta*, the most promising candidates genes for steroidogenic functions are the single copy CYPs from clan 2, CYP376A1 from the mitochondrial clan, CYP3068A1 or CYP3069A1 from clan 26 and CYP3070A1 from clan 46. All of these CYPs should be further examined by *in silico* methods for their potential ability to bind the intermediates of the steroidogenic pathway.

Conclusion

Capitella teleta has an interesting complement of CYPs. CYPs were found in nine of the eleven metazoan CYP clans. There were a total of 24 novel CYP families; careful study will be required to determine their function. The annotation of the *C. teleta* CYPome will make annotating other lophotrochozoan CYPomes easier. With additional annelid and other lophotrochozoan CYP sequences, we will better understand which of the novel CYP families and subfamilies discovered here are specific to annelids. *C. teleta* is known to survive well in polluted environments and two CYP genes CYP331A1 and CYP4AT1 were known to be transcriptionally regulated by PAHs (Li et al., 2004). Indeed, several more closely related homologs were identified in this study. XRE

sequences were found upstream in several of these genes suggesting that CYP331A1, CYP331B1, several CYP4s and the CYP1-like CYP3060A1 and CYP3061A1 genes may be in the AHR gene battery. Empirical testing will be needed to demonstrate this and explore their possible role in PAH metabolism. Functional hypotheses were raised for several of the *C. teleta* CYPs. CYP51A1 is very likely to catalyze the production of cholesterol, due to a ~65% amino acid identity and clear orthology to other CYP51 sequences. Yet, the steroidogenic pathway was not completely identified. Cholesterol side-chain-cleavage has been hypothesized as the function of CYP376A1. Still, there are no obvious candidates for 17α-hydroxylase and aromatase enzymes, which are carried out by CYP17A and CYP19A, respectively, in vertebrates. Considering that *C. teleta* produces *de novo* estradiol, these reactions are likely undertaken by other CYPs. Future studies on invertebrate steroidogenesis should focus on the CYPs with low copy number and phylogenetic positions close to vertebrate steroidogenic CYPs shown in this study. *In silico* protein folding and docking studies may provide important clues to narrow the number of candidates genes for steroidogenic CYPs and direct future functional studies.

Acknowledgements

This research was funded by the Natural Sciences and Engineering Research Council Discovery grant and Accelerator programs and an Early Researcher Award from the Ontario Ministry of Research and Innovation. We thank Jed Goldstone and David Nelson for review of the annotations and nomenclature of the sequences.

References

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics}, 21*(9), 2104-2105. doi: 10.1093/bioinformatics/bti263

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2

Bach, L., Palmqvist, A., Rasmussen, L. J., and Forbes, V. E. (2005). Differences in PAH tolerance between capitella species: Underlying biochemical mechanisms. *Aquatic Toxicology, 74*(4), 307-319.

Bailey, T. L., and Elkan, C. (1994). *Fitting a mixture model by expectation maximization to discover motifs in bipolymers* Department of Computer Science and Engineering, University of California, San Diego.

Baker, M. E. (2011a). Insights from the structure of estrogen receptor into the evolution of estrogens: Implications for endocrine disruption. *Biochemical Pharmacology, 82*(1), 1-8. doi: http://dx.doi.org/10.1016/j.bcp.2011.03.008

Baker, M. E. (2011b). Origin and diversification of steroids: Co-evolution of enzymes and nuclear receptors. *Molecular and Cellular Endocrinology, 334*(1–2), 14-20. doi: 10.1016/j.mce.2010.07.013

Baldwin, W., Marko, P., and Nelson, D. (2009). The cytochrome P450 (CYP) gene superfamily in daphnia pulex. *BMC Genomics, 10*(1), 169. doi: 10.1186/1471-2164-10-169

Blake, J. A., Grassle, J. P., and Eckelbarger, K. J. (2009). Capitella teleta, a new species designation for the opportunistic and experimental capitella sp. I, with a review of the literature for confirmed records.

Bradfield, J. Y., Lee, Y. H., and Keeley, L. L. (1991). Cytochrome P450 family 4 in a cockroach: Molecular cloning and regulation by regulation by hypertrehalosemic hormone. *Proceedings of the National Academy of Sciences of the United States of America, 88*(10), 4558-4562.

Butler, R. A., Kelley, M. L., Powell, W. H., Hahn, M. E., and Van Beneden, R. J. (2001). An aryl hydrocarbon receptor (AHR) homologue from the soft-shell clam, mya arenaria: Evidence that invertebrate AHR homologues lack 2, 3, 7, 8-tetrachlorodibenzo- p-dioxin and β-naphthoflavone binding. *Gene, 278*(1), 223-234.

Callard, G. V., Tarrant, A. M., Novillo, A., Yacci, P., Ciaccia, L., Vajda, S., . . . Cotter, K. A. (2011). Evolutionary origins of the estrogen signaling system: Insights from amphioxus. *The Journal of Steroid Biochemistry and Molecular Biology*, *127*(3–5), 176-188. doi: 10.1016/j.jsbmb.2011.03.022

Denison, M. S., Fisher, J., and Whitlock, J. (1988). The DNA recognition site for the dioxin-ah receptor complex. nucleotide sequence and functional analysis. *Journal of Biological Chemistry, 263*(33), 17221-17224.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research}, 32*(5), 1792-1797. doi: 10.1093/nar/gkh340

Garciaa-Alonso, J., and Rebscher, N. (2005). Estradiol signalling in nereis virens reproduction. *Invertebrate Reproduction and Development, 48*(1-3), 95-100. doi: 10.1080/07924259.2005.9652175

Gardner, W. S., Lee, R. F., Tenore, K. R., and Smith, L. W. (1979). Degradation of selected polycyclic aromatic hydrocarbons in coastal sediments: Importance of microbes and polychaete worms. *Water, Air, and Soil Pollution, 11*(3), 339-347.

Gille, C., and Frommel, C. (2001). STRAP: Editor for STRuctural alignments of proteins. *Bioinformatics (Oxford, England), 17*(4), 377-378.

Goldstone, J. V. (2008). Environmental sensing and response genes in cnidaria: The chemical defensome in the sea anemone nematostella vectensis. *Cell Biology and Toxicology, 24*(6), 483-502.

Goldstone, J. V., McArthur, A., Kubota, A., Zanette, J., Parente, T., Jonsson, M., . . . Stegeman, J. (2010). Identification and developmental expression of the full

complement of cytochrome P450 genes in zebrafish. *BMC Genomics, 11*(1), 643. doi: 10.1186/1471-2164-11-643

Goldstone, J., Hamdoun, A., Cole, B., Howard-Ashby, M., Nebert, D., Scally, M., . . . Stegeman, J. (2006). The chemical defensome: Environmental sensing and response genes in the strongylocentrotus purpuratus genome. *Dev Biol, 300*, 366-384. doi: 10.1016/j.ydbio.2006.08.066

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., . . . White, O. (2003). Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research}, 31*(19), 5654-5666. doi: 10.1093/nar/gkg770

Hahn, M. E., Woodin, B. R., Stegeman, J. J., and Tillitt, D. E. (1998). Aryl hydrocarbon receptor function in early vertebrates:: Inducibility of cytochrome P450 1A in agnathan and elasmobranch fish. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology, 120*(1), 67-75.

Hahn, M. (2002). Aryl hydrocarbon receptors: Diversity and evolution. *Chem Biol Interact, 141*, 131-160. doi: 10.1016/S0009-2797(02)00070-4

Holland, L., Albalat, R., Azumi, K., Benito-Gutierrez, E., Blow, M., Bronner-Fraser, M., . . . et al. (2008). The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Research, 18*(7), 1100-1111.

Janer, G., and Porte, C. (2007). Sex steroids and potential mechanisms of non-genomic endocrine disruption in invertebrates. *Ecotoxicology, 16*(1), 145-160.

Kalsotra, A., Turman, C. M., Kikuta, Y., and Strobel, H. W. (2004). Expression and characterization of human cytochrome P450 4F11: Putative role in the metabolism of therapeutic drugs and eicosanoids. *Toxicology and Applied Pharmacology, 199*(3), 295-304. doi: 10.1016/j.taap.2003.12.033

Keay, J., and Thornton, J. W. (2009). Hormone-activated estrogen receptors in annelid invertebrates: Implications for evolution and endocrine disruption. *Endocrinology, 150*(4), 1731-1738. doi: 10.1210/en.2008-1338

Kozu, S., Tajiri, R., Tsuji, T., Michiue, T., Saigo, K., and Kojima, T. (2006). Temporal regulation of late expression of bar homeobox genes during drosophila leg development by spineless, a homolog of the mammalian dioxin receptor. *Developmental Biology, 294*(2), 497. doi: http://dx.doi.org/10.1016/j.ydbio.2006.03.015"

Lamb, D. C., Kelly, D. E., and Kelly, S. L. (1998). Molecular diversity of sterol 14alpha-demethylase substrates in plants, fungi and humans. *FEBS Letters, 425*(2), 263-265.

Lee, R. F. (1998). Annelid cytochrome P-450. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology, 121*(1), 173-179.

Lee, R. F., and Singer, S. C. (1980). Detoxifying enzymes system in marine polychaetes: Increases in activity after exposure to aromatic hydrocarbons.

Lewis, D. F. V. (2004). 57 varieties: The human cytochromes P450. *Pharmacogenomics, 5*(3), 305-318. doi: 10.1517/phgs.5.3.305.29827

Li, B., Bisgaard, H. C., and Forbes, V. E. (2004). Identification and expression of two novel cytochrome \P450\ genes, belonging to CYP4 and a new CYP331 family, in the polychaete capitella capitata sp.I. *Biochemical and Biophysical Research Communications, 325*(2), 510. doi: http://dx.doi.org/10.1016/j.bbrc.2004.10.066"

Linke-Gamenick, I., Forbes, V. E., and M\'endez, N. (2000). Effects of chronic fluoranthene exposure on sibling species of capitella with different development modes.

Maddison, W., and Maddison D. (2011). *Mesquite: A modular system for evolutionary analysis* (2.75th ed.)

Mount, S. M. (1982). A catalogue of splice junction sequences. *Nucleic Acids Research}, 10*(2), 459-472. doi: 10.1093/nar/10.2.459

Nebert, D. W., and Gonzalez, F. J. (1987). P450 genes: Structure, evolution, and regulation. *Annual Review of Biochemistry, 56*, 945-993. doi: 10.1146/annurev.bi.56.070187.004501

Nebert, D. W., and Russell, D. W. (2002). Clinical importance of the cytochromes P450. *The Lancet, 360*(9340), 1155-1162. doi: 10.1016/S0140-6736(02)11203-7

Nebert, D. W., Wikvall, K., and Miller, W. L. (2013). Human cytochromes P450 in health and disease. *Philosophical Transactions of the Royal Society B: Biological Sciences}, 368*(1612) doi: 10.1098/rstb.2012.0431

Nelson, D. R. (2009). The cytochrome P450 homepage. *Hum Genomics, 4*(1), 59-65.

Nelson, D. R. (2011). Progress in tracing the evolutionary paths of cytochrome P450. *Biochimica Et Biophysica Acta.Proteins and Proteomics, 1814*(1), 14-18.

Nelson, D. R. (1999). Cytochrome \P450\ and the individuality of species. *Archives of Biochemistry and Biophysics, 369*(1), 1. doi: http://dx.doi.org/10.1006/abbi.1999.1352"

Nelson, D. R. (2003). Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. *Archives of Biochemistry and Biophysics, 409*(1), 18-24. doi: 10.1016/S0003-9861(02)00553-2

Nelson, D. R., Goldstone, J. V., and Stegeman, J. J. (2013). The cytochrome P450 genesis locus: The origin and evolution of animal cytochrome P450s. *Philosophical Transactions of the Royal Society B: Biological Sciences}, 368*(1612) doi: 10.1098/rstb.2012.0474

Nelson, D. R., Koymans, L., Kamataki, T., Stegeman, J. J., Feyereisen, R., Waxman, D. J., . . . Nebert, D. W. (1996). P450 superfamily: Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics and Genomics, 6*(1) Retrieved from http://journals.lww.com/jpharmacogenetics/Fulltext/1996/02000/P450_superfamily__update_on_new_sequences,_gene.2.aspx

Oost, R. v. d., Beyer, J., and Vermeulen, N. P. E. (2003). Fish bioaccumulation and biomarkers in environmental risk assessment: A review. *Environmental Toxicology and Pharmacology, 13*(2), 57. doi: http://dx.doi.org/10.1016/S1382-6689(02)00126-6"

Powell-Coffman, J. A., Bradfield, C. A., and Wood, W. B. (1998). Caenorhabditis elegans orthologs of the aryl hydrocarbon receptor and its heterodimerization partner the aryl hydrocarbon receptor nuclear translocator. *Proceedings of the National Academy of Sciences, 95*(6), 2844-2849.

Pearson, W., and Lipman, D. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A, 85*(8), 2444-2448. Retrieved from http://europepmc.org/abstract/MED/3162770

Reitzel, A. M., and Tarrant, A. M. (2010). Correlated evolution of androgen receptor and aromatase revisited. *Molecular Biology and Evolution, 27*(10), 2211-2215. doi: 10.1093/molbev/msq129

Rupasinghe, S., Schuler, M. A., Kagawa, N., Yuan, H., Lei, L., Zhao, B., . . . Lamb, D. C. (2006). The cytochrome P450 gene family CYP157 does not contain EXXR in the K-helix reducing the absolute conserved P450 residues to a single cysteine. *{FEBS} Letters, 580*(27), 6338. doi: http://dx.doi.org/10.1016/j.febslet.2006.10.043"

Rushmore, T. H., and Pickett, C. (1990). Transcriptional regulation of the rat glutathione S-transferase ya subunit gene. characterization of a xenobiotic-responsive element controlling inducible expression by phenolic antioxidants. *Journal of Biological Chemistry, 265*(24), 14648-14653.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M., and Barrell, B. (2000). Artemis: Sequence visualization and annotation. *Bioinformatics, 16*(10), 944-945. Retrieved from http://europepmc.org/abstract/MED/11120685

Sanders, H. L., Grassle, J. F., Hampson, G. R., Morse, L. S., Garner-Price, S., and Jones, C. C. (1980). Anatomy of an oil spill: Long-term effects from the grounding of the barge florida off west falmouth, massachusetts.

Selck, H., Palmqvist, A., and Forbes, V. E. (2003). Biotransformation of dissolved and sediment-bound fluoranthene in the polychaete, capitella sp. I. *Environmental Toxicology and Chemistry, 22*(10), 2364-2374.

Sezutsu, H., Le Goff, G., and Feyereisen, R. (2013). Origins of P450 diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences, 368*(1612)

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics}, 22*(21), 2688-2690. doi: 10.1093/bioinformatics/btl446

Sun, Y. V., Boverhof, D. R., Burgoon, L. D., Fielden, M. R., and Zacharewski, T. R. (2004). Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences. *Nucleic Acids Research, 32*(15), 4512-4523. doi: 10.1093/nar/gkh782

Tijet, N., Helvig, C., and Feyereisen, R. (2001). The cytochrome P450 gene superfamily in drosophila melanogaster: Annotation, intron-exon organization and phylogeny. *Gene, 262*(12), 189. doi: http://dx.doi.org/10.1016/S0378-1119(00)00533-3"

Twan, W., Hwang, J., and Chang, C. (2003). Sex steroids in scleractinian coral, euphyllia ancora: Implication in mass spawning. *Biology of Reproduction, 68*(6), 2255-2260.

Verslycke, T., Goldstone, J. V., and Stegeman, J. J. (2006). Isolation and phylogeny of novel cytochrome P450 genes from tunicates (ciona spp.): A CYP3 line in early deuterostomes? *Molecular Phylogenetics and Evolution, 40*(3), 760. doi: http://dx.doi.org/10.1016/j.ympev.2006.04.017"

Werck-Reichhart, D., and Feyereisen, R. (2000). Cytochromes P450: A success story. *Genome Biology, 1*(6). doi: 10.1186/gb-2000-1-6-reviews3003

Wu, M., Chatterji, S., and Eisen, J. A. (2012). Accounting for alignment uncertainty in phylogenomics. *PloS One, 7*(1), e30288. doi: 10.1371/journal.pone.0030288; 10.1371/journal.pone.0030288

Tables and Figures

Table 2.1. Conserved motifs across the *Capitella teleta* CYPome. Four motifs (I-helix, K-helix, meander coil, and heme loop) are represented in an aligned format to show conservation across the *C. teleta* CYPs. Bolded letters represent conserved residues. AA is the amino acid number where the motif begins in each gene. The expected motif sequence is given in each heading for comparison. The glutamic acid and arginine residues in the meander coil, and the cysteine residue in the heme loop are conserved across the entire CYPome. Note the lack of conservation in CYP372B1 (I-helix) and CYP3067A1 (I-helix and heme loop).

CYP	I-helix		K-helix		Meander Coil		Heme Loop	
	AA	[A/G]GX[D/E]T[T/S]	AA	EXXR	AA	FDPER	AA	PFXXGXRXCXG
CYP10B1	284	GAVETT	341	ETFR	393	FKPER	415	PFGHGARMCIG
CYP20A1	281	AGFHTT	338	ESLR	390	FDPER	410	PFGFGKRKCLG
CYP26D1	289	AGYETT	348	EVLR	400	FDPDR	421	PFGSGSRSCAG
CYP26E1	317	AGHLPT	378	EVLR	430	FNPDR	459	PFGSGQRSCVA
CYP3052A1	307	AGTATT	362	ELLR	415	FEPER	440	PFGAGPRVCLG
CYP3052A10	304	AGTATT	365	ELLR	418	FQPER	443	PFGAGPRVCLG
CYP3052A11	305	AGTATT	352	ELLR	405	FQPER	430	PFGAGPRVCLG
CYP3052A12	306	AGTATT	362	ELLR	415	FQPER	440	PFGAGLRVCLG
CYP3052A13	306	AGTATT	362	ELLR	415	FHPER	440	PFGAGPRVCLG
CYP3052A2	307	AGTATT	361	ELLR	414	FEPER	439	PFGAGPRVCMG
CYP3052A3	308	GGTATT	361	ELLR	414	FEPER	439	PFGAGPRVCLG
CYP3052A4	295	AGTSTT	349	ELLR	378	..PER	401	PFGAGPRVCLG
CYP3052A5	305	AGTSTT	363	ELLR	416	FLPER	441	PFGAGPRVCLG
CYP3052A6	303	AGTATT	363	ELLR	416	FQPER	441	PFSAGPRVCLG
CYP3052A7	304	AGTATT	360	ELLR	413	FQPER	438	PFSAGPRVCLG
CYP3052A8	305	AGTATT	364	ELLR	417	FQPER	442	PFGAGPRVCLG
CYP3052A9	305	AGTSTT	364	ELLR	417	FQPER	442	PFGAGPRVCLG
CYP3052B1	303	AGTTTT	360	ELLR	413	FQPER	438	PFGAGPRLCIG
CYP3052B2	308	GATTTT	365	ELLR	418	FRPER	443	PFGAGTRVCLG
CYP3052B3	307	AGTGTT	364	ELLR	417	FRPER	442	PFGAGTRVCIG
CYP3052B4	308	AGTGTT	365	ELLR	418	FRPER	443	PFGAGTRVCIG
CYP3052B5	300	AGTSTT	357	EILR	410	FKPER	435	PFGAGPRVCVG
CYP3052B6	300	AGTSTT	357	EILR	410	FKPER	435	PFGAGPRVCVG
CYP3052C1	299	AGTGTT	356	ELLR	409	FRPDR	434	AFGAGARVCIG
CYP3052C2	303	AGTSTT	360	ELLR	413	FKPER	438	TFGGGQRKCIG
CYP3052D1	300	AGTSTT	357	ELLR	410	FRPER	435	PFGAGTRVCLG
CYP3052D2	300	AGTSTT	357	ELLR	410	FRPER	435	PFGAGTRVCLG

CYP	I-helix		K-helix		Meander Coil		Heme Loop	
	AA	[A/G]GX[D/E]T[T/S]	AA	EXXR	AA	FDPER	AA	PFXXGXRXCXG
CYP3053A1	299	AGVGTF	356	ELTR	409	FKPER	434	AYGAGQRVCLG
CYP3054A1	304	AGVLTT	360	EIMR	413	FRPER	438	PFGMGPRICAG
CYP3054A2	303	AGVLST	359	EIMR	412	FRPER	437	PFGMGPRICAG
CYP3054A3	303	AGVLTT	359	EIMR	412	FRPER	437	PFGMGPRICAG
CYP3054A4	303	AGVLTT	359	EIMR	412	FRPER	437	PFGMGPRICAG
CYP3054A5	303	AGVLTT	359	EIMR	412	FRPER	437	PFGLGPRICAG
CYP3055A1	296	AGIVTT	352	ELLR	405	FIPER	430	AFGAGPRMCVG
CYP3055B1	297	AGSVST	354	ELFR	407	FNPER	432	PFSAGPRVCMG
CYP3056A1	301	SGTLTS	358	ETLR	411	FNPNR	436	PFGAGRRMCLG
CYP3057A1	286	AGIDTI	342	ENQR	395	FRPER	416	PFAGGRRVCLG
CYP3058A1	312	AGAETT	368	EVQR	421	FNPER	441	PFSVGPRMCAG
CYP3058A2	301	AGAETT	357	EVQR	410	FNPER	430	PFSVGPRMCAG
CYP3058A3	295	AGGETT	351	EVQR	404	FNPER	424	PFGVGPRMCAG
CYP3058B1	310	AGADTS	366	EIQR	419	FRPGR	439	NFGIGKWSCPG
CYP3058C1	313	AGSVTT	369	EIQR	422	FRPER	442	PYGIGPRACAG
CYP3059A1	300	AGTESS	356	EIQR	409	FDPSR	429	PFGIGRRVCLG
CYP3059A2	300	AGTETT	356	EIQR	409	FDPSR	429	PFGIGRRVCLG
CYP3059A3	293	AGTETT	349	EIQR	402	FNPSR	422	PFGIGRRLCLG
CYP3060A1	299	AGLDIV	356	ELLR	409	FKPER	432	PYGMGRRRCIG
CYP3061A1	289	AGADTV	351	EVAR	404	FRPER	426	MFGYGKRRCIG
CYP3062A1	306	AGVESM	359	EVYR	412	FNPDN	434	PFGYGMRRCPG
CYP3062A2	319	AGTESM	372	EVYR	425	FNPNR	447	PFGAGMRRCPG
CYP3063A1	301	AGTETS	358	EIMR	411	FNPDR	437	PFGAGKRKCIG
CYP3064A1	281	GVSDGS	335	EVLR	388	FNPSR	410	PFSTGQRSCVG
CYP3065A1	322	DSLDTL	379	ETYR	432	FRPER	452	PFGVGPRSCPG
CYP3065A2	279	DSLDTL	336	ETYR	389	FRPER	409	PFGVGPRSCPG
CYP3065A3	279	DSLDTL	336	ETYR	389	FRPER	409	PFGVGPRSCVG
CYP3065A4	320	DALDSL	377	ECYR	430	FKPER	450	PFGLGPRACAG

CYP	I-helix		K-helix		Meander Coil		Heme Loop	
	AA	[A/G]GX[D/E]T[T/S]	AA	EXXR	AA	FDPER	AA	PFXXGXRXCXG
CYP3065B1	318	DSADTL	375	ETFR	428	FKPER	448	PFGLGPRACLG
CYP3066A1	317	NAFSTI	374	ESLR	426	FDPER	446	PFGFGPRNCVG
CYP3066A2	317	AAFGTI	374	ESLR	426	FDPER	446	PFGFGPRNCVG
CYP3066A3	321	AAYGTI	378	ESLR	430	FIPER	450	PFGQGPRHCIA
CYP3066B1	302	AGYSTI	358	ESLR	410	FDPDR	430	PFGFGPRHCIG
CYP3066C1	323	SGHSTV	380	ESLR	432	FNPKR	452	PFGMGPRSCIG
CYP3067A1	286	....NT	344	ESFR	400	FKYDR	424	AFGS...LCPG
CYP3068A1	278	ASQETL	336	EMLR	388	FDPYR	411	PFGAGNRTCVG
CYP3069A1	264	GAQETL	312	ETLR	364	FNPDQ	383	PFGGGAHACVG
CYP3070A1	309	AGQETT	366	ETLR	418	FNPDR	436	PFSLGQRSCLG
CYP3071A1	314	AGHETT	372	EVQR	424	FDPGR	442	PFSTGPHKCLG
CYP3072A1	306	AGHETT	363	ETLR	416	FNPDR	435	PFLIGPRMCLG
CYP331A1	348	AGYETT	405	ETLR	460	FEPER	480	PFGAGPRNCIG
CYP331A2	306	AGYDTT	363	ETLR	418	FEPER	438	PFGAGPRNCIG
CYP331A3	344	AGFETS	401	ETLR	456	FEPER	476	PFGVGPRNCMG
CYP331B1	350	AGFETT	407	ETLR	462	FEPER	482	PFGAGPRNCVG
CYP362B1	288	AGVDTT	345	ESQR	397	FIPER	421	PFGHGARSCIG
CYP371B1	346	GAVDTT	403	EALR	455	FIPER	476	PFGFGARSCIG
CYP372A1	281	AGIDST	346	ESFR	391	FIPER	414	PFGYGPRMCIG
CYP372B1	280	PNIEIEDRST	341	ESFR	394	YHPER	421	PFSHGLRACPG
CYP376A1	277	AAVDTT	334	EVLR	386	FKPER	409	PFGFGTRMCLG
CYP39B1	282	ASLANA	347	ESIR	398	FKPDR	419	PFGGGRFQCPG
CYP44C1	301	DGMITT	359	EGFR	411	FIPER	432	PFSCGPRMCPG
CYP4AT1	299	EGHDTT	356	ESLR	409	YDPER	429	PFSAGPRNCIG
CYP4BK4	245	EGHDTT	304	ESMR	356	FRPDR	376	PFSAGPRNCIG

CYP	I-helix		K-helix		Meander Coil		Heme Loop	
	AA	[A/G]GX[D/E]T[T/S]	AA	EXXR	AA	FDPER	AA	PFXXGXRXCXG
CYP4ED1	340	EGHDTT	399	ESLR	452	YNPER	472	PFSAGPRNCIG
CYP4V25	308	EGHDTT	367	ETLR	419	FIPDR	439	PFSAGLRNCIG
CYP51A1	306	AGQHTS	364	ETLR	416	FNPDR	437	PFGAGRHRCIG

Table 2.2. Xenobiotic response elements upstream of *C. teleta* CYPs. Each gene was searched for the consensus xenobiotic response element (XRE) sequence TNGCGTG, (Sun, et al., 2004)10kb and 20 kb upstream of the start site. Genes were chosen based on homology to vertebrate CYP1s and genes from clan 3 and 4. Genes were searched in the 10kb and in the 20kb upstream region (inclusive of the 10kb region). The asterisks mark the genes which were transcriptionally upregulated with exposure to the PAHs benzo[α]pyrene, 3-methylcholanthrene, or fluoranthene (Li, et al., 2004).

CYP	Number of XREs 10kb upstream	Number of XREs 20kb upstream
CYP331A1*	3	3
CYP331A2	0	0
CYP331A3	0	3
CYP331B1	1	3
CYP4V25	2	2
CYP4AT1*	1	2
CYP4BK4	4	6
CYP4ED1	0	1
CYP3060A1	1	1
CYP3061A1	1	2

Figure 2.1. Phylogenetic tree of Cytochrome P450s in metazoa. The tree was completed on RaxML using non-parametric bootstrapping with a gamma distribution. The tree was rooted with CYP51. The black names are the *Capitella teleta* sequences. The tree is colour coded by clan: clan 2 orange, clan 3 dark green, clan 4 teal, clan 7 salmon, clan 19 light blue, clan 20 dark blue, clan 26 red, clan 46 lime green, mitochondrial clan yellow, and the two sequences that do not fit into a clan (CYP3071A1 and CYP3072A1) are purple.

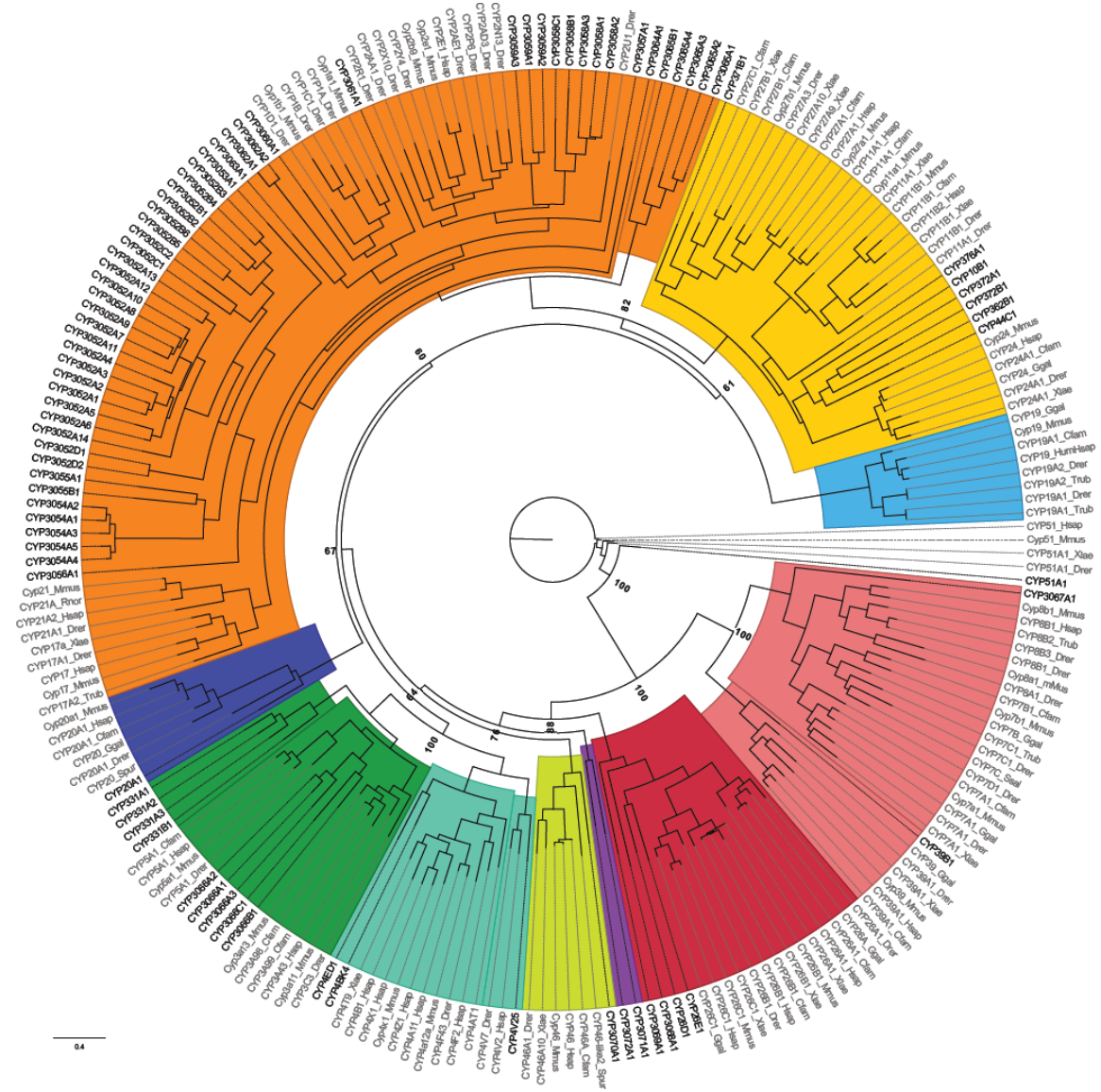


Figure 2.2. Distribution of the major Cytochrome P450 clans in five different species. *Capitella teleta*, *Strongylocentrotus purpuratus* (Goldstone et al., 2006), *Nematostella vectensis* (Goldstone, 2008), *Drosophila melanogaster* (Tijet, Helvig, and Feyereisen, 2001), and *Homo sapiens* (Lewis, 2004) are compared.

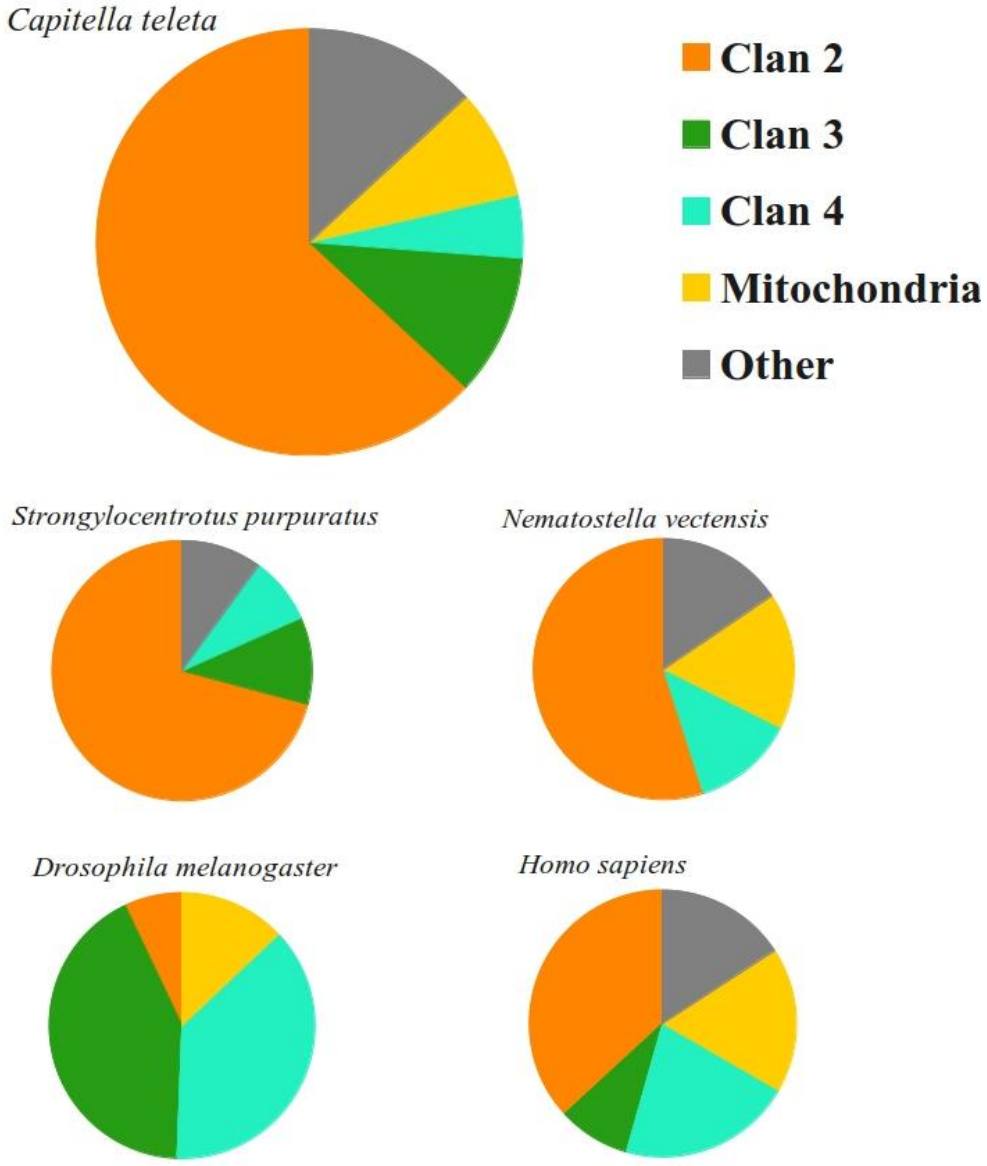
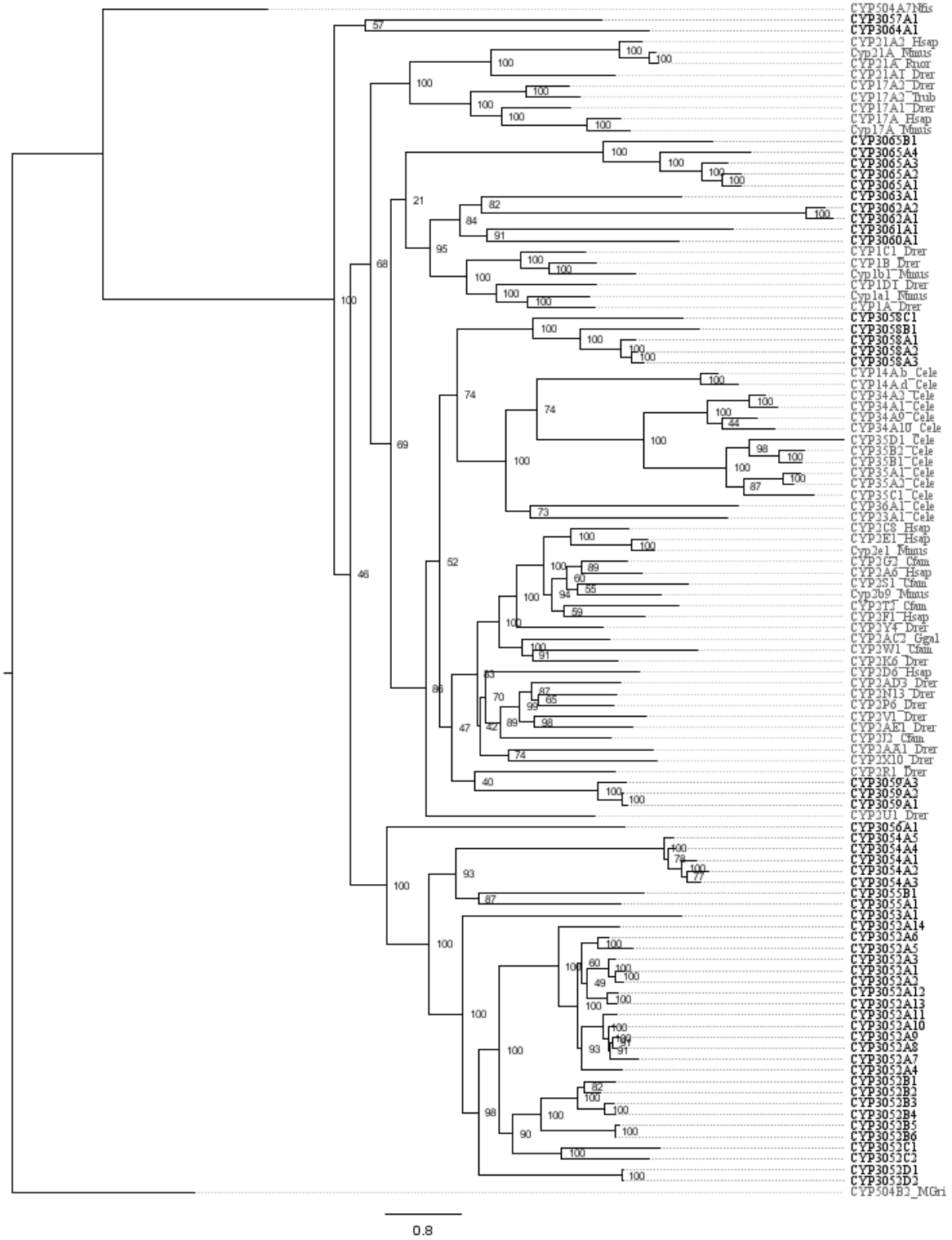
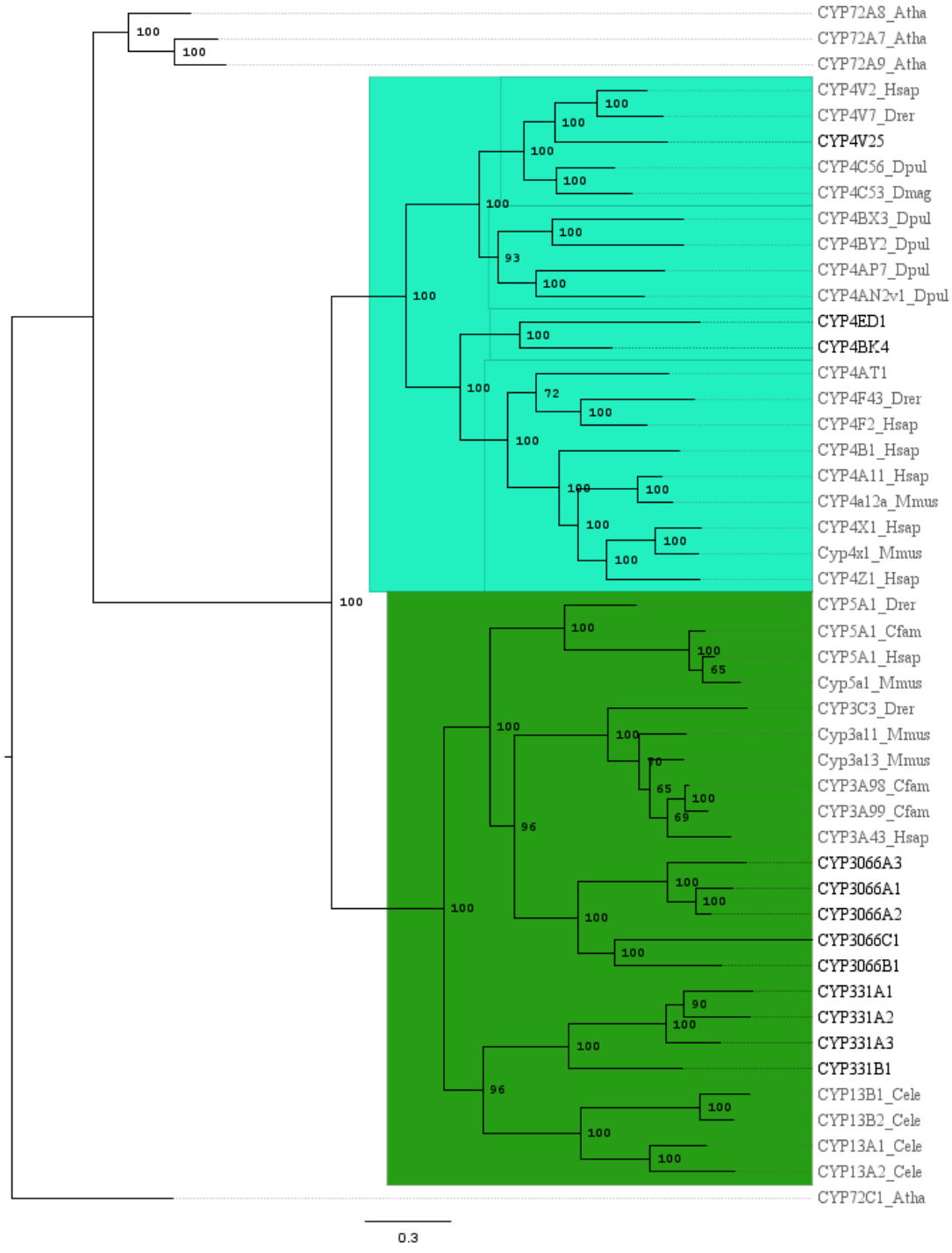

Capitella teleta
Clan 2
Clan 3
Clan 4
Mitochondria
Other
Strongylocentrotus purpuratus
Nematostella vectensis
Drosophila melanogaster
Homo sapiens

Figure 2.3. Phylogenies of Cytochrome P450 clan 2 (A), clan 3 and 4 (B), and the mitochondrial clan (C). The sequences are identical to those in Figure 1 with added invertebrate sequences to increase internal node resolution. The tree was completed on RaxML using non-parametric bootstrapping with a gamma distribution. *C. teleta* sequences are in black, all other sequences are in gray. The phylogenies were rooted using fungal CYP86s (A), *Arabidopsis thaliana* CYP72s (B), or CYP19s and *Arabidopsis thaliana* CYP86s (C).

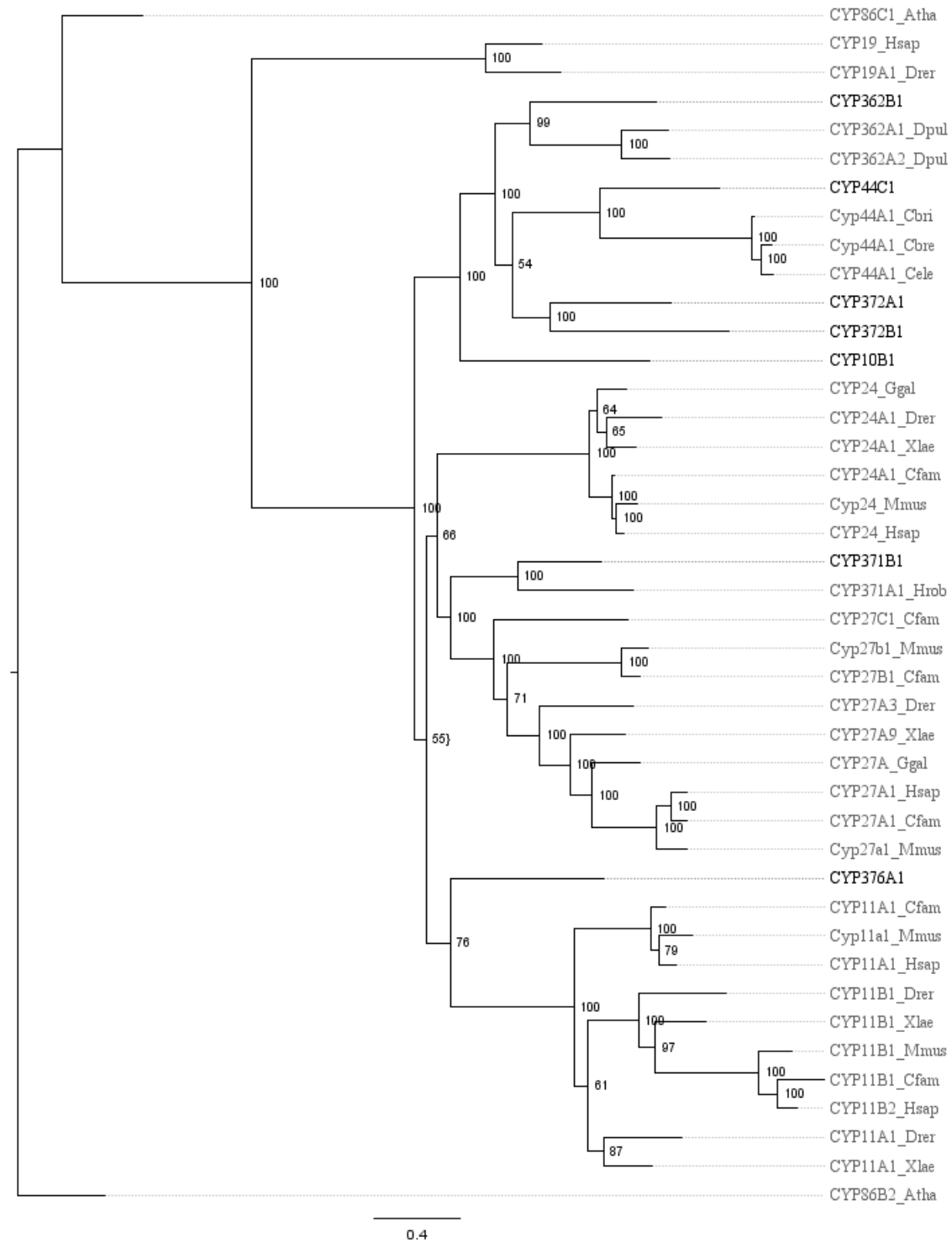
(A)


CYP504A7Nfis
CYP3057A1
CYP3064A1
CYP21A2_Hsap
Cyp21A_Mmus
CYP21A_Fruor
CYP21A1_Drer
CYP17A2_Drer
CYP17A2_Trub
CYP17A1_Drer
CYP17A_Hsap
Cyp17A_Mmus
CYP3065B1
CYP3065A4
CYP3065A3
CYP3065A2
CYP3065A1
CYP3063A1
CYP3062A2
CYP3062A1
CYP3061A1
CYP3060A1
CYP1C1_Drer
CYP1B_Drer
Cyp1b1_Mmus
CYP1D1_Drer
Cyp1a1_Mmus
CYP1A_Drer
CYP3058C1
CYP3058B1
CYP3058A1
CYP3058A2
CYP3058A3
CYP14Ab_Cele
CYP14Ad_Cele
CYP34A2_Cele
CYP34A1_Cele
CYP34A9_Cele
CYP34A10_Cele
CYP35D1_Cele
CYP35B2_Cele
CYP35B1_Cele
CYP35A1_Cele
CYP35A2_Cele
CYP35C1_Cele
CYP36A1_Cele
CYP23A1_Cele
CYP2C9_Hsap
CYP2E1_Hsap
Cyp2e1_Mmus
CYP2G2_Cfam
CYP2A6_Hsap
CYP2S1_Cfam
Cyp2b9_Mmus
CYP2T3_Cfam
CYP2F1_Hsap
CYP2Y4_Drer
CYP2AC2_Ggal
CYP2W1_Cfam
CYP2K6_Drer
CYP2D6_Hsap
CYP2AD3_Drer
CYP2N13_Drer
CYP2P6_Drer
CYP2V1_Drer
CYP2AE1_Drer
CYP2J2_Cfam
CYP2AA1_Drer
CYP2X10_Drer
CYP2R1_Drer
CYP3059A3
CYP3059A2
CYP3059A1
CYP2U1_Drer
CYP3056A1
CYP3054A5
CYP3054A4
CYP3054A1
CYP3054A2
CYP3054A3
CYP3055B1
CYP3055A1
CYP3053A1
CYP3052A14
CYP3052A6
CYP3052A5
CYP3052A3
CYP3052A1
CYP3052A2
CYP3052A12
CYP3052A13
CYP3052A11
CYP3052A10
CYP3052A9
CYP3052A8
CYP3052A7
CYP3052A4
CYP3052B1
CYP3052B2
CYP3052B3
CYP3052B4
CYP3052B5
CYP3052B6
CYP3052C1
CYP3052C2
CYP3052D1
CYP3052D3
CYP504B3_MGri
0.8

(B)


CYP72A8_Atha
CYP72A7_Atha
CYP72A9_Atha
CYP4V2_Hsap
CYP4V7_Drer
CYP4V25
CYP4C56_Dpul
CYP4C53_Dmag
CYP4BX3_Dpul
CYP4BY2_Dpul
CYP4AP7_Dpul
CYP4AN2v1_Dpul
CYP4ED1
CYP4BK4
CYP4AT1
CYP4F43_Drer
CYP4F2_Hsap
CYP4B1_Hsap
CYP4A11_Hsap
CYP4a12a_Mmus
CYP4X1_Hsap
Cyp4x1_Mmus
CYP4Z1_Hsap
CYP5A1_Drer
CYP5A1_Cfam
CYP5A1_Hsap
Cyp5a1_Mmus
CYP3C3_Drer
Cyp3a11_Mmus
Cyp3a13_Mmus
CYP3A98_Cfam
CYP3A99_Cfam
CYP3A43_Hsap
CYP3066A3
CYP3066A1
CYP3066A2
CYP3066C1
CYP3066B1
CYP331A1
CYP331A2
CYP331A3
CYP331B1
CYP13B1_Cele
CYP13B2_Cele
CYP13A1_Cele
CYP13A2_Cele
CYP72C1_Atha
0.3

(C)


CYP86C1_Atha
CYP19_Hsap
CYP19A1_Drer
CYP362B1
CYP362A1_Dpul
CYP362A2_Dpul
CYP44C1
Cyp44A1_Cbri
Cyp44A1_Cbre
CYP44A1_Cele
CYP372A1
CYP372B1
CYP10B1
CYP24_Ggal
CYP24A1_Drer
CYP24A1_Xlae
CYP24A1_Cfam
Cyp24_Mmus
CYP24_Hsap
CYP371B1
CYP371A1_Hrob
CYP27C1_Cfam
Cyp27b1_Mmus
CYP27B1_Cfam
CYP27A3_Drer
CYP27A9_Xlae
CYP27A_Ggal
CYP27A1_Hsap
CYP27A1_Cfam
Cyp27a1_Mmus
CYP376A1
CYP11A1_Cfam
Cyp11a1_Mmus
CYP11A1_Hsap
CYP11B1_Drer
CYP11B1_Xlae
CYP11B1_Mmus
CYP11B1_Cfam
CYP11B2_Hsap
CYP11A1_Drer
CYP11A1_Xlae
CYP86B2_Atha
0.4

Supplementary Information

Supplementary Table 2.1. Cytochrome P450 superfamily complement in *Capitella teleta*. Temporary names were based off the scaffold they were found on. Length is in amino acids. CYPs were named by the CYP nomenclature committee. There are a total of 83 full length CYPs. Only complete CYPs are listed.

Name	Scaffold	Region on Scaffold	length	Number of exons
CYP10B1	aa_51b	278063-281135	493	10
CYP20A1	aa_28	374918-381170	471	11
CYP26D1	aa_827a	39365-41018	491	4
CYP26E1	aa_251	155002-156959	533	6
CYP3052A1	aa_61	510662-512185	507	1
CYP3052A10	aa_335	206104-207633	509	1
CYP3052A11	aa_598	84331-85867	497	2
CYP3052A12	aa_1060	35426-36940	504	1
CYP3052A13	aa_179	255296-256810	504	1
CYP3052A2	aa_3152	3560-5086	508	1
CYP3052A3	aa_919	41783-43301	506	1
CYP3052A4	aa_1290	5262-6776	488	3
CYP3052A5	aa_280	32611-35327	505	2
CYP3052A6	aa_561	125254-126894	508	1
CYP3052A7	aa_604	50890-55648	504	4
CYP3052A8	aa_145a	261379-262905	508	1
CYP3052A9	aa_145b	143553-145079	508	1
CYP3052A14	aa_35	12512-14020	502	1
CYP3052B1	aa_427	73928-75447	509	1
CYP3052B2	aa_458a	67689-69230	513	1
CYP3052B3	aa_277	171440-172979	512	1
CYP3052B4	aa_10	911289-912828	513	1
CYP3052B5	aa_1387	11388-12905	506	1
CYP3052B6	aa_71	368819-370336	505	1
CYP3052C1	aa_102	349221-350733	503	1
CYP3052C2	aa_34a	228965-230860	515	2
CYP3052D1	aa_203	276541-278055	504	1
CYP3052D2	aa_541	32185-33699	504	1
CYP3053A1	aa_241	131474-132970	498	1
CYP3054A1	aa_306a	15283-16800	505	1
CYP3054A2	aa_221	244988-246502	504	1
CYP3054A3	aa_94	349171-350685	505	1
CYP3054A4	aa_1366	11373-12887	504	1
CYP3054A5	aa_41	433769-435283	504	1
CYP3055A1	aa_134	44412-45900	495	1
CYP3055B1	aa_169	72172-73686	503	1
CYP3056A1	aa_20	467436-469516	508	7
CYP3057A1	aa_216	11442-14215	481	10
CYP3058A1	aa_281	187395-189507	503	8
CYP3058A2	aa_274	1964-5573	492	8
CYP3058A3	aa_857	6369-11980	486	8
CYP3058B1	aa_617	72607-75461	499	8

Name	Scaffold	Region on Scaffold	Length	Number of exons
CYP3059A1	aa_126a	226426-228142	492	5
CYP3059A2	aa_126b	246609-248314	491	5
CYP3059A3	aa_233a	212903-215197	485	5
CYP3061A1	aa_327	179333-180830	497	1
CYP3062A1	aa_271a	4824-6314	496	1
CYP3062A2	aa_330a	51294-52824	509	1
CYP3063A1	aa_374	176599-178107	502	1
CYP3064A1	aa_8a	516886-517601	475	3
CYP3065A1	aa_665	57225-58965	515	4
CYP3065A2	aa_330b	185342-187077	472	1
CYP3065A3	aa_18	341290-343531	472	4
CYP3065A4	aa_282	98421-100109	514	5
CYP3065B1	aa_1146	34070-35813	510	4
CYP3066A1	aa_146	10161-12924	507	2
CYP3066A2	aa_4	920137-922924	506	2
CYP3066A3	aa_415	3449-6653	501	3
CYP3066B1	aa_92	409425-411007	496	2
CYP3066C1	aa_233b	80125-86410	513	2
CYP3067A1	aa_173	214821-217549	495	4
CYP3068A1	aa_228	248774-250443	488	4
CYP3069A1	aa_516	98123-103824	445	4
CYP3070A1	aa_225	219819-222601	497	13
CYP3071A1	aa_194	119969-122196	503	12
CYP3072A1	aa_222	119556-122903	496	13
CYP331A1	aa_75	93823-97978	549	13
CYP331A2	aa_1440	32948-36714	503	13
CYP331A3	aa_17	298396-302478	540	13
CYP331B1	aa_91b	197700-204902	544	13
CYP362B1	aa_390	61374-66248	481	9
CYP371B1	aa_34b	572826-579170	537	11
CYP372A1	aa_127	292128-294739	465	9
CYP372B1	aa_658	67989-70914	484	7
CYP376A1	aa_306b	86126-88100	470	8
CYP39B1	aa_1112	28923-39283	490	12
CYP44C1	aa_85	212967-216603	494	9
CYP4AT1	aa_255	217215-220519	490	12
CYP4BK4	aa_296	153301-158233	438	11
CYP4ED1	aa_627	38851-44587	534	11
CYP4V25	aa_827b	59849-65360	503	11
CYP51A1	aa_1071	30313-33814	500	7

Supplementary Table 2.2. Incomplete cytochrome P450s in *Capitella teleta*.

Temporary names are based off the scaffold they were found on. The listed CYPs are not full length and are missing exons but have EST support.

Temporary Name	Region on Scaffold	Number of exons
I_51a	113563-114816	1
I_91a	178629-180233	7
I_145c	321636-322748	1
I_226	133627-136799	10
I_271b	279012-281304	5
I_396	103216-106894	5
I_446	96010-97844	10
I_458b	77247-78132	2
I_520	835384-85469	3
I_881	61663-63036	2
I_897	34082-35605	1
I_3603	6001-7564	3

Supplementary Table 2.3. Cytochrome P450 fragments in *Capitella teleta*.

Temporary names are based off the scaffold they were found on. None of these fragments have EST support, except for p_342, suggesting they may be pseudogenes. P_342 had an early stop codon and is a pseudogene.

Temp Name	Region on Scaffold
P_5	212521-213095
P_342	50262-54509
P_371	38074-39314
P_720	57387-56719
P_1095	8791-7856
P_2211	8299-8873
P_5575	3510-4550
P_7309	871-13
P_8508	5096-4716, 3186-2401
P_10760	1-306
P_10990	334-1402
P_16088	846-2089
P_36404	2-505

Chapter 3: General Discussion

CYPome Annotations

The annotations of the *Capitella teleta* (Chapter 2) and sand goby (*Promatoschistus minutus*, Appendix 1) CYPomes were completed in different ways because the genome of each species had different data and limitations. The maturity of the genome assembly, size of the EST database and availability of well curated CYPomes in closely related species are three important parameters influencing the limitations in annotating a new CYPome. The quality of each parameter determines the methods used for the annotation as well as the overall integrity of the final annotation.

Species with a mature genome assembly with very large scaffolds, or even entire chromosomes, reduces the chance that a gene will be located on the end of a scaffold. Genes located on the end of a scaffold are often truncated and confidence in the annotation is lower when a gene spans scaffolds or contigs. The sand goby genome (Appendix 1) had a genome assembled into contigs with an average contig size of 1534bp. Though some of the larger contigs contained full length CYPs, the majority of the CYPs had exons on several different contigs; this made the annotations a very laborious process as each exon needed to be assembled and inspected carefully to ensure that exons were not mismatched across closely related homologs. Since the *C. teleta* genome was more mature and the scaffolds were large there were only a few cases where a CYP was on a scaffold end. Mature genomes may have fewer unknown regions in the

assembly; unknown regions can interfere with annotation. Both genomes had unknown regions that interfered with several annotations.

CYPomes from closely related species were very helpful in annotation. For the sand goby annotations, well annotated CYPomes from zebrafish and fugu (Goldstone et al., 2010; Nelson, 2003) were available. Multiple fish CYPomes provided more confidence in the expected CYP families and subfamilies and a good estimation for the overall number of CYPs expected in the sand goby CYPome. Homology searches with the zebrafish and fugu sequences provided coverage over each exon, even with the less sensitive BLAST algorithm. More rigorous sequence similarity cut offs can be applied during gene searching strategies if there are sequences from closely related species. Most importantly, however, is the conservation in exon boundaries between species. The exon boundaries are very similar between humans and fugu (Nelson, 2003), suggesting that species within the same phylum may be sufficiently related to provide useful information on exon boundaries. Since there were other bony fish species available for the sand goby annotations, the exon boundaries were trivial to assemble in sand goby. The amino acids near the boundaries were highly conserved and the splice site locations were almost always identical. In contrast, *C. teleta* had no closely related species with a complete CYPome as it was the first CYPome annotated in a lophotrochozoan. The exon boundaries and number of exons proved to be very different from the vertebrate sequences available and used for annotations. For example, CYP51 in vertebrates had 10 exons, but *C. teleta* had 7 exons.

An EST database is helpful for the automated annotation of CYPomes. *C. teleta* had an EST database of 138,404 reads. This EST database was used to find regions on the genome that were transcribed, so when ESTs were aligned to a genome, I could use these as indicators of gene locations. These gene locations were matched with the homology hits and determined to be potential CYPs. With a large EST database, theoretically all expressed genes could be located and exon boundaries can be determined with high precision. Since *C. teleta* does not have closely related species with a CYPome, exon boundaries were based off of the ESTs and homology. When sequence similarity between the *C. teleta* and the reference sequence was too low, and there did not seem to be any gaps that could be fixed by shifting exon boundaries, the boundaries determined by the ESTs were kept to avoid bias. Without ESTs, annotation of *C. teleta* exon boundaries would be very difficult to complete with accuracy. There would be many more missing exons without EST data assisting in extending the BLAST hits. ESTs were also useful to determine whether some of the partial sequences or fragments were likely functional genes. Any partial sequence or fragment with EST support was considered a functional gene, unless there was a premature stop codon. Lack of EST support for a partial sequence or fragment must be interpreted with caution. The *C. telata* EST library is small and without extensive EST library databases, it is equally plausible that the data is simply missing from the library at this time.

C. teleta had a relatively small EST database. With 138,404 reads and 32,415 predicted genes on JGI, coverage was on average just over 4 ESTs per gene. It is hard to determine how large an EST database needs to be before saturation is reached. Data from

the human genome, where the EST libraries are very large, suggests that >80% of all known genes in genomic sequences at least 5 kb long can be identified by aligning ESTs (Bailey et al., 1998). Approximately 1.4 million EST sequences were for this study (Bailey et al., 1998), a size that is an order of magnitude larger than the number of ESTs available in *C. telata.* Currently, zebrafish had 1.8m ESTs and humans had 8.9m ESTs available on NCBI.

The sand goby genome did not have any EST data available. Since there are closely related species to help annotate the CYPome there is little need to use ESTs to help annotate exon boundaries, though they will help to confirm the assembly of data from multiple contigs into one gene. The ESTs will be useful in addressing the incomplete CYPs by aligning and extending the annotations created through homology.

Each of the CYPomes annotated in this study have their own shortcomings. The confidence in the exon boundaries of the genes from the *C. teleta* genome was weaker than with the sand goby CYPome. This is primarily because there is much lower homology between the *C. telata* CYPome and the CYPome from other species and because the EST database is modest. If the *C. telata* EST database expands or if the CYPs are cloned and sequenced in functional studies, the annotations will be become more reliable. The sand goby genome was primarily limited by the small contig size, which is the primary cause of the incomplete annotations of CYPs. Either an improved genome assembly or an EST database will help to complete this CYPome.

CYP Nomenclature based on Phylogenetic Support

Nomenclature of CYP genes is typically based on amino acid sequence identity. Yet, this is inherently problematic if the species being annotated is divergent from those containing the known genes in the superfamily. Because of the saturation of CYP families in vertebrates, naming newly annotated sequences in any vertebrate becomes trivial (Nelson, 2011). There has been only one new CYP family found in vertebrates in 2011, CYP16, the first in 11 years, and only rarely are there new subfamilies (Nelson, 2011). It is easy to name these sequences from sequence identity alone. Outside of vertebrates there is much less sequence information available and when CYPomes are analysed, there is limited data from closely related species. During the *C. telata* annotations, there were only a few CYP sequences available that fell in the 40% and 55% cut offs for naming. There are a few molluscan CYP sequences, such as the CYP10s, but these were not yet publicly available (David Nelson, personal communication) although they helped in final nomenclature assignment.

One big questions remains then, how to apply nomenclature rules when the similarity to known sequences do not meet the guidelines? This is where phylogeny may play an important role in nomenclature. First, all sequences that cluster within the small clans are examined, particularly clans 19, 20, 26, 46 and 51. These clans are expected to have a small number of CYPs in each species, divided into a few families, and *C. teleta* follows this trend. *C. teleta* CYP20A1 only has up to 45% identity with other CYP20A1s but was still named so because its phylogeny infers orthology, with high bootstrap

support to other CYP20A1s including BLAST support from other invertebrate (*H. robusta,* amphioxus and *S. purpuratus*) CYP20A1s. Once the smaller clans have been named, the larger clans are examined. The larger clans are more difficult to determine if there is orthology down to the family level because the phylogeny is more complex, so all sequences that do not pass the 40/55% cut off and do not cluster clearly with other sequences will be placed into new families. Clan 2 *C. telata* sequences had no clear orthology with existing clan 2 CYP families. There were two sequences that clustered outside CYP1s, similar to the phylogenetic arrangement of *S. purpuratus* 'CYP1-like' genes(Goldstone et al., 2006). Another group of sequences clustered with CYP2s in the large phylogeny (Figure 1, Chapter 2), yet were more closely related to *C. elegans* sequences in the clan 2 phylogeny (Figure 2A, Chapter 2). Thus, none of these sequences could be placed in the CYP1 or CYP2 families and were given novel families. Interestingly, in clan 4, CYP4V25 had only ~48-52% identity with other CYP4Vs but was still placed in the CYP4V family based on 3 main points. (1) CYP4V25 clusters only with other CYP4Vs on the phylogenies. (2) All the top BLAST hits are CYP4Vs with an obvious drop in identity with other families. (3) There have been CYP4Vs found in other invertebrates including molluscs and arthropods (David Nelson, personal communication).

Synteny is helpful when looking for duplicated and homologous genes during annotation. Synteny works well when there are other related genomes, generally in the same class and diminishes drastically at the phyla level (Nelson, 2011). Insects have highly rearranged genomes, even compared to each other, so synteny may not be helpful

at the class level in invertebrates (Nelson, 2011). Syntenic regions were searched for around a few sample CYPs in *C. teleta* but no regions were found with species with annotated CYPomes. This may not be surprising since this synteny was examined with species outside lophotrochozoans. Synteny may be more helpful in naming other lophotrochozoan CYPomes such as *Lotta gigantea* and *H. robusta*, which are now available on JGI.

CYPs in Metazoans

In my *C. teleta* CYPome study there were 83 full length CYPs annotated with 23 new families. A majority of these new families were in clan 2. *C. telata* sequences in Clan 3 were from two families, both were novel. The abundance of new families in these clans means that there is very poor sequence saturation across metazoans (Baldwin, Marko, and Nelson, 2009; Nelson, 2011; Tarrant et al., 2009; Tijet et al., 2001) in these clan in particular and many more CYPomes will be necessary before we start seeing any saturation of families. All clan 4 sequences were in the CYP4 family, which was the same in *Daphnia* (Baldwin et al., 2009), but is expanded in *Drosophila* (Tijet et al., 2001). The expansion of clan 4 in insects but not in *Daphnia* (Baldwin et al., 2009), *N. vectensis* (Goldstone, 2008), or *C. teleta* suggests that future expansion of clan 4 may be restricted to insects.

Future CYPome analyses should focus on Platyzoa since there are no named sequences in this superphyla. There have been initial searches done in *Trichoplax adhaerens* utilizing the JGI pipeline by David Nelson and he has posted the results on the

Cytochrome P450 webpage (Nelson, 2009). Clan 3, 4, and mitochondria were all expanded in this species, with 46% of the annotated *C. teleta* CYPome in clan 3, 25% in clan 4 and 20% in the mitochondrial clan. Without further analysis, such as sequence identity comparisons and phylogeny, we will not know whether these *T. adhaerens* CYPs fall into known families. Yet CYPomes from other invertebrate species indicate that when there are large clusters of CYPs, it is likely that these clans will have many new families (e.g. *D. melanogaster* clan 4 in insects; Tijet et al., 2001; and clan 2 in *C. teleta*).

With the *C. elegans, D. pulex and D. melanogaster* CYPomes in Ecdysozoa along with many other insect CYPomes available in varying degrees of completeness on the CYP webpage (Nelson, 2009) there is a fair amount of data available, especially in insects. A more comprehensive examination of Ecdysozoan CYPomes may now be possible. Sequence availability in lophotrochozoa remains minimal and CYPome data in a molluscan would be particularly helpful since there is much interest in molluscan xenobiotic metabolism and their ability to create steroids similar to vertebrates (Fernandes, et al., 2011). The *Aplysia californica* CYPome is being analyzed (Jed Goldstone, personal communication) and may provide important data for understanding lophotrochozoan CYPomes. Of interest is to determine where the clan expansions occurred on the tree of life and how clans arose and diversified from fungi CYP51s (Nelson, 1999).

Phylogenetics as a Tool for Raising Functional Hypotheses

In this study, phylogenetic trees were constructed so the evolutionary relationships may examined between the *C. teleta* CYPs and with CYPs in other species. The evolutionary relationships allow us to think of how relatedness, or position on a phylogeny, may provide evidence for function. CYPs responsible for metabolism of endogenous substrates are generally single copy and cluster alone on a tree. Several clans are restricted to endogenous substrate metabolism (i.e. have no known role in xenobiotic metabolism). Therefore, the phylogenetic placement of CYP3052A4, CYP39B1, CYP51A1, and CYP20A1 on the large phylogeny (Figure 2 in Chapter 2) suggests that they may be involved in metabolism of endogenous substrates since they are single copy and are also the only CYPs in their clan (with exception to the two sequences in clan 7, but they cluster separately). Beyond this, CYPs that cluster with well known sequences, such as CYP51A1, have better support for raising functional hypotheses based on phylogenetic placement.

CYPs involved in exogenous metabolism are generally found in clans with a high copy number in large cluster in the phylogenies. Two good examples are the large groups in clan 2 and 3. Clans 1-4 are well known for their roles in exogenous metabolism in vertebrates and insects (Lewis, 2004) and when there are large clusters of newly identified sequences found in these clans, we typically would hypothesize that they are involved in exogenous metabolism to some degree.

In Silico Protein 3D Analyses as a Tool for Raising Functional Hypotheses

The most powerful *in silico* method for predicting function of proteins is 3D analyses. The most reliable way to fold proteins is to use homologous templates to start a rough fold (threading) and then finding the lowest energy structure (refining; Zhang, 2008b). With a 3D structure, we can analyze the similarity of the folded protein to the overall structure of other CYPs and, more importantly, the conformation of the substrate binding site.

There are two approaches to analyzing the binding site of a protein. First, the folded protein can be compared to other known CYPs and the most similar binding sites determined, as with COFACTOR (Roy, et al., 2012). The search set for COFACTOR is the entire Protein Data Bank (PDB; Bernstein et al., 1977; Gille and Frommel, 2001; Roy et al., 2012). PDB has 549 CYP structures available, of which 163 are eukaryote CYPs (Bernstein et al., 1977). Second, *in silico* docking of candidate substrates can determine how well they fit the binding pocket, measured in binding free energy, as with Auto-Dock(Trott and Olson, 2010). The advantage of looking for the most similar binding sites is that you do not need to choose a particular substrate. While the substrate choice may be obvious for *in silico* docking of some proteins (i.e. lanosterol for putative CYP51A1 genes), it is not clear which substrates or how many substrates should be selected for CYPs thought to be involved in xenobiotic metabolism. Instead, identification of a similar binding pocket may provide broader functional information to help define substrates for *in silico* docking. Molecular docking is a laborious process when many substrates are tested per protein as each must be docked individually (Trott and Olson, 2010) and may

be docked in hundreds to thousands of different combinations of models and docking poses (Prasad, et al., 2007).

The aromatase docking study is an example of testing a CYP involved in endogenous metabolism (Callard et al., 2011). A candidate aromatase gene was found in amphioxus and in order to determine whether androstenedione bound to the candidate gene they folded the protein using MODELLER, which utilizes threading and refinement (Eswar et al., 2006). After a 3D structure was predicted they used Auto-Dock to dock androstenedione to the model binding pocket 100 times using a genetic algorithm for conformational sampling on each run (Callard et al., 2011; Morris et al., 2009). Binding energy of androstenedione to the amphioxus protein model (-10.7 kcal/mol) was very similar to human (-11.3 kcal/mol) suggesting that androstenedione binds with similar affinity between the two proteins (Callard et al., 2011).

Modeling and docking of 3,3',4,4'-tetrachlorobipohenyl, 2,3,7,8-tetrachloro-dibenzo-*p*-dioxin, and benzo[*a*]pyrene has been examined for vertebrate CYP1 enzymes (Prasad et al., 2007). Many models and conformations were tested for each protein allowing for a larger distributions of potential interactions during docking than the aromatase study (1400 compared to 100; Callard et al., 2011; Prasad et al., 2007). Testing a wider array of different docking conformations in proteins that have flexible binding sites, such as those involved in xenobiotic metabolism, is important because this captures the flexibility of the binding pocket. The disadvantage of binding a wide array of conformations is that it is more intensive and requires much greater computational time.

We have generated preliminary data on active site topology for the C. *teleta* CYPs annotated in Chapter 2. CYP51A1 and CYP376A1 were examined using the I-TASSER protein folding suite (Roy, et al., 2010). This approach used homology to initially fold the protein and then refined the model's global topology and removed steric clashes; the lowest energy states were chosen for the final models. This approach was completed twice for each protein, utilizing all homologous sequences as templates for folding, and a benchmark run that only used templates with <30% identity. In the case of CYPs, the templates used in a benchmark run would be in different families and likely have very different functions providing less bias in the initial homology folding stages. After a 3D model was generated it was passed through the COFACTOR server (Roy et al., 2012; Zhang, 2008a), which compares the overall structure to known proteins, predicts overall similarity and predicts binding ligands based on binding site similarity. CYP51A1, of all proteins on PDB, had the most similar substrate binding site to mammalian CYP51A1 from both I-TASSER top models (regular run, and benchmark) with no hits outside other CYP51A1s. While CYP376A1 was most similar in 3D structure to CYP11A, CYP11B, but no similar binding sites with other proteins were found. The functional implications of this data are discussed in the section below.

C. teleta is the first protostome analyzed to have representation in clan 46. CYP46 is the only clan 46 CYP gene in vertebrates and functions as a cholesterol 24-hydroxylase in the brain (Nebert, et al., 2013). Since *C. teleta* CYP3070A1 had only 35% identity with human CYP46A1, it is difficult to predict whether the function is conserved in the *C.*

telata ortholog. *In silico* molecular docking of the protein may help support or refute the possibility that cholesterol is a substrate of CYP3070A1.

There are many limitations when doing *in silico* protein modeling on the *C. telata* CYPs identified in this study. The *C. teleta* annotations will need more EST data or other closely related species CYPomes to support the exon boundaries. There are inevitably errors in some of the annotations, and these errors may impact protein models. The errors in the annotations and the error prone nature of *in silico* folding reduce the reliability of the final 3D structures and may reduce the reliability of any future docking studies. This does not mean this type of analysis should not be done, but rather it should be looked at critically and emphasised as predictions for future *in vitro* and *in vivo* studies.

Future work in this area should focus on two general categories of CYPs in *C. teleta*, CYPs that are thought to have a role in (1) endogenous substrate metabolism and (2) xenobiotic metabolism.

(1) Of those thought to be involved in endogenous substrate metabolism, protein folding studies on CYP51A1, CYP376A1, CYP371B1 and CYP3064A1 are most important. Cholesterol docking studies should be tested in CYP51A1 as this is its function in vertebrates (Goldstone et al., 2010; Lewis, 2004) and the protein folding work completed so far supports this functional hypothesis(Yoshida, et al., 2000). The remaining three proteins should be tested for several potential steroidogensis substrates (e.g. CYP376A1, CYP371B1 and CYP3064A1) since function is much harder to predict in these proteins.

(2) There are many CYPs, phylogenetically placed with CYP families known for xenobiotic metabolism in Chapter 2. Initially, one CYP from each subfamily could be analysed and depending on the diversity of predicted ligands more may be done. Testing one CYP per family/subfamiliy will reduce the number of sequences from 62 to 26 sequences between clan 2 and 3. Using COFACTOR, as well as binding pocket amino acid similarity to other known CYPs, a short-list of potential ligands may be possible to identify for each CYP for ligand docking. Initial tests may be done on the CYP1-like *C. teleta* CYPs to gauge the feasibility (quality of findings and timeframe) of this experiment.

The docking studies will be a start in understanding the function of these CYPs but *in silico* protein studies are only predictions and the most interesting findings will have to be followed up with *in vivo* and *in vitro* studies.

Steroidogensis in Annelids

Functional studies have identified annelids as capable of steroidogenesis, including the production of estradiol (see review in Chapter 1), though which enzymes are responsible have yet to be discovered. CYP51A1 (lanosterol 14 alpha-demethylase) is well known to be involved in early stages of steroidogensis (lanosterol to cholesterol) in a wide range of metazoan species. CYP51A1 was tested *in silico* using I-TASSER and COFACTOR (described above). CYP51A1 had the highest similarity and similar binding ligands, lanosterol, as vertebrate CYP51A1 in both models. This is not surprising because of the high identity with other CYP51A1s in vertebrates (>60%), the clear orthology to

vertebrate CYP51A1 in the phylogeny, and how the function of CYP51 genes across metazoans has typically been conserved. It is very likely that CYP51A1 functions as a lanosterol 14 alpha-demethylase enzyme in *C. teleta*.

Finding other enzymes responsible for steroidogenesis is more difficult. The next step is the synthesis of pregnenolone from cholesterol which is completed by CYP11A1 in vertebrates (Baker, 2011). CYP376A1 phylogenetically clustered with the CYP11s (with poor bootstrap support, only 39/100), which includes CYP11A and CYP11B1,the enzyme responsible the synthesis of cortisol in vertebrates (Baker, 2011). CYP11B1 function can be ruled out since there is no evidence of cortisol production in *C. teleta*. Interestingly, CYP376A1 had the highest sequence identity to vertebrate CYP27A, which is involved in bile acid biosynthesis (Goldstone et al., 2010), but when CYP376A1 was folded, its overall structure was most similar to CYP11A1, CYP11B2 and even CYP24A1 but not CYP27A. COFACTOR could not find any binding sites on PDB that were sufficiently similar to the CYP376A1binding site. Thus, evidence for cholesterol as the putative substrate of CYP376A1 is not clear; docking of cholesterol, and other steroid precursors will likely help clarify which substrate is the best candidate(s) for functional testing.

The next step in steroidogenesis is DHEA synthesis, which is completed by CYP17A in vertebrates. CYP17A is located in clan 2 and there are many *C. teleta* CYPs located in this clan but none were clear orthologs of CYP17As. If we narrow down candidates based on copy number and divergence in clan 2, our best candidates are CYP3064A1 and CYP3057A1 which group distantly outside of other clan 2 sequences.

CYP3064A1 and CYP3057A1 group closest with both CYP17A and CYP21A (which cluster together, Figure 2A, Chapter 2). CYP21A is involved in cortisol synthesis in vertebrates (Baker, 2011), so the *C. teleta* CYPs are not expected to have this function. The last CYP involved in the synthesis of estradiol in chordates is CYP19A, aromatase. There are no particularly good candidate for aromatase in *C. teleta* because there is no representation in clan 19. Thus, any low copy CYP in this annotation could be considered equally likely as candidate genes for this function. Yet, it may be possible to fold and dock a single androgen (e.g. androstenedione) in all single copy CYPs from this genome to determine a short list of candidate genes for aromatase function.

References

Bailey, L. C., Searls, D. B., and Overton, G. C. (1998). Analysis of EST-driven gene annotation in human genomic sequence. *Genome Research, 8*(4), 362-376.

Baker, M. E. (2011). Origin and diversification of steroids: Co-evolution of enzymes and nuclear receptors. *Molecular and Cellular Endocrinology, 334*(1–2), 14-20. doi: 10.1016/j.mce.2010.07.013

Baldwin, W., Marko, P., and Nelson, D. (2009). The cytochrome P450 (CYP) gene superfamily in daphnia pulex. *BMC Genomics, 10*(1), 169. doi: 10.1186/1471-2164-10-169

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., . . . Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology, 112*(3), 535-542.

Callard, G. V., Tarrant, A. M., Novillo, A., Yacci, P., Ciaccia, L., Vajda, S., . . . Cotter, K. A. (2011). Evolutionary origins of the estrogen signaling system: Insights from amphioxus. *The Journal of Steroid Biochemistry and Molecular Biology, 127*(3–5), 176-188. doi: 10.1016/j.jsbmb.2011.03.022

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M., . . . Sali, A. (2006). Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics, ,* 5-6.

Fernandes, D., Loi, B., and Porte, C. (2011). Biosynthesis and metabolism of steroids in molluscs. *The Journal of Steroid Biochemistry and Molecular Biology, 127*(3–5), 189-195. doi: 10.1016/j.jsbmb.2010.12.009

Gille, C., and Frommel, C. (2001). STRAP: Editor for STRuctural alignments of proteins. *Bioinformatics (Oxford, England), 17*(4), 377-378.

Goldstone, J. V. (2008). Environmental sensing and response genes in cnidaria: The chemical defensome in the sea anemone nematostella vectensis. *Cell Biology and Toxicology, 24*(6), 483-502.

Goldstone, J. V., McArthur, A., Kubota, A., Zanette, J., Parente, T., Jonsson, M., . . . Stegeman, J. (2010). Identification and developmental expression of the full complement of cytochrome P450 genes in zebrafish. *BMC Genomics, 11*(1), 643. doi: 10.1186/1471-2164-11-643

Goldstone, J. V., Hamdoun, A., Cole, B., Howard-Ashby, M., Nebert, D., Scally, M., . . . Stegeman, J. (2006). The chemical defensome: Environmental sensing and response genes in the strongylocentrotus purpuratus genome. *Dev Biol, 300*, 366-384. doi: 10.1016/j.ydbio.2006.08.066

Lewis, D. F. V. (2004). 57 varieties: The human cytochromes P450. *Pharmacogenomics, 5*(3), 305-318. doi: 10.1517/phgs.5.3.305.29827

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry, 30*(16), 2785-2791.

Nebert, D. W., Wikvall, K., and Miller, W. L. (2013). Human cytochromes P450 in health and disease. *Philosophical Transactions of the Royal Society B: Biological Sciences}, 368*(1612) doi: 10.1098/rstb.2012.0431

Nelson, D. R. (2009). The cytochrome P450 homepage. *Hum Genomics, 4*(1), 59-65.

Nelson, D. R. (2011). Progress in tracing the evolutionary paths of cytochrome P450. *Biochimica Et Biophysica Acta.Proteins and Proteomics, 1814*(1), 14-18.

Nelson, D. R. (1999). Cytochrome \P450\ and the individuality of species. *Archives of Biochemistry and Biophysics, 369*(1), 1. doi: http://dx.doi.org/10.1006/abbi.1999.1352"

Nelson, D. R. (2003). Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. *Archives of Biochemistry and Biophysics, 409*(1), 18-24. doi: 10.1016/S0003-9861(02)00553-2

Prasad, J. C., Goldstone, J. V., Camacho, C. J., Vajda, S., and Stegeman, J. J. (2007). Ensemble modeling of substrate binding to cytochromes P450:â€‰ analysis of

catalytic differences between CYP1A orthologsâ€ ,â€¡. *Biochemistry, 46*(10), 2640-2654. doi: 10.1021/bi062320m

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols, 5*(4), 725-738. doi: 10.1038/nprot.2010.5; 10.1038/nprot.2010.5

Roy, A., Yang, J., and Zhang, Y. (2012). COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research, 40*(Web Server issue), W471-7. doi: 10.1093/nar/gks372; 10.1093/nar/gks372

Tarrant, A. M., Reitzel, A. M., Blomquist, C. H., Haller, F., Tokarz, J., and Adamski, J. (2009). Steroid metabolism in cnidarians: Insights from nematostella vectensis. *Molecular and Cellular Endocrinology, 301*(1â€“2), 27. doi: http://dx.doi.org/10.1016/j.mce.2008.09.037"

Tijet, N., Helvig, C., and Feyereisen, R. (2001). The cytochrome \P450\ gene superfamily in drosophila melanogaster: Annotation, intron-exon organization and phylogeny. *Gene, 262*(12), 189. doi: http://dx.doi.org/10.1016/S0378-1119(00)00533-3"

Trott, O., and Olson, A. J. (2010). AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry, 31*(2), 455-461.

Yoshida, Y., Aoyama, Y., Noshiro, M., and Gotoh, O. (2000). Sterol 14-demethylase P450 (CYP51) provides a breakthrough for the discussion on the evolution of cytochrome P450 gene superfamily. *Biochemical and Biophysical Research Communications, 273*(3), 799-804.

Zhang, Y. (2008a). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics, 9*(1), 40.

Zhang, Y. (2008b). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology, 18*(3), 342-348.

Appendix I: Annotation of the Sand Goby CYPome

Introduction

The sand goby, *Pomatoschistus minutus*, is a recently sequenced teleost fish species under the IMAGO Marine Genome Project (Center for Marine Evolutionary Biology, University of Gothenburg, Sweden). The sand goby is found on European shores (Larmuseau, et al., 2010) particularly in the NE Atlantic and Baltic Sea. This species has exclusive paternal care; males build nests under empty muscle shells that they cover with sand and prepare with mucus for the females to lay a clutch (Jones, et al., 2001). Males are found in two size morphs and reproductive sneaking behaviour is common. Sand gobies are euryhaline and are found throughout the Baltic Sea over large salinity gradients. They are a small, experimentally tractable fish that is an interesting species from reproductive biology, ecology and evolutionary perspectives.

The overall predicted size of the genome is 1Gbp, The current genome assembly (10/10/2012) has a median contig size of 1534bp, with the total assembly size covering just over half of the genome (568Mbp) and 74Gbp of total sequence data, based on a library with a fragment size of 300 bp. The coverage is 53 fold after filtering of the sequencing reads leaving a genome with a contig N50 of 1534bp (CeMEB, personal communication). The coverage over the non-repetitive portion of the genome has not been documented. Further information can be retrieved from: http://www.cemeb.science.gu.se/research/target-species-imago+/pomatoschistus-minutus/. The genome itself is still private, though it has been made available for use in

this study. There is currently no available EST or RNAseq information available for this species, but should be available before the end of 2013 (Ola Svensson, personal communication).

Sand gobies are a well researched species, with diverse studies focused on the sexual behaviour (Forsgren, et al., 1996; Forsgren, et al., 1996; Jones et al., 2001; Nyman, et al., 2006), evolution (Gysels, et al., 2004; Larmuseau, et al., 2010; M. H. Larmuseau, et al., 2010; Pampoulie et al., 2004) and reproductive toxicology (Saaristo, et al., 2010; Waring, et al., 1996) to name a few. The following provides an annotation of the sand goby CYPome based on the current genome assembly.

Methods

Since there was no EST or RNAseq data available for the sand goby genome at the time of annotation, the CYPome was completed entirely through homology with other teleost species. First, the entire CYPome of zebrafish and fugu, two teleost fish with well curated CYPomes and mature genome assemblies (Goldstone et al., 2010; Nelson, 2003), were used in iterative BLAST using the tBLASTn algorithm (Altschul, et al, 1990). All of the contigs that were hit by at least one of the CYPs was recorded and used in reciprocal BLAST to the zebrafish and fugu CYPomes to record the best hits for each fragment. Each fragment was sorted by family, and when possible, into subfamilies. Since the contig sizes were small compared to the expected gene size (CYPs average 1500bp for the coding sequence), a single CYP gene was often split across multiple contigs. For CYP families or subfamilies with a single gene, this did not present a problem. However, with

CYP families or subfamilies with multiple genes, it was more difficult to determine which fragment belonged to a specific CYP gene. Homology with the CYPs from the other fish CYPomes became very important for annotating these genes. Exon length, exon phase, and amino acid conservation near the exon boundaries were used to determine CYP identity and for correct annotation of exon boundaries. FASTA (Pearson and Lipman, 1988), since it is more sensitive than BLAST, was used to find missing exons that were expected to be on a contig due to inferred intron size but was not detected using BLAST due to poor homology. Visualization, adjustment of exons, and creation of 'embl' files was completed using Artemis (Rutherford et al., 2000).

Results and Discussion

The tBLASTn searches using zebrafish and fugu CYP genes provided a large number of high scoring pairs (HSPs) with varying sizes as shown in Figure 1. The median of ~60aa is expected, as this is roughly the most common exon size from the teleost fish used. Thus, most HSPs were thought to identify a single exon of a CYP gene. There were several very large hits that seem to contain whole CYPs (HSP size >400aa, Figure 1); these genes are CYP1C, and CYP8B, which only have one exon in teleost species (Goldstone et al., 2010; Nelson, 2009).

The sand goby CYPome coverage is shown in Table 1. There were 12 fully annotated CYPs, including two CYP1s, two CYP2s, CYP3B, CYP7A, CYP8B, CYP20 and CYP27C genes. There were 21 annotated CYPs with missing segments, generally full exons. The missing regions are likely due to the small contig sizes found in the current

assembly. Some exons were difficult, and in many cases impossible, to find; these exons may have spanned across contigs, having small amounts of sequence on each and were not detected during homology searches. Overall, the sand goby CYPs have nearly identical exon boundaries with very high sequence similarity and exon phase to zebrafish or fugu CYP genes, which increased the confidence in boundary annotations. Many fragments of CYPs, including exons and parts of exons were not assembled into genes and properly annotated because of their high sequence identity to multiple zebrafish or fugu CYP genes. Typically, these were in CYP subfamilies where multiple genes were expected. Several of the CYP2 family genes could not be annotated fully; however, the individual exons were collected and a rough gene counts were deduced. There were exons identified from the CYP7C, CYP8A, CYP27A and CYP46 families, with an inferred total of 17-19 fragmented CYPs. With the assembly in its current condition it is very difficult to determine whether these fragments are genes, with missing exons, or pseudogenes. Better scaffolding of the genome and EST data will be pivotal tools to improving the annotations of the sand goby CYPome, determining which fragments are pseudogenes and resolving partial gene sequences identified through this annotation process.

The sand goby CYPome has approximately 50-52 sequences when including the partial and fragmented annotations. This is comparable to the fugu CYPome, which has 53 CYPs (D. Nelson, 2009; D. R. Nelson, 2003), but is markedly lower than the zebrafish CYPome of 94 CYPs(Goldstone et al., 2010). Known vertebrate CYPomes range from 43 CYPs in *Gallus gallus* to 103 CYPs in *mus musculus* (Nelson, et al., 2013). The zebrafish, similar to the mouse, has some families (e.g. CYP2s, CYP4s) where there are species or

lineage specific amplifications of genes, which accounts for the larger CYPomes in these species (Goldstone et al., 2010; Kirischian and Wilson, 2012).

The sand goby CYPome includes 17 families; a similar family composition as fugu and zebrafish (Goldstone et al., 2010; Nelson, 2003). The only significant difference between the sand goby and fugu CYPomes is the reduction of the CYP3 family. Fugu has five CYP3s, three 3CYPAs, and 2 CYP3Bs, while only one gene in each family of the sand goby genome was identified. Neither fugu nor sand goby have a CYP39, which is commonly absent in fish (D. R. Nelson, 2003). CYP39 converts cholesterol to bile acids in vertebrates (Goldstone et al., 2010) though CYP7A1, present in sand goby, has the same function (Nelson, 2003). The zebrafish CYP2 family is widely expanded, having over triple the number of CYP2 genes compared to either sand goby and fugu (Goldstone et al., 2010; Nelson, 2003). Interestingly, the only subfamily that is not found in sand goby, but is found in both fugu and zebrafish is CYP26B (Goldstone et al., 2010; Nelson, 2003).

The annotation of the CYPome from sand goby is now partially complete. We have annotated 12 full length CYP genes, 21 partial genes, and 27 gene fragments. The confidence in the full length gene annotations was high, as these genes were highly similar in sequence, exon number, exon boundary location, and exon boundary sequences, when compared to other teleost CYP gene sequences. The confidence in the partial gene sequences was likewise high; these genes were typically missing only 1-2 exons representing a small portion of the total gene sequence. The gene fragments may

have included some pseudogenes but with an improved genome assembly, and RNAseq data, the annotation of these fragments can be resolved.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2

Forsgren, E., Karlsson, A., and Kvarnemo, C. (1996). Female sand gobies gain direct benefits by choosing males with eggs in their nests. *Behavioral Ecology and Sociobiology, 39*(2), 91-96.

Forsgren, E., Kvarnemo, C., and Lindstrom, K. (1996). Mode of sexual selection determined by resource abundance in two sand goby populations. *Evolution, 50*(2), 646-654.

Goldstone, J. V., McArthur, A., Kubota, A., Zanette, J., Parente, T., Jonsson, M., . . . Stegeman, J. (2010). Identification and developmental expression of the full complement of cytochrome P450 genes in zebrafish. *BMC Genomics, 11*(1), 643. doi: 10.1186/1471-2164-11-643

Gysels, E., Hellemans, B., Pampoulie, C., and Volckaert, F. (2004). Phylogeography of the common goby, pomatoschistus microps, with particular emphasis on the colonization of the mediterranean and the north sea. *Molecular Ecology, 13*(2), 403-417.

Jones, A. G., Walker, D., Kvarnemo, C., Lindström, K., and Avise, J. C. (2001). How cuckoldry can decrease the opportunity for sexual selection: Data and theory from a

genetic parentage analysis of the sand goby, pomatoschistus minutus. *Proceedings of the National Academy of Sciences, 98*(16), 9151-9156. doi: 10.1073/pnas.171310198

Kirischian, N. L., and Wilson, J. Y. (2012). Phylogenetic and functional analyses of the cytochrome P450 family 4. *Molecular Phylogenetics and Evolution, 62*(1), 458-471.

Larmuseau, M. H., Vancampenhout, K., Raeymaekers, J. A. M., Van Houdt, J. K. J., and Volkaert, F. A. M. (2010). Differential modes of selection on the rhodopsin gene in coastal baltic and north sea populations of the sand goby, pomatoschistus minutus. *Molecular Ecology, 19*(11), 2256-2268. doi: 10.1111/j.1365-294X.2010.04643.x

Larmuseau, M. H., Huyse, T., Vancampenhout, K., Van Houdt, J. K., and Volckaert, F. A. (2010). High molecular diversity in the rhodopsin gene in closely related goby fishes: A role for visual pigments in adaptive speciation? *Molecular Phylogenetics and Evolution, 55*(2), 689-698.

Nelson, D. R. (2009). The cytochrome P450 homepage. *Hum Genomics, 4*(1), 59-65.

Nelson, D. R. (2003). Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. *Archives of Biochemistry and Biophysics, 409*(1), 18-24. doi: 10.1016/S0003-9861(02)00553-2

Nelson, D. R., Goldstone, J. V., and Stegeman, J. J. (2013). The cytochrome P450 genesis locus: The origin and evolution of animal cytochrome P450s. *Philosophical*

Transactions of the Royal Society B: Biological Sciences}, 368(1612) doi: 10.1098/rstb.2012.0474

Nyman, A., Kvarnemo, C., and Svensson, O. (2006). The capacity for additional matings does not affect male mating competition in the sand goby. *Animal Behaviour, 71*(4), 865-870.

Pampoulie, C., Gysels, E., Maes, G., Hellemans, B., Leentjes, V., Jones, A., and Volckaert, F. (2004). Evidence for fine-scale genetic structure and estuarine colonisation in a potential high gene flow marine goby (pomatoschistus minutus). *Heredity*, *92*(5), 434-445.

Pearson, W., and Lipman, D. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A, 85*(8), 2444-2448. Retrieved from http://europepmc.org/abstract/MED/3162770

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M., and Barrell, B. (2000). Artemis: Sequence visualization and annotation. *Bioinformatics, 16*(10), 944-945. Retrieved from http://europepmc.org/abstract/MED/11120685

Saaristo, M., Craft, J. A., Lehtonen, K. K., and Lindstrom, K. (2010). An endocrine disrupting chemical changes courtship and parental care in the sand goby. *Aquatic Toxicology*, *97*(4), 285-292.

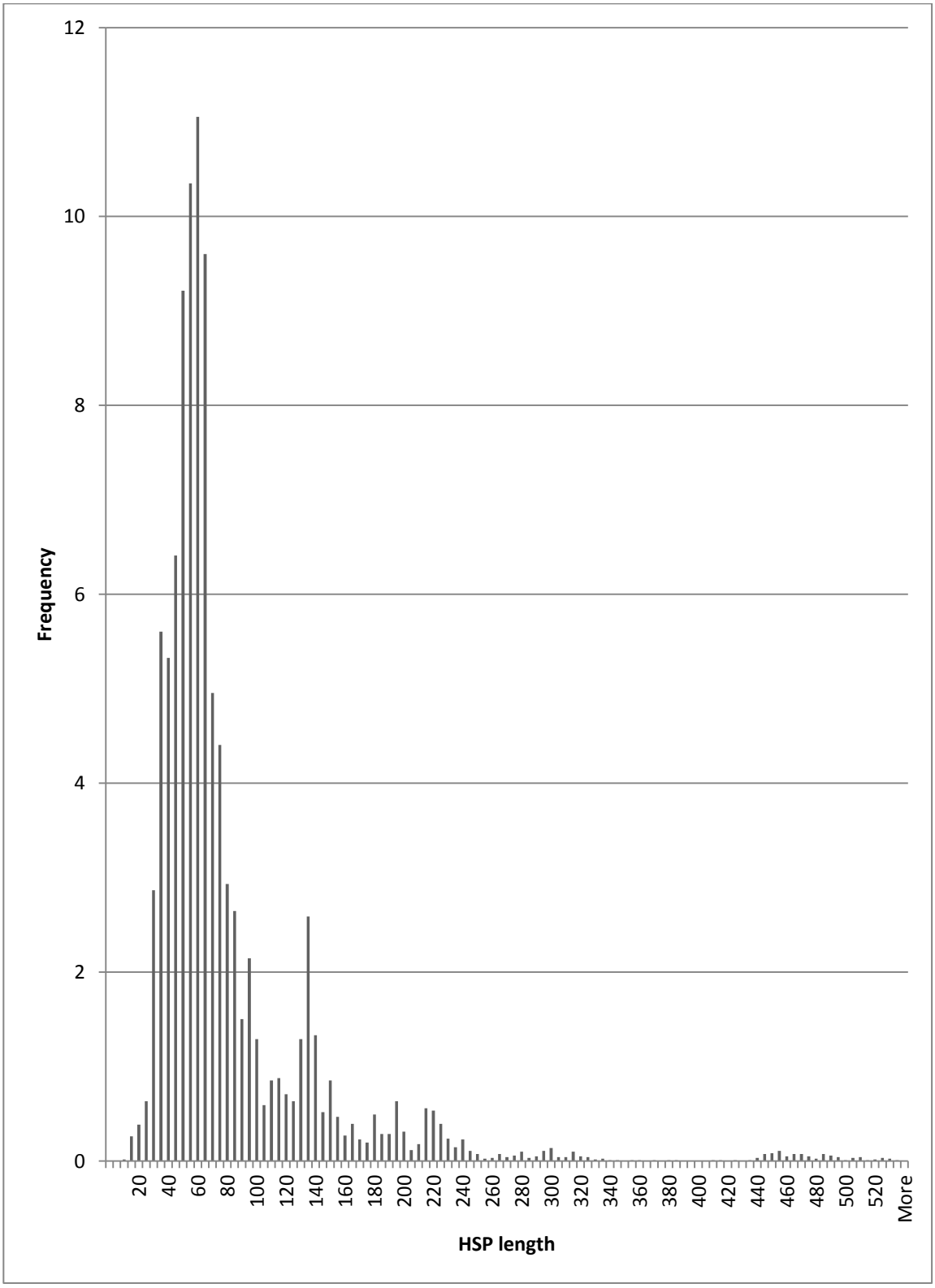
Waring, C., Stagg, R., Fretwell, K., McLay, H., and Costello, M. (1996). The impact of sewage sludge exposure on the reproduction of the sand goby. *Pomatoschistus Minutus,* , 17-25.

Tables and Figures

Table 3.1. The sand goby CYPome composition by CYP gene family. CYP genes were annotated using BLAST, FASTA and a manual curation process (see materials and methods for details). Full length CYPs include those genes with all exons identified in the annotation. Partial CYPs are missing at least part of the gene, typically 1-2 exons. Fragments had high sequence similarity to a zebrafish or fugu CYP but without further annotation; these were typically single exons. The letter represents the subfamily and the bracket contains the number of CYPs found in that family. Partial CYPs are considered estimated numbers only.

CYP family	Full	Partial	Fragment
1	A(1), C(1)	B(1)	
2	R(1), U(1)	N(1), P(1), X(1), Y(2)	K(4), N(2-3), X(3) Y(1)
3	B(1)	A(1)	
4	V(2)	F(1), T(1)	
5	A(1)		
7	A(1)		C(1)
8	B(1)		A(2)
11		A(1), B(1)	
17		A(2)	
19		A(2)	
20	A(1)		
21		A(1)	
24		A(1)	
26		A(1), C(1)	
27	C(1)	B(1)	A(3)
46			A(1-2)
51		A(1)	
Total	**12**	**21**	**17-19**

Figure 1. Histogram of the lengths of high scoring pairs from the sand goby genome. The high scoring pairs (HSPs) were identified in BLAST searches using zebrafish and fugu CYP genes. The length of HSPs is shown in amino acids (aa). Median length is ~60aa with another local peak at ~135aa. Note that there are several HSPs above 400aa in length.


12
10
8
6
4
2
0
Frequency
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
420
440
460
480
500
520
More
HSP length