Comparative Analysis of ABC Transporter Genes in Pathogenic and Non-pathogenic Nematodes

by

Xinyin Zhao

B.Sc, Beijing Forestry University, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Fall 2015

Approval

Name:	**Xinyin Zhao**
Degree:	**Master of Science**
Title:	**Comparative Analysis of ABC Transporter Genes in Pathogenic and Non-pathogenic Nematodes**
Examining Committee:	**Chair:** Dr. William Davidson Professor

Jack Chen
Senior Supervisor
Professor

David Baillie
Supervisor
Professor

Ryan Morin
Supervisor
Assistant Professor

Jonathan Sheps
Supervisor
Research Scientist

By written consultation, Vancouver

Fiona Brinkman
Internal Examiner
Professor
Department of Molecular Biology and Biochemistry

Date Defended/Approved: November 20th, 2015

Abstract

The ATP-binding cassette (ABC) transporter gene superfamily is a large protein family with diverse physiological functions in different organisms. Recent genome sequencing projects have reported expansion of ABC transporter gene family in parasitic nematodes and hypothesized that such expansion may enable the parasites to become pathogenic or have increased virulence. Some of these reported expansions may reflect the completeness of sequenced genomes, use of bioinformatics programs, and parameters and criteria used in these projects. The goal of this thesis research is to develop a robust bioinformatics pipeline for annotating high-quality ABC transporter genes so that we can reduce the contribution of technical errors. Our comparative analysis of 29 nematode genomes suggests that pathogenic nematodes generally contain fewer ABC transporter genes than non-pathogenic nematodes, suggesting that expansion in ABC superfamily may not be a mechanism for pathogenic nematodes to survive in their host environment. However, many pathogenic nematodes have genome-specific ABC transporter genes.

Keywords: Non-pathogenic nematode; pathogenic nematodes; ABC transporter gene; bioinformatics; comparative genomics; phylogeny

Dedication

All my love to my parents, for finding me the light,
whenever it was far away

Acknowledgements

I wish to express my deepest appreciation to my senior supervisor, Dr. Jack Chen, who provides creative ideas and valuable insights for my research. His wisdom and passion for research has influenced me a lot.

My gratitude also goes to my supervisors, Dr. David Baillie, Dr. Ryan Morin and Dr. Jonathan Sheps for providing helpful suggestion and reviewing the thesis. I am also grateful to Dr. Fiona Brinkman and Dr. William Davidson for serving on my examining committee.

A special thanks goes to Dr. Jiarui Li, Zhaozhao Qin, Dr. Maja Tarailo-Graovac, Dr. Christian Frech, Dr. Jeffrey Chu, Dr. Xi Chen, Jun Wang, Timothy Warrington, Shirley Yin for their kind help during my study at SFU.

Moreover, my sincerest gratitude goes to my parents for their endless love and support throughout all these years.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ABC	ATP-binding cassette
AEDs	Antiepileptic drugs
BCRP	Breast cancer resistance protein
BmCN	A strain of *B. mucronatus,* obtained from Zhejiang Province, China
BxC	The R-form *B. xylophilus* strain, obtained from Zhejiang Province, China
BxCA	The M-form *B. xylophilus* strain, obtained from Canada
BxJP	The R-form *B. xylophilus* strain, obtained from Kikuchi group
CF	Cystic fibrosis
CFTR	Cystic fibrosis transmembrane conductance regulator
HUGO	Human genome organization
IVM	Ivermectin
MDR	Multidrug resistance
MHC	Major histocompatibility complex
MOX	Moxidectin
MRP	Multidrug resistance-associated protein
MXR1	Mitoxantrone resistance protein
NBD	Nucleotide-binding domain
NCBI	National center for biotechnology information
PGP	P-glycorprotein
PWD	Pine wilt disease
PWN	Pinewood nematode
PXE	Pseudoxanthoma elasticum
STGD	Stargardt disease
TMD	Transmembrane domain
X-ALD	X-linked adrenoleukodystrophy
XLSA/A	X-linked sideroblastic anemia and ataxia

Glossary

BLASTN

A program that searches nucleotide databases using a nucleotide query.

BLASTP

A program that searches protein databases using a protein query.

Contig

A contig is a set of overlapping DNA segments that together represent a consensus region of DNA (Kutil et al. 2004).

FASTA

FASTA is a suite of programs for searching nucleotide or protein databases with a query sequence (Pearson 1994).

Homolog

Homology is the existence of shared ancestry between a pair of structures, or genes, in different species.

genBlastG

A homology-based gene finders using protein sequences as queries to search for genomic sequences (She et al. 2011).

Inparalog

Paralogs in a given lineage that all evolved by gene duplications that happened after the speciation event that separated the given lineage from the other lineage under consideration (Sonnhammer and Koonin 2002).

InterProScan

InterProScan is a tool that scans given protein sequences against the protein signatures of the InterPro member databases (Jones et al. 2014).

JDotter

A platform-independent Java interactive interface for the Linux version of Dotter, a widely used program for generating dotplots of large DNA or protein sequences (Brodie et al. 2004).

KEGG

Kyoto Encyclopedia of Genes and Genomes is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (Kanehisa and Goto 2000).

MEGA6

A software that contains facilities for building sequence alignments, inferring phylogenetic histories, and conducting molecular evolutionary analysis (Tamura et al. 2013).

Orthologs

Genes in different species that evolved from a common ancestral gene by speciation (Koonin 2005).

OrthoMCL

OrthoMCL is an algorithm for grouping proteins into ortholog groups based on their sequence similarity (Li et al. 2003).

Outparalog

Paralogs in the given lineage that evolved by gene duplications that happened before the speciation event (Sonnhammer and Koonin 2002).

Paralogs

Genes related by duplication within a genome (Koonin 2005).

Pfam

A comprehensive collection of protein domains and families, with a range of well-established uses including genome annotation. Each family in Pfam is represented by two multiple sequence alignments and two profile-Hidden Markov Models (Finn et al. 2014).

Pathogenicity

Pathogenicity is the ability to produce disease in a host organism (Casadevall and Pirofski 1999).

Pseudogene

Pseudogenes, defined as non-functional copies of gene fragments incorporated into the genome by either retro-transposition of mRNA or duplication of genomic DNA, are found throughout the genomes of most eukaryotic organisms (Karro et al. 2007).

RNA-seq

RNA-seq (RNA sequencing), also called whole transcriptome shotgun sequencing, is a technology that uses the capabilities of next-generation sequencing to reveal a snapshot of RNA presence and quantity from a genome at a given moment in time (Morin et al. 2008; Chu and Corey 2012).

TBLASTN

A program that searches translated nucleotide databases using a protein query.

Virulence

This is a quantitative trait, representing the extent of the pathology. Virulence is therefore a trait expressing the interaction between a pathogen and its host (Casadevall and Pirofski 2001).

Chapter 1. Introduction

1.1. Overview of the structure and function of ABC transporter genes

ATP-binding cassette (ABC) transporter genes are also known as ABC systems. They constitute one of the largest gene families in different living organisms on earth (Higgins 1992). ABC transporter genes were first identified and characterized in prokaryotes (Ferenci et al. 1977). The first eukaryotic ABC transporter (P-glycoprotein) was identified in human and it showed high similarity to bacterial ABC transporters. (Gerlach et al. 1986; Gros et al. 1986). All ABC transporters identified so far can be classified into three main functional categories: importers, exporters (Figure 1.1) and non-transporters. Importers are unique to prokaryotes and mediate the uptake of substrates including saccharides, ions, amino acids, peptides, metals, polyamine cations, opines, and vitamins (Davidson et al. 2008). Exporters have been found in all domains of life and they are involved in the secretion of various molecules, such as, lipids, hydrophobic drugs, polysaccharides, and proteins including toxins (Saurin et al. 1999; Davidson et al. 2008). The ABC-containing non-transporters are involved in several non-transport-related processes, such as translation elongation and DNA repair (Chakraburtty 2001; Goosen and Moolenaar 2001; Zhao et al. 2007).

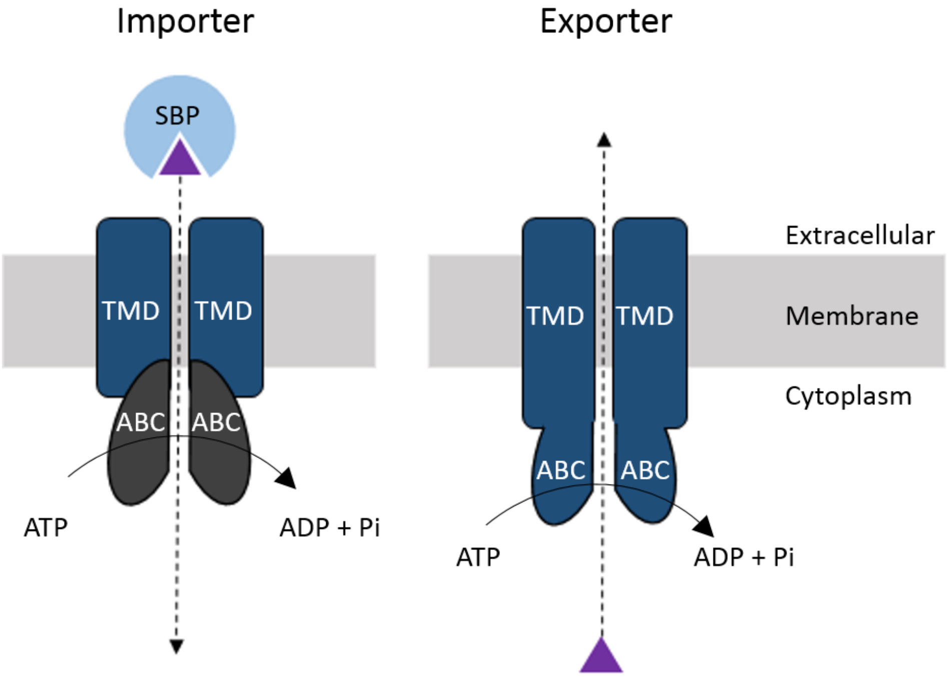


Figure 1.1: Schematic of ABC transporter function
ABC importers transport substrates from extracellular environment to cytoplasm, which requires a substrate binding protein (SBP) that binds and delivers substrates to the TMDs. The ABC domains and TMDs of ABC importer are separate subunits. ABC exporters transport substrates from cytoplasm to extracellular environment. The ABC domains and TMDs are fused to each other in ABC exporters.

All ABC transporters share a highly conserved domain, called ABC domain (also referred to as nucleotide-binding domain [NBD]), which is responsible for coupling transport to ATP hydrolysis. ABC domain is characterized by three conserved motifs, Walker A, ABC Signature and Walker B motifs (Figure 1.2). Walker A and Walker B are indicative of the presence of a nucleotide binding site while the ABC Signature is located between Walker A and Walker B motifs (Schneider and Hunke 1998; Jones and George 2004). An ABC transporter also harbors a transmembrane domain (TMD), consisting of several transmembrane α-helices (in most cases of six membrane spanning α-helices). TMD is responsible for translocating a variety of substrates across cellular membrane (Hyde et al. 1990). In contrast to the high conservation of ABC domain, TMD is loosely conserved. TMDs in different ABC transporters have rather different sequences and lengths. The core functional unit of an ABC transporter constitutes of two ABC domains and two TMDs. This apparent diversification of TMD has been used to explain the ability of ABC transporters to transport diverse substrates (Holland 2011; Kang et al. 2011). The core functional unit of an ABC transporter constitutes of two ABC domains and two TMDs. ABC domains and TMDs of importers are encoded as separate polypeptide chains (Figure 1.1) (Biemans-Oldehinkel et al. 2006). In contrast, ABC domains and TMDs of exporters can be encoded as a single polypeptide, referred to as full transporter, or can be encoded as two separate polypeptides, referred to as half transporters, each containing one ABC domain and one TMD (Figure 1.3). Thus, half transporters require to form either homo- or hetero-dimers to be functional (Kispal et al. 1999; Xu et al. 2004).

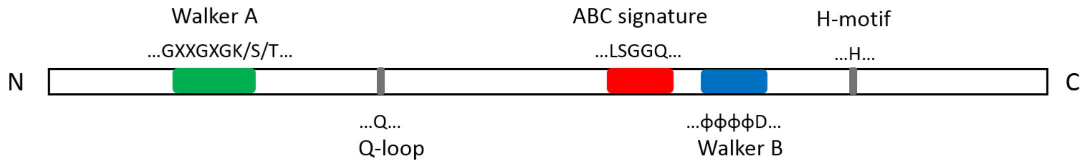


Figure 1.2: A linear representation of ABC domain, illustrating the relative positions of the conserved motifs.

Walker A and Walker B are indicative of the presence of a nucleotide binding site while the ABC Signature, unique to ABC proteins, is located upstream of Walker B and downstream of Walker A. The Q loop is believed to be involved in the interaction of the ABC domain and TMD and the H motif contains a highly conserved histidine residue, functioning in the interaction of the ABC domain with ATP.

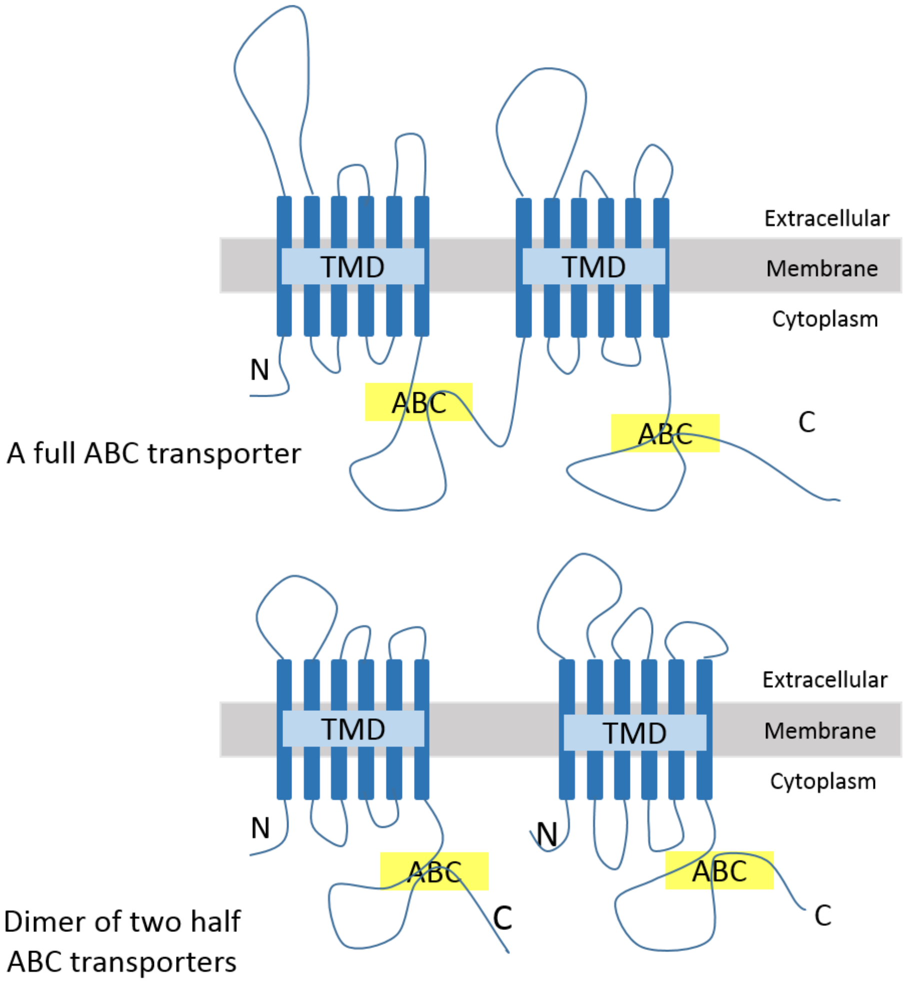


Figure 1.3: A working model of the membrane topology for functional ABC exporters
Exporters can be encodes as a single polypeptide, referred to as full transporter, or can be encodes as two separate polypeptides, referred to as half transporters, each containing one ABC domain and one TMD.

1.2. ABC transporter genes in human

Human Genome Organization (HUGO) classified 49 ABC transporter genes into seven subfamilies (ABCA to ABCG) based on the sequence similarity of their ABC domains (Dean and Allikmets 2001; Dean et al. 2001; Dean 2005) (Table 1.1). Many human ABC transporter genes were identified to be medically relevant (Allikmets et al. 1996; Dean et al. 2001).

Table 1.1: Subfamily information of ABC transporter genes in human.

Domain organization is indicated in the bottom color boxes

Subfamily	Full transporter	Half transporter	Not real transporter
ABCA	12 🟦		
ABCB	4 🟦	7 🟥	
ABCC	13 🟦		
ABCD		4 🟥	
ABCE			1 🟩
ABCF			3 🟩
ABCG		5 🟧	

🟦 TM – ABC – TM – ABC

🟥 TM – ABC

🟩 ABC – ABC

🟧 ABC – TM

1.2.1. Human ABC transporter genes in multidrug resistance and cancer therapy

Multidrug resistance (MDR) is a serious problem that hampers the success of cancer therapy (Chang 2003; Wu et al. 2008). The most common mechanism underlying MDR is the overexpression of ABC efflux transporter genes in cancer cells (Chang 2003; Wu et al. 2008; Choi and Yu 2014). ABCB1 (PGP1/MDR1) was identified to confer a MDR phenotype to cancer cells that had developed resistance to chemotherapy drugs (Kartner et al. 1985). ABCB1 is expressed in the kidney, liver, gastrointestinal tract, and blood brain barrier (Thiebaut et al. 1987; Cordon-Cardo et al. 1989; Schinkel et al. 1997). The likelihood of failure in cancer treatment is increased when the expression of ABCB1 is upregulated during the treatment. About half of human cancers develop MDR because of ABCB1 (Gottesman et al. 2002).

Multidrug resistance-associated protein 1 (MRP1/ABCC1) was first found to be responsible for developing the multidrug resistance phenotype in a drug-selected human lung cancer cell line (Cole et al. 1992). MRP1 is expressed nearly in all kinds of tissues and cell types, with high levels in lungs, testicles, kidney, skeletal and cardiac muscle (Cole et al. 1992; Flens et al. 1996; St-Pierre et al. 2000). Later, MRP1-mediated resistance to drugs was demonstrated in cell lines of various types of solid tumors, such as lung, breast, ovarian, prostate and colon tumors (Hipfner et al. 1999). Recently, researchers found up-regulation of MRP1 is responsible for the resistance of brain cells to antiepileptic drugs (AEDs) in the amygdale kindling rats, suggesting that MRP1 is involved in the mechanism of brain cell resistance to AEDs in refractory epilepsy (Chen et al. 2013).

ABCG is a more recently identified drug transporter. It is also known as breast cancer resistance protein (BCRP) (Doyle et al. 1998) or mitoxantrone resistance protein (MXR1) (Miyake et al. 1999). Unlike ABCB1 and ABCC1, ABCG2 as a half ABC transporter must function as a homodimer or heterodimer (Xu et al. 2004). Similar to ABCB1 and ABCC1, ABCG2 transports a variety of drugs (Allen et al. 1999; Brangi et al. 1999; Maliepaard et al. 1999; Robey et al. 2001; Janvilisri et al. 2003; van Herwaarden et al. 2007) and protects our tissues, such as intestine, placenta, liver, and the blood–brain barrier against various xenobiotics (Sarkadi et al. 2004).

In addition to ABCB1, ABCC1 and ABCG2, other ABC transporters including ABCA2, MDR2, ABCC2, ABCC4 and ABCC11 that were found overexpression in cell lines resistant to drug (Borst et al. 1993; Schuetz et al. 1999; Borst et al. 2000; Turriziani et al. 2002; Mack et al. 2008), suggesting that many ABC transporters have drug resistance capacity.

1.2.2. ABC transporter genes related human genetic disease

In addition to causing drug resistance in cancer, ABC transporters have been found to be responsible for many human diseases when they are mutated. To date, 14 ABC transporter genes have been associated with genetic disorders such as neurological disease, retinal degeneration, cholesterol and bile transport defects, cystic fibrosis, anemia (Klein et al. 1999; Dean et al. 2001; Vasiliou et al. 2009). Mutations in ABCA1 has been identified to be responsible for Tangier disease, a disorder of cholesterol transport between tissues and the liver (Remaley et al. 1999; Rust et al. 1999). Mutations in ABCA4 leads to an accumulation of retinoids in the outer segment or the retinal pigment epithelium, causing a genetic eye disorder called Stargardt disease (STGD) (Maugeri et al. 2000; Battu et al. 2015). Mutations in ABCB2 (TAP1) and ABCB3 (TAP2) have been linked to an immune disorder (a loss of cell surface expression of MHC class I molecules) (Lankat-Buttgereit and Tampe 2002). Mutations in ABCC3 cause a human disorder of organic ion transport called Dubin–Johnson syndrome, an increase of conjugated bilirubin in the serum without elevation of liver enzymes (Toh et al. 1999; Tsujii et al. 1999). Mutations in the ABCC6 (MRP6) have been established as the cause of pseudoxanthoma elasticum (PXE) characterized by soft tissue calcification affecting the skin, eyes and cardiovascular system (Wang et al. 2001; Trip et al. 2002; Ronchetti et al. 2013). Mutations in ABCC7 (CFTR) cause cystic fibrosis (CF), one of the most common fatal childhood diseases in Caucasian populations characterized by abnormal exocrine activity of the lung, pancreas, sweat ducts, and intestine (Figure 1.4) (Dean et al. 2001; Gadsby et al. 2006). Loss-of-function mutations in ABCD1 (ALD) cause a severe neurodegenerative disease, X-linked adrenoleukodystrophy (X-ALD) with accumulation of very long chain fatty acids in organs, serum and central demyelination (Mosser et al. 1993; Pujol et al. 2004). Mutations in half transporters ABCG5 and ABCG8 are associated with sitosterolemia, a disease characterized by defective transport sterols and cholesterol (Gregg et al. 1986; Patel et al. 1998).

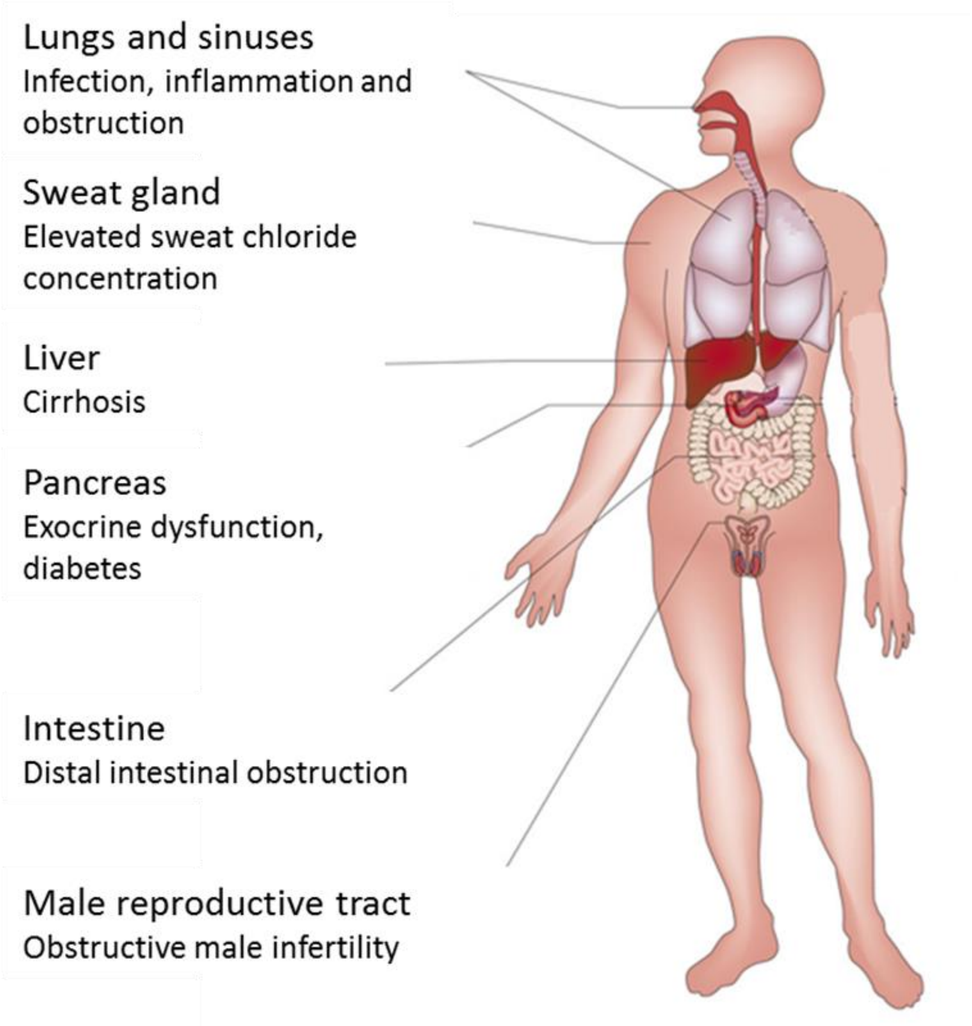

Lungs and sinuses
Infection, inflammation and obstruction
Sweat gland
Elevated sweat chloride concentration
Liver
Cirrhosis
Pancreas
Exocrine dysfunction, diabetes
Intestine
Distal intestinal obstruction
Male reproductive tract
Obstructive male infertility

Figure 1.4: Symptoms of cystic fibrosis

Cystic fibrosis is a human genetic disease caused by the mutation in CFTR. It affects mostly the lungs causing cause obstructions that lead to inflammation, tissue damage and destruction, but also affects the pancreas, liver, kidneys and intestine. Figure obtained from (Cutting 2015).

In summary, ABC transporter genes are involved in diverse biological processes and the mutations of them could cause severe human disease. It suggests that ABC transporter genes are extremely important to living organisms and more efforts should be made to identify ABC transporter genes as well as understand the mechanisms underlying ABC transporter genes in different biological processes.

1.3. ABC transporter genes in other representative organisms

1.3.1. ABC transporter genes in bacteria

The complete genome of *E. coli* K-12 serotype enabled the genome wide identification of ABC transporter genes (Blattner et al. 1997). The collection of 79 ABC transporter genes constitutes the largest family in the *E. coli* K-12 genome, comprising 5% of the total genome (Linton and Higgins 1998).

Pathogenic bacteria exhibit a smaller genome size than free-living bacteria (Ochman and Davalos 2006). Genome reduction can be associated with increased virulence, as many of the most virulent bacterial pathogens have smaller genomes than closely related species (Fournier et al. 2014). There exists a linear relationship between the total number of ABC transporter genes and the size of genome in different prokaryotic organisms. *E. coli* (with a genomes size of 4.6 Mb) and *Bacillus subtilis* (with a genomes size of 4.2 Mb) have 79 and 84 ABC transporter genes, respectively, which are typical numbers for this size of genome. Most bacteria of small size (0.5-1.5Mb) are intracellular parasites, which have rendered some inessential ABC transporter genes, leading to the disruption or deletion of these genes. The remained ABC transporter genes in those intracellular parasites could constitute the minimal requirement of ABC transporter genes for their life. In contrast, bacteria found in soil, such as *Agrobacterium tumefaciens* and *Mesorhizobium loti* (with a genome size of 5.67 and 7.6 Mb, respectively) contain larger numbers of ABC transporter genes (more than 200). This dramatic expansion could be caused by highly competitive environmental conditions that those bacteria have to confront in the soil (Davidson et al. 2008).

1.3.2. ABC transporter genes in yeast

The budding yeast, *Saccharomyces cerevisiae*, was the first organism that had its complete inventory of ABC transporters identified. 30 ABC transporter gene proteins were originally characterized based on homology searches (Decottignies and Goffeau 1997). 28 of these ABC proteins were classified based on their homology to mammalian ABC transporter gene subfamilies, whereas two (CAF16 and YDR061w) failed to be classified into HUGO subfamilies and were therefore categorized as “other” (Paumi et al. 2009). Intriguingly, yeast does not contain any ABC transporter genes in ABCA subfamily. There was an expansion of 10 members in subfamily G in yeast compared to only 5 members in human. Most (nine of 10) genes in subfamily G were full ABC transporters in comparison to only half ABC transporters existing in subfamily G in human genome (Dean et al. 2001), suggesting that ABC transporter genes in subfamily G have been through dramatic changes during evolution.

The functions of many ABC transporters in yeast have been successfully characterized. The only full-size ABC transporter STE6 in subfamily B (the first ABC transporter identified in yeast) is required for secretion of the lipopeptide mating pheromone α-factor, which is essential for mating of haploid yeast cells (Kuchler et al. 1989; Kuchler et al. 1993). The half ABC transporter in subfamily B ATM1 is located in the mitochondrial inner membrane and performs an essential function in the generation of cytosolic Fe/S proteins by mediating the export of Fe/S cluster precursors (Kispal et al. 1999). Cells lacking a functional ATM1 gene showed an unstable mitochondrial genome that completely lacked cytochromes, suggesting that AMT1 is necessary for normal cell development (Leighton and Schatz 1995). Deletion of the subfamily C member YCF1 (full ABC transporter) causes high sensitivity to cadmium (Szczypka et al. 1994). Along with YCF1, BPT1 is involved in the transport of unconjugated bilirubin and in heavy metal detoxification via glutathione conjugates (Klein et al. 2002), suggesting these two transporters play a role in cellular detoxification. Two half ABC transporters of subfamily D, PXA1 and PXA2, are peroxisomal membrane proteins that function as a heterodimer in fatty acid transport (Shani and Valle 1996). Subfamily G members PDR5 and SNQ2 are responsible for drug resistance (Servos et al. 1993; Kolaczkowski et al. 1998). PDR12 can confer resistance to weak organic acids (Piper et al. 1998). The expression of AUS1 or PDR11 in subfamily G was confirmed to be required for anaerobic growth and sterol uptake (Wilcox et al. 2002). Non-transporter member of subfamily E, RLI1, is an essential yeast protein which may play an important role in both translation initiation and ribosome biogenesis (Dong et al. 2004). Deletion of elongation factor 3, a subfamily F member, is lethal in yeast, indicating its essential role in fungal translational process (Chakraburtty and Triana-Alonso 1998).

ABC transporter genes have been identified in 27 fungal species representing five phyla and eighteen orders of fungi (Kovalchuk and Driessen 2010). The number of ABC transporter genes varied by more than five times between different species. Interestingly, an uneven distribution of ABCA proteins (including both half and full transporters) among fungi demonstrates multiple loss events during fungal evolution. The absence of ABCA members in many of the analyzed fungi genomes suggests that fungal ABCA transporters are not essential for survival. While the numbers of half ABC transporter genes in subfamily B were similar among different fungal genomes, full-size ABCB proteins have apparently undergone an extensive amplification. In general, it shows that after divergence of fungal phyla, a significant diversification occurred in ABC transporter genes, suggesting that ABC superfamily is a dynamic gene family during evolution (Kovalchuk and Driessen 2010).

1.3.3. ABC transporter genes in Drosophila

The fruit fly *Drosophila melanogaster* is a model organism that has been widely used by researches for many decades to study a wide range of phenomena (Beckingham et al. 2005). In total, 56 ABC transporter genes in *D.* melanogaster were identified based on homology searches and at least one representative ABC transporter gene was found in each of the known mammalian subfamilies (Dreesen et al. 1988; Mackenzie et al. 1999; Dean et al. 2001). The eighth subfamily H which is most closely related to subfamily ABCG was first characterized in the *D. melanogaster* genome (Dean et al. 2001). Surprisingly, compared to human and mouse genomes which only contain five and six known ABCG genes, respectively, the *D. melanogaster* genome contains15 ABCG genes, making ABCG the most abundant subfamily in the fly genome (Dean et al. 2001). In addition, these ABCG genes are phylogenetically divergent, suggesting that there were many independent and ancient gene duplication events.

The best studied ABC transporter genes in *Drosophila* are the eye pigment precursor transporter genes, *white*, *scarlet*, and *brown* in subfamily G. These genes encode half-transporters that can heterodimerize and have been proposed to transport guanine and tryptophan (precursors of the red and brown eye pigments) into pigment cells (Ewart et al. 1994; Campbell and Nash 2001). For example, White can form a full ABC transporter with either of Brown or Scarlet (Mackenzie et al. 1999). It has been suggested that the *white* and *brown* genes may influence neural function outside of the eye as well (Wu et al. 1991; Ewart et al. 1994). In subfamily B, several ABC transporter genes were reported to be able to confer drug resistance. For instance, sensitivity to dietary colchicine and tumor progression significantly increased in fly with *mdr49* deletion (Wu et al. 1991; Buss et al. 2002). *mdr65* was proved to be responsible for the α-amanitin resistance (Begun and Whitley 2000) and also has been suggested to play a role in regulating cadmium toxicity in *Drosophila* as well (Tapadia and Lakhotia 2005). Besides drug resistance capacity, *mdr65* has been shown to function as an ortholog of human ABCB1 and is necessary for chemical protection within fruit fly brain (Mayer et al. 2009). In subfamily C, dSUR gene which shares orthologous relationship with human ABCC8 (known to cause the human disease hyperinsulinemic hypoglycemia of infancy), has been suggested a role in protecting the heart from anoxic damage (Akasaka et al. 2006). In subfamily C, overexpression of dMRP4 can increase oxidative stress resistance and extend lifespan (Huang et al. 2014a). Another ABC transporter gene, CG10505, in subfamily C was indicated to be involved in biochemical detoxification of zinc and copper (Yepiskoposyan et al. 2006).

Phylogenetic analysis revealed very few Drosophila ABC transporter genes sharing orthologous relationship with those of human due to the high rate of birth and death of ABC transporter genes (Figure 1.5), indicating the genes have evolved with functions that are specialized to either insects or mammals. For example, the eye pigment transporters in flies have no ortholog in vertebrates. Similarly, the vertebrate ABCA4 (photoreceptor-specific transporter) and CFTR genes are not identified in insects and nematodes (Dean et al. 2001). Therefore, ABC transporter genes have undergone dramatic changes during evolution, which made it more interested to study those genes in different organisms.

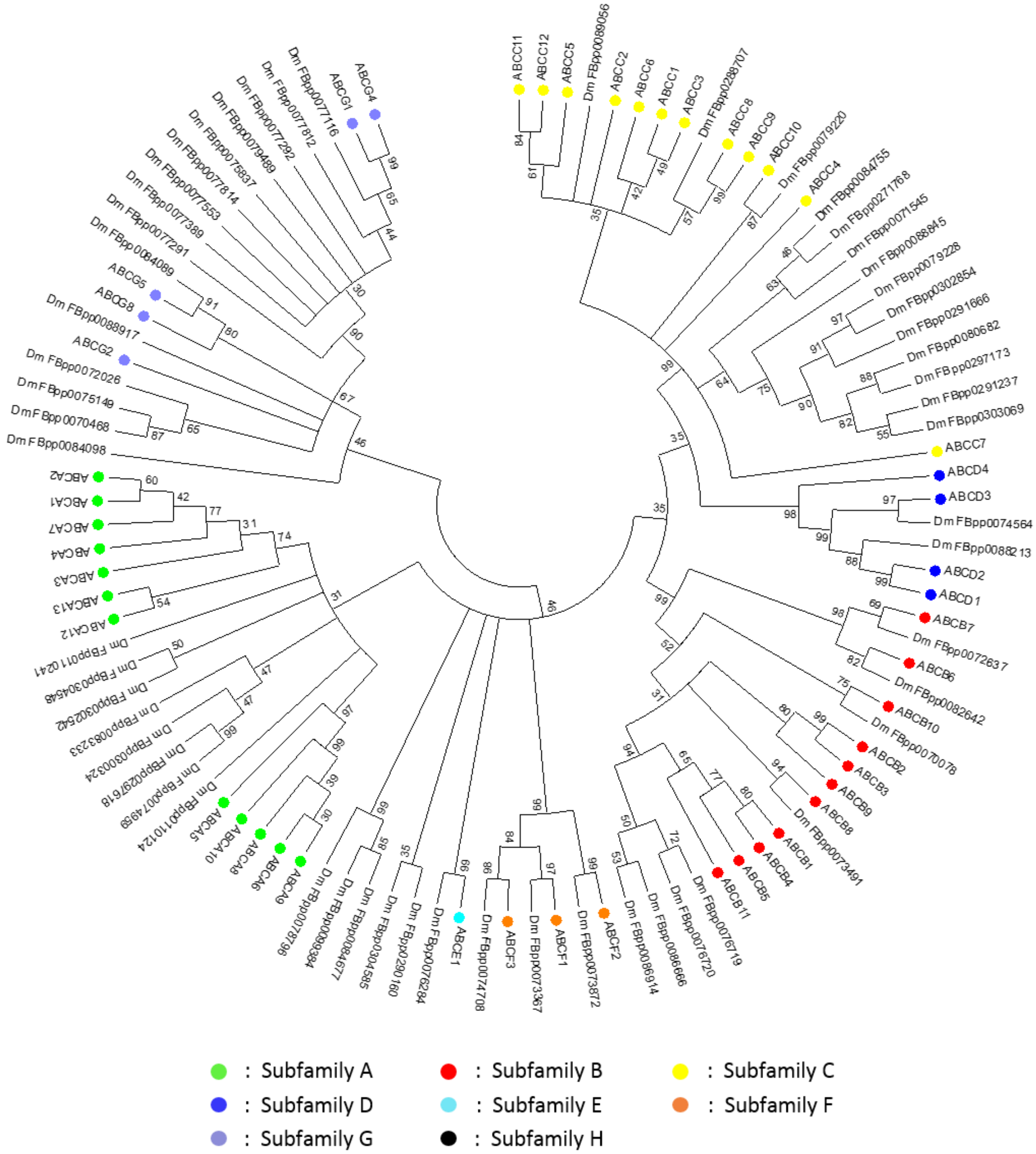


Figure 1.5: Phylogenetic analysis of ABC transporter genes in human and D. melanogaster

ABC transporter genes were first characterized in human and *D. melanogaster*. Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in human and *D. melanogaster*. ABC transporter genes in human were highlighted by different color representing for different subfamilies. Very few Drosophila ABC transporter genes share orthologous relationship with those of human.

1.3.4. ABC transporter genes in fish

The zebrafish (*Danio rerio*) embryo has become an important vertebrate model in many fields, such as genetics and human disease, toxicology and pharmacology (Scholz et al. 2008). 41 ABC transporter genes are found in zebrafish (Dean and Annilo 2005). All ABC subfamilies found in zebrafish are also found in the mammalian subfamilies, except the ABCH subfamily. Zebrafish is the only vertebrate that contains ABCH1 (Popovic et al. 2010). Although the function of ABCH1 in zebrafish remains unknown. Tissue distribution pattern revealed the highest ABCH1 expression in brain, gills and kidney, followed by lower expression in intestine, gonads, skeletal muscle and liver. Because ABCH1 is closely related to ABCG subfamily, it has been hypothesized ABCH1 is either involved in sterol transport similar to ABCG1, or is a part of the multidrug defence system like ABCG2 (Popovic et al. 2010).

The functions of some ABC transporter genes in zebrafish have been studied. ABCB4 in zebrafish acted as the multixenobiotic transporter and functionally similar to human ABCB1 (Fischer et al. 2013). ABCB5 plays a role in biliary excretion (Bard 2000; Luckenbach et al. 2014). ABCB11, which has been shown to be highly conserved across vertebrate taxa, functions as bile salt exporter in liver of zebrafish (Ballatori et al. 2000). ABCC2 is expressed in excretory organs of zebrafish, including kidney, liver and intestine and an up-regulation of ABCC1 and ABCC5 gene in embryos was observed when the zebrafish was exposed to heavy metals (Long et al. 2011). ABCC6a is essential for normal development of the zebrafish and knockdown the expression of ABCC6a after fertilization showed shortening of the body, delay of the head development, and decreased tail length (Li et al. 2010).

1.4. ABC transporter genes in nematodes

To date, ABC transporter genes have been annotated for eight nematode genomes, including free-living nematodes *Caenorhabditis elegans* (Sheps et al. 2004), *Caenorhabditis briggsae, Caenorhabditis remanei* (Zhao et al. 2007) and *Panagrellus redivivus* (Srinivasan et al. 2013), a necromenic nematode *Pristionchus pacificus* (Dieterich et al. 2008; Sommer and McGaughran 2013), a human parasite *Brugia malayi*

(Ardelli et al. 2010), a ruminant parasite *Haemonchus contortus* (Laing et al. 2013), and the pinewood nematode *Bursaphelenchus xylophilus* (Kikuchi et al. 2011).

1.4.1. ABC transporter genes in *C. elegans*

The model organism *C. elegans* was the first nematode whose complete inventory of ABC transporter genes was identified. Through homology searches, 61 candidate ABC transporter genes were identified in the *C. elegans* genome (Sheps et al. 2004; Zhao et al. 2007). These ABC transporter genes were assigned into eight subfamilies (A-H) based on their orthologous relationship to 49 annotated human ABC transporter genes (Dean et al. 2001). Among these ABC transporter genes in these two organisms, eight pairs of one-to-one orthologous relationships were identified. Compared to humans, *C. elegans* had a dramatic expansion (15 members) in the PGP subgroup of subfamily B, compared to only four PGP genes in human. Subfamily G in *C. elegans* also experienced a slight expansion with nine members, compared to five in human. These observations suggest that while some ABC transporter subfamilies were highly conserved, while others were highly dynamic in evolution.

Functions of some ABC transporter genes in *C. elegans* have been characterized. Subfamily A member *ced-7* is widely expressed during embryogenesis and worms with mutations in *ced-7* showed defect in engulfment process, indicating that *ced-7* functions in both dying cells and engulfing cells (Wu and Horvitz 1998). A half ABC transporter gene *abtm-1* in subfamily B is expressed in mitochondria. Worms with depleted *abtm-1* had morphogenetic defects and putative apoptotic events. Besides, worms with *abtm-1* (RNAi) showed accumulation of ferric iron and increased oxidative stress, indicating that *abtm-1* contribute to establishing iron homeostasis (Gonzalez-Cabo et al. 2011). Mutation in ABCB7 gene (the ortholog of *abmt-1* in human) can cause a rare inherited disorder, X-linked sideroblastic anemia and ataxia (XLSA/A) (Pondarre et al. 2007), suggesting that functional studies of *abtm-1* in *C. elegans* could help us to understand the mechanism underlying XLSA/A (Gonzalez-Cabo et al. 2011). Another half ABC transporter gene in subfamily B, *hmt-1*, is expressed in coelomocytes, head neurons and intestinal cells, conferring tolerance to multiple heavy metals (Schwartz et al. 2010). Considering that ABCB6, *hmt-1* ortholog of humans, is expressed in similar cell types that are also affected by heavy metals (Bressler et al. 2007), further studies of the *hmt-1* in *C. elegans* can contribute to the development for understanding heavy metal-caused diseases. A number of full transporters in subfamily B, PGPs (PGP-1, PGP-2, PGP-3), are expressed within intestinal cells, playing a role in gut granule biogenesis and protecting the animal against dietary (Lincke et al. 1993; Schroeder et al. 2007). It is apparently an evolutionarily conserved feature of these genes as PGPs are widely present in organs associated with the digestive tract in mammals. In addition to the normal function of PGP subgroup, drug resistance of *C. elegans* have also been shown to be related to this subgroup. Mutant with deleted *pgp-3* is sensitive to both colchicine and chloroquine (Broeks et al. 1995). Deletion of either *pgp-1* or *pgp-3* in *C. elegans* shows fast killing of nematodes by phenazine toxin secreted from the bacterial pathogen *Pseudomonas aeruginosa* (Mahajan-Miklos et al. 1999). Compared to wild-type, inactivation of *pgp-2*, *pgp-5*, *pgp-6*, *pgp-7*, *pgp-12* and *pgp-13* lead to higher sensitivity of nematodes to ivermectin (IVM) (Ardelli and Prichard 2013). Other than PGP subgroup, subfamily B contains another subgroup, HAF, with half ABC transporter genes. HAF-1, a mitochondria-localized ABC transporter, functions in regulating the stress of unfolded protein within mitochondrial (Haynes et al. 2010). Whereas another two genes in subfamily B, *haf-4* and *haf-9,* are involved in the formation of intestinal granule (Kawai et al. 2009). In addition to PGP subgroup, drug resistance ability have also been examined in a number of MRP knockout *C. elegans* strains following treatment with IVM and moxidectin (MOX). The results shows that strains with *mrp-3* and *mrp*-4 deletion are more sensitive to IVM, whereas the ones with *mrp-6* and *mrp-8* deletion are affected more severe by MOX (Ardelli 2008), suggesting that MRPs are involved in detoxification of IVM and MOX. In subfamily G, a half ABC transporter gene, *wht-2,* inhibition of which can cause the delayed birefringent contents formation in intestine, suggesting its essential role gut granule formation (Currie et al. 2007).

In addition to the role of export substrates, four non-transporters encoded by ABC transporter genes in subfamily E and F are not related to transporting molecules. Instead, *abce-1* is involved in gene transcription and translation and *abcf-1, abcf-2, abcf-3* are generally regarded as forming ribosome associated proteins involved in regulation of mRNA translation (Zhao et al. 2004b). Moreover, RNAi defects are observed in the nematodes with defective ABC transporter genes as well. At least ten ABC transporter genes from different subfamilies (i.e., *haf-6*, *abt-1*, *pgp-4*) reported in the *C. elegans* genome are required for efficient RNAi, which may help to explain evolutionary conservation of this diverse group of genes (Sundaram et al. 2006; Sundaram et al. 2008).

1.4.2. ABC transporter genes in *C. briggsae* and *C. remanei*

Homology-based searches identified 58 and 59 ABC transporter genes in *C. briggsae* and *C. remanei*, respectively (Zhao et al. 2007). The comparative analysis of ABC transporter gene families among *C. elegans*, *C. briggsae* and *C. remanei* showed that, 53 ABC transporter genes in *C. elegans* were found to have one-to-one orthologs in *C. briggsae* and *C. remanei*, suggesting high conservation of ABC transporter genes in these three closely related nematodes. Of the 53 ABC orthologous trios, 39 *C. briggsae* and *C. remanei* ABC transporter genes cluster with each other, with the *C. elegans* ABC gene as an outgroup, suggesting that *C. briggsae* and *C. remanei* ABC transporter genes are more closely related to each other than to those of *C. elegans*. These 53 strongly conserved ABC transporter genes belong to half ABC transporter genes in subfamily B, subfamily C, subfamily D, subfamily E, subfamily F, subfamily G and subfamily H. Species-specific expansions or loss of ABC transporter genes were rare and were seen primarily in the subfamily A and full ABC transporter genes in subfamily B. Interestingly, 16 ABC transporter genes form two four-gene clusters (*pgp-12*, *pgp-13*, *pgp-14* and *pgp-15*; *pgp-5*, *pgp-6*, *pgp-7* and *pgp-8*) and four two-gene clusters(*pgp-3* and *pgp-4*; *mrp-1* and *mrp-2*; *abch-1* and *abcx-1*; *pmp-1* and *pmp-2*), organized in tandem in the *C. elegans* genome, the majority of which were also present within all three species, suggesting that they were duplicated before speciation.

1.4.3. ABC transporter genes in *P. redivivus*

Another free-living nematode, *P. redivivus* (the “microworm”), has been used as a model system considering its phylogenetic distance to *C. elegans* (Srinivasan et al. 2013). By applying InterProScan for searching for ABC transporter genes, a much larger set of 94 putative ABC transporters were reported in *P. redivivus* compared to that in *C. elegans*. 52 of ABC transporter genes in *C. elegans* showed orthologous (not necessarily one to one) relationship with those in *P. redivivus*, indicating that ABC transporter genes are generally conserved among these two genomes. Interestingly, *hmt-1*-like and *pgp*-like

ABC transporters functional in heavy metal tolerance and other toxins showed expansion in *P. redivivus*, which might explain the higher level of copper tolerance reported in *P. reivivus* than *C. elegans*. In conclusion, despite the general conservation, species-specific ABC transporter genes reflect the diversity of ABC transporter during evolution.

1.4.4. ABC transporter genes in *P. pacificus*

Compared to other non-pathogenic nematodes, *P. pacificus,* known as a necromenic species associated with beetles, is used as a model system in evolutionary developmental biology (Sommer and McGaughran 2013). *P. pacificus* resembles *C. elegans* in many traits, such as hermaphroditic propagation, but contains a substantially larger genome (169 Mb) as well as more predicted genes (26000) (Dieterich et al. 2008) than those of *C. elegans* (100Mb and 19735) (Hillier et al. 2005). Previous study in 2008 found a large number (129) of putative ABC transporter genes in *P. pacificus* through KEGG pathway annotation. Thus, it was hypothesized that the relatively higher number of ABC transporter genes in *P. pacificus* is consistent with its necromenic lifestyle and could contribute to the preadaptation for parasitism of *P. pacificus* (Dieterich et al. 2008). However, a recent study reported a much smaller number (65) of ABC transporter genes in *P. pacificus* by the same group but different method (Pfam annotation) (Markov et al. 2015) and the authors did not explain the reason for such difference. ABC transporter annotation in *P. pacificus* highlights the importance of high-quality annotation before any meaningful conclusions can be drawn regarding the relevance of ABC transporter genes in evolution.

1.4.5. ABC transporter genes in *B. malayi*

B. malayi is a human parasite that causes lymphatic filariasis or elephantiasis. It has a relatively small genome (71Mb) (Ghedin et al. 2007) than *C. elegans*. PCR-based cloning approach combined with genomic mining (TBLASTN and ExPASy) identified 33 putative ABC transporter genes in *B. malayi*, 31 of which were divided into subfamilies (Liu et al. 2011). The remaining two were classified as class 3, which was suggested to function in DNA repair (Ardelli et al. 2010). The low number of ABC transporter genes in *B. malayi* is partly due to the many gaps in draft genome. For instance, seven of the 33 putative ABC transporter genes contained TMD but lacked ABC domain. Functional analysis showed that a significant increase in the transcriptional profiles of a number of ABC transporter genes mostly within the PGP and MRP subgroups was observed when the worms were exposed to IVM, suggesting that PGPs and MRPs play a role in drug resistance in *B. malayi* (Tompkins et al. 2011). In addition to subfamily B and C, members in subfamily A and G may also have a role in resistance, based on their overexpression following treatment with IVM and MOX (Stitt et al. 2011; Tompkins et al. 2011), suggesting that the majority of ABC transporters in *B. malayi* are important in drug resistance.

1.4.6. ABC transporter genes in *H. contortus*

H. contortus, known as the barber's pole worm, is one of the most pathogenic nematodes of ruminants (Gilleard 2006). 46 ABC transporter genes were identified in *H. contortus*, assigned into mammalian subfamilies except for subfamily E (Laing et al. 2013). According to previous analysis, there were some differences in ABC transporter genes in *H. contortus* and those in *C. elegans*. Significant expansion of *ced-7* was found in *H. contortus* when compared to *C. elegans.* A reduced complement of HAF subgroup were found in the parasite. A cluster of four *C. elegans* genes, *pgp-5*, *pgp-6*, *pgp-7* and *pgp-8*, did not have orthologs in *H. contortus*.

Most of the functional studies of ABC transporters in *H. contortus* have focused on the relationship between anthelmintics sensitivity and the expression of PGP subgroup (Molento and Prichard 1999; Sangster et al. 1999; Kerboeuf et al. 2002; Blackhall et al. 2008; Bartley et al. 2009; Bartley et al. 2012), suggesting that PGP subgroup share a conserved function in drug resistance and can be involved in the detoxification of host products.

1.4.7. ABC transporter genes in *B. xylophilus*

Pine wilt disease (PWD) destroys pine forests in many Asian and European countries (Futai 2013; Shinya et al. 2013). PWD was first reported in 1905 in Japan and in 1971, the nematode *B. xylophilus* (known as pinewood nematode, PWN) was identified to be the causing pathogen (Mamiya 1988). Although PWN was first identified in Japan, its origin has been traced to North America including Canada and the USA (Nickle et al. 1981). In the 1980s, the pathogen was found to spread to East Asian regions including Hong Kong, mainland China, Taiwan, and Korea (Futai 2013). It was recently found causing PWD in Portugal in 1999, and in Spain in 2008, suggesting that PWN is quickly spreading worldwide to harm more pine trees (Futai 2013; Shinya et al. 2013). The mechanisms underlying the pathogenicity of PWN remain unknown. Recent whole genome sequencing and InterProScan analysis of PWN suggested that the PWN genome harbors an unusually large family (106) of the ATP-binding cassette (ABC) transporter genes (Kikuchi et al. 2011). Because ABC transporter genes have been implicated in detoxification (Ardelli 2013), it has been hypothesized that the highly expanded family of ABC transporter genes facilitate the invasion and pathogenicity of PWN (Kikuchi et al. 2011).

1.5. Thesis aim and organization

Since ABC transporters are involved in diverse biological processes in different living organisms, identification and comparative analysis of ABC transporter genes can give us a clue of how these genes evolve. Although ABC superfamily is generally conserved, gene loss and gain did happen all the time, making ABC transporter gene set varied from species to species. As mentioned above, different numbers of ABC transporter genes have been reported in different nematode genomes (Ardelli 2013), indicating that the sizes of ABC transporter gene family are dynamic in evolution and ABC transporter genes may be important for organisms to adapt and survive in evolution. Such changes in the ABC transporter gene family do not simply reflect the differences of genome sizes. For example, the barber's pole worm *H. contortus*, which has a much larger genome (370 Mb) than most other nematodes, harbors only 46 ABC transporter genes (Laing et al. 2013), while the free-living nematode *P. redivivus*, which has a small genome, was identified to possess 94 ABC transporter genes (Srinivasan et al. 2013). Thus, these differences in the ABC transporter gene family in different nematode species may reflect their differential functional contribution to the physiology and survival.

However, some of these reported differences, however, could be the result of trivial technical reasons such as incomplete genome sequencing, problematic genome assembly, and mis-annotation. Indeed, genome assembly of various nematode genomes have different quality, which may be responsible for some of the reported differences of the sizes of ABC transporter gene family. Some genomes, such as the model organism *Caenorhabditis elegans* genome, have been fully or nearly fully sequenced, others may contain extensive gaps that could harbor ABC transporter genes. Additionally, the quality of genome annotation can also cause differences. For example, for genome assembly that contains large numbers of small contigs, ABC transporter genes could be truncated into multiple fragments, which could lead to inflated ABC transporter gene family sizes. Furthermore, different methods were applied to annotate gene models and to define ABC transporter gene families in different studies, which could contribute to the differences observed in ABC gene family of these nematode genomes (Sheps et al. 2004; Zhao et al. 2007; Dieterich et al. 2008; Kikuchi et al. 2011; Liu et al. 2011; Laing et al. 2013; Srinivasan et al. 2013). Therefore, a high-quality annotation is essential before any meaningful conclusions can be drawn regarding the relevance of ABC transporter genes in evolution.

The rest of the thesis is organized as follows. In Chapter 2, we will provide a detailed description how we developed our bioinformatics pipeline for annotating ABC transporter genes in nematode genomes and what results we obtained after testing the pipeline to *Bursaphelenchus* genomes. In Chapter 3, we will review the main results after applying our annotation pipeline to each selected nematode genomes as well as the high-quality ABC transporter genes that we finally obtained. Then, the result of comparative analysis among high-quality ABC transporter genes in all 29 nematode genomes will be introduced in Chapter 4, mainly focusing on some highly conserved ABC transporter genes as well as some species specific ABC transporter genes. Finally, we will conclude the thesis and propose the future directions in Chapter 5.

Chapter 2. Developing bioinformatics pipeline for annotating ABC transporter genes

2.1. Introduction

To address technical issues and obtain high-quality annotation of ABC transporter genes, in this study, a bioinformatics pipeline will be developed and applied to uniformly annotate ABC transporter genes in selected nematode genomes to ensure that the annotation results are robust and the comparison is meaningful. Additionally, for each candidate ABC transporter gene, the completeness of the gene model will be examined and revised by applying a homology-based gene finding program.

In this study, considering that the *C. elegans* genome has been completed sequenced and assembled without any gaps (Consortium 1998) and ABC transporter genes in *C. elegans* have been well annotated (Sheps et al. 2004), the model organism *C. elegans* was used as a test case to develop a ABC transporter gene annotation pipeline. The pipeline will be further tested by using it to search for ABC transporter genes in the genome of the pinewood nematode, *B. xylophilus,* which has recently be sequenced and was characterized to harbor an unusually large number of ABC transporter genes (Kikuchi et al. 2011). The tested pipeline will be used to annotate ABC transporter genes in set of sequenced nematode genomes, which will be described in Chapter 3.

2.2. Developing an ABC transporter gene annotation pipeline using *C. elegans* as a test case

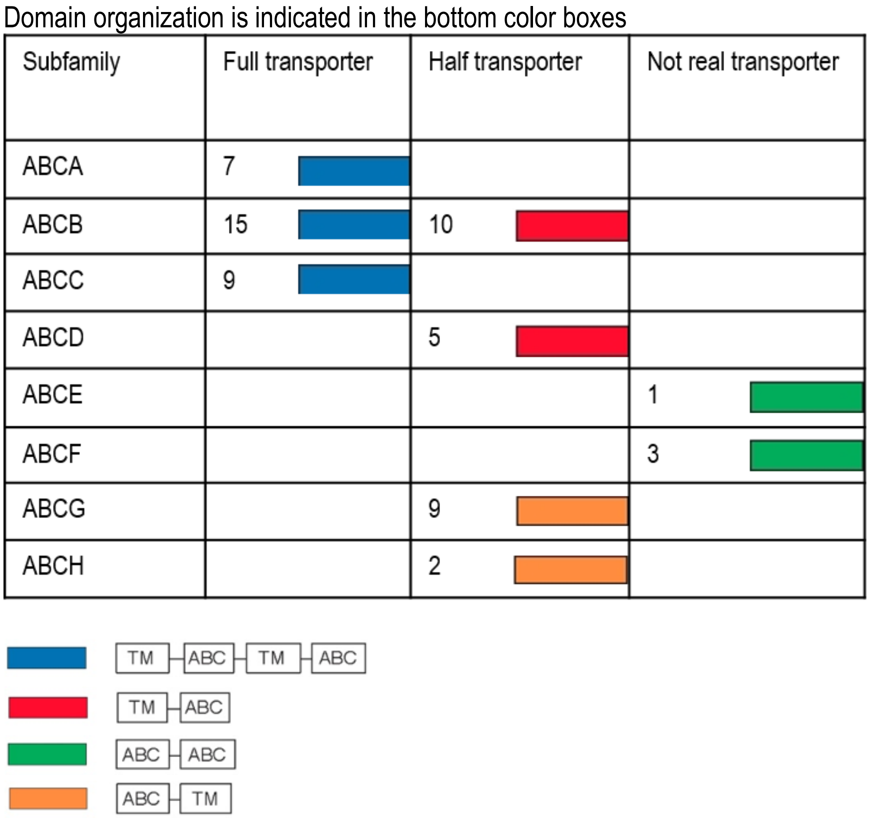
2.2.1. Background: ABC transporter genes in *C. elegans* have been annotated

The model organism *C. elegans* was the first animal whose genome was subjected to whole genome sequencing (Consortium 1998) and is the currently completely sequenced and assembled (Hillier et al. 2005). The ABC transporter genes in *C. elegans* were identified through similarity searches using FASTA search using initial query sequences that were those of known *C. elegans* ABC proteins (for example, PGP-1). Only those with highly significant matches to annotated ABC protein in the sequence database were retained. After that, representative members of different ABC transporter subfamilies were used as query sequences to search the updated WormPep81 file using BLAST. Initially, 60 ABC transporter genes were found and classified into eight ABC transporter taxonomy on the basis of amino acid sequence and domain organization (Sheps et al. 2004). Later on, with the release of the genome sequences of both *C. briggsae* (Stein et al. 2003) and *C. remanei* (http://genome.wustl.edu), PFAM and homology-based analysis enabled researchers to identify 58 and 59 ABC transporter genes in the genomes of *C. briggsae* and *C. remanei*, respectively (Zhao et al. 2007). Despite some patterns of divergence among ABC transporter genes in the three nematode species, frequent one-to-one orthology was apparent.

2.2.2. Molecular features of ABC transporter genes in *C. elegans*

In *C. elegans*, there are 61 annotated ABC transporter genes so far and they are divided into eight subfamilies (A-H). Among these 61 genes, three (F55G11.9, F22E10.4 and Y49E10.9) of them are confirmed to be pseudogenes and do not have corresponding protein sequences in WormBase (http://www.wormbase.org/species/c_elegans/gene/). F56F4.6 is truncated with a short protein length (252 aa), which could be a pseudogene as well (Sheps et al. 2004). To focus on the functional ABC transporter genes, we excluded it from our further analysis. In addition to F56F4.6, C56E6.1 in subfamily H was also excluded from our analysis since it does not contain any putative ABC domain. In total, 56 ABC transporter genes were used in this analysis, including 25 half transporters, 27 full transporters and four non-transporters that do not contain TMD (Table 2.1).

Table 2.1: Subfamily information of ABC transporter genes in *C. elegans*.



Domain organization is indicated in the bottom color boxes

Subfamily	Full transporter	Half transporter	Not real transporter
ABCA	7 🟦		
ABCB	15 🟦	10 🟥	
ABCC	9 🟦		
ABCD		5 🟥	
ABCE			1 🟩
ABCF			3 🟩
ABCG		9 🟧	
ABCH		2 🟧	

🟦	TM – ABC – TM – ABC
🟥	TM – ABC
🟩	ABC – ABC
🟧	ABC – TM

To annotate high-quality ABC transporter genes, it is important to evaluate the key molecular features of a candidate gene by comparing these features with those of its ortholog in *C. elegans*, assuming that key molecular features of orthologs remain similar. The first molecular feature to consider is the number of ABC domains contained in an ABC transporter gene. The second molecular feature is the length of ABC domain, which should correspond to the range of lengths of ABC domains in *C. elegans*. We analyzed 87 ABC domains contained in the 56 annotated ABC transporter proteins in *C. elegans*. The lengths of these *C. elegans* ABC domains range from 125 aa to 165 aa (Figure 2.1).We also examined the InterProScan e-value of ABC domain in annotated *C. elegans* ABC transporter, which ranged from 2.40 x 10^{-10} to 1.70 x 10^{-39}(Figure 2.2). The fourth molecular feature is the number of TM helices within each TM domain. We applied SCAMPI (http://scampi.cbr.su.se/) to predict the TM helices within each ABC transporter gene. In terms of half transporter, TM helices, of which the number ranged from four to 11, formed one cluster whereas the TM helix number of full ABC transporters ranged from eight to 17, usually formed two clusters (Figure 2.3).

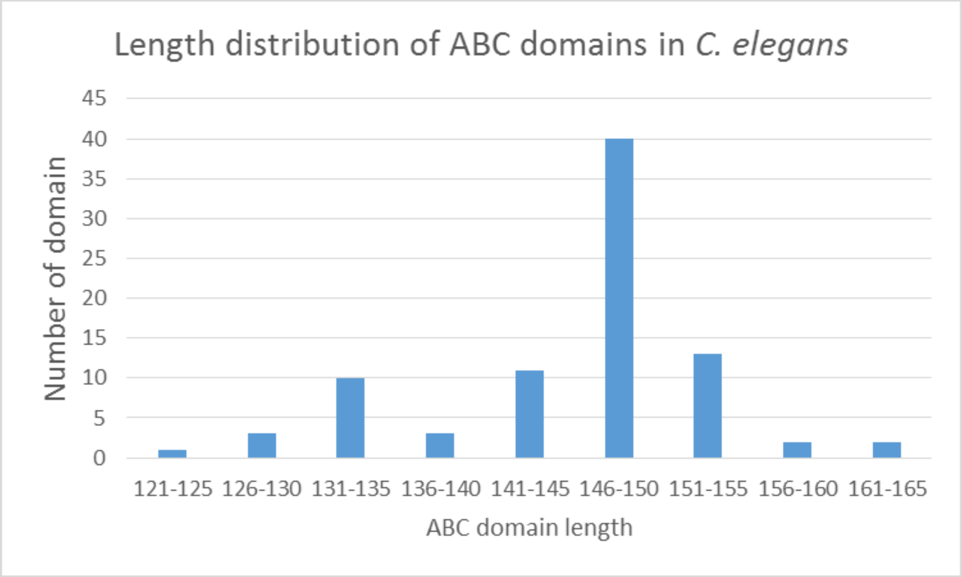


Figure 2.1: ABC domain length distribution in C. elegans ABC transporter proteins

ABC domain sequences of 56 ABC transporter proteins in *C. elegans* were extracted to draw the ABC domain length distribution.

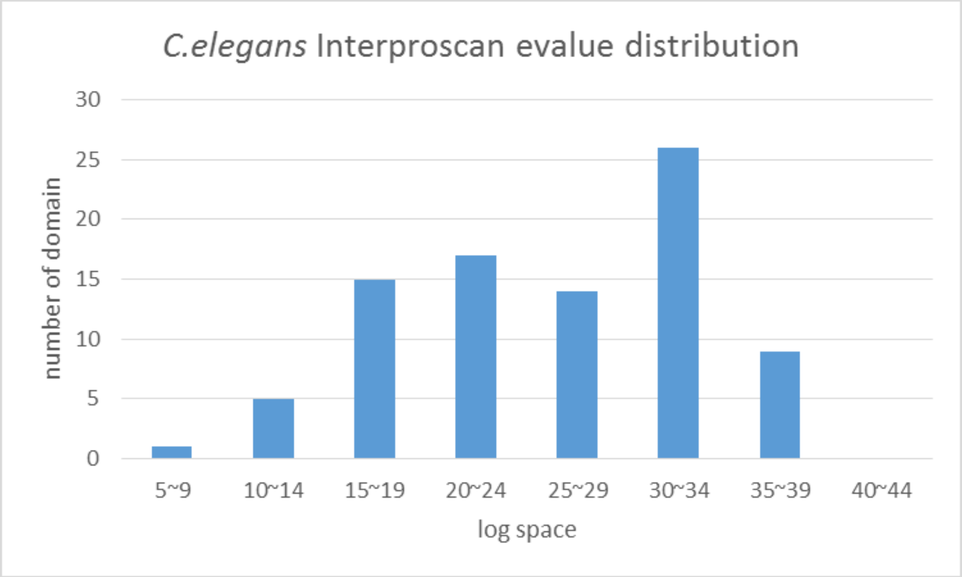


Figure 2.2: ABC domain InterProScan e-value distribution in *C. elegans* ABC transporter proteins

InterProscan e-value for each ABC domain of 56 ABC transporter proteins in *C. elegans* were extracted to draw the ABC domain InterProScan e-value distribution

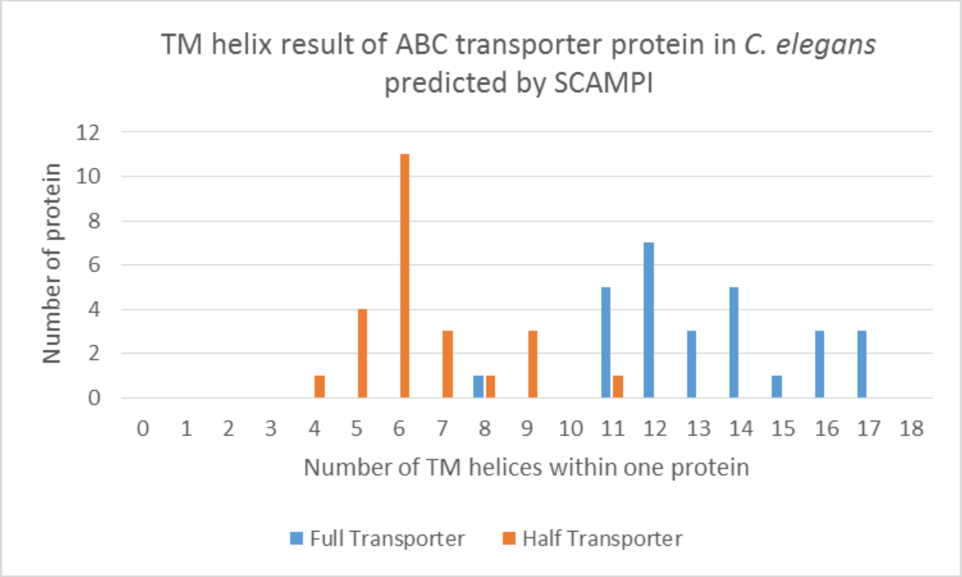


Figure 2.3: Distribution of TM helix number in ABC transporters in *C. elegans*
The number of TM helices within each ABC transporters in C. elegans were predicted by SCAMPI

2.2.3. Developing a bioinformatics pipeline for searching for ABC transporter genes in *C. elegans*

Because the ABC domain is shared by all ABC transporter genes in *C. elegans*, to develop a simple but effective bioinformatics pipeline (Figure 2.4) for searching for ABC transporter genes, we first tried to search for ABC transporter genes in *C. elegans* genome by searching for the presence of the ABC domain in *C. elegans* proteins using InterProScan. After removing the redundant isoforms of the same gene, we got 56 ABC transporter genes, which were exactly the curated ones without any false positive or false negative. ABC transporter related molecular features in *C. elegans* are listed in Table 2.1. The analysis of annotated ABC transporter genes in *C. elegans* suggested that InterProScan-based method can allow us to identify all ABC transporter genes if the genome is fully assembled and well annotated without any contamination.

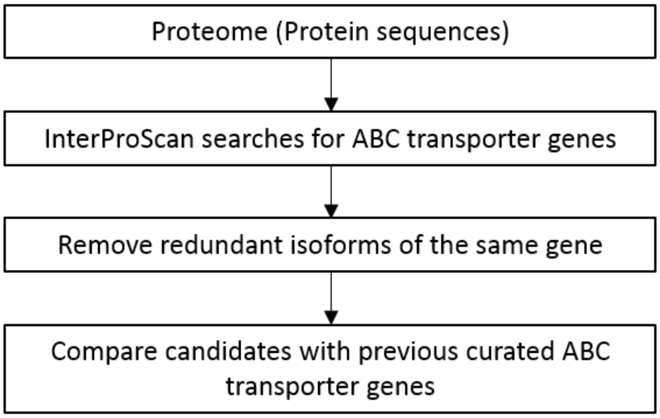

Proteome (Protein sequences)
InterProScan searches for ABC transporter genes
Remove redundant isoforms of the same gene
Compare candidates with previous curated ABC transporter genes

Figure 2.4: Bioinformatics pipeline for searching for ABC transporter gene in *C. elegans* genome

Table 2.2: List of ABC transporter related molecular features in *C. elegans*

Gene	Domain organization	ABC domain length	ABC domain e-value	ABC domain length	ABC domain e-value
Y39D8C.1	6TM-ABC-8TM-ABC	142	1.60E-27	144	5.20E-20
C48B4.4d	7TM-ABC-7TM-ABC	136	4.70E-19	144	1.40E-26
F12B6.1a	9TM-ABC-7TM-ABC	147	1.30E-24	144	2.40E-30
Y53C10A.9	7TM-ABC-7TM-ABC	145	7.30E-27	137	1.10E-22
C24F3.5a	6TM-ABC-6TM-ABC	133	2.10E-15		
C30H6.6	6TM-ABC	153	2.50E-34		
Y48G8AL.11a	6TM-ABC	150	2.10E-33		
F57A10.3	6TM-ABC	149	6.70E-37		
Y50E8A.16	9TM-ABC	150	9.60E-37		
ZK484.2a	9TM-ABC	150	2.20E-32		
W09D6.6	9TM-ABC	149	5.00E-34		
F43E2.4	8TM-ABC	150	3.30E-33		
W04C9.1a	9TM-ABC	149	1.60E-32		
Y57G11C.1	6TM-ABC	149	6.70E-37		
Y74C10AR.3a	7TM-ABC	150	1.60E-30		
F22E10.2	6TM-ABC-6TM-ABC	150	2.70E-34	149	5.40E-31
F42E11.1a	6TM-ABC-5TM-ABC	150	1.60E-34	150	1.00E-33
C47A10.1	6TM-ABC-6TM-ABC	152	4.30E-38	150	8.50E-36
T21E8.3	5TM-ABC-3TM-ABC	150	1.60E-36	151	2.70E-33
T21E8.1a	6TM-ABC-6TM-ABC	150	2.00E-34	151	6.30E-35
DH11.3	6TM-ABC-5TM-ABC	148	7.40E-35	149	2.10E-34

ZK455.7	6TM-ABC-5TM-ABC	150	8.90E-34	150	1.80E-32
C54D1.1	6TM-ABC-7TM-ABC	150	8.70E-35	152	6.20E-30
C34G6.4	7TM-ABC-6TM-ABC	150	5.00E-37	151	4.60E-33
F22E10.1	6TM-ABC-5TM-ABC	150	1.00E-33	149	4.20E-32
K08E7.9	6TM-ABC-6TM-ABC	150	1.40E-36	152	5.20E-35
F22E10.3	6TM-ABC-6TM-ABC	149	5.30E-30	150	9.00E-34
T21E8.2	6TM-ABC-5TM-ABC	151	3.40E-35	150	2.00E-34
C05A9.1a	6TM-ABC-6TM-ABC	151	1.80E-33	150	9.30E-36
C18C4.2	6TM-ABC-8TM-ABC	134	2.40E-10	146	1.30E-15
F20B6.3	7TM-ABC-6TM-ABC	134	9.60E-19	148	3.20E-30
F21G4.2	11TM-ABC-6TM-ABC	135	4.00E-22	148	2.60E-27
F14F4.3b	8TM-ABC-6TM-ABC	130	3.10E-18	149	4.80E-30
F57C12.5b	10TM-ABC-5TM-ABC	135	9.30E-23	148	7.00E-27
E03G2.2	11TM-ABC-6TM-ABC	149	2.80E-25	134	9.60E-15
F57C12.4	11TM-ABC-5TM-ABC	148	3.90E-28	135	4.00E-25
Y75B8A.26	10TM-ABC-7TM-ABC	135	4.10E-19	149	1.30E-28
Y43F8C.12	11TM-ABC-5TM-ABC	135	2.90E-18	149	7.30E-28
C44B7.8	4TM-ABC	143	7.40E-18		
C44B7.9	5TM-ABC	143	9.50E-18		
T02D1.5	6TM-ABC	144	2.60E-18		
C54G10.3b	7TM-ABC	145	7.20E-20		
T10H9.5b	6TM-ABC	144	4.30E-15		
Y39E4B.1	ABC-ABC	143	5.80E-20		
F42A10.1	ABC-ABC	164	1.30E-20		
F18E2.2	ABC-ABC	161	2.90E-22		
T27E9.7	ABC-ABC	156	2.20E-23		
C16C10.12	ABC-5TM	152	1.50E-20		
F19B6.4	ABC-6TM	153	2.80E-23		
Y42G9A.6a	ABC-5TM	146	6.00E-25		
T26A5.1	ABC-6TM	151	1.00E-22		
C05D10.3	ABC-6TM	156	5.60E-23		
Y47D3A.11	ABC-6TM	153	9.10E-21		
C10C6.5	ABC-6TM	151	1.30E-21		
F02E11.1	ABC-5TM	138	4.30E-19		
C56E6.5	ABC-7TM	147	6.70E-12		

2.3. Searching for ABC transporter genes in *B. xylophilus*

2.3.1. Introduction

Pine wilt disease caused by *B. xylophilus* is quickly spreading worldwide to harm more pine trees (Futai 2013; Shinya et al. 2013). An unusually large family of ABC transporter genes were identified in previous study and were proposed to facilitate the invasion and pathogenicity of PWN (Kikuchi et al. 2011). In this study, we aimed to further test the bioinformatics pipeline for annotating ABC transporter genes by annotating ABC transporter genes in in a newly sequenced genome of a *B. xylophylus* strain that was isolated in China (Zhejiang Province, BxCN). We also would like to compare the number of ABC transporter genes in this strain against the number of ABC transporter gene candidates in the *B. xylophylus* strain (BxJP) reported recently (Kikuchi et al. 2011).

2.3.2. Searching for ABC transporter genes in BxCN

Applying the above bioinformatics pipeline, we found 60 protein-coding genes in BxCN encoded at least one ABC domain and we took these 60 genes as ABC transporter candidates.

2.3.3. Genomic contamination detection and filtration

Genomic contamination could be introduced during sample collection or sequencing. It is particularly common to observe bacterial genomic contamination of eukaryotic samples. In this case, because *B. xylophilus* worms were fed with fungi (Kikuchi et al. 2011), it is likely that the genome sequences were contaminated with fungal DNA sequences as well. In order to detect and filter contamination, protein sequence of each ABC transporter candidate was used as query to run TBLASTN against NCBI Nucleotide Collection database. If the best hit was a species of bacteria or fungi, suggesting that the corresponding query was obtained from contamination and then the query was filtered out. Based on our TBLASTN result, all our 60 candidate ABC transporter genes were retained.

2.3.4. Annotation and quality assessment

We classified 60 ABC transporter gene candidates in BxCN based on their homology to annotated ABC transporter genes in *C. elegans*. The quality of each ABC transporter candidate was then evaluated by examining key attributes of ABC domain, including the number of ABC domains they possess, the length of each ABC domain, and the e-value of the InterProScan result. Based on the key attributes of annotated ABC transporter proteins in *C. elegans* mentioned in section 2.2.2, we defined a high-quality ABC transporter as one that has (1) appropriate number of ABC domains; (2) the length of each ABC domain should be longer than 130 aa and shorter than 165 aa; and (3) The InterProScan e-value of predicted ABC domain should be better than (*i.e.*, lower than) 1.0 x 10^{-10}.

After applying these three criteria to each candidate gene, we found 43 ABC transporter candidate satisfying all criteria, which were called high-quality ABC transporter genes. For the 17 candidates that did not satisfy the criteria, we next examined the nature of the defects and whether it was possible to improve their gene models (Table 2.3).

Table 2.3: Improvement of defective ABC gene models based on InterProScan searches

ID	Improved or not	ID after improvement	Notes
BxCN09178	Yes	BxCN09178	
BxCN10724	Yes	BxCN10724	Merged with BxCN10725
BxCN10725	Yes		
BxCN12661	Yes	BxCN12661	
BxCN13281	Yes	BxCN13281	
BxCN14853	Yes	BxCN14853a BxCN14853b	Split into two genes
BxCN16314	Yes	BxCN16314	Merged with BxCN16315
BxCN16315	Yes		
BxCN13779	Yes	BxCN13779	
BxCN08056	No	BxCN08056	Keep the original model
BxCN04424	No		
BxCN12000	No		
BxCN13603	No		
BxCN14290	No		
BxCN14292	No		

2.3.5. Improvement of defective gene models

We attempted to improve all defective gene models. For each defective candidate, its ortholog in *C. elegans* was used as query to run the homology-based gene finder genBlastG (She et al. 2011) against the BxCN genomic sequences to define a potential full-length high-quality ABC transporter gene. The new gene model generated by genBlastG was then examined to check whether it now satisfies the same set of criteria described above (section 2.2.4). Through this effort, we successfully constructed eight new high-quality ABC transporter gene models, among which two gene models were the result of splitting one candidate ABC gene model (BxCN14853) (Figure 2.5), four gene models were the result of merging two separate pairs of neighboring gene models (Figure 2.6, Figure 2.7 and Figure 2.8), and two gene models were improved without affecting neighboring gene models (Figure 2.9). Among those eight newly constructed ABC transporter gene models, the revisions of six were supported by RNA-seq data in BxCN. Revisions in the remaining two ABC gene models that were not directly supported by RNA-seq data in BxCN. However, they were supported by RNA-seq data of their orthologs in BxCA (a *B. xylophilus* strain isolated in Canada) (Figure 2.7 and Figure 2.8). For all improved gene models and the original gene model before improvement, their gene structures, location of ABC domains as well as three motifs within each domain were displayed in Figure 2.10.

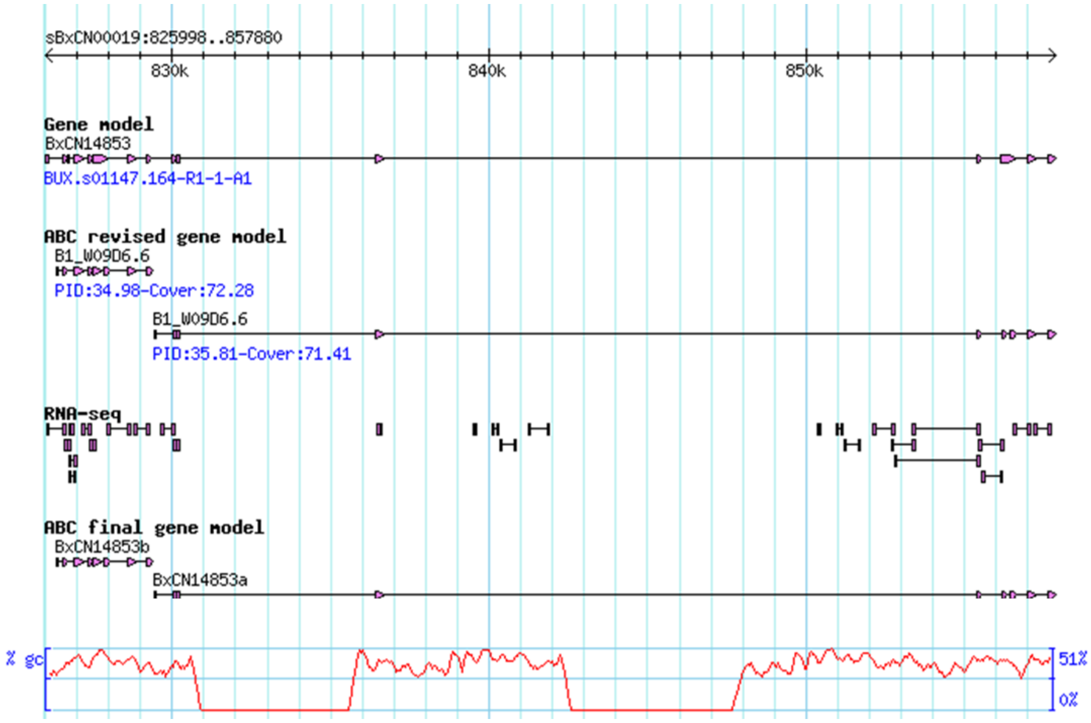


Figure 2.5: A representative case in which a candidate ABC gene model was be split into two separate ABC gene models.

"Gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Based on the ortholog in *C. elegans*, BxCN14853 should be split into two new gene models, each of which is a half ABC transporter gene with a single ABC domain. No RNA-seq support the hypothetical intron between these two gene models, further suggesting that these are two separate gene models. BxCN14853a had two sequencing gap in its genomic region, but considering it has a high-quality ABC domain, we kept this gene model into our final set.

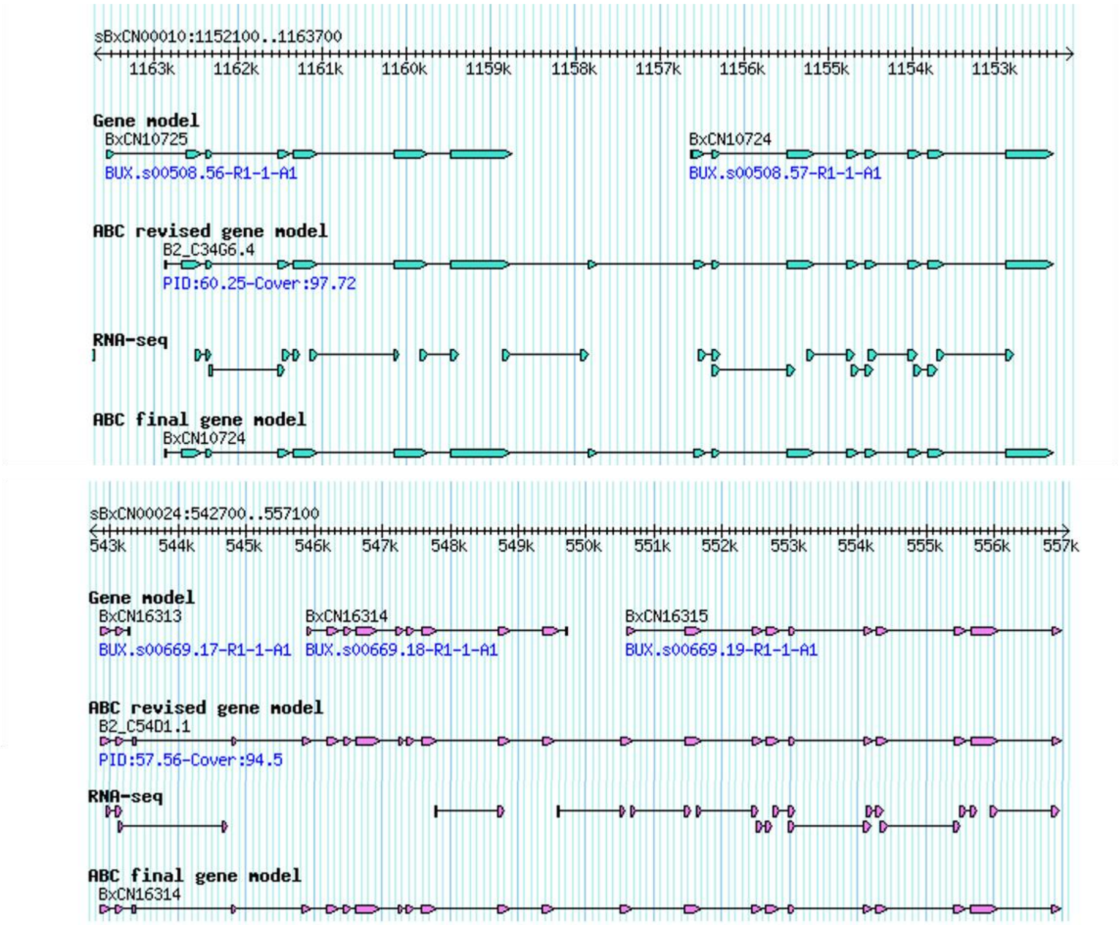


Figure 2.6: Two representative cases in which multiple adjacent candidate genes were merged to form a high-quality ABC transporter gene.

"Gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Both of the improved gene was annotated as a full ABC transporter genes in subfamily B. BxCN10725 and BxCN10724 were merged into a high-quality ABC transporter gene, annotated as a full ABC transporter gene in subfamily B based on genBlastG revision. Similarly, BxCN16313, BxCN16314 and BxCN16315 were merged into a full transporter gene in subfamily B. RNA-seq data supports the hypothetical introns, further suggesting that these are two separate gene models

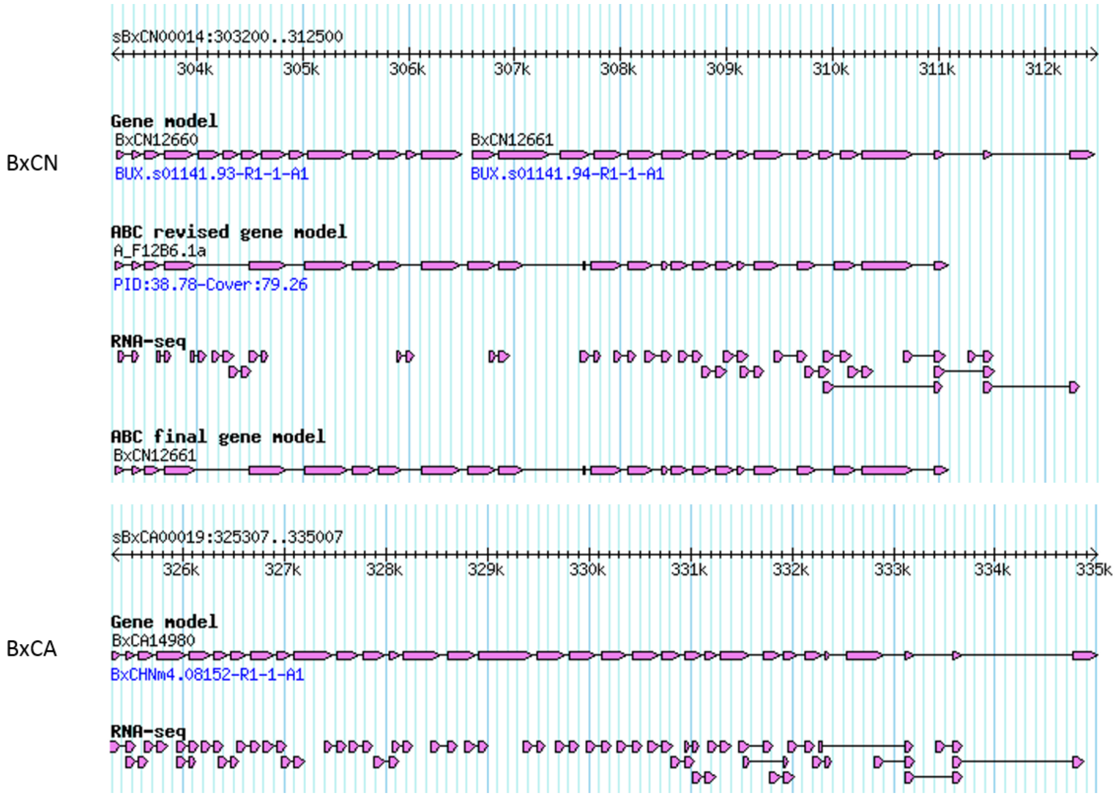


Figure 2.7: A representative case in which one candidate ABC gene model was merged with its neighboring gene model to form a larger gene model

"Gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. BxCN12661 and BxCN12660 should be merged into one new gene model, which is a full ABC transporter with two ABC domains in subfamily A. Although there is no RNA-seq support to one hypothetical intron in the junction region but its ortholog in BxCA, BxCA14980, has its intron supported by RNA-seq, suggesting that BxCN12661 and BxCN12660 should be merged.

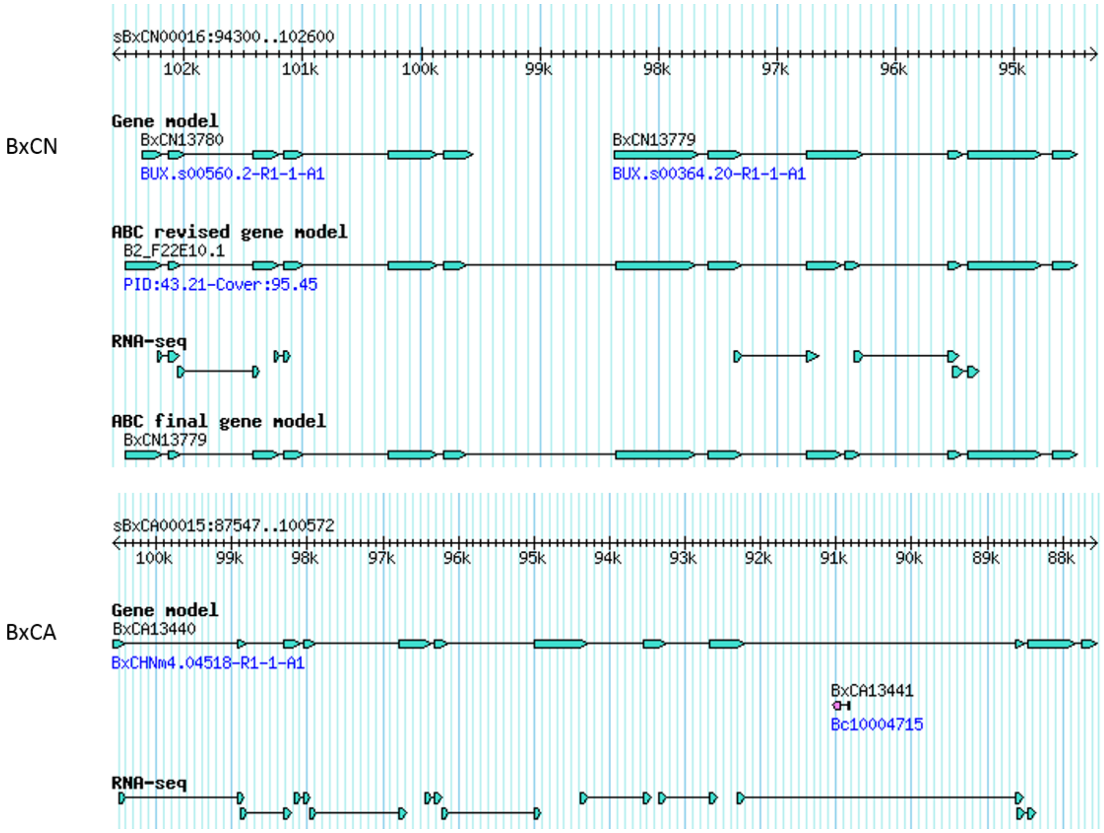


Figure 2.8: A representative case in which one candidate ABC gene model was merged with its neighboring gene model to form a larger gene model.

“Gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “RNA-seq” track includes introns predicted by RNA-seq data; “ABC final gene model” track includes high-quality ABC transporter gene models. BxCN13779 and BxCN13780 should be merged into one new gene model, which is a full ABC transporter with two ABC domains in subfamily B. Although, there is no RNA-seq support to one hypothetical intron in the junction region but its ortholog in BxCA, BxCA13440, has almost its intron supported by RNA-seq, suggesting that BxCN13779 and BxCN13780 should be merged.

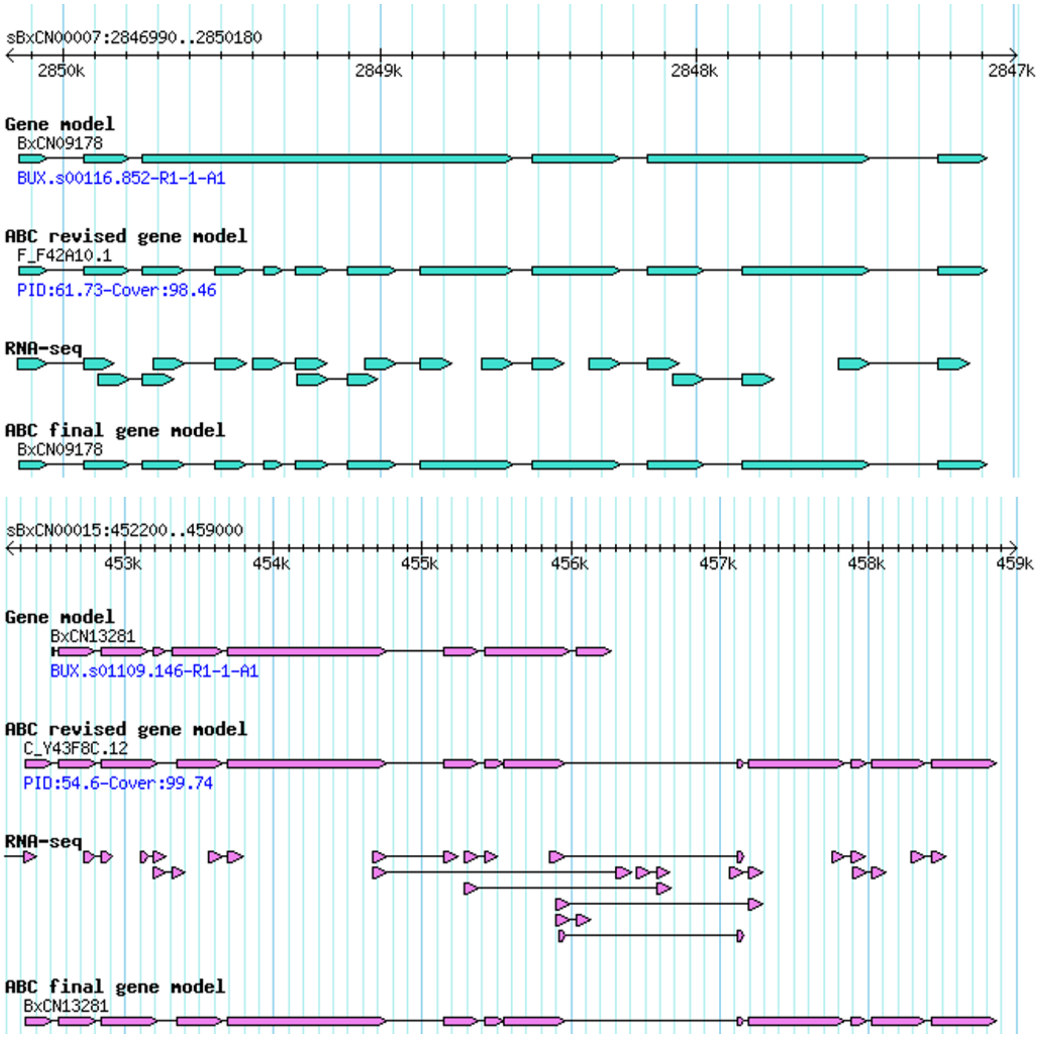


Figure 2.9: Two cases in which the exons of one candidate gene model were improved.

“Gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “RNA-seq” track includes introns predicted by RNA-seq data; “ABC final gene model” track includes high-quality ABC transporter gene models. Both of the newly constructed gene model has RNA-seq data supported.

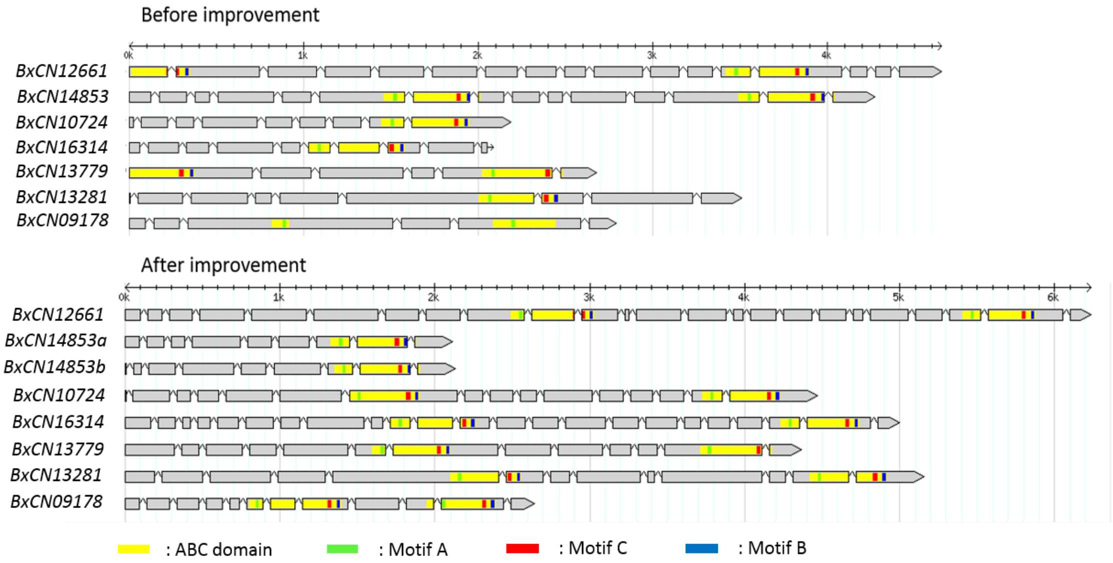


Figure 2.10: Gene structure, ABC domain and motifs within ABC domain before and after improvement in BxCN

Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Compared to original gene models, all eight newly constructed ABC gene models contain proper ABC domain (s) and motifs within each ABC domain.

Despite much effort, eight gene models could not be improved to satisfy all criteria. Two of them, BxCN04424 and BxCN12000, had unfavorable InterProScan e-values (5.20 x 10^{-6} and 1.50 x 10^{-5}), very short ABC domain lengths (33 aa and 31aa), and missed motifs in their ABC domains (BxCN04424 only had Walker A motif and BxCN12000 only had Walker B motif). Thus, they were most likely pseudogenes. The third candidate, BxCN13603, showed high similarity only to a small portion of its neighboring gene BxCN13602 (a high-quality ABC transporter gene), likely due to a partial duplication, or a genome assembly error. Additionally, the orthologs of BxCN13603 in BxJP (BxJP11047), BxCA (BxCA17905), BmCN (BmCN15800) all had defects: BmCN15800 was located in genomic regions with stretches of Ns and had one domain with a length of 88 aa. BxCA17905 had one domain and was located in a short contig and BxJP11047 had one ABC domain with short length of 55 aa. Therefore, BxCN13603 is more likely to be a deteriorating pseudogene. Two candidates BxCN14290 and BxCN14292 were located in genomic regions in BxCN with stretches of Ns, suggesting that these two genomic regions were badly sequenced and assembled. The ortholog of BxCN14290 in BxCA, BxCA12681, was a high-quality ABC transporter gene that satisfied all three criteria described above, suggesting that BxCN might also have a high-quality ABC transporter gene model in this genomic region. In contrast, BxCN14292 might be pseudogene because none of its ortholog in BxJP and BxCA was high-quality ABC transporter gene models, and its ortholog in BmCN was not found. The sixth gene model BxCN08056 could not be further improved. However, the only criterion it did not meet was the length of one of its predicted ABC domains, which was 124 aa, slightly shorter than the lower threshold of 130 aa described above. Nevertheless, this ABC domain does contain all three motifs (Figure 2.11), suggesting that BxCN08056 was a high-quality ABC transporter gene. Thus we included it in the high-quality ABC transporter gene set.

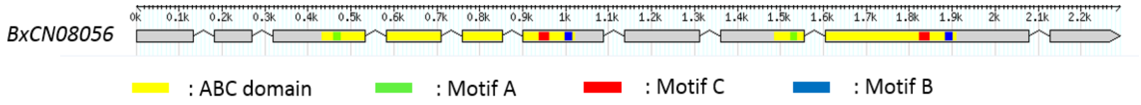


Figure 2.11: Gene structure, ABC domain and motifs within BxCN08056
Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. BxCN08056 annotated as an ABC transporter gene in subfamily E with one of its ABC domain slightly shorter than our criterion. Considering that both ABC domains within BxCN08056 contained all three motifs, we included this gene in our final set.

Two defective candidate genes, BxCN14848 and BxCN14849, had much longer domain length (191 aa, 182 aa) than the upper limit of 165 aa, which is the longest length of ABC domains in ABC transporter in *C. elegans*. Almost all introns of the BxCN14848 gene model were supported by RNA-seq data (Figure 2.12). Although not all introns in BxCN14849 were supported by RNA-seq data, all introns of its ortholog in BxCA (BxCA15805), which also had longer domain length (182 aa) as well, were supported by RNA-seq data (Figure 2.13), suggesting that the BxCN14849 gene model is of high-quality. In addition, ABC domains in both BxCN14848 and BxCN14849 contain all three key motifs (Figure 2.14). Thus, we included BxCN14848 and BxCN14849 in our final set of ABC transporter genes in BxCN.

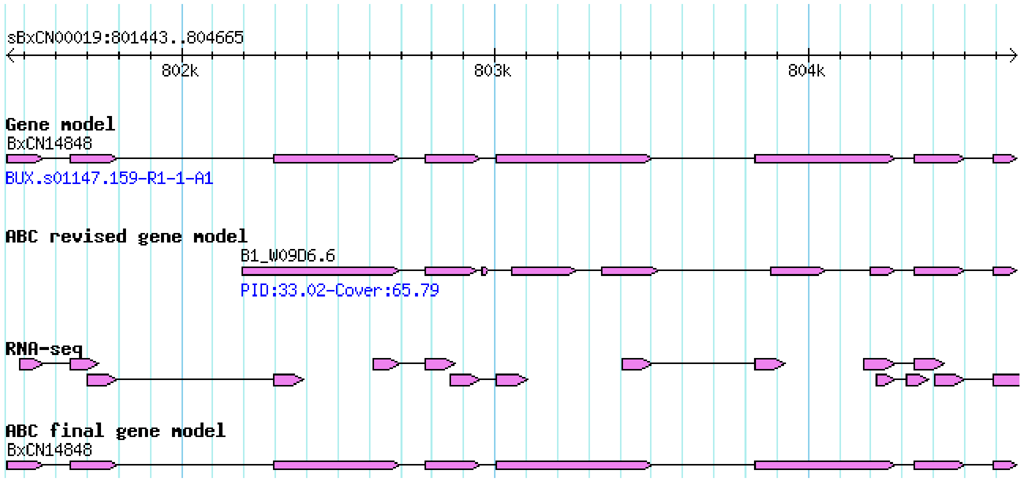


Figure 2.12: Original gene model of BxCN14848 was supported by RNA-seq data "Gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. BxCN14848, annotated as a half ABC transporter in subfamily B had a longer ABC domain (191 aa) than our criterion. However, compared to the revised gene model, the original gene model had almost its intron supported by RNA-seq data. So, we included the original gene model of BxCN14848 in our final set.

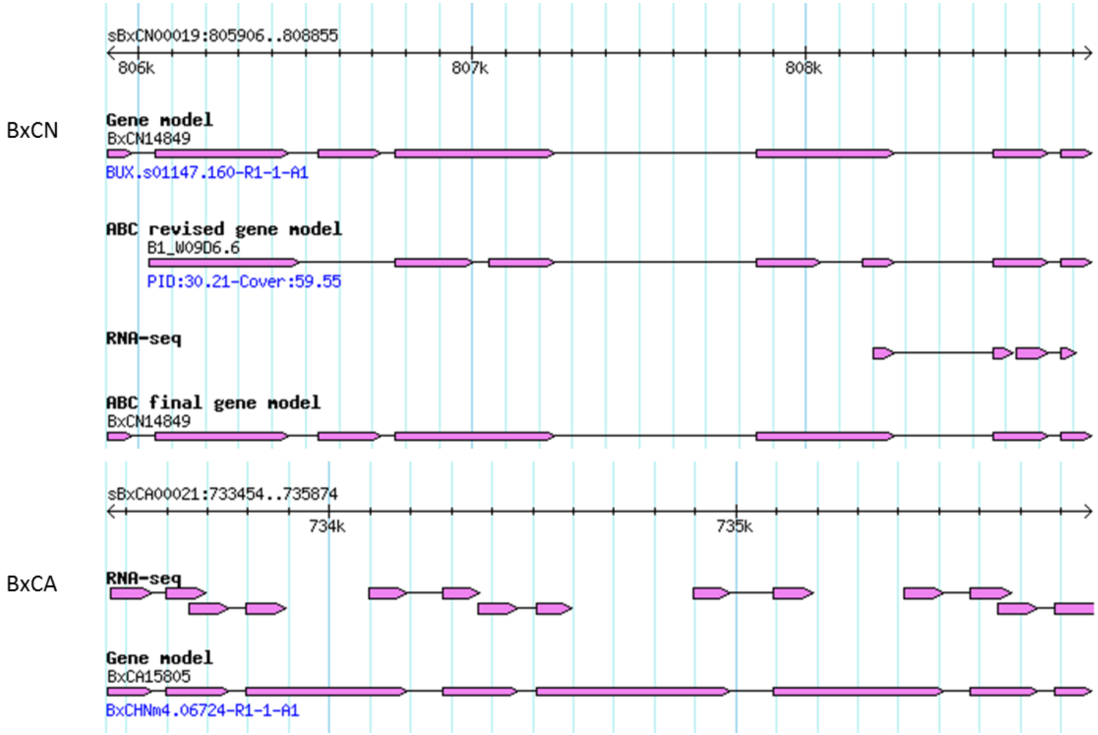


Figure 2.13: Original gene model of BxCN14849 was indirectly supported by its ortholog in BxCA

"Gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. BxCN14849, annotated as a half ABC transporter in subfamily B had a longer ABC domain (182 aa) than our criterion. RNA-seq data for this region was not sufficient. However, the ortholog of BxCN14849 in BxCA, BxCA15805, had all its intron supported, indirectly supporting the original model of BxCN14849. So, we included the original gene model of BxCN14849 in our final set.

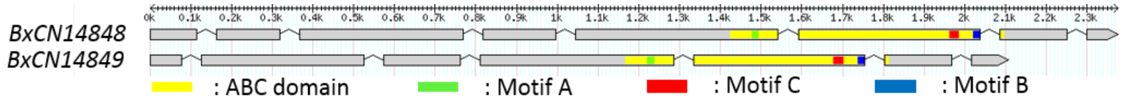

BxCN14848
BxCN14849
0k 0.1k 0.2k 0.3k 0.4k 0.5k 0.6k 0.7k 0.8k 0.9k 1k 1.1k 1.2k 1.3k 1.4k 1.5k 1.6k 1.7k 1.8k 1.9k 2k 2.1k 2.2k 2.3k
: ABC domain
: Motif A
: Motif C
: Motif B

Figure 2.14: Gene structure, ABC domain and motifs within BxCN14848 and BxCN14849

Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The original gene model of BxCN14848 and BxCN14849 both had three motifs within ABC domain.

In summary, of the eight defective candidate ABC gene models, three (BxCN08056, BxCN14848 and BxCN14849) were included in the high-quality ABC transporter gene set. These gene models are more diversified from *C. elegans* ABC transporter genes. The status of one gene (BxCN14290) could not be determined because its defect was due to sequencing or assembly errors. The remaining four candidate gene models (BxCN04424, BxCN12000, BxCN14292 and BxCN13603) were most likely pseudogenes.

In total, we have annotated 54 high-quality ABC candidates in BxCN. We expect that BxCN could have 55 high-quality candidate ABC genes when the genome is fully sequenced and assembled.

2.3.6. Evaluating the completeness of the high-quality ABC transporter gene set

Although we have found all ABC transporter genes in the BxCN genome that contain ABC domain (i.e., PF00005 domain), ABC transporter gene models that had defective annotated ABC domains could be missed in the search. To ensure that we had identified all ABC transporter genes in BxCN, we further searched the BxCN protein dataset using BLASTP with all *C. elegans* ABC transporters as queries. All hits with e-value was less than or equal to 10^{-10} were compared to the set of ABC transporter genes obtained through InterProScan search. Seven additional ABC transporter gene candidates (BxCN04601, BxCN06788, BxCN11336, BxCN12660, BxCN13780, BxCN14815 and BxCN16313) were found by the BLASTP search. Among those seven genes, three (BxCN12660, BxCN13780 and BxCN16313) had already been used to merge with adjacent ABC transporter genes to form high-quality ABC transporter genes (Figure 2.6, Figure 2.7 and Figure 2.8).

From the four candidates, two new high-quality ABC transporter gene models were formed by merging two separate pairs of neighboring gene models (BxCN06787 and BxCN06788, BxCN11336 and BxCN11337) (Figure 2.15 and Figure 2.16). Two genes BxCN14815 and BxCN04601 did not have any ABC transporter related Pfam domains and the length of predicted proteins (127 aa and 166 aa) were too short to be ABC transporter genes, suggesting that BxCN14815 and BxCN04601 were false positives from BLASTP search (Figure 2.17).

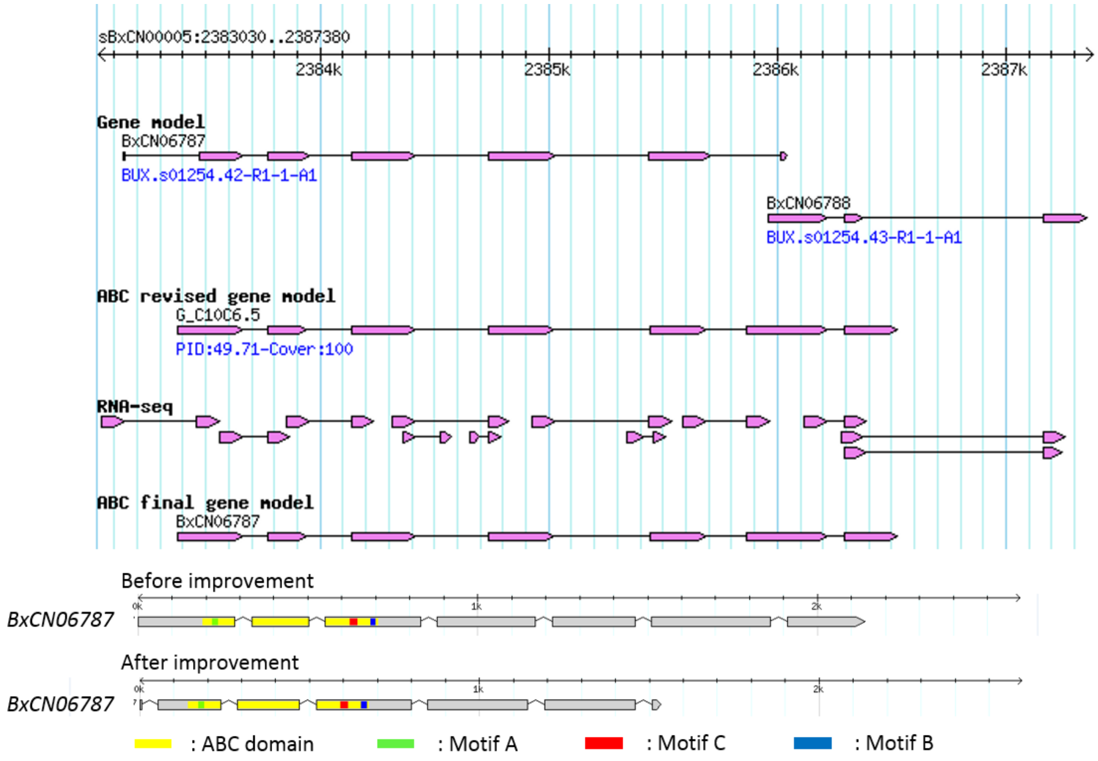


Figure 2.15: A representative case in which one candidate ABC gene model was merged with its neighboring gene model to form a larger gene model

“Gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “RNA-seq” track includes introns predicted by RNA-seq data; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. BxCN06788 was obtained from BLAST searches. After improvement, BxCN06787 and BxCN06788 were merged with each other and the newly constructed gene was annotated as a complete half ABC transporter in subfamily G.

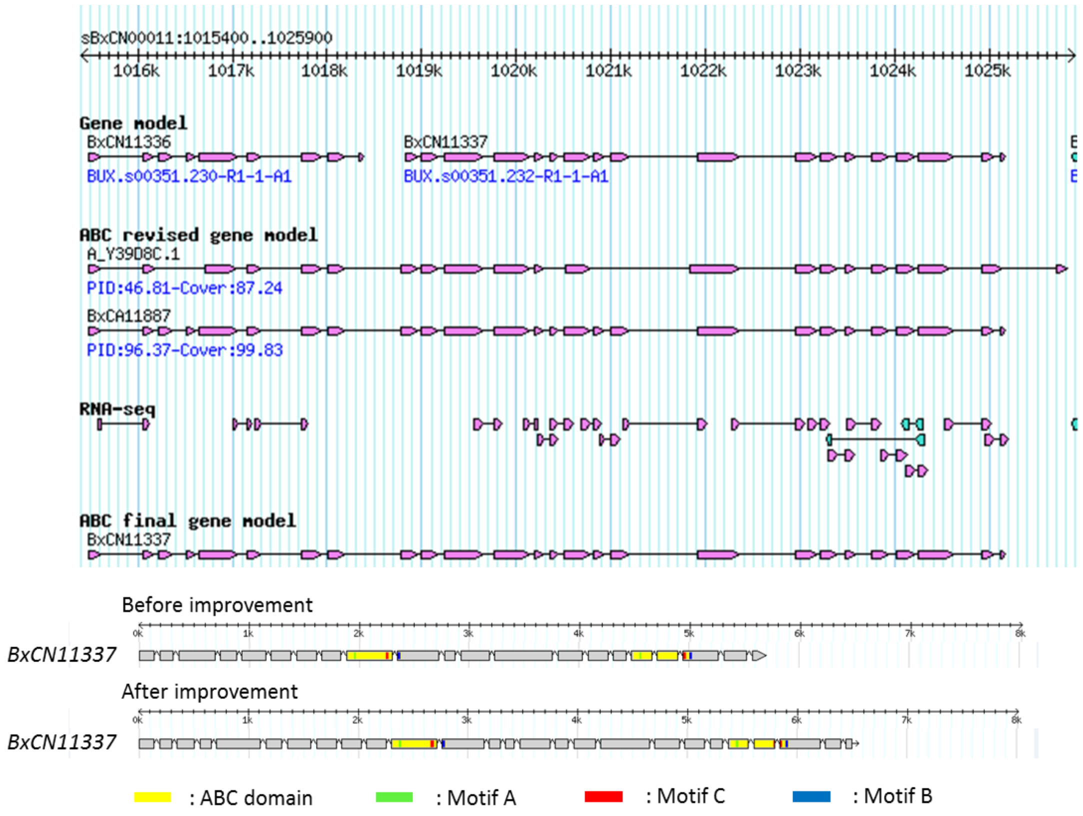


Figure 2.16: A representative case in which one candidate ABC gene model needed to be merged with its neighboring gene model

"Gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and C. elegans orthologs as query proteins; "RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. BxCN11336 was obtained from BLAST searches. After improvement, we found the newly constructed gene model using BxCA11887 (ortholog in BxCA) was supported better than that using Y39D8C.1 (ortholog in C. elegans). Based on the prediction, BxCN06787 and BxCN06788 were merged with each other and the newly constructed gene was annotated as a full ABC transporter in subfamily A.

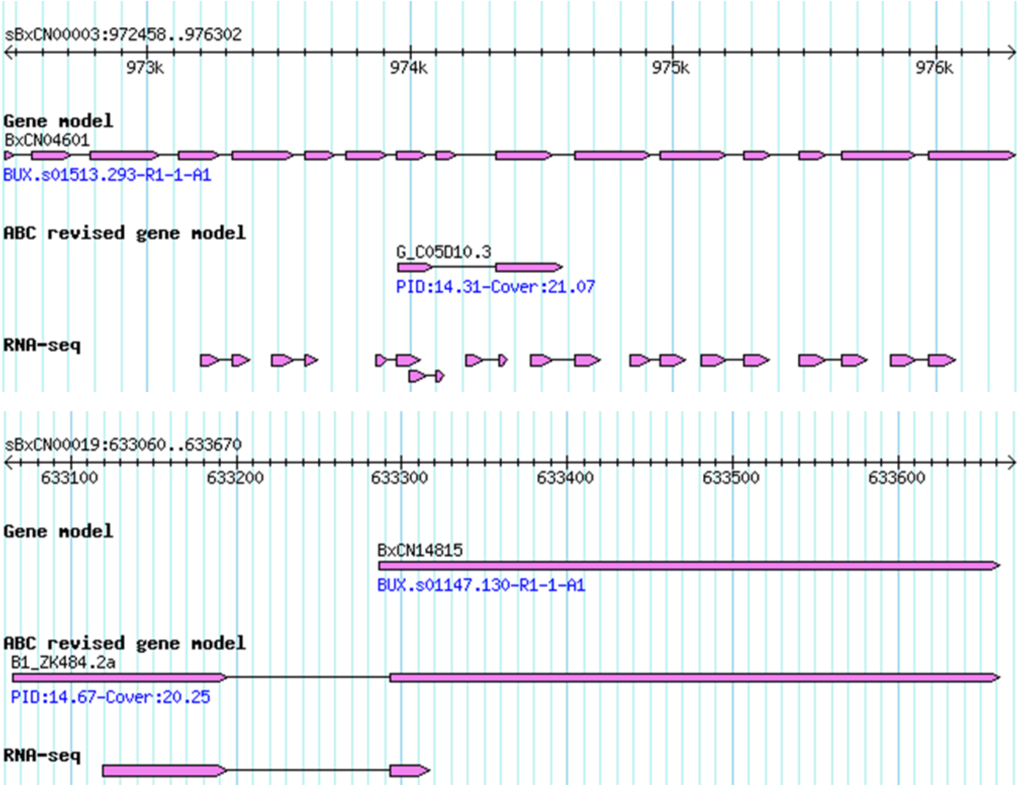


Figure 2.17: Two representative false positive cases in BxCN
"Gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and C. elegans orthologs as query proteins; "RNA-seq" track includes introns predicted by RNA-seq data; Two candidates, BxCN14815 and BxCN04601, did not have any ABC transporter related Pfam domains even after improvement and the length of predicted proteins (127 aa and 166 aa) were too short to be ABC transporter genes, suggesting that BxCN14815 and BxCN04601 were false positives from BLASTP searches

In summary, through BLAST searches and genBlastG improvement, two gene models in the high-quality ABC transporter gene set BxCN11337 and BxCN06787 were replaced by two newly constructed gene models, without changing the total number of high-quality ABC transporter genes in BxCN. Therefore, we confirmed that all the potential ABC transporter genes were identified and examined in the genome of BxCN.

2.3.7. Evaluating the completeness of each ABC transporter gene

To evaluate whether the 54 high-quality ABC transporter genes are full-length genes, the number of TM helices were also examined. For candidates that did not have the appropriate number of TM helices (full transporters from 10 to 17 and half transporters from four to 11), we tried to improve their gene models. However, we did not apply it as a judgement to exclude any candidates whose final number of TM helices was outside of the range. Among 54 high-quality ABC transporter genes, 52 of them had appropriate number of TM helices within TM domain (s) indicating that most ABC transporter genes in BxCN were complete after our examination and improvement. Two of them (BxCN14849 and BxCN14853a) had zero and three TM helices respectively, which were less than our criteria and these two gene models could not be further improved. We propose that a BxCN ABC transporter gene is likely full-length if it encodes a protein with similar length to its ortholog in *C. elegans* and with similar Pfam domains. Overall, the distributions of protein length between ABC transporters in BxCN and *C. elegans* are similar (Figure 2.18), suggesting that our ABC transporter gene set in BxCN had a good quality. More ABC transporter genes in BxCN contain a small number of introns when compared to those in BxCN (Figure 2.19), indicating the species specific gene attributes which may reflect the diversity of evolution between those two species. In the Pfam domain analysis, 50 of these 54 ABC transporter genes in BxCN shared the same Pfam domain with their orthologs (Table 2.5). Among four ABC transporter genes in BxCN that did not share the same Pfam domain with their ortholog, three of them (BxCN14848, BxCN14849, BxCN14853a) show shorter length (677 aa, 604 aa, 573 aa) than their *C. elegans* ortholog (801 aa). In particular, BxCN14853a contained gaps in the genomic region, indicating it was incomplete. The only member of subfamily H in BxCN, BxCN04600, had an additional Pfam domain, PF00089, compared to its ortholog C56E6.5 in *C. elegan*s. Thus BxCN04600 was likely a full-length ABC transporter gene with an extra Pfam domain. In addition, 48 of 54 ABC transporter genes in BxCN had both proper start codon and stop codon (Table 2.5).

Table 2.4: TM information of final set of ABC transporter genes in BxCN

Class	ID	Domain organization	Class	ID	Domain organization
A	BxCN11337	6TM-ABC-8TM-ABC	B (Half transporter)	BxCN01157	4TM-ABC
	BxCN11459	6TM-ABC-8TM-ABC		BxCN03385	5TM-ABC
	BxCN12661	10TM-ABC-6TM-ABC		BxCN05520	9TM-ABC
	BxCN14341	5TM-ABC-6TM-ABC		BxCN05567	8TM-ABC
	BxCN14342	5TM-ABC-8TM-ABC		BxCN12619	11TM-ABC
	BxCN14343	7TM-ABC-7TM-ABC		BxCN14229	11TM-ABC
B (Full transporter)	BxCN02239	6TM-ABC-6TM-ABC		BxCN14796	5TM-ABC
	BxCN08640	6TM-ABC-6TM-ABC		BxCN14811	8TM-ABC
	BxCN08790	6TM-ABC-5TM-ABC		BxCN14816	5TM-ABC
	BxCN09171	6TM-ABC-5TM-ABC		BxCN14818	10TM-ABC
	BxCN10724	6TM-ABC-6TM-ABC		BxCN14820	9TM-ABC
	BxCN13361	6TM-ABC-6TM-ABC		BxCN14847	6TM-ABC
	BxCN13602	6TM-ABC-6TM-ABC		BxCN14848	5TM-ABC
	BxCN13777	5TM-ABC-5TM-ABC		BxCN14849	0TM-ABC
	BxCN13779	6TM-ABC-5TM-ABC		BxCN14852	6TM-ABC
	BxCN16314	6TM-ABC-6TM-ABC		BxCN14853a	3TM-ABC
C	BxCN02679	7TM-ABC-6TM-ABC		BxCN14853b	4TM-ABC
	BxCN07404	8TM-ABC-6TM-ABC		BxCN14854	5TM-ABC
	BxCN09034	10TM-ABC-7TM-ABC		BxCN15675	6TM-ABC
	BxCN11889	10TM-ABC-7TM-ABC		BxCN15852	9TM-ABC
	BxCN13281	11TM-ABC-5TM-ABC	G	BxCN00285	ABC-6TM
D	BxCN05517	4TM-ABC		BxCN01094	ABC-6TM
	BxCN13607	6TM-ABC		BxCN02410	ABC-6TM
E	BxCN08056	ABC-ABC		BxCN04425	ABC-6TM
F	BxCN02823	ABC-ABC		BxCN06787	ABC-5TM
	BxCN07831	ABC-ABC		BxCN13610	ABC-7TM
	BxCN09178	ABC-ABC	H	BxCN04600	ABC-7TM

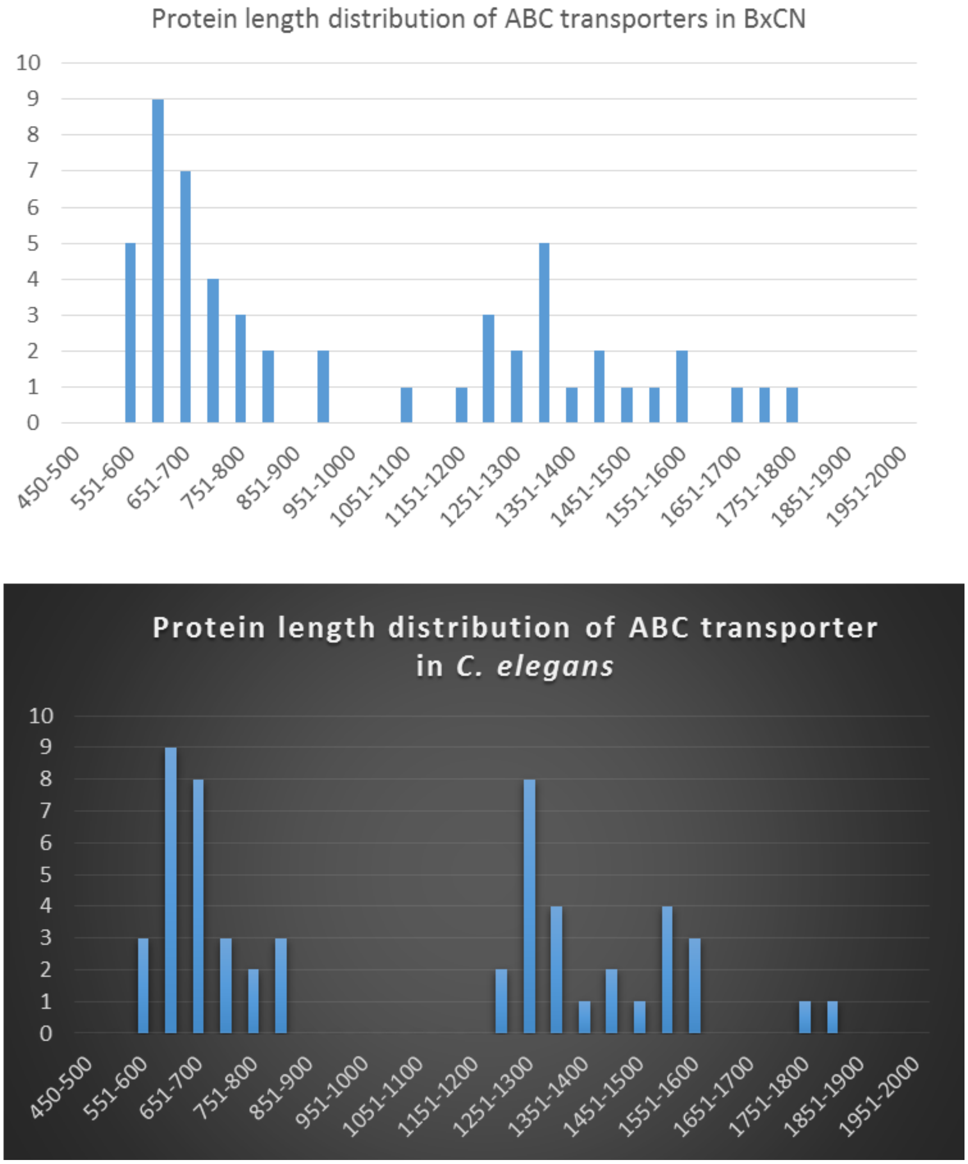
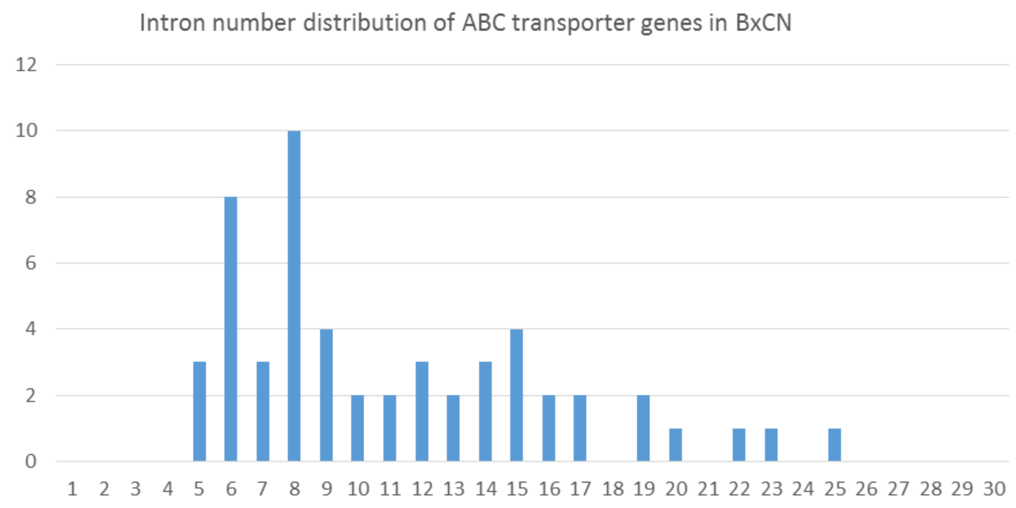


Figure 2.18: Protein length distribution of ABC transporters in BxCN and *C. elegans*

Protein length for each high-quality ABC transporter in BxCN and in *C. elegans* was obtained to draw the protein length distribution of ABC transporter



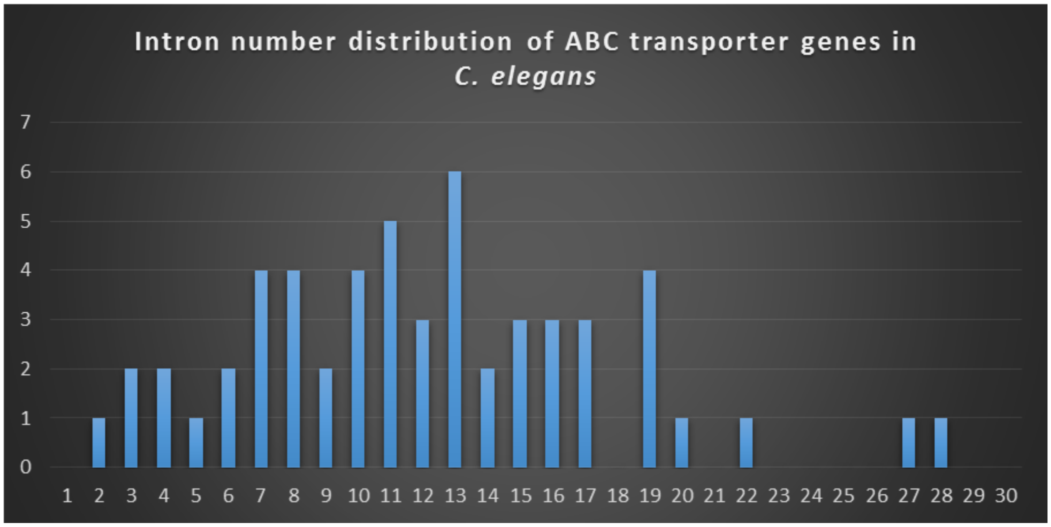


Figure 2.19: Intron number distribution of ABC transporters in BxCN and *C. elegans*

The number of intron for each high-quality ABC transporter in BxCN and in *C. elegans* was obtained to draw the intron number distribution of ABC transporter

Table 2.5: Additional characteristics of ABC transporter genes and proteins in BxCN

ID	Protein length	Ortholog protein length	Genomic span	Ortholog genomic span	Exon number	Number of intron supported	Ortholog exon number	Whether has start and stop codon	Other Pfam domain
BxCN00285	681	567	2633	2618	11	all	8	Yes	PF01061
BxCN01094	650	619	5329	3870	7	all	13	Yes	PF01061
BxCN01157	566	668	3428	12980	6	all	10	No start codon (GCG)	PF00664
BxCN02239	1331	1272	5028	9347	15	6	14	Yes	PF00664
BxCN02410	678	684	2923	7626	12	all	12	Yes	PF01061
BxCN02679	1272	1427	7839	8066	21	all	20	Yes	PF00664
BxCN02823	633	622	3018	2335	10	all	5	Yes	PF12848
BxCN03385	676	666	2309	5725	8	all	14	Yes	PF00664
BxCN04425	608	619	2491	3870	9	all	13	Yes	PF01061
BxCN04600	907	595	3591	2996	15	12	12	Yes	PF00089
BxCN05517	669	661	4270	3716	6	all	9	Yes	PF06472
BxCN05520	803	761	4518	2849	9	all	7	Yes	PF00664
BxCN05567	775	761	3215	2849	6	all	7	Yes	PF00664
BxCN06787	615	619	3150	3870	7	6	13	Yes	PF01061
BxCN07404	1478	1573	5323	5427	16	12	9	Yes	PF00664
BxCN07831	618	622	2128	4565	7	all	6	Yes	PF12848
BxCN08056	616	610	2256	6716	10	all	5	Yes	PF00037 and PF04068
BxCN08640	1301	1321	5245	7474	13	all	14	Yes	PF00664
BxCN08790	1364	1324	7848	5380	14	all	14	Yes	PF00664
BxCN09034	1592	1573	6443	5427	20	16	9	Yes	PF00664
BxCN09171	1238	1280	6580	5645	14	12	14	Yes	PF00664
BxCN09178	702	712	3056	2868	12	10	4	Yes	PF12848
BxCN10724	1244	1272	10536	9347	16	13	14	Yes	PF00664
BxCN11337	1573	1802	10345	6951	21	12	16	Yes	PF12698
BxCN11459	1657	1758	10321	9658	15	all	15	Yes	PF12698
BxCN11889	1418	1415	5197	8686	18	all	18	Yes	PF00664
BxCN12619	1064	815	4504	4181	20	18	16	Yes	PF00059 and PF00664

BxCN12661	1702	2146	7875	16944	24	11	33	No start codon (TCA)	PF12698
BxCN13281	1522	1525	6541	18202	13	9	14	No start codon (CCC)	PF00664
BxCN13361	1229	1268	5279	4955	17	15	13	Yes	PF00664
BxCN13602	1322	1280	5662	5645	16	all	14	Yes	PF00664
BxCN13607	675	734	2559	3168	9	all	12	Yes	PF06472
BxCN13610	702	598	2661	2783	9	all	11	Yes	PF01061
BxCN13777	1192	1280	6767	5645	10	7	14	Yes	PF00664
BxCN13779	1259	1318	8027	5583	13	6	15	Yes	PF00664
BxCN14229	797	801	3489	8399	7	all	12	Yes	PF00664
BxCN14341	1350	1758	9796	9658	16	13	15	No start codon (AAT)	PF12698
BxCN14342	1410	1564	9461	7393	17	13	20	Yes	PF12698
BxCN14343	1564	1758	8457	9658	18	14	15	Yes	PF12698
BxCN14796	594	668	4031	12980	7	4	10	Yes	PF00664
BxCN14811	740	761	3494	2849	8	all	7	Yes	PF00664
BxCN14816	711	761	4998	2849	9	7	7	Yes	PF00664
BxCN14818	918	761	5808	2849	11	8	7	Yes	PF00664
BxCN14820	794	761	3976	2849	9	7	7	Yes	PF00664
BxCN14847	632	801	4042	8399	7	all	12	Yes	PF00664
BxCN14848	677	801	3222	8399	8	all	12	Yes	
BxCN14849	604	801	2949	8399	7	2	12	Yes	
BxCN14852	637	801	2598	8399	7	all	12	Yes	PF00664
BxCN14853a	573	801	28465	8399	9	4	12	No start codon (TTT)	
BxCN14853b	580	801	3075	8399	9	5	12	No start codon (TTA)	PF00664
BxCN14854	587	801	5426	8399	10	7	12	Yes	PF00664
BxCN15675	686	704	2625	12965	9	all	8	Yes	PF00664
BxCN15852	801	787	2877	6171	9	all	10	Yes	PF00664
BxCN16314	1307	1382	14144	6981	23	12	29	Yes	PF00664

In summary, 46 of 54 ABC transporter genes in BxCN had both proper start and stop codon, appropriate TM domain, as well as the same set of Pfam domains to their orthologs in *C. elegans*, indicating that the majority of ABC transporter genes in BxCN are full-length high-quality gene models.

2.3.8. Finalizing the bioinformatics pipeline for annotating high-quality ABC transporter genes

Based on ABC transporter gene annotation in BxCN, we further revised the annotation pipeline (Figure 2.20). In order to demonstrate that our InterProScan and BLAST based analysis can be used as a precise approach to annotate ABC transporter genes, we compared the results against that obtained using another approach, which was based on the comparative gene family classification (Frech and Chen 2010). The comparison (Figure 2.21 and Figure 2.22) showed that the InterProScan & BLAST based bioinformatics pipeline (Figure 2.20) is more effective because it not only found all candidate ABC transporter genes, but also revised gene models that were defective and got high-quality ABC transporter genes.

In conclusion, through the analysis of ABC transporters in BxCN, a robust bioinformatics pipeline was developed to annotate ABC transporter, in which InterProScan and BLAST were applied in parallel to search for ABC transporter gene candidates in the nematode genomes

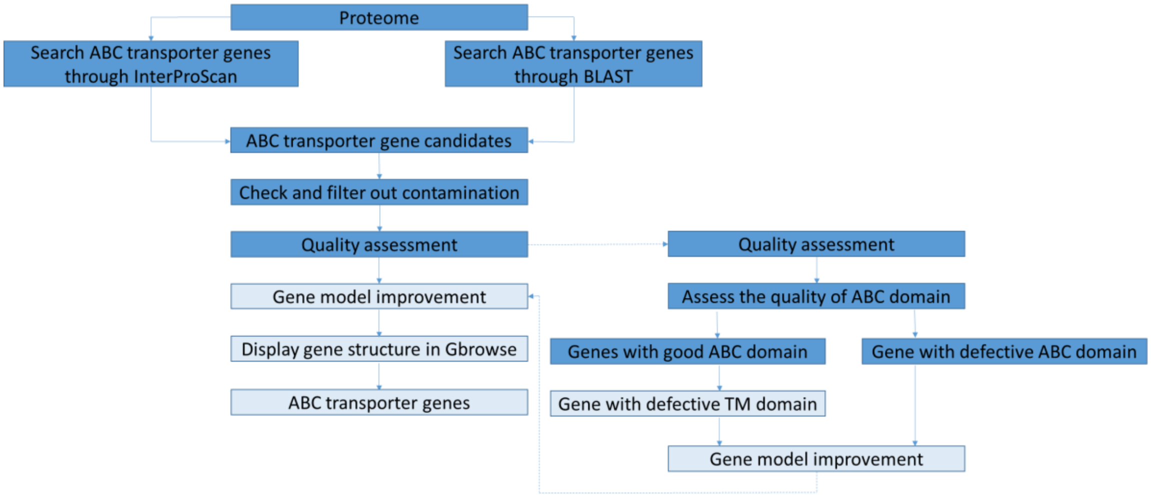

Proteome
Search ABC transporter genes through InterProScan
Search ABC transporter genes through BLAST
ABC transporter gene candidates
Check and filter out contamination
Quality assessment
Gene model improvement
Display gene structure in Gbrowse
ABC transporter genes
Quality assessment
Assess the quality of ABC domain
Genes with good ABC domain
Gene with defective ABC domain
Gene with defective TM domain
Gene model improvement

Figure 2.20: Final version of Bioinformatics pipeline for annotating ABC transporter genes.

Through InterProscan & BLAST searches, ABC transporter gene candidates were obtained. After filtering out contamination, assessing quality, improving gene model and re-assessing, ABC transporter with high-quality can be included in our final set. Analyses performed were fully automated (dark blue boxes) or involved some manual intervention (light blue boxes).

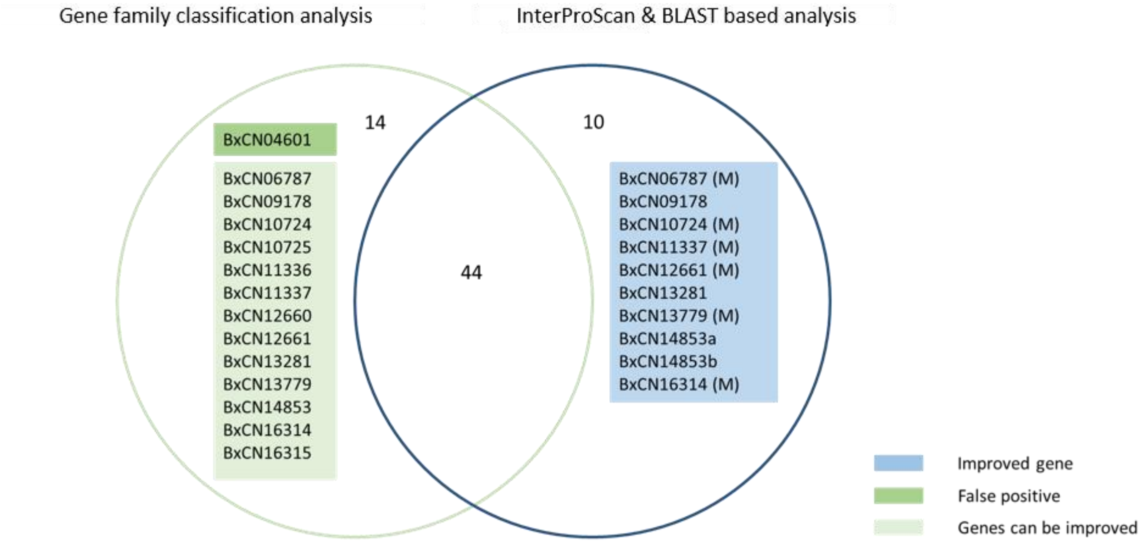


Figure 2.21: Comparing the results from gene family classification analysis and InterProScan & BLAST based analysis

44 high-quality ABC transporter genes were included in the results of both methods. Incomplete ABC transporter genes in the result of gene family classification analysis could be examined and improved by InterProScan & BLAST based analysis. False positive in the result of gene family classification analysis could be examined and excluded by InterProScan & BLAST based analysis.

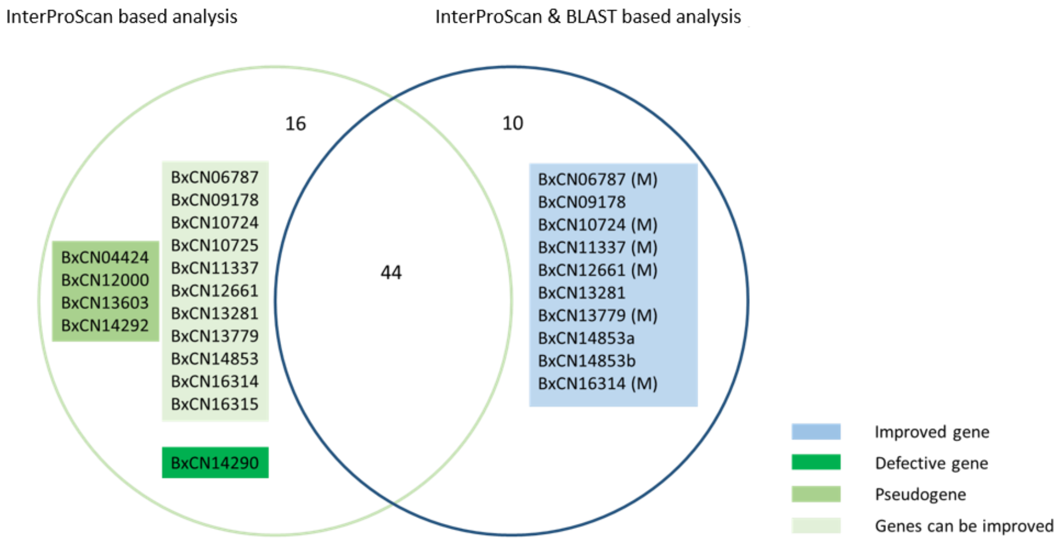


Figure 2.22: Comparing the result from InterProScan and InterProScan & BLAST based analysis

44 high-quality ABC transporter genes were included in the results of both methods. Incomplete ABC transporter genes in the result of InterProScan based analysis could be examined and improved by InterProScan & BLAST based analysis. Pseudogenes and defective genes in the result of InterProScan based analysis could be examined and excluded by InterProScan & BLAST based analysis.

2.3.9. Naming of ABC transporter genes

Gene names that contain functional information have been assigned to *C. elegans* ABC transporter genes. Assigning names to newly annotated ABC transporter genes in other nematode species may be useful and convenient for comparison analysis. Therefore, we constructed a phylogenetic tree for a combined set of 54 annotated high-quality ABC transporter genes in BxCN and 56 annotated ABC transporter genes in *C. elegans* using the neighbor-joining method available at MEGA6 (Tamura et al. 2013). To avoid confusion introduced by including the protein domains that are not shared by different proteins, we only included the ABC domains as proxies for all ABC transporter genes in the phylogenetic analysis. For full transporters that have two ABC domains, we only used the left domains to avoid redundancies. We have attempted to assign gene names to all newly annotated ABC transporter genes based on their orthologous relationship to ABC transporters in *C. elegans* (Figure 2.23). First of all, for ABC transporter gene in BxCN that shared clear one-to-one orthologous relationship with that in *C. elegans*, we simply assigned the same name of its ortholog in *C. elegans* to this

BxCN gene, for example, members in subfamily F. Then, for the sub-trees that contain *C. elegans* genes but show BxCN specific expansion, we first used all the reference gene names present in the sub-tree and then used the smallest number of reference gene name (in the same subfamily) that has not been used yet to name the additional BxCN genes, for example, the sub-tree with *haf-2*. Lastly, for the sub-tree that does not contain any reference genes, we created new name based on which subfamily the genes belonging to, for example, the sub-tree with *abcb-1* genes. After conducting the name rule above, we got a list of name corresponding to the ABC transporter genes in BxCN (Table 2.6).

Table 2.6: ABC transporter gene names in BxCN based on *C. elegans*

ID	Gene name	ID	Gene name	ID	Gene name
BxCN14341	BxCN-abt-1	BxCN14229	BxCN-hmt-1	BxCN07404	BxCN-mrp-3
BxCN12661	BxCN-abt-2	BxCN14848	BxCN-abcb-1	BxCN09034	BxCN-mrp-4
BxCN14342	BxCN-abt-3	BxCN14849	BxCN-abcb-2	BxCN02679	BxCN-mrp-5
BxCN11337	BxCN-abt-4	BxCN14847	BxCN-abcb-3	BxCN11889	BxCN-mrp-6
BxCN11459	BxCN-abt-5	BxCN14852	BxCN-abcb-4	BxCN13281	BxCN-mrp-7
BxCN14343	BxCN-abt-6	BxCN14853b	BxCN-abcb-5	BxCN05517	BxCN-pmp-1
BxCN15675	BxCN-abtm-1	BxCN14853a	BxCN-abcb-6	BxCN13607	BxCN-pmp-4
BxCN14796	BxCN-haf-1	BxCN14854	BxCN-abcb-7	BxCN08056	BxCN-abce-1
BxCN05520	BxCN-haf-10	BxCN13361	BxCN-pgp-1	BxCN02823	BxCN-abcf-1
BxCN14820	BxCN-haf-11	BxCN16314	BxCN-pgp-10	BxCN07831	BxCN-abcf-2
BxCN05567	BxCN-haf-2	BxCN02239	BxCN-pgp-11	BxCN09178	BxCN-abcf-3
BxCN03385	BxCN-haf-3	BxCN08790	BxCN-pgp-12	BxCN13610	BxCN-wht-1
BxCN15852	BxCN-haf-4	BxCN10724	BxCN-pgp-2	BxCN01094	BxCN-wht-2
BxCN14811	BxCN-haf-5	BxCN09171	BxCN-pgp-3	BxCN04425	BxCN-wht-3
BxCN01157	BxCN-haf-6	BxCN13602	BxCN-pgp-4	BxCN00285	BxCN-wht-4
BxCN14816	BxCN-haf-7	BxCN13777	BxCN-pgp-5	BxCN06787	BxCN-wht-5
BxCN14818	BxCN-haf-8	BxCN13779	BxCN-pgp-6	BxCN02410	BxCN-wht-7
BxCN12619	BxCN-haf-9	BxCN08640	BxCN-pgp-7	BxCN04600	BxCN-abch-1

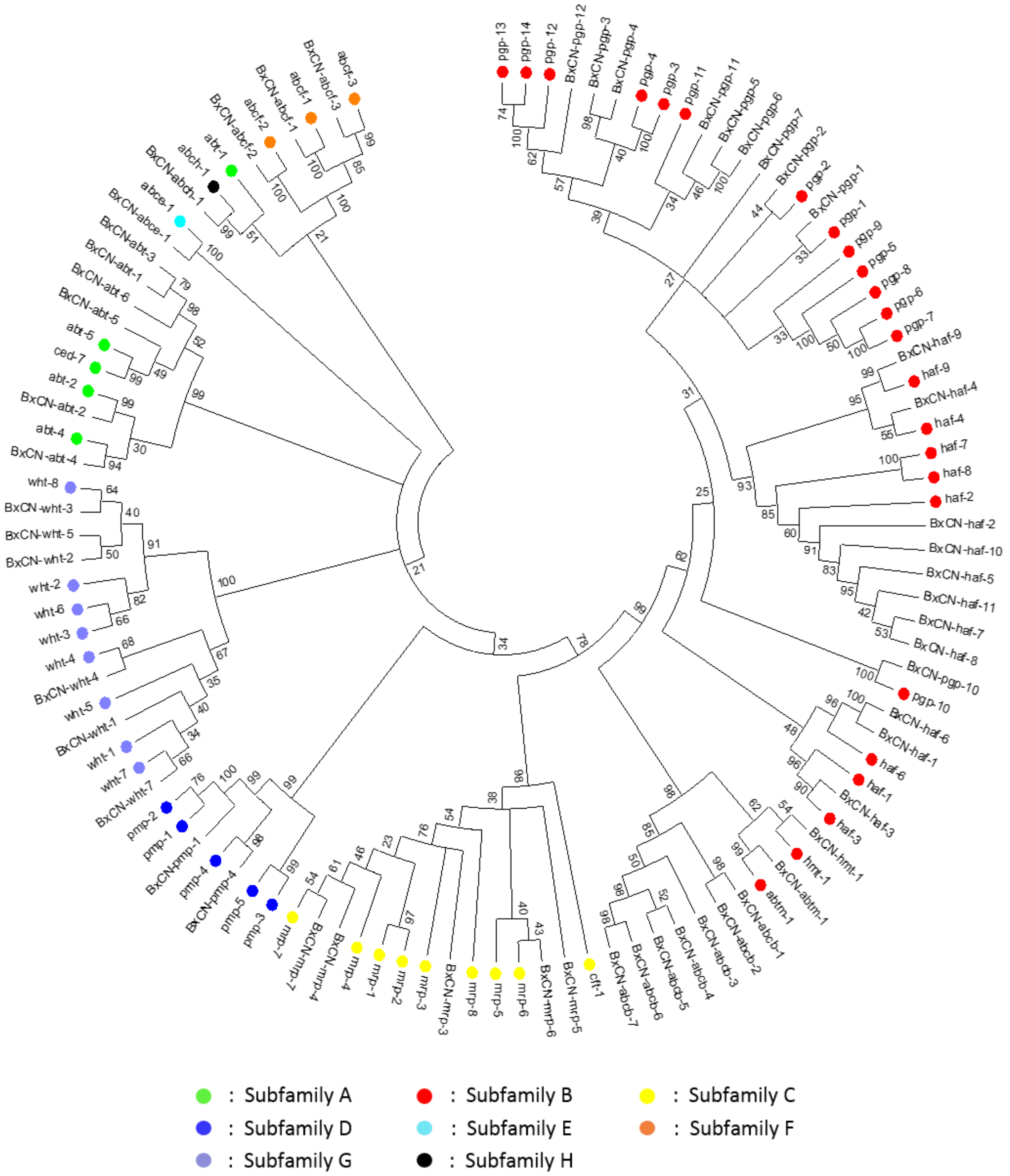


Figure 2.23: Phylogenetic analysis of ABC transporter genes in BxCN and *C. elegans*

Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only domain sequences of half transporters in BxCN and in *C. elegans*. ABC transporter gene names in *C. elegans* were obtained from WormBase and were highlighted by different color representing for different subfamilies. All ABC transporter gene names in BxCN were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

2.3.10. Comparative analysis of ABC transporter genes in *B. xylophilus* and *C. elegans*

Through phylogenetic analysis, we tried to evaluate the evolutionary relationships between ABC transporter genes in *B. xylophilus* (BxCN) and *C. elegans*. The phylogenetic tree showed that subfamilies E (1:1) and F (1:1) are highly conserved, with all members showing clear one-to-one orthologous relationships in *C. elegans* and BxCN, which is consistent with the previous studies (Zhao et al. 2007; Liu et al. 2011; Xie et al. 2012). Although members in subfamily H (1:1) also represented a one-to-one orthologous relationship, previous study demonstrated this subfamily appears to be the most divergent (Sheps et al. 2004). Subfamily B contains both half transporters and full transporters, showing more evolutionary activities than other subfamilies. For instance, in a species-specific expansions in BxCN, 7 ABC transporter genes (*BxCN-abcb-1*, *BxCN-abcb-2*, *BxCN-abcb-3*, *BxCN-abcb-4*, *BxCN-abcb-5*, *BxCN-abcb-6*, *BxCN-abcb-7*), which were half transporters, shared a common ancestor with 2 pairs of ABC transporter genes (*BxCN-abtm-1*, *abtm-1*; *BxCN-hmt-1* and *hmt-1*) in *C. elegans* and BxCN, resulting in a 9: 2 expansion in BxCN. This BxCN specific expansion might originate from tandem duplication. Subfamily A and subfamily C only contained full transporters. In subfamily A, there were two one-to-one orthologous relationship between BxCN and *C. elegans* (*BxCN-abt-2* and *abt-2*, *BxCN-abt-4* and *abt-4*), as well as one BxCN specific expansion (*BxCN-abt-1*, *BxCN-abt-3*, *BxCN-abt-5* and *BxCN-abt-6*). *BxCN-mrp-7* and *mrp-7* was the only pair in subfamily C that showed one-to-one orthologous relationship. Subfamily D and subfamily G both contained half transporters. Similar to subfamily C, there was just one one-to-one orthologous relationship in subfamily D (*BxCN-pmp-4* and *pmp-4*). This small subfamily showed two species-specific expansions in *C. elegans*. Members in subfamily G also represented one-to-one orthologous relationship (*BxCN-wht-3* and *wht-3*, *BxCN-wht-4* and *wht-4*, *BxCN-wht-7* and *wht-7*).

In summary, although there are similar numbers of ABC transporter genes in *C. elegans* (56) and BxCN (54), only 21 pairs of ABC transporter genes in *C. elegans* and BxCN showed clear one-to-one orthologous relationships. These genes might perform important and similar functions in these two species. Species-specific ABC transporter expansion in BxCN might be related to the specific function that needed to interact with the surrounding environment, reflecting the consistency between molecular level of genome and behavior of an organism.

2.4. Annotating ABC transporter genes in *B. xylophilus* (BxJP)

Considering that a recent study reported 106 of ABC transporter genes in a *B. xylophilus* strain isolated in Japan (thus BxJP) (Kikuchi et al. 2011) but our annotation found only 54 ABC transporter genes in BxCN, we would like to re-annotate ABC transporter genes in BxJP using our bioinformatics pipeline. We found 68 ABC transporter gene candidates in BxJP, 57 of which were found though InterProScan searches and 11 additional ones were found through BLAST searches. One candidate, BxJP18132, was obtained from contamination. After excluding BxJP18132, quality assessment identified 30 high-quality ABC transporter and 37 candidates that were defective and needed improvement (Table 2.7). Our improvement procedure generated six revised gene models of high-quality. In total, we annotated 49 high-quality ABC transporter genes, 46 of which also encode appropriate number of TM helices within TM domain (s) (Table 2.8). Thus, in contrast to what was reported (Kikuchi et al. 2011), the size of ABC transporter gene family is not substantially larger than other in other nematodes.

Table 2.7: Improvement of defective ABC transporter gene models in BxJP

ID	Improved or not	ID after improvement	Notes
BxJP06513	No		Keep the original model (with shorter ABC domain length)
BxJP10733	Yes		
BxJP11047	Yes		
BxJP11366	Yes		Merged with BxJP11367
BxJP11367	Yes		
BxJP11659	Yes		
BxJP11658	Yes	BxJP11658	Merged with BxJP11659
BxJP12666	Yes		
BxJP12665	Yes	BxJP12665	Merged with BxJP12665
BxJP14501	Yes		keep original model (based on the ortholog in BxCN)
BxJP14502	Yes		keep original model (based on the ortholog in BxCN)
BxJP14506	Yes	BxJP14506a BxJP14506b	Split into two genes
BxJP15073	Yes		
BxJP15074	Yes		
BxJP15072	Yes	BxJP15072	Merged with BxJP15073, BxJP15074
BxJP16375	Yes	BxJP16375	
BxJP05284	Yes	BxJP05284	
BxJP09873	Yes	BxJP09873	
BxJP16345	No		TM domain could not be improved; keep the original model
BxJP07046	No		
BxJP08124	No		
BxJP08495	No		
BxJP12494	No		
BxJP13552	No		
BxJP14507	No		
BxJP14472	No		
BxJP16233	No		
BxJP16388	No		
BxJP16650	No		
BxJP17299	No		
BxJP17492	No		
BxJP18132	No		

Table 2.8: High-quality ABC transporter genes in BxJP

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*BxJP-abt-1*	BxJP13601	BxJP13601	3TM-ABC-7TM-ABC	8875	15	1340	No start codon
	BxJP-abt-2	BxJP11658	BxJP11658	8TM-ABC-8TM-ABC	7875	23	1717	BxJP11659 was merged BxJP11658; No start codon
	BxJP-abt-4	BxJP09873	BxJP09873	6TM-ABC-8TM-ABC	9918	22	1553	BxJP09874 was merged with BxJP09873; TM helices were improved
	BxJP-abt-5	BxJP09986	BxJP09986	6TM-ABC-8TM-ABC	10585	15	1657	
	BxJP-abt-6	BxJP13604	BxJP13604	8TM-ABC-2TM-ABC	6146	15	1287	
B	*BxJP-abtm-1*	BxJP13943	BxJP13943	6TM-ABC	2625	9	686	
	BxJP-haf-10	BxJP03929	BxJP03929	9TM-ABC	4518	9	803	
	BxJP-haf-11	BxJP14475	BxJP14475	9TM-ABC	3975	9	794	
	BxJP-haf-2	BxJP16345	BxJP16345	0TM-ABC	1566	3	395	No start codon
	BxJP-haf-3	BxJP16197	BxJP16197	4TM-ABC	2309	8	617	
	BxJP-haf-4	BxJP14177	BxJP14177	9TM-ABC	2878	9	801	
	BxJP-haf-5	BxJP14468	BxJP14468	8TM-ABC	3494	8	740	
	BxJP-haf-6	BxJP16253	BxJP16253	4TM-ABC	2115	5	566	No start codon
	BxJP-haf-8	BxJP14473	BxJP14473	7TM-ABC	4334	9	711	
	BxJP-haf-9	BxJP11620	BxJP11620	11TM-ABC	4504	20	1064	
	BxJP-hmt-1	BxJP13498	BxJP13498	11TM-ABC	3489	7	797	
	BxJP-abcb-1	BxJP14501	BxJP14501	5TM-ABC	3219	8	677	
	BxJP-abcb-2	BxJP14502	BxJP14502	0TM-ABC	2667	7	604	
	BxJP-abcb-3	BxJP14500	BxJP14500	6TM-ABC	4045	7	632	
	BxJP-abcb-4	BxJP14505	BxJP14505	6TM-ABC	2606	7	637	
	BxJP-abcb-5	BxJP14506b	BxJP14506	4TM-ABC	3751	9	500	Split from BxJP14506
	BxJP-abcb-6	BxJP14506a	BxJP14506	4TM-ABC	4006	7	579	Split from BxJP14506; No start codon
	BxJP-pgp-1	BxJP10812	BxJP10812	6TM-ABC-6TM-ABC	5279	17	1229	
	BxJP-pgp-10	BxJP15072	BxJP15072	6TM-ABC-6TM-ABC	14486	23	1307	
	BxJP-pgp-11	BxJP02116	BxJP02116	6TM-ABC-6TM-ABC	5028	15	1331	
	BxJP-pgp-2	BxJP11366	BxJP11366	6TM-ABC-6TM-ABC	10571	16	1244	BxJP11367 was merged with BxJP11366
	BxJP-pgp-3	BxJP08488	BxJP08488	6TM-ABC-5TM-ABC	5748	13	1237	
	BxJP-pgp-4	BxJP11047	BxJP11047	5TM-ABC-6TM-ABC	5189	17	1225	Exons were improved; No start codon
	BxJP-pgp-5	BxJP12663	BxJP12663	5TM-ABC-5TM-ABC	6612	10	1192	
	BxJP-pgp-6	BxJP12665	BxJP12665	6TM-ABC-4TM-ABC	7840	12	1222	BxJP12666 was merged with BxJP12665; No start codon
	BxJP-pgp-7	BxJP07975	BxJP07975	6TM-ABC-6TM-ABC	5245	13	1301	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
C	*BxJP-mrp-3*	BxJP05848	BxJP05848	8TM-ABC-6TM-ABC	5323	16	1478	
	BxJP-mrp-4	BxJP08360	BxJP08360	10TM-ABC-7TM-ABC	6443	20	1592	
	BxJP-mrp-5	BxJP02494	BxJP02494	7TM-ABC-6TM-ABC	7860	21	1272	
	BxJP-mrp-6	BxJP12384	BxJP12384	10TM-ABC-7TM-ABC	5197	18	1418	
	BxJP-mrp-7	BxJP10733	BxJP10733	11TM-ABC-3TM-ABC	5381	12	1498	Exons were improved; No start codon
D	*BxJP-pmp-1*	BxJP03926	BxJP03926	4TM--ABC	4271	6	669	
	BxJP-pmp-4	BxJP11051	BxJP11051	6TM--ABC	2563	9	675	
E	*BxJP-abce-1*	BxJP06513	BxJP06513	ABC-ABC	2256	10	616	
F	*BxJP-abcf-1*	BxJP02633	BxJP02633	ABC-ABC	3018	10	633	
	BxJP-abcf-2	BxJP06293	BxJP06293	ABC-ABC	2128	7	618	
	BxJP-abcf-3	BxJP16375	BxJP16375	1TM-ABC-1TM-ABC	2890	12	689	Exons were improved
G	*BxJP-wht-1*	BxJP11054	BxJP11054	ABC-7TM	3806	9	702	
	BxJP-wht-2	BxJP01106	BxJP01106	ABC-6TM	8281	7	650	
	BxJP-wht-3	BxJP06867	BxJP06867	ABC-6TM	2490	9	608	
	BxJP-wht-4	BxJP00275	BxJP00275	ABC-6TM	2633	11	681	
	BxJP-wht-5	BxJP05283	BxJP05283	ABC-7TM	4926	8	582	BxJP05284 was merged with BxJP05283; TM helices were improved; No start codon
	BxJP-wht-7	BxJP02231	BxJP02231	ABC-6TM	2923	12	678	
H	*BxJP-abch-1*	BxJP07045	BxJP07045	ABC-7TM	3592	15	907	

Because BxCN and BxJP are two strains of a same species *B. xylophilus*, we expect that the size of their ABC transporter gene families are very close, if not identical. To compare ABC transporter genes in BxJP and BxCN, a phylogenetic tree containing all ABC transporter genes was constructed. All 49 BxJP ABC transporter genes showed clear one-to-one orthologous relationship with those in BxCN and shared the same name with their ortholog in BxCN (Figure 2.24). The ortholog of *BxCN-pgp-12* in BxJP was fragmented into two genes *BxJP08124* and *BxJP16388* due to genome assembly error (Figure 2.25) so that we did not included this defective gene. Thus, there was no ortholog of *BxCN-pgp-12* in the phylogenetic tree. *BxCN-abcb-7* does not have any ortholog in BxJP, which may also due to assembly error or sequencing error. The ortholog of *BxCN-abt-3* in BxJP was a pseudogene (Figure 2.26), fragmented into *BxJP13602* and *BxJP13603* which contain TM domain but no ABC domain. In comparison to BxJP, BxCN had two additional genes, *BxCN-haf-8* and *BxCN-haf-1*. *BxCN-haf-1* was a result of a local duplication, showing high similarity to *BxCN-haf-6* based on JDotter result (Figure 2.27). Similarly, *BxCN-haf-8* may be duplicated by *BxCN-haf-11*(Figure 2.28). In conclusion, BxJP shared an almost identical set of ABC transporter genes with BxCN with three exceptions and did not contain an expanded number of ABC transporter genes as reported by previous study.

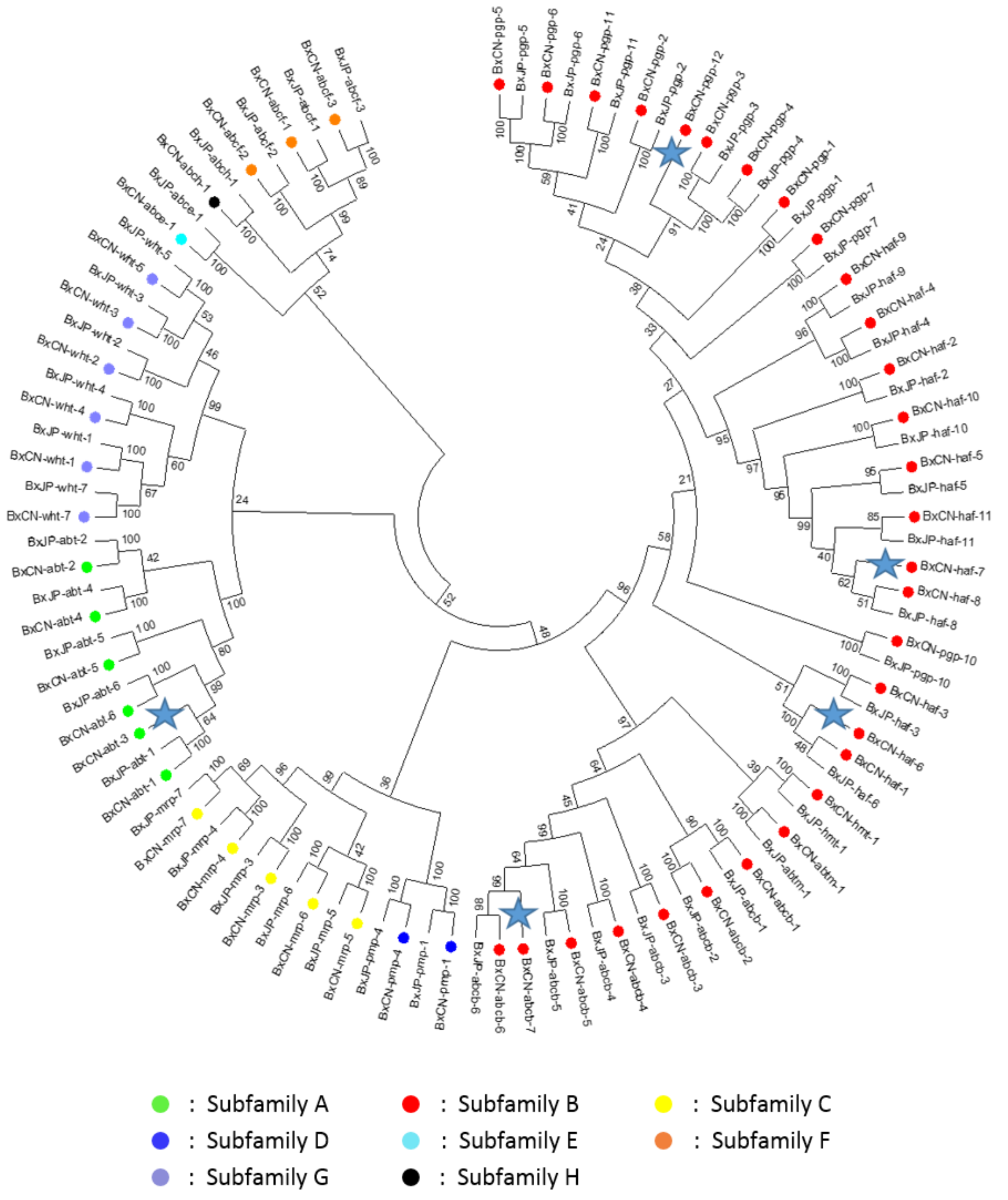


Figure 2.24: Phylogenetic analysis of ABC transporter genes in BxJP and BxCN

Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the ABC only domain sequences of half transporters in BxJP and in BxCN. ABC transporter genes in BxCN were highlighted by different color representing for different subfamilies. All ABC tranpsorter gene names in BxJP wereassigned based on ABC transporter genes in BxCN by applying the name rule. Stars stands for the genes that do not have orthologs in BxJP or BxCN

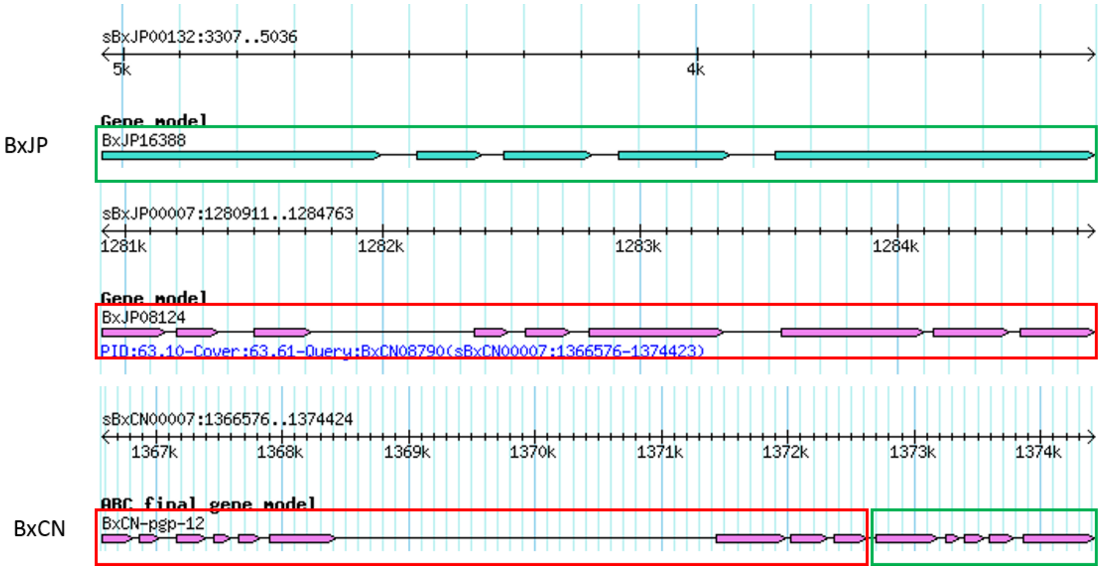


Figure 2.25: *BxJP-pgp-12* was fragmented into two gene BxJP08124 and BxJP16388 due to assembly error.

"Gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. BxJP16388 and BxJP09124 were located in different contigs and were two fragments of BxJP-pgp-12. BxJP16388 shared orthologous relationship to the right part of *BxCN-pgp-12* and BxJP09124 shared orthologous relationship to the left part of *BxCN-pgp-12*. The assembly error made *BxJP-pgp-12* fragmented.

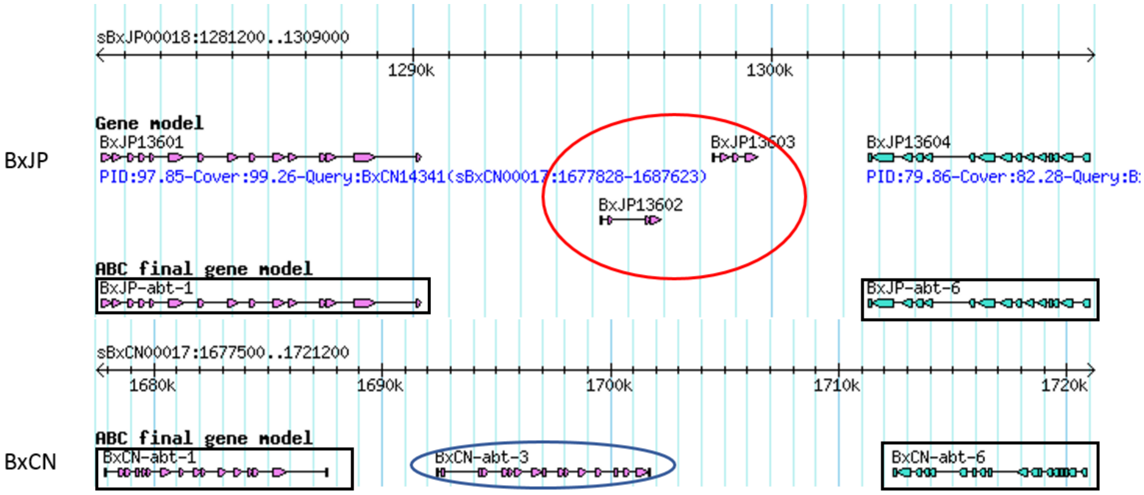
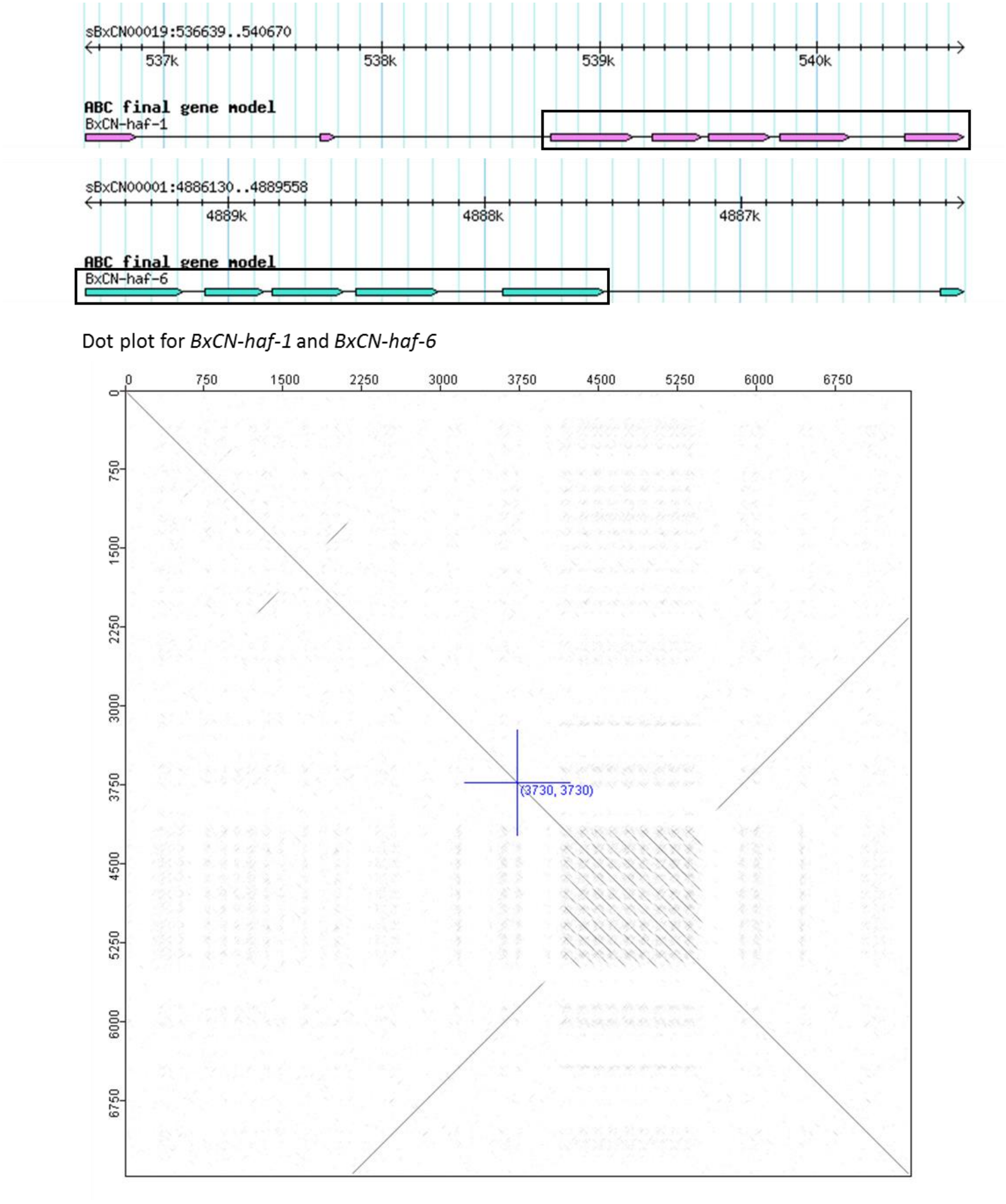


Figure 2.26: *BxJP-abt-3* was a pseudogene, fragmented into BxJP13602 and BxJP13603.

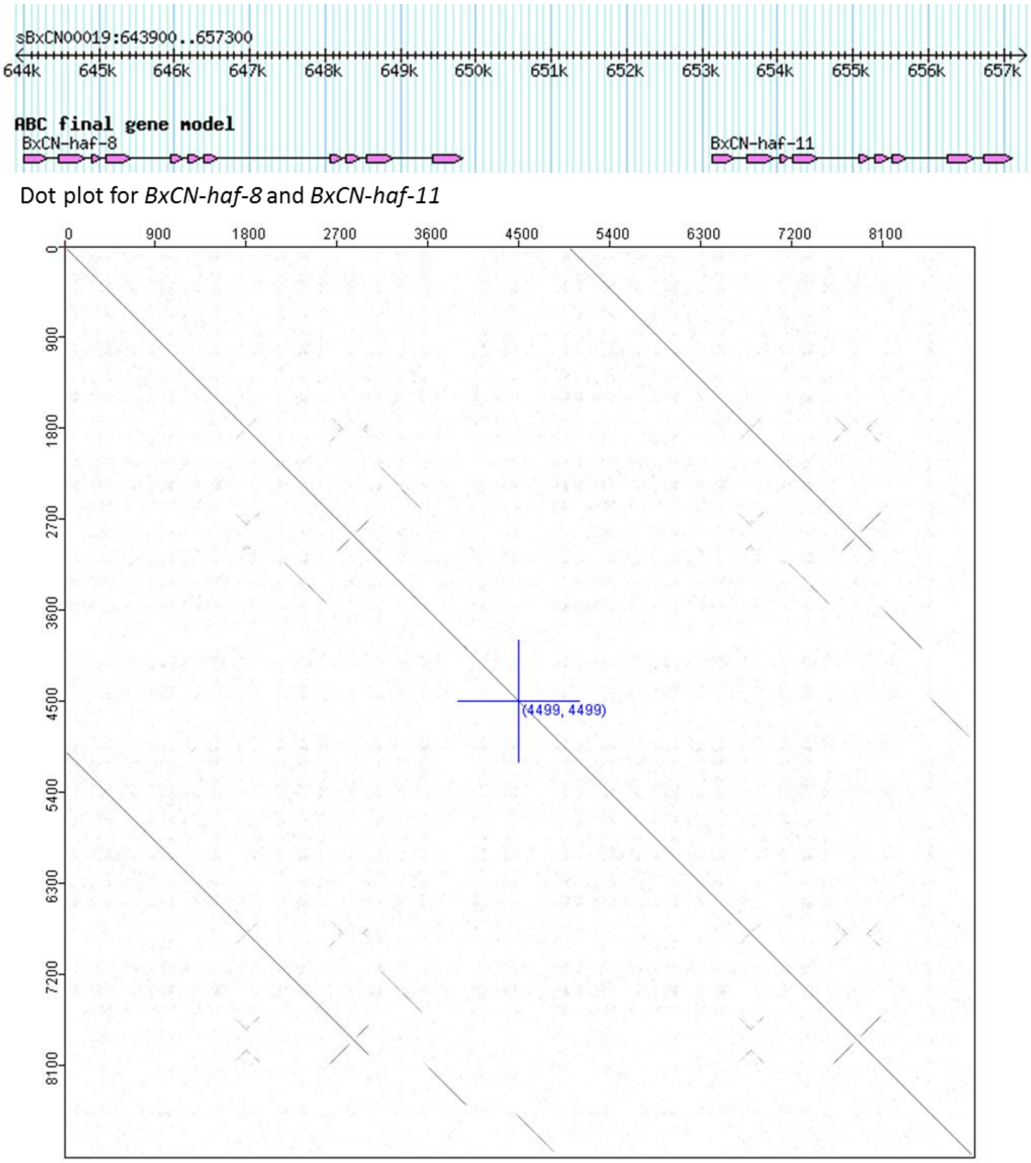
"Gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. Genes in the boxes are conserved between BxJP and BxCN. *BxJP-abt-3* was likely a pseudogene, fragmented into BxJP13602 and BxJP13603 which contained TM domain but no ABC domain



Dot plot for *BxCN-haf-1* and *BxCN-haf-6*

Figure 2.27: A part of *BxCN-haf-1* could be a duplication of the corresponding part of *BxCN-haf-6*

"ABC final gene model" track includes high-quality ABC transporter gene models. DNA sequences of the genomic region for *BxCN-haf-1* and *BxCN-haf-6* were extracted and combined to run JDotter. The result showed the two regions highlighted by black boxes shared some similarity.



Dot plot for *BxCN-haf-8* and *BxCN-haf-11*

Figure 2.28: One BxCN specific gene, *BxCN-haf-8* that could be obtained from local duplication.

"ABC final gene model" track includes high-quality ABC transporter gene models. DNA sequences of the genomic region for *BxCN-haf-8* and *BxCN-haf-11* were extracted and combined to run JDotter. The result showed these two regions shared some similarity.

2.5. Annotating ABC transporter genes in *B. xylophilus* (BxCA)

In order to further confirm that *B. xylophilus* has a similar number of ABC transporter genes to that in *C. elegans,* we annotated ABC transporter genes in BxCA, which was an M-form *B. xylophilus* strain showing mild pathogenicity to pine wood tree. we found 62 ABC transporter gene candidates, of which 58 were found by InterProScan searches and four additional ones were obtained by BLAST searches. According to TBLASTN result, none of the candidates was contamination. Quality assessment identified 39 high-quality ABC transporter genes and 23 candidates that were defective and needed improvement (Table 2.9)

Table 2.9: Improvement of defective ABC transporter gene models in BxCA

ID	Improved or not	ID after improvement	Notes
BxCA06754	Yes	BxCA06754	
BxCA08322	No		Keep the original model (with a shorter ABC domain length)
BxCA09571	Yes		
BxCA12681	Yes		
BxCA12687	Yes		Do not include this model
BxCA12754	No		Keep the original model
BxCA12925	Yes		
BxCA14980	No		Keep the original model
BxCA15803	Yes		
BxCA15804	Yes		Keep the original model
BxCA15805	Yes		Keep the original model
BxCA15809	Yes	BxCA15809a BxCA15809b	Split into two genes
BxCA17262	Yes	BxCA17262	Merged with BxCA17263 and BxCA17264
BxCA17263	Yes		
BxCA17264	Yes		
BxCA01901	No		
BxCA04529	No		
BxCA04710	No		
BxCA07652	No		
BxCA15813	No		
BxCA15810	No		
BxCA17905	No		

Our improvement procedure generated eight revised gene models of high-quality. In total, 53 high-quality ABC transporter genes were annotated in our analysis, the majority (50) of which also encode appropriate number of TM helices within TM domain (s) (Table 2.10). Together with BxCN and BxJP, three *B. xylophilus* genomes share similar number of ABC transporter genes, suggesting that the previous annotation was incorrect.

Table 2.10: High-quality ABC transporter genes in BxCA

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*BxCA-abt-1*	BxCA12754	BxCA12754	5TM-ABC-6TM-ABC	9505	17	1322	No start codon
	BxCA-abt-2	BxCA14980	BxCA14980	10TM-ABC-6TM-ABC	9700	32	2212	
	BxCA-abt-3	BxCA12755	BxCA12755	6TM-ABC-8TM-ABC	12889	18	1584	
	BxCA-abt-4	BxCA11887	BxCA11887	6TM-ABC-8TM-ABC	8934	26	1762	
	BxCA-abt-5	BxCA11828	BxCA11828	6TM-ABC-8TM-ABC	11342	15	1652	
	BxCA-abt-6	BxCA12756	BxCA12756	8TM-ABC-6TM-ABC	8020	18	1563	
B	*BxCA-abtm-1*	BxCA15323	BxCA15323	6TM-ABC	2625	9	686	
	BxCA-haf-10	BxCA16512	BxCA16512	9TM-ABC	6300	10	805	
	BxCA-haf-11	BxCA15780	BxCA15780	5TM-ABC	3647	10	690	
	BxCA-haf-2	BxCA05704	BxCA05704	8TM-ABC	2689	7	742	
	BxCA-haf-3	BxCA03480	BxCA03480	5TM-ABC	2506	9	664	
	BxCA-haf-4	BxCA15488	BxCA15488	9TM-ABC	2879	9	801	
	BxCA-haf-5	BxCA15779	BxCA15779	9TM-ABC	3140	8	822	
	BxCA-haf-6	BxCA01269	BxCA01269	6TM-ABC	2897	8	672	
	BxCA-haf-8	BxCA15782	BxCA15782	9TM-ABC	3492	9	799	
	BxCA-haf-9	BxCA14941	BxCA14941	11TM-ABC	4504	20	1064	
	BxCA-hmt-1	BxCA12604	BxCA12604	11TM-ABC	2949	7	797	
	BxCA-abcb-1	BxCA15804	BxCA15804	6TM-ABC	2835	8	677	
	BxCA-abcb-2	BxCA15805	BxCA15805	3TM-ABC	2420	8	665	
	BxCA-abcb-3	BxCA15803	BxCA15803	5TM-ABC	3822	9	612	Exons were improved; No start codon
	BxCA-abcb-4	BxCA15808	BxCA15808	6TM-ABC	2735	7	637	
	BxCA-abcb-5	BxCA15809a	BxCA15809	3TM-ABC	3277	9	540	Split from BxCA15809; No start codon
	BxCA-abcb-6	BxCA15809b	BxCA15809	4TM-ABC	3627	8	520	Split from BxCA15809; No start codon
	BxCA-abcb-8	BxCA06754	BxCA06754	3TM-ABC	2342	6	431	Exons were improved
	BxCA-pgp-1	BxCA13011	BxCA13011	6TM-ABC-6TM-ABC	5862	18	1254	
	BxCA-pgp-10	BxCA17262	BxCA17262	4TM-ABC-7TM-ABC	16946	20	1141	BxCA17263 and BxCA17264 were merged with BxCA17262
	BxCA-pgp-11	BxCA02348	BxCA02348	6TM-ABC-6TM-ABC	4916	15	1331	
	BxCA-pgp-12	BxCA09060	BxCA09060	6TM-ABC-5TM-ABC	7151	15	1343	
	BxCA-pgp-2	BxCA10561	BxCA10561	6TM-ABC-6TM-ABC	12206	16	1288	
	BxCA-pgp-3	BxCA09564	BxCA09564	6TM-ABC-5TM-ABC	5751	14	1261	
	BxCA-pgp-4	BxCA13256	BxCA13256	6TM-ABC-6TM-ABC	9016	16	1322	
	BxCA-pgp-5	BxCA13437	BxCA13437	6TM-ABC-5TM-ABC	10900	11	1272	
	BxCA-pgp-6	BxCA13440	BxCA13440	6TM-ABC-6TM-ABC	13025	12	1247	
	BxCA-pgp-7	BxCA08904	BxCA08904	6TM-ABC-6TM-ABC	7111	13	1301	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
C	*BxCA-mrp-1*	BxCA12681	BxCA12681	11TM-ABC-5TM-ABC	9302	18	1476	Exons were improved; No start codon
	BxCA-mrp-4	BxCA09426	BxCA09426	10TM-ABC-7TM-ABC	6732	20	1592	
	BxCA-mrp-5	BxCA02791	BxCA02791	7TM-ABC-6TM-ABC	6989	21	1272	
	BxCA-mrp-6	BxCA14016	BxCA14016	10TM-ABC-7TM-ABC	7154	18	1418	
	BxCA-mrp-7	BxCA12925	BxCA12925	11TM-ABC-5TM-ABC	5354	12	1502	Exons were improved; No start codon
D	*BxCA-pmp-1*	BxCA05648	BxCA05648	4TM-ABC	3164	6	669	
	BxCA-pmp-2	BxCA16515	BxCA16515	4TM-ABC	4462	6	669	
	BxCA-pmp-4	BxCA13260	BxCA13260	6TM-ABC	2559	9	675	
E	*BxCA-abce-1*	BxCA08322	BxCA08322	ABC-ABC	2263	10	616	
F	*BxCA-abcf-1*	BxCA02933	BxCA02933	ABC-ABC	3025	10	633	
	BxCA-abcf-2	BxCA08091	BxCA08091	ABC-ABC	2885	7	618	
	BxCA-abcf-3	BxCA09571	BxCA09571	ABC-ABC	3214	13	689	
G	*BxCA-wht-1*	BxCA13263	BxCA13263	ABC-7TM	2669	9	702	
	BxCA-wht-2	BxCA01209	BxCA01209	ABC-6TM	7433	7	649	
	BxCA-wht-3	BxCA04527	BxCA04527	ABC-6TM	2473	9	608	
	BxCA-wht-4	BxCA00335	BxCA00335	ABC-6TM	3806	11	681	
	BxCA-wht-5	BxCA07084	BxCA07084	ABC-6TM	6225	9	638	
	BxCA-wht-7	BxCA02523	BxCA02523	ABC-6TM	2860	12	678	
H	*BxCA-abch-1*	BxCA04709	BxCA04709	ABC-7TM	3587	15	907	

Although BxCA is a M-from *B.xylophilus* strain, we expect that ABC transporter gene families are very similar in BxCA and BxCN. To compare ABC transporter genes in BxCA and BxCN, a phylogenetic tree containing all ABC transporter genes was constructed. Among 53 BxCA ABC transporter genes, 50 of them showed one-to-one orthologous relationship and shared the same name with their ortholog in BxCN (Figure 2.29). One difference between BxCA and BxCN was that the ortholog of *BxCN-abcb-7* and that of *BxCN-mrp-3* were both defective in BxCA. More specifically, *BxCA-abct-7* (*BxCA15810*) had one shorter ABC domain with a length of 60 aa due to assembly error. Similarly, *BxCA-mrp-3* (*BxCA07652*) had two shorter ABC domains with lengths of 40 aa. In subfamily D, one extra gene in BxCA, *BxCA-pmp-2* (*BxCA16515*), was a result of duplication, showing in Figure 2.30. In subfamily B, one extra gene in BxCA, *BxCA-abcb-8*, was improved by using its ortholog in BmCN, a strain of a related species *Bursaphelenchus mucronatus* (Zhejiang, China). However neither BxCN nor BxJP had it. As shown in Figure 2.31, BxJP and BxCN did not contain any *abcb-8* candidates in the syntenic genomic region after using *BmCN-abcb-8* as query to run genBlastG. Another difference was caused by defective *BxCN-mrp-1* (*BxCN14290*) due to assembly error. To conclude, BxCA shows a similar number of ABC transporter genes in BxCN with a slight difference.

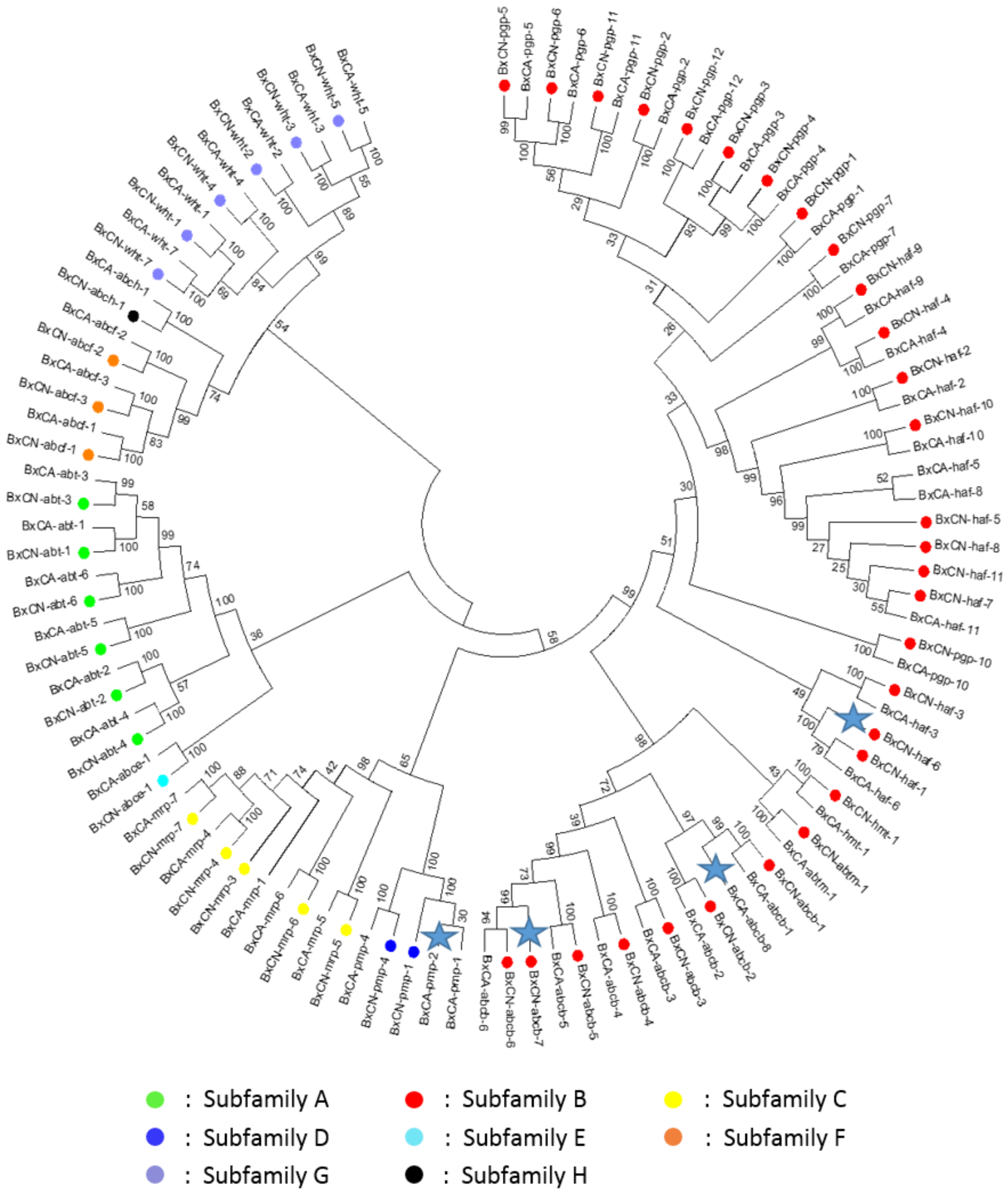


Figure 2.29: Phylogenetic analysis of ABC transporter genes in BxCA and BxCN
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in BxCA and in BxCN. ABC transporter genes in BxCN were highlighted by different color representing for different subfamilies. All ABC tranpsorter gene names in BxCA were assigned based on ABC transporter genes in BxCN by applying the name rule. Stars stands for the genes that do not have orthologs in BxCA or BxCN

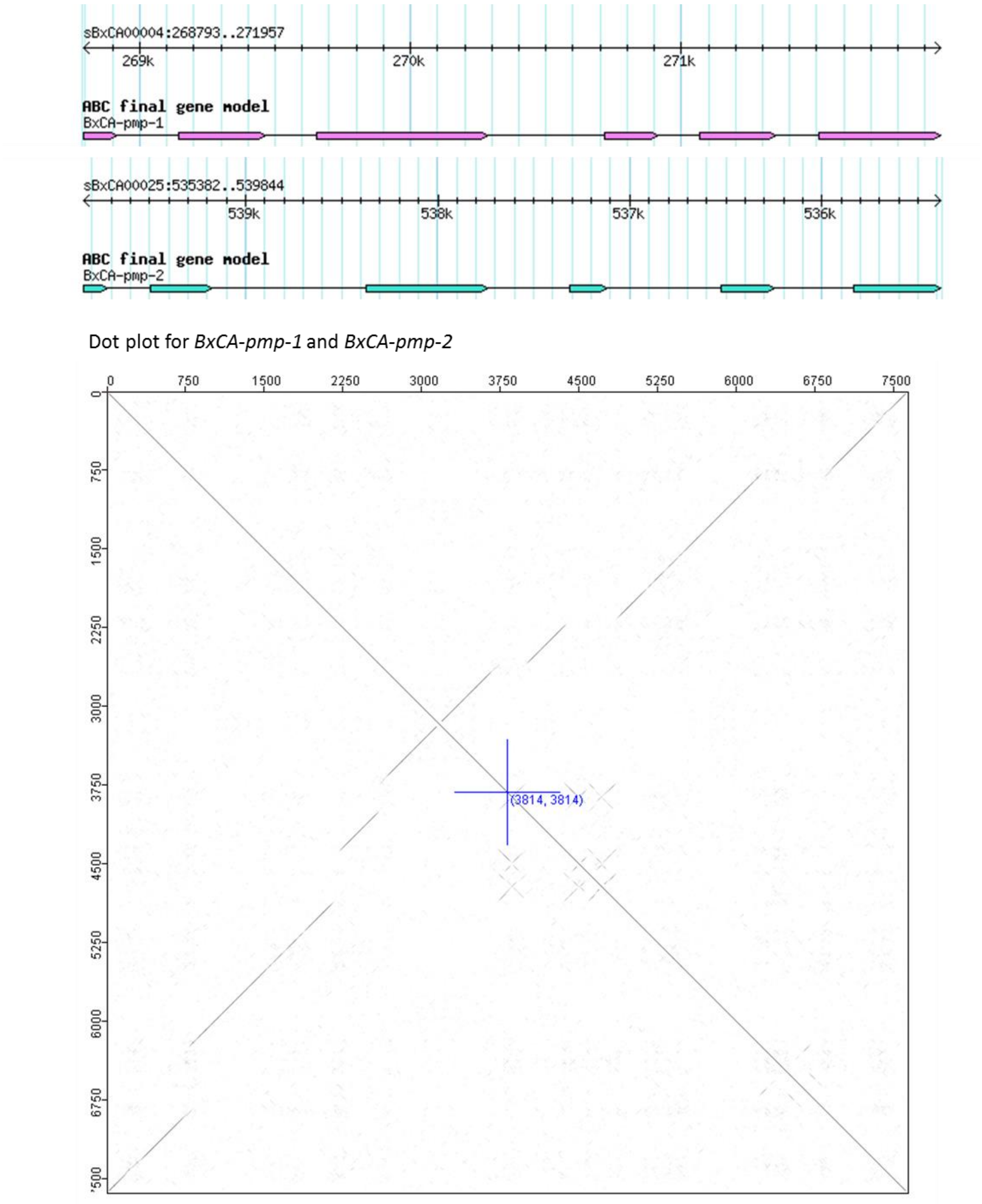


Figure 2.30: Duplication of ABC transporter gene in subfamily D in BxCA

"ABC final gene model" track includes high-quality ABC transporter gene models. BxCA-pmp-1 and *BxCA-pmp-2* both shared orthologous relationship with *pmp-1* in *C. elegans* based on the phylogenetic analysis. DNA sequences of the genomic region for *BxCA-pmp-1* and *BxCA-pmp-2* were extracted and combined to run JDotter. The result showed these two regions shared some similarity.

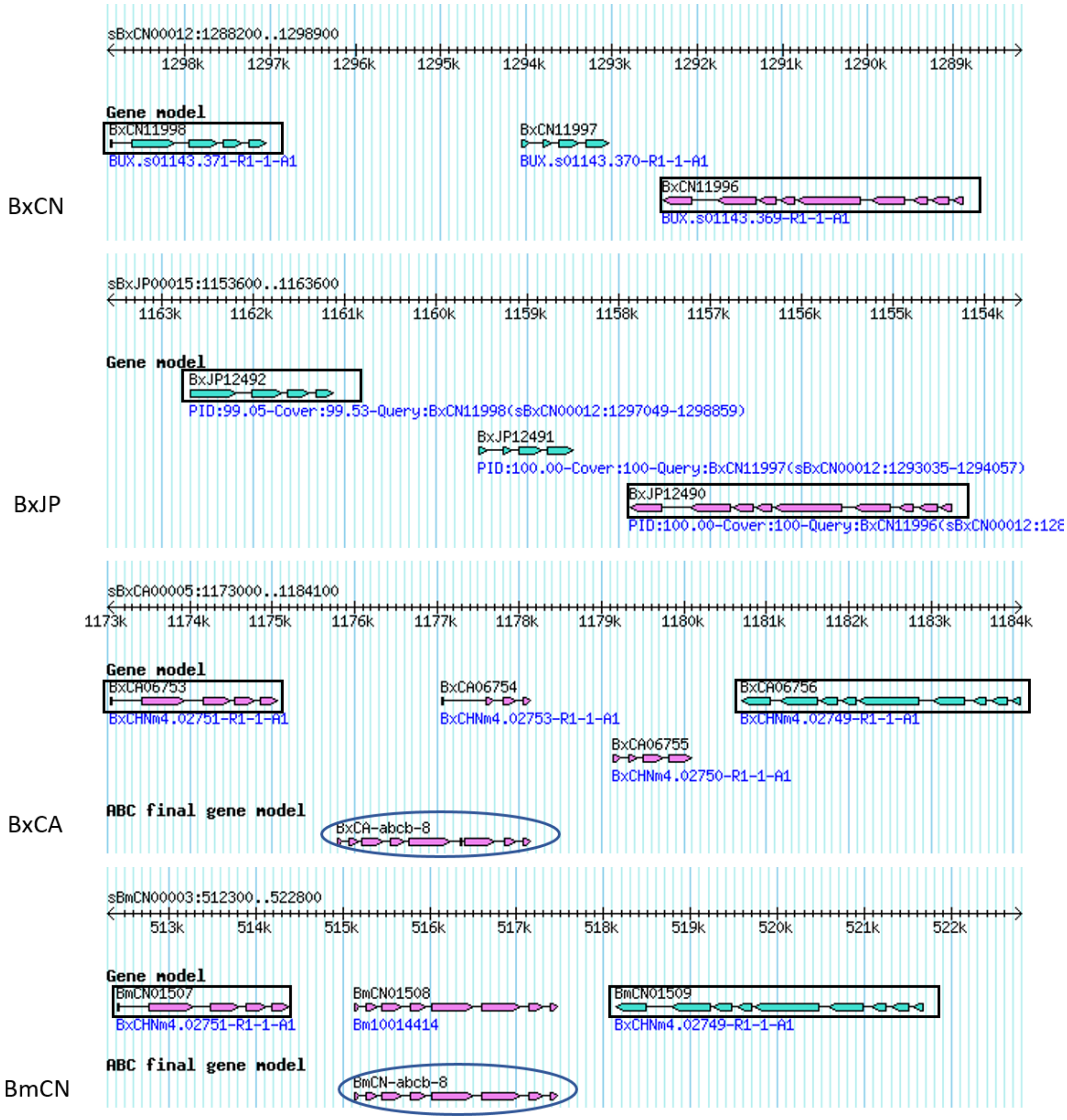


Figure 2.31: The ortholog of *BxCN-abcb-8* can be found in BmCN but not in BxJP and BxCN

"Gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. The genes in the black boxes are conserved within four *Bursaphelechus* genomes. Within the conserved region, we cannot find any putative ABC transporter genes in BxCN and BxJP.

2.6. Annotating ABC transporter genes in *B. mucronatus* (BmCN)

B. mucronatus is a species closely related to *B. xylophilus* but does not kill pine trees (Mamiya 1979). By annotating ABC transporter genes in *B. mucronatus* (BmCN, one of B.mucronatus obtained from Zhejiang Province, China) and comparing ABC transporter genes between *B. mucronatus* and *B. xylophilus*, we can understand how conserved the complete inventory of ABC transporter genes between these two closed nematode species. In total, we found 61 ABC transporter candidates in BmCN, of which 57 were found by InterProScan searches and four additional ones were found by BLAST searches. None of these candidates was contamination. After quality assessment, we identified 40 high-quality ABC transporter genes and 17 defective candidates that needed improvement (Table 2.11). Our improvement procedure generated five revised gene models of high-quality, two of which were TM domain improved gene models. The original gene models (BmCN11415 and BmCN11416), both annotated as a half ABC transporter in subfamily, had 14 and two TM helices, respcetively. After improvement, the new gene models have 10 and seven TM helices (Figure 2.32). These two genes were similar, likely due to gene duplication. In total, 53 high-quality ABC transporter genes were annotated in our analysis, 51 of which also encode appropriate number of TM helices within TM domain (s) (Table 2.12).

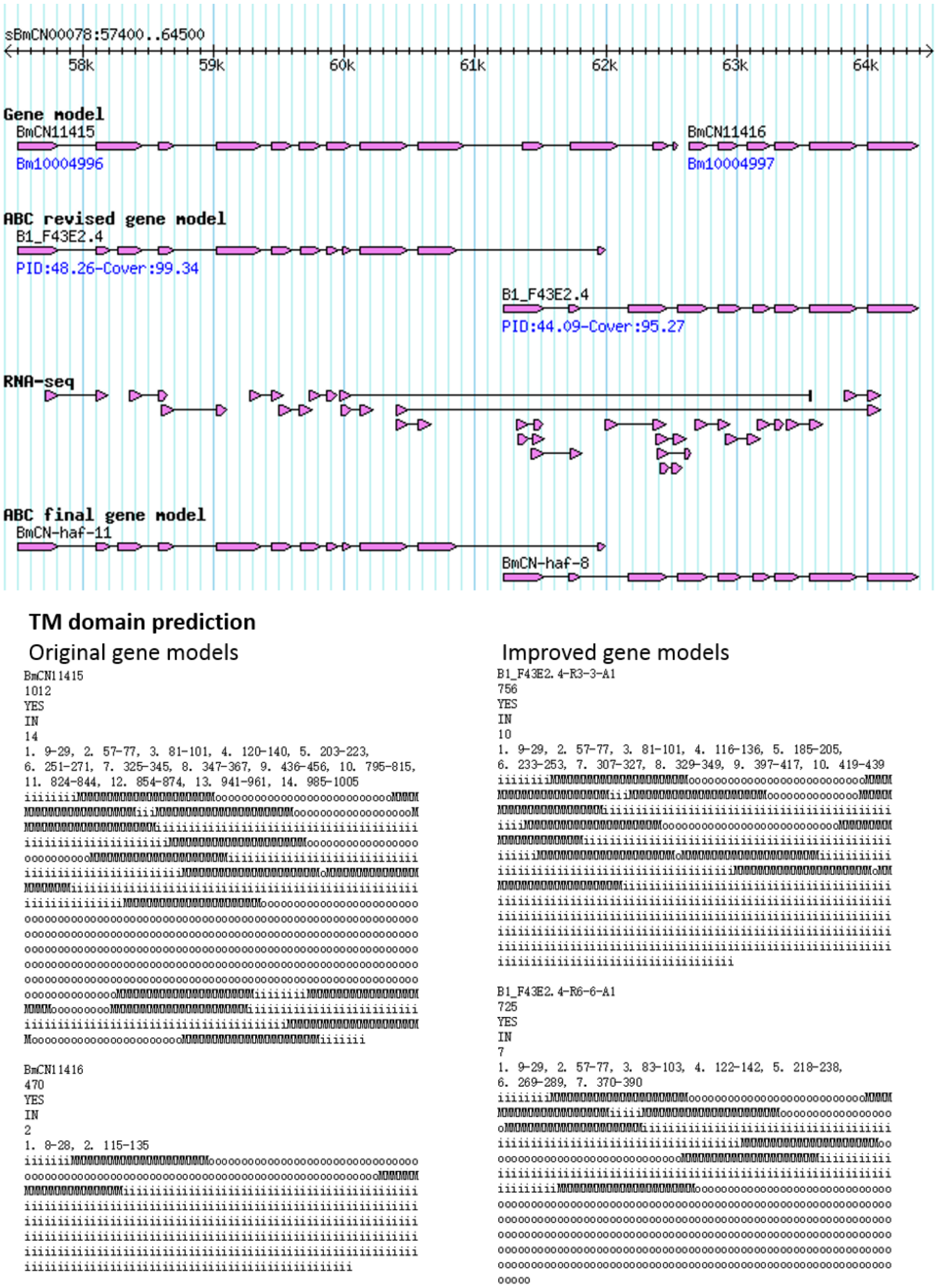
Table 2.11: Improvement of defective ABC transporter gene models in BmCN

ID	Improved or not	ID after improvement	Notes
BmCN01508	No		Keep original model (with longer ABC domain length)
BmCN02262	Yes	BmCN02262	
BmCN09849	Yes	BmCN09849	
BmCN11441	Yes		Keep original model (based on the ortholog in BxCN)
BmCN11442	Yes		keep original model
BmCN11784	Yes	BmCN11784	
BmCN06377	No		Keep original model
BmCN11415	Yes	BmCN11415	TM domain was improved
BmCN11416	Yes	BmCN11416	TM domain was improved
BmCN15800	No		
BmCN16482	No		
BmCN16481	No		
BmCN04047	No		
BmCN11412	No		
BmCN11440	No		
BmCN09760	No		
BmCN11004	No		

Table 2.12: High-quality ABC transporter genes in BmCN

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*BmCN-abt-1*	BmCN08166	BmCN08166	7TM-ABC-8TM-ABC	9174	19	1622	
	BmCN-abt-2	BmCN05801	BmCN05801	8TM-ABC-6TM-ABC	14517	31	2233	
	BmCN-abt-3	BmCN08167	BmCN08167	5TM-ABC-9TM-ABC	8398	17	1364	
	BmCN-abt-4	BmCN07741	BmCN07741	6TM-ABC-8TM-ABC	8434	26	1763	
	BmCN-abt-5	BmCN08219	BmCN08219	6TM-ABC-8TM-ABC	9623	15	1657	
	BmCN-abt-6	BmCN08168	BmCN08168	7TM-ABC-7TM-ABC	6550	19	1538	
B	*BmCN-abtm-1*	BmCN03720	BmCN03720	6TM-ABC	2634	9	686	
	BmCN-abtm-2	BmCN03725	BmCN03725	6TM-ABC	2634	9	686	
	BmCN-haf-10	BmCN16021	BmCN16021	9TM-ABC	4519	9	803	
	BmCN-haf-11	BmCN11415	BmCN11415	10TM-ABC	5043	13	1013	
	BmCN-haf-2	BmCN15170	BmCN15170	8TM-ABC	3282	6	770	
	BmCN-haf-3	BmCN11160	BmCN11160	5TM-ABC	2308	9	664	
	BmCN-haf-4	BmCN03554	BmCN03554	9TM-ABC	2736	9	801	
	BmCN-haf-5	BmCN11413	BmCN11413	10TM-ABC	3685	8	819	
	BmCN-haf-6	BmCN10234	BmCN10234	4TM-ABC	2553	6	566	No start codon
	BmCN-haf-8	BmCN11416	BmCN11416	7TM-ABC	1756	6	471	
	BmCN-haf-9	BmCN05843	BmCN05843	11TM-ABC	7285	20	1064	
	BmCN-hmt-1	BmCN14028	BmCN14028	11TM-ABC	3345	7	797	
	BmCN-abcb-1	BmCN11441	BmCN11441	5TM-ABC	3285	9	652	
	BmCN-abcb-2	BmCN11442	BmCN11442	1TM-ABC	2331	6	576	
	BmCN-abcb-3	BmCN11437	BmCN11437	6TM-ABC	3784	7	632	
	BmCN-abcb-4	BmCN11444	BmCN11444	6TM-ABC	2426	7	637	
	BmCN-abcb-6	BmCN11445	BmCN11445	5TM-ABC	3892	11	769	TM helices were improved
	BmCN-abcb-7	BmCN11446	BmCN11446	5TM-ABC	4413	10	600	TM helices were improved
	BmCN-abcb-8	BmCN01508	BmCN01508	3TM-ABC	2323	8	586	
	BmCN-pgp-1	BmCN02180	BmCN02180	6TM-ABC-6TM-ABC	7326	18	1256	
	BmCN-pgp-10	BmCN11989	BmCN11989	4TM-ABC-6TM-ABC	18815	19	1168	
	BmCN-pgp-11	BmCN02533	BmCN02533	6TM-ABC-6TM-ABC	5117	15	1331	
	BmCN-pgp-12	BmCN05530	BmCN05530	6TM-ABC-5TM-ABC	6286	14	1362	
	BmCN-pgp-3	BmCN09842	BmCN09842	6TM-ABC-5TM-ABC	5263	14	1238	
	BmCN-pgp-4	BmCN01927	BmCN01927	6TM-ABC-6TM-ABC	5658	16	1322	
	BmCN-pgp-5	BmCN15801	BmCN15801	6TM-ABC-5TM-ABC	9105	11	1276	
	BmCN-pgp-6	BmCN15799	BmCN15799	6TM-ABC-6TM-ABC	8955	11	1274	
	BmCN-pgp-7	BmCN12097	BmCN12097	6TM-ABC-6TM-ABC	4669	14	1290	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
C	*BmCN-mrp-2*	BmCN11784	BmCN11784	7TM-ABC-5TM-ABC	8529	17	1370	Exons were improved; No start codon
	BmCN-mrp-3	BmCN06947	BmCN06947	8TM-ABC-6TM-ABC	5311	16	1478	
	BmCN-mrp-4	BmCN06572	BmCN06572	10TM-ABC-7TM-ABC	6403	20	1592	
	BmCN-mrp-5	BmCN04823	BmCN04823	7TM-ABC-6TM-ABC	6297	21	1282	
	BmCN-mrp-6	BmCN01618	BmCN01618	10TM-ABC-6TM-ABC	5257	18	1418	
	BmCN-mrp-7	BmCN02262	BmCN02262	9TM-ABC-5TM-ABC	9750	11	1473	Exons were improved; No start codon
D	*BmCN-pmp-1*	BmCN16025	BmCN16025	4TM-ABC	3290	6	670	
	BmCN-pmp-4	BmCN01923	BmCN01923	6TM-ABC	4135	9	675	
E	*BmCN-abce-1*	BmCN06377	BmCN06377	ABC-ABC	2273	10	616	
F	*BmCN-abcf-1*	BmCN00051	BmCN00051	ABC-ABC	3752	11	665	
	BmCN-abcf-2	BmCN03550	BmCN03550	ABC-ABC	2328	8	667	
	BmCN-abcf-3	BmCN09849	BmCN09849	ABC-ABC	3453	12	695	Exons were improved
G	*BmCN-wht-1*	BmCN01921	BmCN01921	ABC-7TM	2654	9	707	
	BmCN-wht-2	BmCN13442	BmCN13442	ABC-6TM	5621	7	649	
	BmCN-wht-3	BmCN05305	BmCN05305	ABC-6TM	2541	9	608	
	BmCN-wht-4	BmCN07269	BmCN07269	ABC-6TM	2700	11	681	
	BmCN-wht-5	BmCN16262	BmCN16262	ABC-6TM	4236	8	567	
	BmCN-wht-7	BmCN02698	BmCN02698	ABC-6TM	2773	12	678	
H	*BmCN-abch-1*	BmCN04046	BmCN04046	ABC-8TM	3845	16	890	



TM domain prediction

Original gene models

```
BmCN11415
1012
YES
IN
14
1. 9-29, 2. 57-77, 3. 81-101, 4. 120-140, 5. 203-223,
6. 251-271, 7. 325-345, 8. 347-367, 9. 436-456, 10. 795-815,
11. 824-844, 12. 854-874, 13. 941-961, 14. 985-1005
iiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooMMMM
MMMMMMMMMMMMMMMMMiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooM
MMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooo
ooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMM
MMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiMMMMMMMMMMMMMMMMM
MMMMoooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMM
MoooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiii

BmCN11416
470
YES
IN
2
1. 8-28, 2. 115-135
iiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooooMMMMMM
MMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

Improved gene models

```
B1_F43E2.4-R3-3-A1
756
YES
IN
10
1. 9-29, 2. 57-77, 3. 81-101, 4. 116-136, 5. 185-205,
6. 233-253, 7. 307-327, 8. 329-349, 9. 397-417, 10. 419-439
iiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooMMMM
MMMMMMMMMMMMMMMMMiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooMMMMM
MMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooMMMMMMMM
MMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMM
MMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii

B1_F43E2.4-R6-6-A1
725
YES
IN
7
1. 9-29, 2. 57-77, 3. 83-103, 4. 122-142, 5. 218-238,
6. 269-289, 7. 370-390
iiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooMMMM
MMMMMMMMMMMMMMMMMiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooo
oMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoo
ooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooo
```

Figure 2.32: TM domain improvement of two adjacent genes in BmCN
"Gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. BmCN11415 and BmCN11416 were both annotated as a half ABC transporter gene in subfamily B. After improvement, BmCN11415 (*BxCN-haf-11*) had 10 TM helices, compared to 14 before improvement and BmCN11416 (*BmCN-haf-8*) had seven TM helices, compared to two before improvement.

We expect that the ABC transporter genes are similar between two closely related species, *B. mucronatus* and *B. xylophilus*. Through phylogenetic analysis, we found 50 out of 53 ABC transporter genes in BmCN showed one-to-one orthologous relationship and shared the same name with their ortholog in BxCN (Figure 2.33). Considering the differences, BxCN had two specific duplications, *BxCN-haf-1* and *BxCN-haf-7*, as mentioned in section 2.4. *BmCN-mrp-2* had good quality compared to the defective *BxCN-mrp-2* (*BxCN1429*). While *BmCN-pgp-2* (*BmCN11004*), located in the end of a contig, was incomplete with just one ABC domain (supposed to have two) due to assembly error.

In subfamily B, BmCN did not have the ortholog of *BxCN-abcb-5* which was obtained by splitting a single ABC transporter genes into two. As mentioned in section 2.5, BmCN had one extra ABC transporter gene, *BmCN-abcb-8* (Figure 2.31). Additionally, in BmCN, there was one specific duplication event in which five adjacent genes duplicated tandem during evolution (Figure 2.34). *BmCN-abtm-1* was one of the five genes, which contributed one extra ABC transporter gene in BmCN compared to BxCN. Through this analysis, we found ABC transporter genes were well conserved between *B. mucronatus* and *B. xylophilus*, both of which also shared a similar number of ABC transporter genes to *C. elegans*.

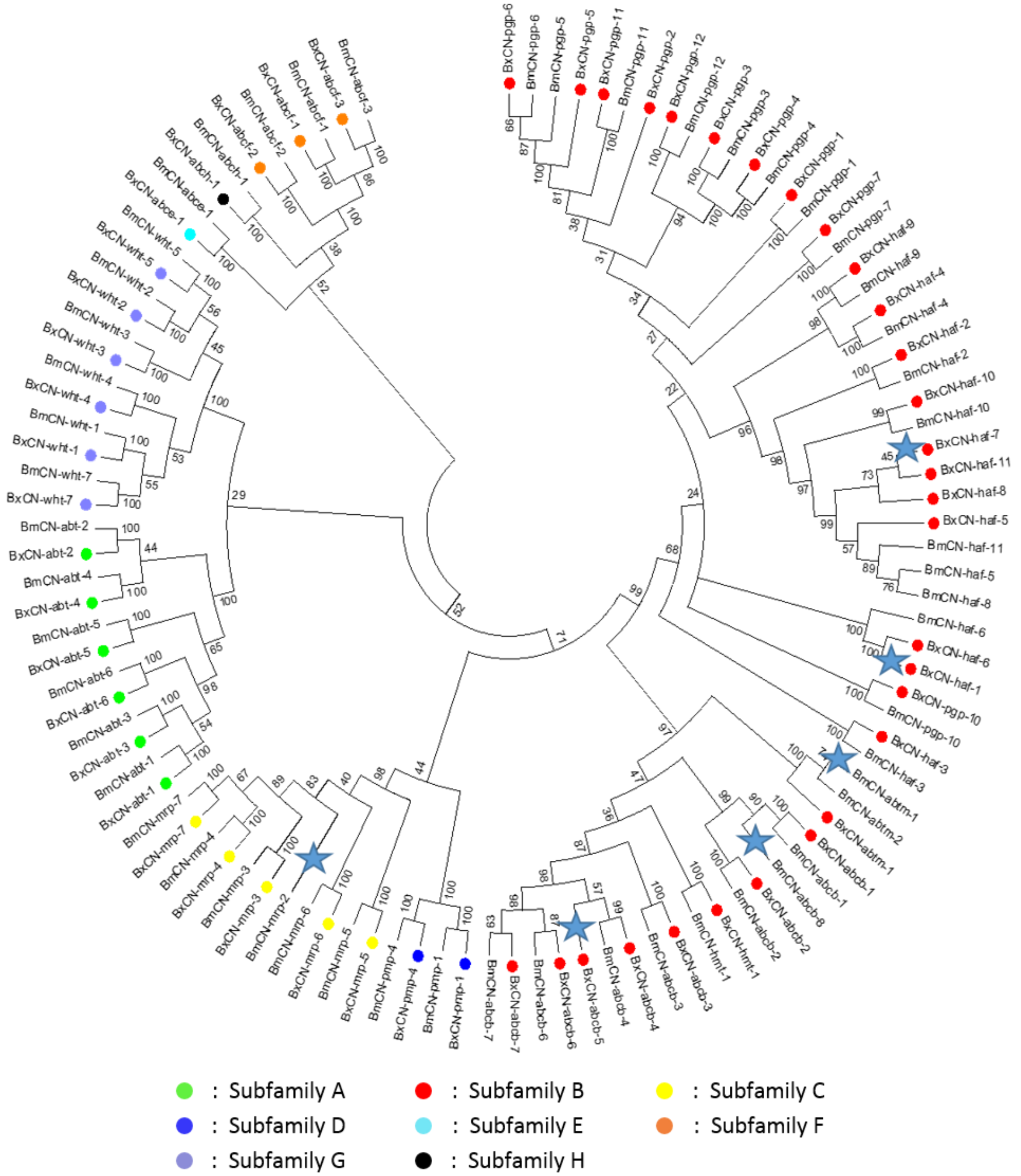
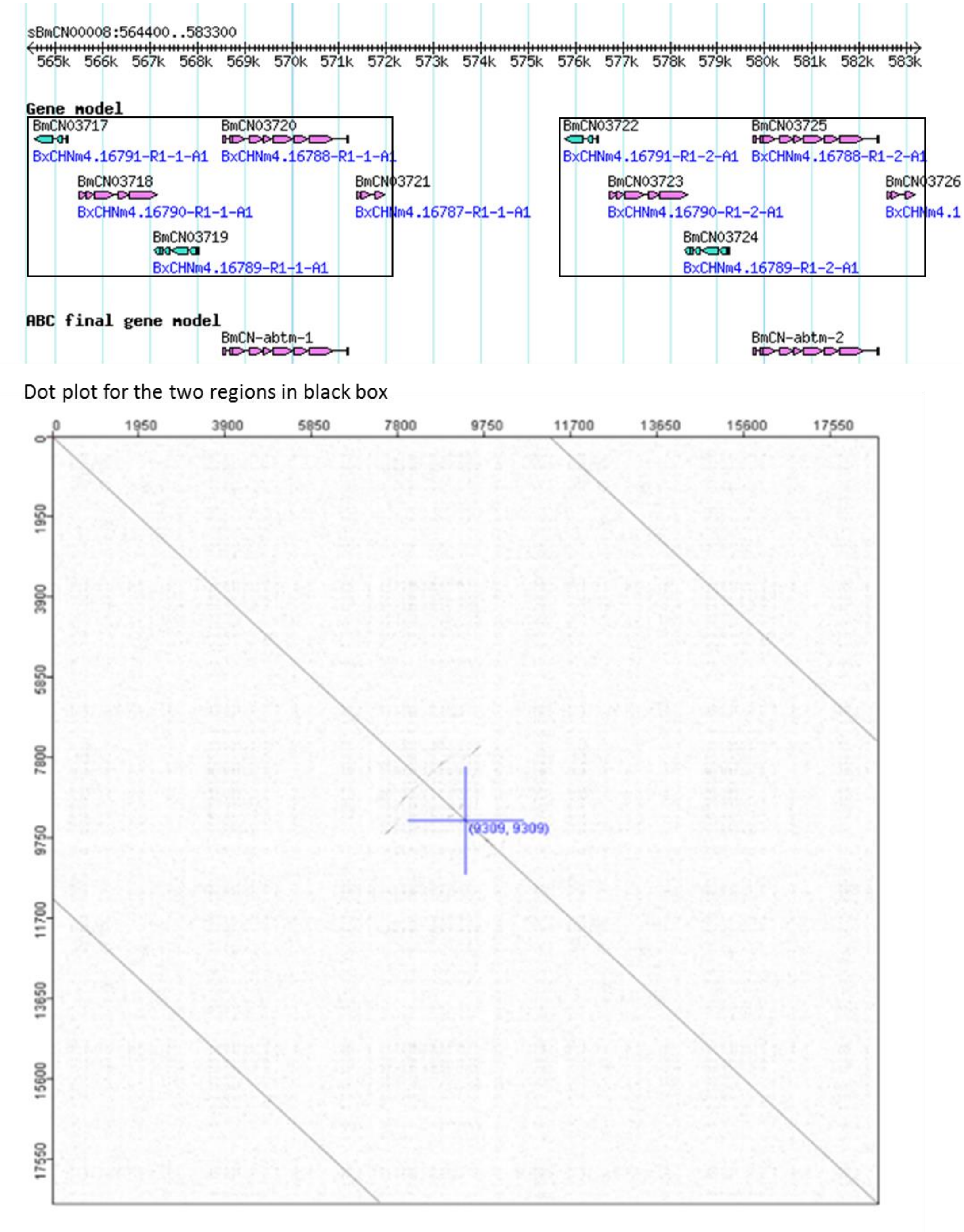


Figure 2.33: Phylogenetic analysis of ABC transporter genes in BmCN and BxCN
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in BmCN and in BxCN. ABC transporter genes in BxCN were highlighted by different color representing for different subfamilies. All ABC tranpsorter gene names in BmCN were assigned based on ABC transporter genes in BxCN by applying the name rule. Stars stands for the genes that do not have orthologs in BmCN or BxCN



Dot plot for the two regions in black box

Figure 2.34: BmCN specific ABC transporter gene caused by duplication event in which five adjacent genes duplicated tandem

"Gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. There was one BmCN specific duplication event in which five adjacent genes in the black box duplicated tandem. *BmCN-abtm-1* was one of the five genes. Therefore, the duplication contributed one more gene (*BmCN-abtm-2*) of subfamily B in BmCN genome compared to BxCN genome.

2.7. Discussion

In this study, domain-based search and homology-based improvement were applied to annotate and improve ABC transporter genes in *B. xylophilus*. Compared to 106 putative ABC transporter genes *B. xylophilus* identified in the previous study, our study found that *B. xylophilus* has a similar number of ABC transporter to that in *C. elegans* (Table 2.13). Therefore, the previous hypothesis that the highly expanded family of ABC transporter genes may facilitate the invasion and pathogenicity of PWN were proved to be incorrect (Kikuchi et al. 2011). The large number of ABC transporter genes reported by previous study was caused largely by the defects of the genome assembly and gene annotation (Kikuchi et al. 2011). In addition, through phylogenetic analysis, we identified subfamily E and subfamily F were highly conserved, with one-to-one orthologous relationship between *Bursaphelenchus* and *C. elegans*. In contrast, members in subfamily B showed more species-specific expansion, suggesting that they were actively evolving. Most ABC transporter gene models annotated in four *Bursaphelenchus* genomes show clear orthologous relationships, suggesting ABC gene family is very well conserved in these genomes. However, turnover in evolution was also observed. In conclusion, this study provided a robust bioinformatics method to identified high-quality ABC transporter genes in nematode genomes, which may contribute to understand the evolution of nematodes and how different inventory of ABC transporters could affect the interaction between nematodes and their surrounding environment.

Table 2.13: Subfamily information for high-quality ABC transporter genes in four *Bursaphelenchus* genomes.

	BxCN	BxJP	BxCA	BmCN	*C. elegans*
Subfamily A	6	5	6	6	5
Subfamily B	30	26	28	28	24
Subfamily C	5	5	5	6	9
Subfamily D	2	2	3	2	5
Subfamily E	1	1	1	1	1
Subfamily F	3	3	3	3	3
Subfamily G	6	6	6	6	8
Subfamily H	1	1	1	1	1
Total	54	49	53	53	56

Chapter 3. Systematic annotation of ABC transporter genes in pathogenic and non-pathogenic nematode genomes

3.1. Introduction

The phylum Nematoda is an ecologically diverse clade with free-living species, as well as parasites of animals and plants (Park et al. 2011). In order to further understand the relationship between the complement of ABC transporter genes and the pathogenicity of various nematodes, we annotated ABC transporter genes in additional 24 nematode species, including seven free-living nematodes, *Caenorhabditis briggsae* (Stein et al. 2003), *Caenorhabditis tropicalis*, *Caenorhabditis sinica*, *Caenorhabditis brenneri*, *Caenorhabditis remanei*, *Caenorhabditis japonica*, *Caenorhabditis angaria* and *Panagrellus redivivus* (Srinivasan et al. 2013), two necromenic nematodes, *Pristionchus pacificus* (Dieterich et al. 2008) and *Pristionchus exspectatus* (Rodelsperger et al. 2014), two plant parasites, *Meloidogyne hapla* (Opperman et al. 2008) and *Meloidogyne incognita* (Caillaud et al. 2008), 12 animal parasites, *Necator americanus* (Tang et al. 2014), *Haemonchus contortus* (Laing et al. 2013), *Ancylostoma ceylanicum* (Schwarz et al. 2015), *Ascaris suum* (Jex et al. 2011), *Brugia malayi* (Ghedin et al. 2007), *Loa loa* (Desjardins et al. 2013), *Onchocerca volvulus* (Unnasch and Williams 2000), *Dirofilaria immitis* (Godel et al. 2012), *Trichinella spiralis* (Mitreva et al. 2011), *Trichuris trichiura* (Foth et al. 2014) and *Trichuris suis* (Jex et al. 2014) and one insect parasite, *Heterorhabditis bacteriophora* (Bai et al. 2013). We downloaded genomic DNA, protein sets and annotation gff3 file for each species from published databases (Table 3.1).

Table 3.1: Data sources for each nematode species

Species	Data resources
Caenorhabditis briggsae	ftp://ftp.wormbase.org/pub/wormbase/species/c_briggsae/ PRJNA10731.WS245
Caenorhabditis tropicalis	ftp://ftp.wormbase.org/pub/wormbase/species/c_sp11/ PRJNA53597.WS246
Caenorhabditis sinica	ftp://ftp.wormbase.org/pub/wormbase/species/c_sp5/ PRJNA194557.WS246
Caenorhabditis brenneri	ftp://ftp.wormbase.org/pub/wormbase/species/c_brenneri/ PRJNA20035.WS245
Caenorhabditis remanei	ftp://ftp.wormbase.org/pub/wormbase/species/c_remanei/ PRJNA53967.WS245
Caenorhabditis elegans	ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/ PRJNA13758.WS244
Caenorhabditis japonica	ftp://ftp.wormbase.org/pub/wormbase/species/c_japonica/ PRJNA12591.WS245
Caenorhabditis angaria	ftp://ftp.wormbase.org/pub/wormbase/species/c_angaria/ PRJNA51225.WS245
Pristionchus exspectatus	ftp://ftp.wormbase.org/pub/wormbase/species/p_exspectatus/ PRJEB6009.WS246
Pristionchus pacificus	ftp://ftp.wormbase.org/pub/wormbase/species/p_pacificus/ PRJNA12644.WS245
Haemonchus contortus	ftp://ftp.wormbase.org/pub/wormbase/species/h_contortus/ PRJNA205202.WS245
Ancylostoma ceylanicum	ftp://ftp.wormbase.org/pub/wormbase/species/a_ceylanicum/ PRJNA231479.WS247
Necator americanus	ftp://ftp.wormbase.org/pub/wormbase/species/n_americanus/ PRJNA72135.WS246
Heterorhabditis bacteriophora	ftp://ftp.wormbase.org/pub/wormbase/species/h_bacteriophora/ PRJNA13977.WS246
Panagrellus redivivus	ftp://ftp.wormbase.org/pub/wormbase/species/p_redivivus/ PRJNA186477.WS245
Meloidogyne incognita	ftp://ftp.wormbase.org/pub/wormbase/species/m_hapla/ PRJNA28837.WS245
Meloidogyne hapla	ftp://ftp.wormbase.org/pub/wormbase/species/m_hapla/ PRJNA29083.WS245
Ascaris suum	ftp://ftp.wormbase.org/pub/wormbase/species/a_suum/ PRJNA80881.WS245
Loa loa	ftp://ftp.wormbase.org/pub/wormbase/species/l_loa/

	PRJNA60051.WS246
Brugia malayi	ftp://ftp.wormbase.org/pub/wormbase/species/b_malayi/ PRJNA10729.WS245
Onchocerca volvulus	ftp://ftp.wormbase.org/pub/wormbase/species/o_volvulus/ PRJEB513.WS246
Dirofilaria immitis	ftp://ftp.wormbase.org/pub/wormbase/species/d_immitis/ PRJEB1797.WS246
Trichinella spiralis	ftp://ftp.wormbase.org/pub/wormbase/species/t_spiralis/ PRJNA12603.WS245
Trichuris trichiura	ftp://ftp.sanger.ac.uk/pub/pathogens/Trichuris/GenomePaper2014/
Trichuris suis	ftp://ftp.wormbase.org/pub/wormbase/species/t_suis/ PRJNA208416.WS245

3.2. Phylogenetic analysis provide a general insight to the evolutionary relationship for nematodes

To determine the evolutionary relationship of these 27 species, we constructed a phylogenetic analysis tree based on the sequence similarities of three conserved genes, *tef1* (translation elongation factor 1-α), *cal1* (calmodulin) and *chi18-5* (endochitinase) (Xie et al. 2014). We characterized the protein sequences of these three conserved genes in 29 nematode genomes (we included three genomes for *B. xylophilus* in this analysis) and for each genome, we concatenated the three protein sequences. More specifically, we first identified the *C. elegans* orthologs of these three conserved genes, *eef-1A.1*, *cmd-1* and *cht-1.* Then, we used these *C. elegans* genes as reference to search for their orthologs in the rest of the 29 proteomes using BLASTP. To ensure high-quality of the candidate genes, we further revised the gene models by using genBlastG. Finally, alignment was done with the concatenated protein sequences in each genomes and phylogenetic tree (Figure 3.1) was constructed using Neighbor-joining method in MEGA6 (Tamura et al. 2013).

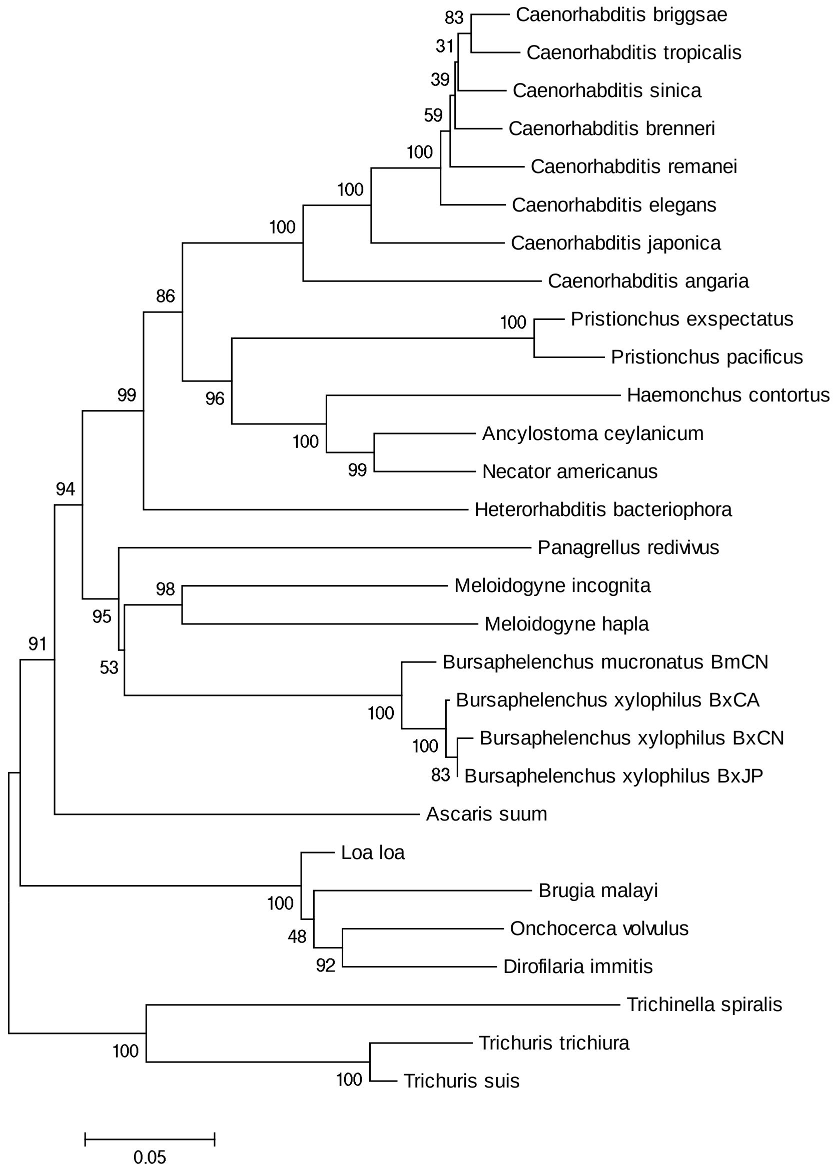


Figure 3.1: Evolutionary relationship for nematodes including in our analysis
Phylogenetic tree was obtained from the concatenated protein sequences of the orthologs of three conserved genes, *tef1*, *cal1* and *chi18-5* in all 29 nematode genomes.

3.3. Annotation of ABC transporter genes in *C. briggsae*

C. briggsae is a free-living nematode that is, closely related to the model organism *C. elegans*. As a matter of fact, *C. briggsae* has been studied extensively along with *C. elegans* for comparative analysis (Stein et al. 2003). After applying the annotation pipeline to *C. briggsae*, we identified 73 ABC transporter gene candidates (65 candidates from InterProScan searches, eight additional ones from BLAST searches), none of which was due to contamination. Among these 73 candidates, 43 were high-quality ABC transporter genes according to our quality checking criteria. For the defective candidates, our improvement procedure generated 18 revised gene models of high-quality, four of which with only TM domain improved. The TM domain improvement was suported by RNA-seq data (Table 3.2). Among the revised gene models, CBG05664 was annotated as a full ABC transporter in subfamily B with two predicted ABC domains, one of which was longer than expected (195 aa). After improvement, the new gene model fully met our criteria and the revision was supported by the RNA-seq data (Figure 3.2). CBG17574, was annotated as a full ABC transporter in subfamily B. After improvement, it was split into two seperate genes, each of which had two typical ABC domains and the revision was supported by RNA-seq data (Figure 3.3). Another example was TM domain improvement procedure revised the TM domains (from 18 TM helices down to 16) in CBG20290, which was annotated as a full ABC transporter in subfamily A. Another example was TM domain improvement in CBG20290 which was characterized as a full ABC transporter in subfamily A. The revised gene model had RNA-seq supported (Figure 3.4). Seven candidates could not be further improved and were likely random hits. In total, we annoated 64 high-quality ABC transporter genes in *C. briggsae*, 59 of which had appropriate TM domain(s) (Table 3.2).

Table 3.2: High-quality ABC transporter genes in *C. briggsae* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Cbr-abt-1*	CBG03394	CBG03394	6TM-ABC-6TM-ABC	6936	28	1553	CBG03395 was merged with CBG03394
	Cbr-abt-2	CBG03891	CBG03891	8TM-ABC-8TM-ABC	9599	34	2328	
	Cbr-abt-3	CBG13094	CBG13094	8TM-ABC-6TM-ABC	8812	23	1558	
	Cbr-abt-4	CBG01265	CBG01265	6TM-ABC-8TM-ABC	6620	9	1814	
	Cbr-abt-5	CBG20290	CBG20290	9TM-ABC-5TM-ABC	9043	18	1506	TM helices were improved
	Cbr-ced-7	CBG10045	CBG10045	7TM-ABC-7TM-ABC	6404	14	1784	
B	*Cbr-abtm-1*	CBG03989	CBG03989	0TM-ABC	3111	8	878	
	Cbr-haf-1	CBG00403	CBG00403	5TM-ABC	2330	7	675	
	Cbr-haf-10	CBG24214b	CBG24214	4TM-ABC	2149	5	584	Split from CBG24214
	Cbr-haf-2	CBG13145	CBG13145	9TM-ABC	2648	6	793	
	Cbr-haf-3	CBG23985	CBG23985	6TM-ABC	10008	13	863	
	Cbr-haf-4	CBG22668	CBG22668	9TM-ABC	7837	10	843	
	Cbr-haf-5	CBG24214a	CBG24214	4TM-ABC	2023	4	585	Split from CBG24214
	Cbr-haf-6	CBG08495	CBG08495	6TM-ABC	10508	8	661	Exons were improved
	Cbr-haf-7	CBG11616	CBG11616	10TM-ABC	2695	4	806	
	Cbr-haf-8	CBG21809	CBG21809	9TM-ABC	5446	4	785	
	Cbr-haf-9	CBG20243	CBG20243	9TM-ABC	6240	16	825	
	Cbr-hmt-1	CBG13182	CBG13182	8TM-ABC	3460	9	700	
	Cbr-pgp-1	CBG13514	CBG13514	6TM-ABC-6TM-ABC	5463	10	1319	
	Cbr-pgp-10	CBG10984	CBG10984	6TM-ABC-5TM-ABC	6176	26	1373	Exons were improved; No stop codon
	Cbr-pgp-11	CBG13356	CBG13356	6TM-ABC-5TM-ABC	7520	13	1209	
	Cbr-pgp-12	CBG00086	CBG00086	6TM-ABC-5TM-ABC	4733	13	1360	
	Cbr-pgp-13	CBG00083	CBG00083	6TM-ABC-6TM-ABC	4417	10	1334	
	Cbr-pgp-14	CBG00079	CBG00079	7TM-ABC-6TM-ABC	4507	9	1371	
	Cbr-pgp-15	CBG00078	CBG00078	6TM-ABC-5TM-ABC	4481	11	1309	Exons were improved
	Cbr-pgp-16	CBG12969	CBG12969	4TM-ABC-7TM-ABC	4840	24	1231	
	Cbr-pgp-17	CBG25498	CBG25498	5TM-ABC-4TM-ABC	4609	12	1295	
	Cbr-pgp-2	CBG04013	CBG04013	6TM-ABC-6TM-ABC	9121	13	1305	
	Cbr-pgp-3	CBG17357	CBG17357	6TM-ABC-6TM-ABC	6044	16	1269	Exons were improved
	Cbr-pgp-4	CBG17356	CBG17356	6TM-ABC-6TM-ABC	11775	14	1402	
	Cbr-pgp-5	CBG17574a	CBG17574	6TM-ABC-5TM-ABC	5271	15	1216	Split from CBG17574; No start codon
	Cbr-pgp-6	CBG17574b	CBG17574	2TM-ABC-6TM-ABC	4849	14	1024	Split from CBG17574
	Cbr-pgp-7	CBG17569	CBG17569	7TM-ABC-6TM-ABC	5573	17	1271	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
B	*Cbr-pgp-8*	CBG17570	CBG17570	4TM-ABC-7TM-ABC	9251	18	1525	
	Cbr-pgp-9	CBG05664	CBG05664	6TM-ABC-6TM-ABC	4465	9	1293	Exons were improved
C	*Cbr-cft-1*	CBG09374	CBG09374	6TM-ABC-8TM-ABC	4637	20	1272	
	Cbr-mrp-1	CBG08145	CBG08145	11TM-ABC-5TM-ABC	8169	21	1528	TM helices were improved
	Cbr-mrp-2	CBG08146	CBG08146	11TM-ABC-5TM-ABC	8182	16	1578	
	Cbr-mrp-3	CBG15993	CBG15993	10TM-ABC-6TM-ABC	6082	23	1505	Exons were improved
	Cbr-mrp-4	CBG01916	CBG01916	11TM-ABC-6TM-ABC	5681	10	1561	
	Cbr-mrp-5	CBG07659	CBG07659	8TM-ABC-6TM-ABC	5414	15	1439	
	Cbr-mrp-6	CBG14361	CBG14361	6TM-ABC-6TM-ABC	7247	17	1332	
	Cbr-mrp-7	CBG23578	CBG23578	5TM-ABC-5TM-ABC	9688	8	924	
	Cbr-mrp-8	CBG00493	CBG00493	9TM-ABC-6TM-ABC	27580	21	1444	CBG00494 and CBG00495 were merged with CBG00493; TM helices were improved
	Cbr-mrp-9	CBG08354	CBG08354	13TM-ABC-6TM-ABC	5581	9	1559	
D	*Cbr-pmp-2*	CBG11176	CBG11176	5TM-ABC	4772	9	662	
	Cbr-pmp-3	CBG04568	CBG04568	6TM-ABC	4680	6	660	
	Cbr-pmp-4	CBG00387	CBG00387	6TM-ABC	2748	10	747	
	Cbr-pmp-5	CBG18988	CBG18988	6TM-ABC	2657	12	599	
	Cbr-abce-1	CBG22999	CBG22999	ABC-ABC	4963	6	612	
F	*Cbr-abcf-1*	CBG08583	CBG08583	ABC-ABC	2057	6	604	
	Cbr-abcf-2	CBG21198	CBG21198	ABC-ABC	5288	6	622	
	Cbr-abcf-3	CBG21100	CBG21100	ABC-ABC	6404	5	712	
G	*Cbr-wht-9*	CBG21060	CBG21060	ABC-2TM	1846	7	337	Exons were improved; No start codon
	Cbr-wht-1	CBG21143	CBG21143	ABC-7TM	5256	12	577	Exons were improved
	Cbr-wht-2	CBG06151	CBG06151	ABC-6TM	2359	10	625	
	Cbr-wht-3	CBG20172	CBG20172	ABC-6TM	4822	9	606	
	Cbr-wht-4	CBG04332	CBG04332	ABC-5TM	4345	12	615	
	Cbr-wht-5	CBG11767	CBG11767	ABC-6TM	7210	9	655	
	Cbr-wht-6	CBG19685	CBG19685	ABC-6TM	3820	10	613	
	Cbr-wht-7	CBG21114	CBG21114	ABC-5TM	20265	10	666	CBG21117 merged with CBG21114; No start codon
	Cbr-wht-8	CBG13191	CBG13191	ABC-6TM	8472	11	953	No start codon
	Cbr-wht-10	CBG09020	CBG09020	ABC-5TM	3903	10	677	
H	*Cbr-abch-1*	CBG02685	CBG02685	ABC-6TM	3351	12	595	

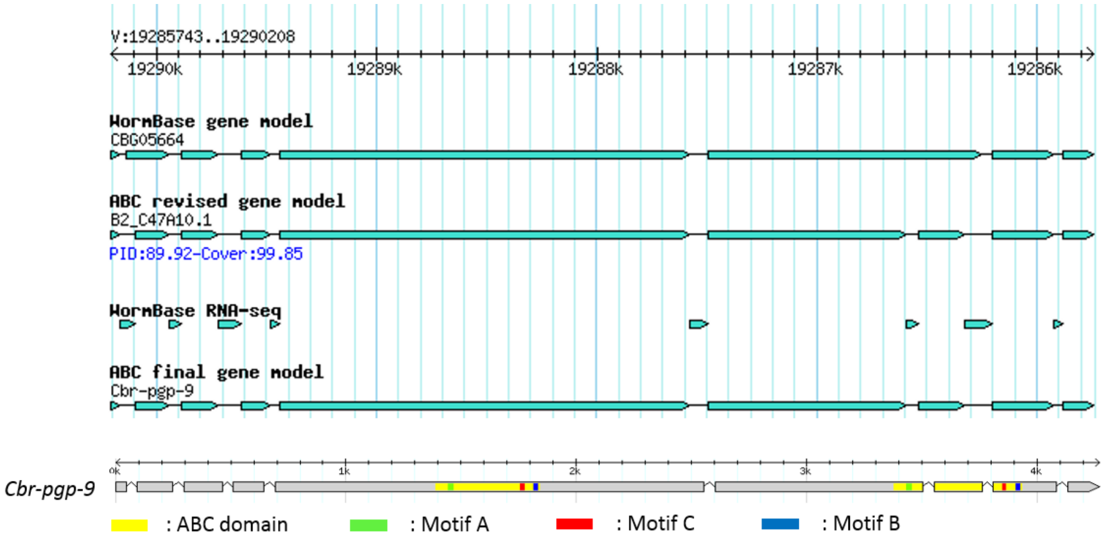


Figure 3.2: A representative case that the exons of one defective ABC transporter gene in *C. briggsae* were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "WormBase RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. CBG05664 had two predicted ABC domains, one of which was longer than expected (195 aa). The revised gene model was supported by the RNA-seq data and was annotated as a high-quality full ABC transporter gene in subfamily B.

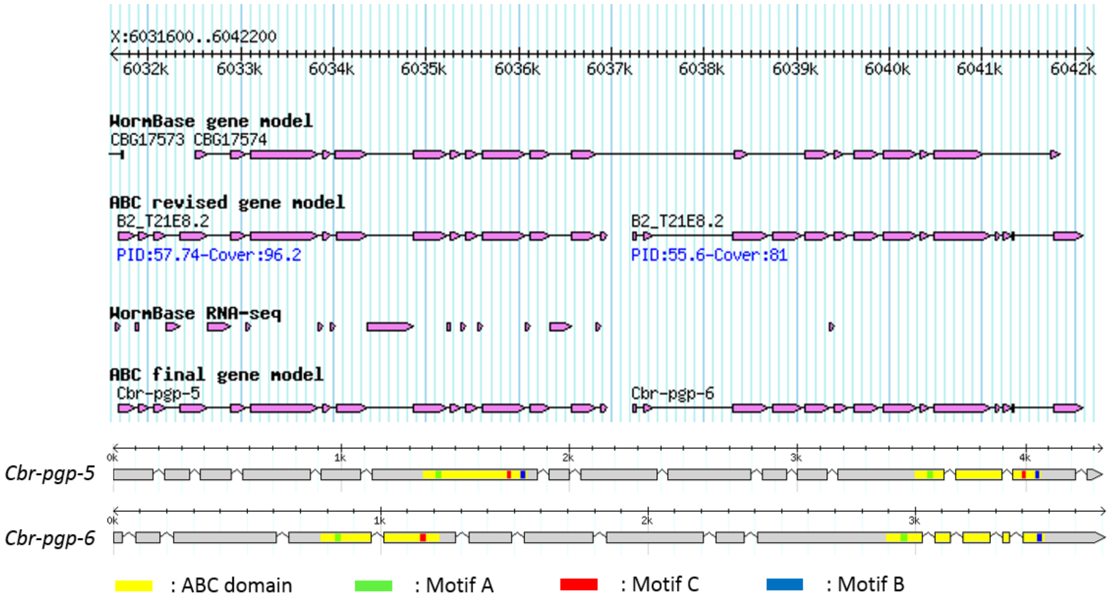
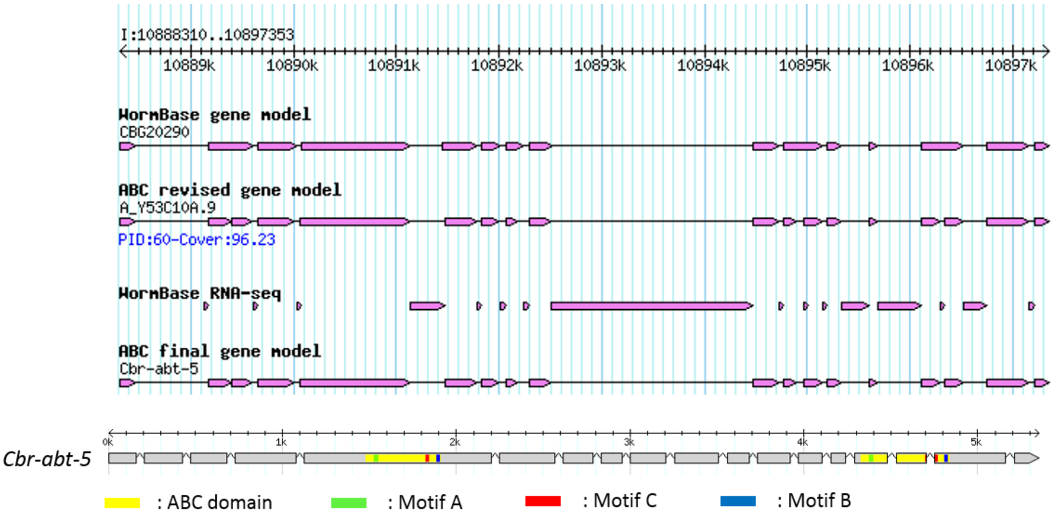
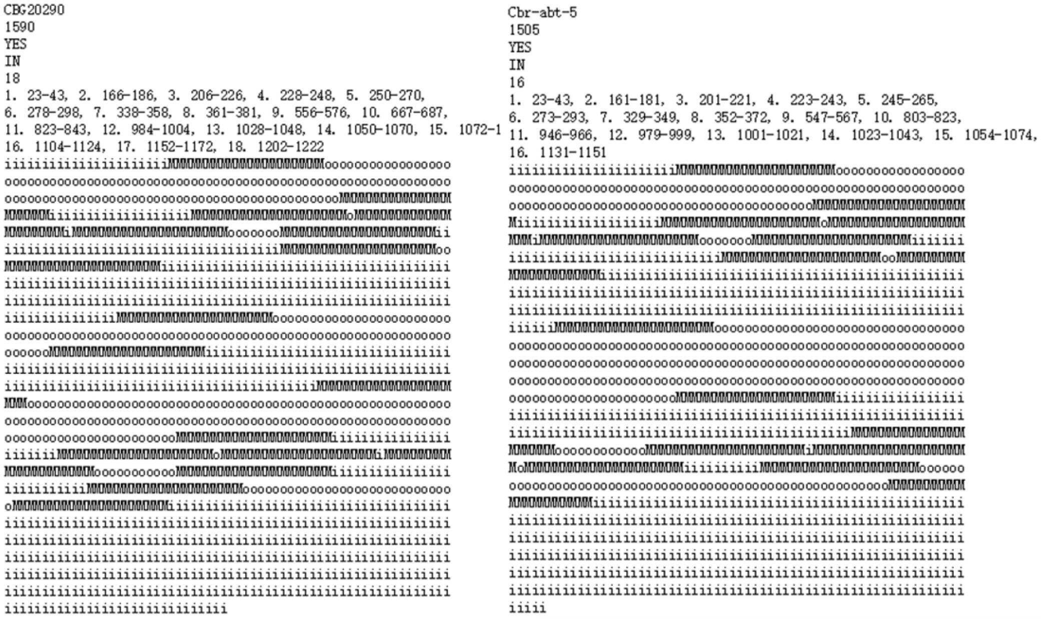


Figure 3.3: A representative case that one candidate was split into two high-quality ABC transporter genes

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “WormBase RNA-seq” track includes introns predicted by RNA-seq data; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. CBG17574, was split into two genes based on the prediction. Each of the revised gene models (*Cbr-pgp-5* and *Cbr-pgp-6*) had two typical ABC domains and the revision was supported by RNA-seq data.



TM domain prediction



```
CBG20290
1590
YES
IN
18
1. 23-43, 2. 166-186, 3. 206-226, 4. 228-248, 5. 250-270,
6. 278-298, 7. 338-358, 8. 361-381, 9. 556-576, 10. 667-687,
11. 823-843, 12. 984-1004, 13. 1028-1048, 14. 1050-1070, 15. 1072-1
16. 1104-1124, 17. 1152-1172, 18. 1202-1222
iiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMM
MMMMMMiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMM
MMMMMMMMiMMMMMMMMMMMMMMMMMMMMMoooooooMMMMMMMMMMMMMMMMMMMMMii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoo
MMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMM
MMMooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiii
iiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiMMMMMMMMM
MMMMMMMMMMMMoooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiii
iiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooo
oMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

```
Cbr-abt-5
1505
YES
IN
16
1. 23-43, 2. 161-181, 3. 201-221, 4. 223-243, 5. 245-265,
6. 273-293, 7. 329-349, 8. 352-372, 9. 547-567, 10. 803-823,
11. 946-966, 12. 979-999, 13. 1001-1021, 14. 1023-1043, 15. 1054-1074,
16. 1131-1151
iiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMM
MiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMM
MMMiMMMMMMMMMMMMMMMMMMMMMoooooooMMMMMMMMMMMMMMMMMMMMMiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooMMMMMMMMM
MMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMM
MMMMMMooooooooooooMMMMMMMMMMMMMMMMMMMMMiMMMMMMMMMMMMMMMMMMMM
MoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooMMMMMMMMMM
MMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiii
```

Figure 3.4: A representative case that the TM domain of an ABC transporter gene candidate was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "WormBase RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. CBG20290 was characterized as a full ABC transporter in subfamily A with 18 TM helices. After improvement, the number of TM helices decreased to 16 and the new gene model had RNA-seq data support.

Considering the close evolutionary relationship between *C. elegans* and *C.briggsae*, we expected to see that the ABC transporter genes from these two species are highly similar. Through phylogenetic analysis, we found 51 out of 64 ABC transporter genes in *C. briggsae* showed one-to-one orthologous relationship with the ABC transporter genes in *C. elegans*. Thus, we assigned the ABC transporter genes in *C. briggsae* share the same gene names as their ortholog in *C. elegans* (Figure 3.5). For the rest of the ABC transporter genes in *C. briggsae*, we first assigned the existed names reported in previous stuides and then followed our rule for naming. According to the phylogenetic tree, there were several expansions in *C. briggsae*: one case (*Cbr-abt-3*) in subfamily A; two cases (*Cbr-haf-5* and *Cbr-haf-10*) in half ABC transporter subgroup of subfamily B; three cases in full transporter subgroup of subfamily B (*Cbr-pgp-15*, *Cbr-pgp-16* and *Cbr-pgp-17*); one case in subfamily C (*Cbr-mrp-9*); two cases in subfamily G (*Cbr-wht-9* and *Cbr-wht-10*). There was only one expansion in *C. elegans, pmp-1 and pmp-2,* which were corresponding to *Cbr-pmp-2* in *C. briggsae.* In short, ABC transporter genes in *C. elegans* and *C. briggsae* showed high conservation with only a few species specific expansion.

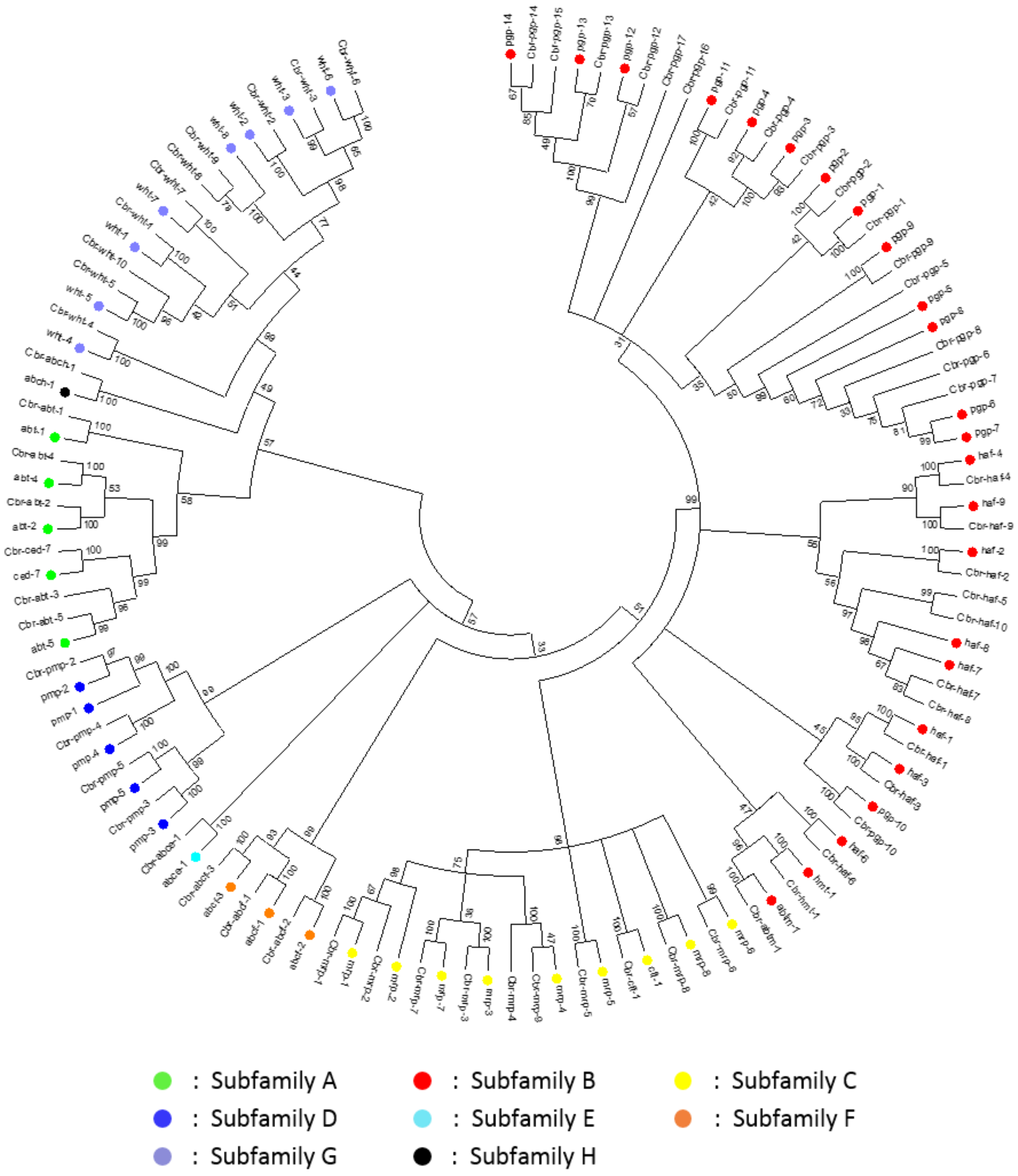


Figure 3.5: Phylogenetic analysis between *C. briggsae* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *C. briggsae* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *C. briggsae* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.4. Annotation of ABC transporter genes in *C. tropicalis*

C. tropicalis (ex-sp. 11) is a species of the *Elegans* group of *Caenorhabditis* genus (Felix et al. 2014). After applying the annotation pipeline to *C. tropicalis*, we identified 74 ABC transporter gene candidates, 64 candidates from InterProScan searches, 10 additional ones from BLAST searches. None of these candidates was due to contamination. Then, we examined the quality of all candidates and found that 41 of them were high-quality ABC transporter genes. All these 41 genes also encoded appropriate TM domain(s) (Table 3.3). For the defective candidates, we tried to improve each of their gene models. After examining the quality of new gene models, we generated 10 improved gene models with high-quality, six of which with only TM domain improved. For instance, four adjacent gene models (Csp11.Scaffold630.g18160, Csp11.Scaffold630.g18162, Csp11.Scaffold630.g18163 and Csp11.Scaffold630.g18164) were identified to be the fragments of one single ABC transporter gene in subfamily C. After improvement, we obtained a revised gene model which encoded a full ABC transporter with two high-quality ABC domains (Figure 3.6). Another example is for TM domain improvement. Csp11.Scaffold629.g11557 was annotated as a half ABC transporter gene in subfamily H, but encoded a protein without predicted TM domain. After improvement, we obtained a longer protein with six predicted TM helices in the TM domain (Figure 3.7). For remaining10 defective candidates which could not be further improved to be high-quality ABC transporter gene, we excluded them from the candidate list. In summary, we annotated 56 high-quality ABC transporter genes in *C. tropicalis*, 52 of which had appropriate number of TM domain (s) (Table 3.3).

Table 3.3: High-quality ABC transporter genes in *C. tropicalis* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Ctr-abt-1*	Csp11.Scaffold585.g4771.t1	Csp11.Scaffold585.g4771.t1	7TM-ABC-8TM-ABC	6024	29	1552	Csp11.Scaffold585.g4773 was merged with Csp11.Scaffold585.g4771; TM helices were improved
	Ctr-abt-2	Csp11.Scaffold80.g498.t1	Csp11.Scaffold80.g498.t1	6TM-ABC-8TM-ABC	8335	35	2220	
	Ctr-abt-3	Csp11.Scaffold629.g12941.t1	Csp11.Scaffold629.g12941.t1	7TM-ABC-8TM-ABC	5932	14	1784	
	Ctr-abt-4	Csp11.Scaffold559.g3862.t1	Csp11.Scaffold559.g3862.t1	6TM-ABC-8TM-ABC	5597	10	1729	
	Ctr-abt-5	Csp11.Scaffold58.g340.t1	Csp11.Scaffold58.g340.t1	6TM-ABC-7TM-ABC	5715	20	1608	
	Ctr-abt-6	Csp11.Scaffold629.g9402.t1	Csp11.Scaffold629.g9402.t1	4TM-ABC-6TM-ABC	5652	15	1151	Csp11.Scaffold629.g9403 was merged with Csp11.Scaffold629.g9402 TM helices were improved
	Ctr-ced-7	Csp11.Scaffold630.g18359.t2	Csp11.Scaffold630.g18359.t2	6TM-ABC-7TM-ABC	5207	11	1555	
B	*Ctr-abtm-1*	Csp11.Scaffold630.g17132.t1	Csp11.Scaffold630.g17132.t1	0TM-ABC	3161	8	704	
	Ctr-haf-1	Csp11.Scaffold441.g1160.t2	Csp11.Scaffold441.g1160.t2	5TM-ABC	2399	9	679	
	Ctr-haf-2	Csp11.Scaffold526.g3078.t1	Csp11.Scaffold526.g3078.t1	9TM-ABC	2826	8	752	Csp11.Scaffold526.g3079.t1 was merged with Csp11.Scaffold526.g3078.t1; TM helices were improved
	Ctr-haf-3	Csp11.Scaffold522.g2869.t1	Csp11.Scaffold522.g2869.t1	6TM-ABC	2460	11	666	
	Ctr-haf-4	Csp11.Scaffold629.g9931.t2	Csp11.Scaffold629.g9931.t2	8TM-ABC	7022	11	755	
	Ctr-haf-6	Csp11.Scaffold629.g16065.t2	Csp11.Scaffold629.g16065.t2	4TM-ABC	2687	6	519	
	Ctr-haf-7	Csp11.Scaffold629.g15902.t1	Csp11.Scaffold629.g15902.t1	9TM-ABC	2733	4	798	
	Ctr-haf-9	Csp11.Scaffold595.g5251.t1	Csp11.Scaffold595.g5251.t1	9TM-ABC	3476	15	810	Csp11.Scaffold595.g5252 was merged with Csp11.Scaffold595.g5251; TM helices were improved
	Ctr-hmt-1	Csp11.Scaffold596.g5300.t1	Csp11.Scaffold596.g5300.t1	11TM-ABC	2849	10	803	
	Ctr-pgp-1	Csp11.Scaffold629.g10654.t1	Csp11.Scaffold629.g10654.t1	6TM-ABC-6TM-ABC	5299	12	1321	
	Ctr-pgp-10	Csp11.Scaffold578.g4494.t3	Csp11.Scaffold578.g4494.t3	6TM-ABC-7TM-ABC	5645	28	1372	
	Ctr-pgp-11	Csp11.Scaffold630.g18476.t2	Csp11.Scaffold630.g18476.t2	6TM-ABC-5TM-ABC	7738	16	1241	
	Ctr-pgp-12	Csp11.Scaffold630.g18809.t1	Csp11.Scaffold630.g18809.t1	6TM-ABC-5TM-ABC	4534	14	1312	
	Ctr-pgp-13	Csp11.Scaffold630.g18813.t2	Csp11.Scaffold630.g18813.t2	6TM-ABC-5TM-ABC	4267	12	1242	Exons were improved; TM helices were improved
	Ctr-pgp-14	Csp11.Scaffold630.g18812.t1	Csp11.Scaffold630.g18812.t1	6TM-ABC-6TM-ABC	4498	8	1319	
	Ctr-pgp-15	Csp11.Scaffold630.g18810.t3	Csp11.Scaffold630.g18810.t3	5TM-ABC-6TM-ABC	4376	11	1214	
	Ctr-pgp-2	Csp11.Scaffold630.g17102.t1	Csp11.Scaffold630.g17102.t1	6TM-ABC-6TM-ABC	5723	14	1265	
	Ctr-pgp-3	Csp11.Scaffold31.g135.t1	Csp11.Scaffold31.g135.t1	6TM-ABC-6TM-ABC	4710	16	1283	
	Ctr-pgp-4	Csp11.Scaffold31.g134.t1	Csp11.Scaffold31.g134.t1	6TM-ABC-5TM-ABC	4456	14	1262	
	Ctr-pgp-5	Csp11.Scaffold629.g13816.t1	Csp11.Scaffold629.g13816.t1	4TM-ABC-4TM-ABC	4007	13	1035	
	Ctr-pgp-6	Csp11.Scaffold629.g13815.t1	Csp11.Scaffold629.g13815.t1	5TM-ABC-6TM-ABC	4323	17	1170	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
B	*Ctr-pgp-7*	Csp11.Scaffold629.g13817.t1	Csp11.Scaffold629.g13817.t1	5TM-ABC-6TM-ABC	4261	15	1203	
	Ctr-pgp-8	Csp11.Scaffold504.g2387.t3	Csp11.Scaffold504.g2387.t3	4TM-ABC-4TM-ABC	5130	22	1078	Exons were improved
	Ctr-pgp-9	Csp11.Scaffold630.g21863.t1	Csp11.Scaffold630.g21863.t1	6TM-ABC-6TM-ABC	4195	8	1292	
C	*Ctr-cft-1*	Csp11.Scaffold460.g1459.t1	Csp11.Scaffold460.g1459.t1	5TM-ABC-7TM-ABC	5627	21	1248	Exons were improved; No start codon
	Ctr-mrp-1	Csp11.Scaffold630.g18160.t1	Csp11.Scaffold630.g18160.t1	11TM-ABC-5TM-ABC	7255	22	1516	Csp11.Scaffold630.g18162, Csp11.Scaffold630.g18163 and Csp11.Scaffold630.g18164 were merged with Csp11.Scaffold630.g18160
	Ctr-mrp-2	Csp11.Scaffold630.g18159.t2	Csp11.Scaffold630.g18159.t2	11TM-ABC-5TM-ABC	7681	18	1547	
	Ctr-mrp-3	Csp11.Scaffold629.g13503.t1	Csp11.Scaffold629.g13503.t1	9TM-ABC-6TM-ABC	5538	21	1464	
	Ctr-mrp-4	Csp11.Scaffold629.g13626.t1	Csp11.Scaffold629.g13626.t1	10TM-ABC-5TM-ABC	5057	9	1570	
	Ctr-mrp-5	Csp11.Scaffold629.g10018.t1	Csp11.Scaffold629.g10018.t1	8TM-ABC-6TM-ABC	4944	16	1418	
	Ctr-mrp-6	Csp11.Scaffold630.g18287.t1	Csp11.Scaffold630.g18287.t1	7TM-ABC-6TM-ABC	5169	16	1369	
	Ctr-mrp-7	Csp11.Scaffold630.g21901.t1	Csp11.Scaffold630.g21901.t1	11TM-ABC-5TM-ABC	5705	12	1499	
	Ctr-mrp-8	Csp11.Scaffold629.g14095.t1	Csp11.Scaffold629.g14095.t1	9TM-ABC-6TM-ABC	6340	9	1564	
D	*Ctr-pmp-1*	Csp11.Scaffold629.g12104.t1	Csp11.Scaffold629.g12104.t1	4TM-ABC	2835	9	663	
	Ctr-pmp-2	Csp11.Scaffold629.g12105.t1	Csp11.Scaffold629.g12105.t1	5TM-ABC	2851	9	659	
	Ctr-pmp-3	Csp11.Scaffold629.g15926.t1	Csp11.Scaffold629.g15926.t1	6TM-ABC	4204	7	660	
	Ctr-pmp-4	Csp11.Scaffold70.g413.t1	Csp11.Scaffold70.g413.t1	5TM-ABC	2684	10	746	
	Ctr-pmp-5	Csp11.Scaffold601.g5429.t1	Csp11.Scaffold601.g5429.t1	5TM-ABC	2913	13	598	
E	*Ctr-abce-1*	Csp11.Scaffold629.g14485.t2	Csp11.Scaffold629.g14485.t2	ABC-ABC	2533	7	630	
F	*Ctr-abcf-2*	Csp11.Scaffold84.g529.t1	Csp11.Scaffold84.g529.t1	ABC-ABC	3327	7	612	Csp11.Scaffold84.g530 was merged with Csp11.Scaffold84.g529
	Ctr-abcf-3	Csp11.Scaffold596.g5273.t1	Csp11.Scaffold596.g5273.t1	ABC-ABC	2410	5	712	
G	*Ctr-wht-1*	Csp11.Scaffold619.g6102.t1	Csp11.Scaffold619.g6102.t1	ABC-5TM	3673	12	652	
	Ctr-wht-2	Csp11.Scaffold507.g2458.t1	Csp11.Scaffold507.g2458.t1	ABC-7TM	2327	12	613	
	Ctr-wht-4	Csp11.Scaffold630.g16973.t2	Csp11.Scaffold630.g16973.t2	ABC-6TM	3767	10	637	
	Ctr-wht-5	Csp11.Scaffold629.g16005.t2	Csp11.Scaffold629.g16005.t2	ABC-6TM	6482	9	823	
	Ctr-wht-6	Csp11.Scaffold630.g17480.t1	Csp11.Scaffold630.g17480.t1	ABC-1TM	3092	6	438	
	Ctr-wht-7	Csp11.Scaffold629.g15996.t1	Csp11.Scaffold629.g15996.t1	ABC-5TM	5371	9	497	No start codon
	Ctr-wht-8	Csp11.Scaffold596.g5308.t2	Csp11.Scaffold596.g5308.t2	ABC-6TM	3101	12	693	
H	*Ctr-abch-1*	Csp11.Scaffold629.g11557.t1	Csp11.Scaffold629.g11557.t1	ABC-6TM	2297	12	596	Exons were improved; TM helices were improved

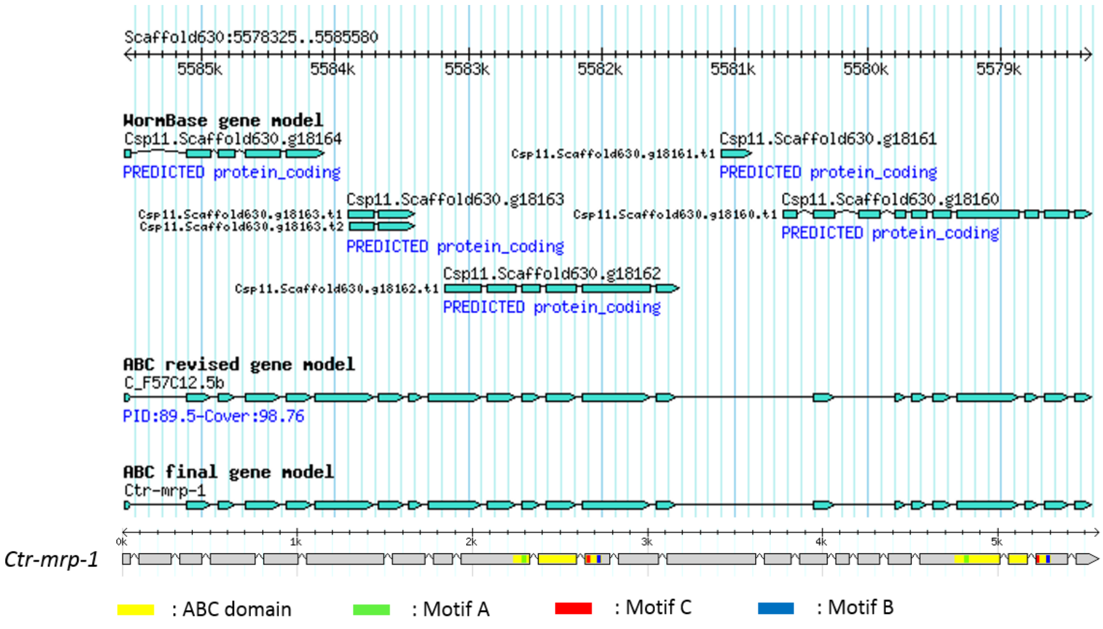
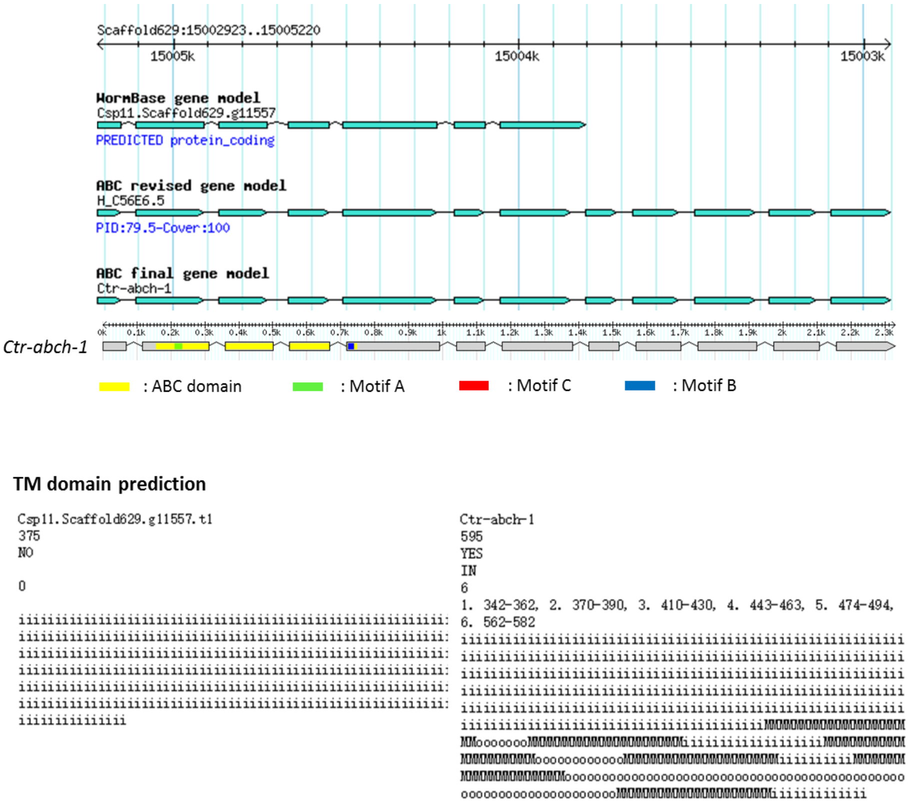


Figure 3.6: A representative case that four adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and C. elegans orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Csp11.Scaffold630.g18160, Csp11.Scaffold630.g18162, Csp11.Scaffold630.g18163 and Csp11.Scaffold630.g18164) were identified to be the fragments of an ABC transporter gene in subfamily C. After improvement, the revised gene encoding a transporter with two high-quality ABC domains.



TM domain prediction

```
Csp11.Scaffold629.g11557.t1
375
NO

0

iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii:
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii:
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii:
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii:
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii:
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii:
iiiiiiiiiiiiiii
```

```
Ctr-abch-1
595
YES
IN
6
1. 342-362, 2. 370-390, 3. 410-430, 4. 443-463, 5. 474-494,
6. 562-582
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMM
MMoooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiMMMMMMMMMMM
MMMMMMMMMMooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiMMMMMMM
MMMMMMMMMMMMMMoooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiii
```

Figure 3.7: A representative case that the TM domain of an ABC transporter gene candidate was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and C. elegans orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Csp11.Scaffold629.g11557 was annotated as a half ABC transporter gene in subfamily H, but no predicted TM domain. After improvement, we obtained a longer gene model with a TM domain, containing six predicted TM helices

Through phylogenetic analysis, we found 46 out of 56 ABC transporter genes in *C. tropicalis* showed clear one-to-one orthologous relationship with the ABC transporter genes in *C. elegans* and we assigned the gene names for ABC transporter genes in *C. tropicalis* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.8). Although the total number of ABC transporter genes in *C. tropicalis* was exactly the same to that in *C. elegans*, there were some slight differences. For instance, in subfamily G, *wht-3* did not have an ortholog in *C. tropicalis* while in subfamily A, there were two expansions (*Ctr-abt-3* and *Ctr-abt-6*) in *C. tropicalis.* In general, the ABC transporter genes between these two speices were quite conserved during evolution.

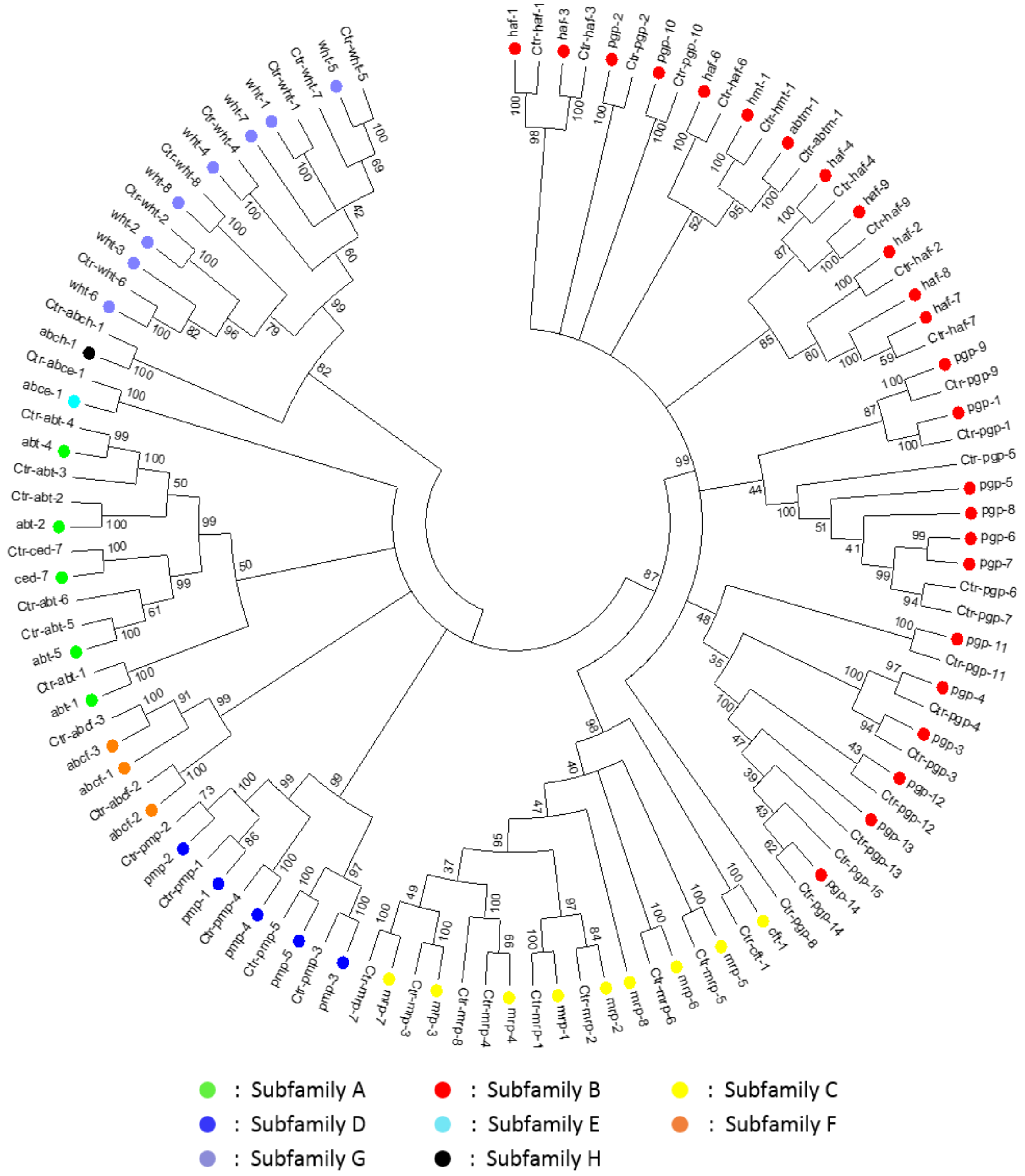


Figure 3.8: Phylogenetic analysis between *C. tropicalis* and *C. elegans*.
Phylogenetic tree was constructed via using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *C. tropicalis* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *C. tropicalis* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.5. Annotation of ABC transporter genes in *C. sinica*

C. sinica (ex-sp. 5) is another member of the Elegans group, with overall morphologically resembling *C. elegans* itself (Huang et al. 2014b). After applying the annotation pipeline to *C. sinica*, we identified in total 90 ABC transporter gene candidates, with 81 candidates from InterProScan searches and nine additional ones from BLAST searches. Two of these candidates, Csp5_scaffold_05607.g35279 and Csp5_scaffold_02584.g27122 were due to contamination and were excluded in our further analysis. Among the 88 candidates, 59 were high-quality ABC transporter genes. All of these genes encode proteins with appropriate TM domain (s). For the defective candidates, we tried to improve. After examining the quality of new gene models, seven were improved with high-quality, three of which with only TM domain improved (Table 3.4). For instance, Csp5_scaffold_02177.g25247 and Csp5_scaffold_02177.g25249 both were annotated as an ABC transporter gene in subfamily F but each of them encoded only one predicted ABC domain. We generated a high-quality gene model, by merging these two candidates together (Figure 3.9). Another candidate, Csp5_scaffold_02097.g24832, was annotated as a half ABC transporter gene in subfamily B and did not encode proper TM domain (only three TM helices). The revised gene model for this candidate was longer and the new region encoding an additional part of TM domain (Figure 3.10), making TM domain complete. After improvement and re-examination process, 16 candidates could not be further improved to be high-quality ABC transporter genes. In summary, we annotated 70 high-quality ABC transporter genes in *C. sinica,* 68 of which had proper TM domain (s) (Table 3.4).

Table 3.4: High-quality ABC transporter genes in *C. sinica* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Csi-abt-1*	Csp5_scaffold_00369.g10291.t2	Csp5_scaffold_00369.g10291.t2	8TM-ABC-6TM-ABC	5908	27	1532	
	Csi-abt-2	Csp5_scaffold_00275.g8485.t2	Csp5_scaffold_00275.g8485.t2	7TM-ABC-5TM-ABC	9666	31	2255	
	Csi-abt-3	Csp5_scaffold_01437.g20922.t1	Csp5_scaffold_01437.g20922.t1	7TM-ABC-7TM-ABC	5922	25	1614	
	Csi-abt-4	Csp5_scaffold_03656.g30878.t1	Csp5_scaffold_03656.g30878.t1	6TM-ABC-10TM-ABC	6663	11	1810	
	Csi-abt-5	Csp5_scaffold_00069.g3252.t1	Csp5_scaffold_00069.g3252.t1	8TM-ABC-8TM-ABC	6632	25	1443	TM helices were improved
	Csi-abt-6	Csp5_scaffold_05607.g35279.t4	Csp5_scaffold_05607.g35279.t4	1TM-ABC-6TM-ABC	4073	10	891	Exons were improved
	Csi-ced-7	Csp5_scaffold_00091.g4006.t2	Csp5_scaffold_00091.g4006.t2	7TM-ABC-7TM-ABC	6614	15	1769	
B	*Csi-abtm-1*	Csp5_scaffold_05369.g34866.t1	Csp5_scaffold_05369.g34866.t1	7TM-ABC	2140	5	649	
	Csi-haf-1	Csp5_scaffold_00272.g8430.t2	Csp5_scaffold_00272.g8430.t2	5TM-ABC	2418	9	674	
	Csi-haf-10	Csp5_scaffold_01193.g19134.t1	Csp5_scaffold_01193.g19134.t1	9TM-ABC	3255	5	772	
	Csi-haf-11	Csp5_scaffold_01193.g19135.t1	Csp5_scaffold_01193.g19135.t1	6TM-ABC	3003	4	740	
	Csi-haf-2	Csp5_scaffold_00343.g9785.t2	Csp5_scaffold_00343.g9785.t2	7TM-ABC	3292	8	806	
	Csi-haf-3	Csp5_scaffold_01508.g21357.t1	Csp5_scaffold_01508.g21357.t1	5TM-ABC	5326	11	666	No start codon
	Csi-haf-4	Csp5_scaffold_04314.g32617.t1	Csp5_scaffold_04314.g32617.t1	6TM-ABC	4603	11	755	No stop codon
	Csi-haf-5	Csp5_scaffold_01471.g21124.t2	Csp5_scaffold_01471.g21124.t2	9TM-ABC	4483	13	823	
	Csi-haf-6	Csp5_scaffold_00234.g7628.t1	Csp5_scaffold_00234.g7628.t1	8TM-ABC	6646	8	673	Exons were improved
	Csi-haf-7	Csp5_scaffold_00381.g10456.t1	Csp5_scaffold_00381.g10456.t1	9TM-ABC	2804	4	792	
	Csi-haf-8	Csp5_scaffold_02097.g24832.t1	Csp5_scaffold_02097.g24832.t1	10TM-ABC	3351	6	764	TM helices were improved
	Csi-haf-9	Csp5_scaffold_00375.g10385.t1	Csp5_scaffold_00375.g10385.t1	10TM-ABC	7071	17	815	
	Csi-hmt-1	Csp5_scaffold_02528.g26869.t1	Csp5_scaffold_02528.g26869.t1	11TM-ABC	2850	10	802	
	Csi-pgp-1	Csp5_scaffold_00024.g1431.t1	Csp5_scaffold_00024.g1431.t1	6TM-ABC-6TM-ABC	6136	11	1319	
	Csi-pgp-10	Csp5_scaffold_00059.g2906.t1	Csp5_scaffold_00059.g2906.t1	6TM-ABC-5TM-ABC	6125	28	1405	
	Csi-pgp-11	Csp5_scaffold_00167.g6131.t1	Csp5_scaffold_00167.g6131.t1	6TM-ABC-5TM-ABC	5828	15	1284	
	Csi-pgp-12	Csp5_scaffold_00019.g1172.t1	Csp5_scaffold_00019.g1172.t1	6TM-ABC-5TM-ABC	4793	16	1300	
	Csi-pgp-13	Csp5_scaffold_00019.g1171.t1	Csp5_scaffold_00019.g1171.t1	6TM-ABC-6TM-ABC	4671	13	1330	
	Csi-pgp-14	Csp5_scaffold_00019.g1168.t1	Csp5_scaffold_00019.g1168.t1	6TM-ABC-5TM-ABC	4491	11	1330	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
B	*Csi-pgp-15*	Csp5_scaffold_00019.g1169.t1	Csp5_scaffold_00019.g1169.t1	6TM-ABC-6TM-ABC	4552	8	1327	
	Csi-pgp-16	Csp5_scaffold_00044.g2330.t1	Csp5_scaffold_00044.g2330.t1	6TM-ABC-6TM-ABC	4426	16	1242	
	Csi-pgp-17	Csp5_scaffold_00106.g4444.t1	Csp5_scaffold_00106.g4444.t1	6TM-ABC-5TM-ABC	4925	24	1252	
	Csi-pgp-18	Csp5_scaffold_00124.g4968.t1	Csp5_scaffold_00124.g4968.t1	5TM-ABC-5TM-ABC	4943	13	1309	
	Csi-pgp-19	Csp5_scaffold_00297.g8920.t1	Csp5_scaffold_00297.g8920.t1	6TM-ABC-6TM-ABC	4539	18	1247	
	Csi-pgp-2	Csp5_scaffold_03252.g29561.t1	Csp5_scaffold_03252.g29561.t1	4TM-ABC-6TM-ABC	6311	14	1288	
	Csi-pgp-20	Csp5_scaffold_03907.g31579.t2	Csp5_scaffold_03907.g31579.t2	6TM-ABC-5TM-ABC	8777	11	1194	
	Csi-pgp-3	Csp5_scaffold_00458.g11649.t1	Csp5_scaffold_00458.g11649.t1	6TM-ABC-6TM-ABC	4690	16	1268	
	Csi-pgp-4	Csp5_scaffold_00458.g11648.t1	Csp5_scaffold_00458.g11648.t1	6TM-ABC-5TM-ABC	5181	15	1278	
	Csi-pgp-5	Csp5_scaffold_00044.g2331.t1	Csp5_scaffold_00044.g2331.t1	6TM-ABC-6TM-ABC	4340	15	1235	
	Csi-pgp-6	Csp5_scaffold_00044.g2326.t3	Csp5_scaffold_00044.g2326.t3	6TM-ABC-4TM-ABC	5041	16	1128	
	Csi-pgp-7	Csp5_scaffold_00044.g2328.t1	Csp5_scaffold_00044.g2328.t1	6TM-ABC-4TM-ABC	5168	16	1128	
	Csi-pgp-8	Csp5_scaffold_00044.g2329.t1	Csp5_scaffold_00044.g2329.t1	6TM-ABC-4TM-ABC	4402	16	1236	
	Csi-pgp-9	Csp5_scaffold_00355.g9985.t1	Csp5_scaffold_00355.g9985.t1	6TM-ABC-6TM-ABC	4301	10	1293	
C	*Csi-cft-1*	Csp5_scaffold_00484.g11971.t1	Csp5_scaffold_00484.g11971.t1	5TM-ABC-7TM-ABC	4658	21	1245	
	Csi-mrp-1	Csp5_scaffold_00837.g15932.t1	Csp5_scaffold_00837.g15932.t1	10TM-ABC-5TM-ABC	9539	20	1535	TM helices were improved
	Csi-mrp-10	Csp5_scaffold_02584.g27122.t2	Csp5_scaffold_02584.g27122.t2	1TM-ABC-5TM-ABC	3469	6	948	
	Csi-mrp-2	Csp5_scaffold_02111.g24904.t1	Csp5_scaffold_02111.g24904.t1	11TM-ABC-5TM-ABC	5886	19	1502	
	Csi-mrp-3	Csp5_scaffold_00031.g1743.t1	Csp5_scaffold_00031.g1743.t1	11TM-ABC-6TM-ABC	5738	25	1529	
	Csi-mrp-4	Csp5_scaffold_00072.g3367.t1	Csp5_scaffold_00072.g3367.t1	11TM-ABC-6TM-ABC	5114	9	1574	
	Csi-mrp-5	Csp5_scaffold_00203.g6958.t2	Csp5_scaffold_00203.g6958.t2	8TM-ABC-6TM-ABC	6782	18	1452	
	Csi-mrp-6	Csp5_scaffold_00195.g6767.t1	Csp5_scaffold_00195.g6767.t1	7TM-ABC-6TM-ABC	6071	17	1417	
	Csi-mrp-7	Csp5_scaffold_01959.g24081.t1	Csp5_scaffold_01959.g24081.t1	10TM-ABC-5TM-ABC	5108	9	1500	
	Csi-mrp-8	Csp5_scaffold_00582.g13135.t2	Csp5_scaffold_00582.g13135.t2	9TM-ABC-5TM-ABC	15955	24	1620	
	Csi-mrp-9	Csp5_scaffold_02111.g24905.t1	Csp5_scaffold_02111.g24905.t1	11TM-ABC-5TM-ABC	6455	19	1502	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
D	*Csi-pmp-1*	Csp5_scaffold_00307.g9127.t1	Csp5_scaffold_00307.g9127.t1	4TM-ABC	2842	9	663	
	Csi-pmp-2	Csp5_scaffold_00307.g9128.t1	Csp5_scaffold_00307.g9128.t1	6TM-ABC	2912	8	673	
	Csi-pmp-3	Csp5_scaffold_00361.g10108.t1	Csp5_scaffold_00361.g10108.t1	6TM-ABC	5462	8	686	
	Csi-pmp-4	Csp5_scaffold_01130.g18623.t1	Csp5_scaffold_01130.g18623.t1	5TM-ABC	3661	12	713	
	Csi-pmp-5	Csp5_scaffold_00013.g886.t1	Csp5_scaffold_00013.g886.t1	6TM-ABC	2676	13	599	
E	*Csi-abce-1*	Csp5_scaffold_02175.g25237.t1	Csp5_scaffold_02175.g25237.t1	ABC-ABC	6231	6	610	
F	*Csi-abcf-1*	Csp5_scaffold_01575.g21824.t2	Csp5_scaffold_01575.g21824.t2	ABC-ABC	1777	5	500	No start codon
	Csi-abcf-2	Csp5_scaffold_02177.g25247.t1	Csp5_scaffold_02177.g25247.t1	ABC-ABC	7917	6	623	Csp5_scaffold_02177.g25249 was merged with Csp5_scaffold_02177.g25247
	Csi-abcf-3	Csp5_scaffold_00814.g15718.t1	Csp5_scaffold_00814.g15718.t1	ABC-ABC	2665	5	712	
G	*Csi-wht-1*	Csp5_scaffold_00387.g10557.t2	Csp5_scaffold_00387.g10557.t2	ABC-5TM	2910	12	654	
	Csi-wht-2	Csp5_scaffold_00384.g10501.t1	Csp5_scaffold_00384.g10501.t1	ABC-6TM	2285	11	610	
	Csi-wht-3	Csp5_scaffold_01565.g21743.t1	Csp5_scaffold_01565.g21743.t1	ABC-6TM	2450	8	545	
	Csi-wht-4	Csp5_scaffold_01809.g23266.t2	Csp5_scaffold_01809.g23266.t2	ABC-5TM	2845	12	651	
	Csi-wht-5	Csp5_scaffold_01771.g23031.t1	Csp5_scaffold_01771.g23031.t1	ABC-6TM	4565	10	700	
	Csi-wht-6	Csp5_scaffold_00051.g2612.t1	Csp5_scaffold_00051.g2612.t1	ABC-6TM	2534	11	613	
	Csi-wht-7	Csp5_scaffold_00387.g10548.t2	Csp5_scaffold_00387.g10548.t2	ABC-5TM	10831	10	675	Csp5_scaffold_00387.g10553.t1 was merged with Csp5_scaffold_00387.g10548.t2
	Csi-wht-8	Csp5_scaffold_04864.g33912.t1	Csp5_scaffold_04864.g33912.t1	ABC-6TM	2295	8	627	
	Csi-wht-9	Csp5_scaffold_00229.g7531.t1	Csp5_scaffold_00229.g7531.t1	ABC-6TM	2992	12	667	
H	*Csi-abch-1*	Csp5_scaffold_02106.g24880.t1	Csp5_scaffold_02106.g24880.t1	ABC-6TM	2309	12	595	

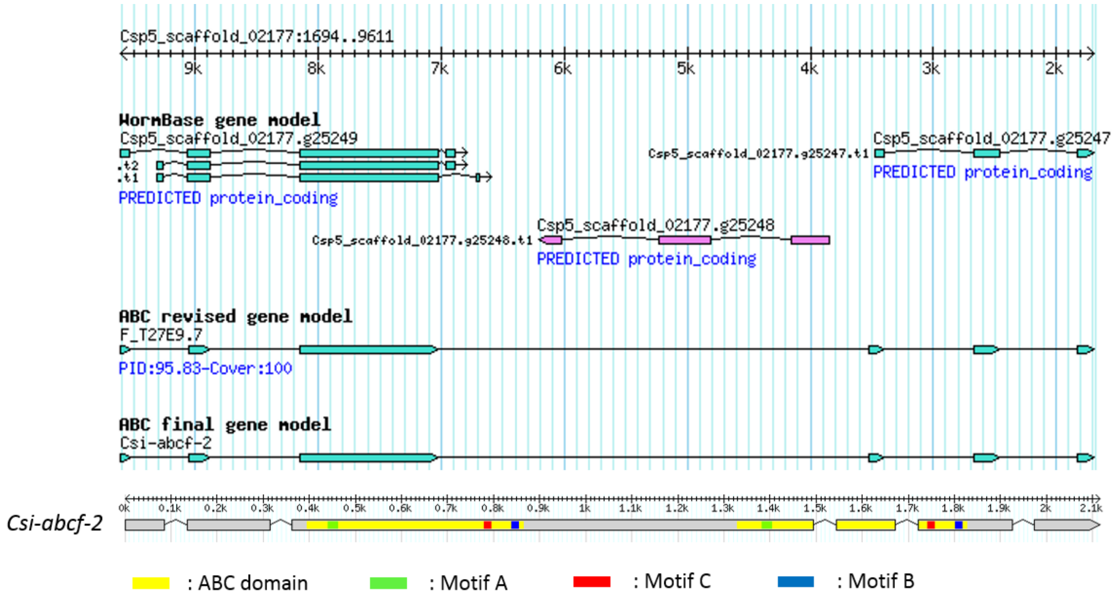
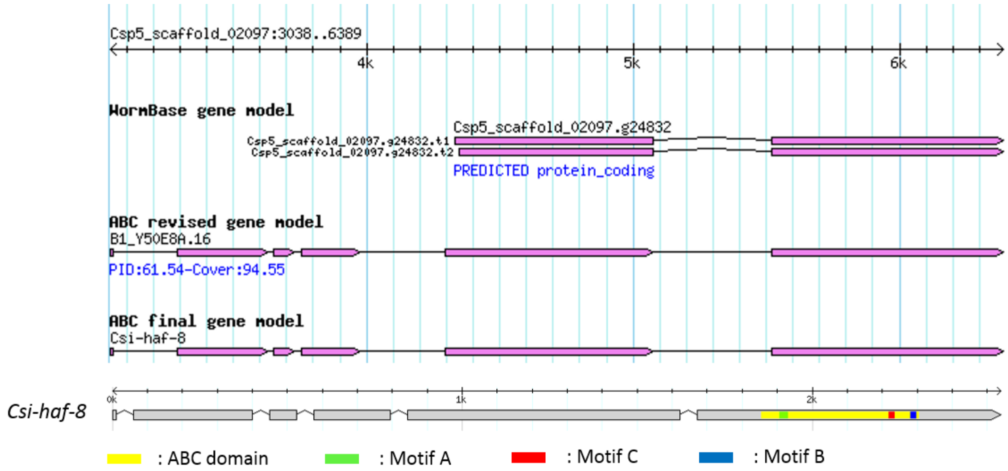


Figure 3.9: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Csp5_scaffold_02177.g25247 and Csp5_scaffold_02177.g25249 were fragments of an ABC transporter gene in subfamily F, each of which encoded one predicted ABC domain. After improvement, two candidates were merged together to be a high-quality ABC transporter gene.



TM domain prediction

```
Csp5_scaffold_02097.g24832.t1
536
YES
IN
3
1. 94-114, 2. 116-136, 3. 185-205
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMM
MMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooo
```

```
Csi-haf-8
763
YES
OUT
8
1. 47-67, 2. 75-95, 3. 109-129, 4. 199-219, 5. 239-259,
6. 321-341, 7. 343-363, 8. 412-432
ooooooooooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMM
MMMMMMMiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooMMMMMMMMMMMM
MMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooMM
MMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMM
MMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMM
MMMMMMMMMMMMoooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooo
```

Figure 3.10: A representative case that the TM domain of an ABC transporter gene candidate was improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Csp5_scaffold_02097.g24832, annotated as a half ABC transporter gene in subfamily B encoding only three TM helices. The revised gene model for this candidate was longer and encoded a complete TM domain.

Through phylogenetic analysis, we found 46 out of 70 ABC transporter genes in *C. sinica* showed one-to-one orthologous relationship with the ABC transporter genes in *C. elegans*. We assigned the gene names for the ABC transporter genes in *C. sinica* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.11). In comparison with ABC transporter genes in *C. elegans*, those in *C. sinica* showed both high level conservation and divergence. On one hand, we found that ABCD, ABCE, ABCF and ABCH subfamily were well conserved between *C. elegans* and *C. sinica*. On the other hand, there were some small expansions of subfamily A (*Csi-abt-3* and *Csi-abt-6*), C (*Csi-mrp-9* and *Csi-mrp-10*) and G (*Csi-wht-9*) in *C. sinica.* The largest expansion occurred in subfamily B, which could also be recognized from the total number of ABC transporter gene in this subfamily, 24 in *C. elegans* and 33 in *C. sinica.* Especicaly, for the full ABC transporter genes in subfamily B, there were obviously expansions in the cluster that had *pgp-12*. In conclusion, most of the subfamilies were general conserved between *C. elegans* and *C. sinica*, with most expansions of subfamily B in *C. sinica.*

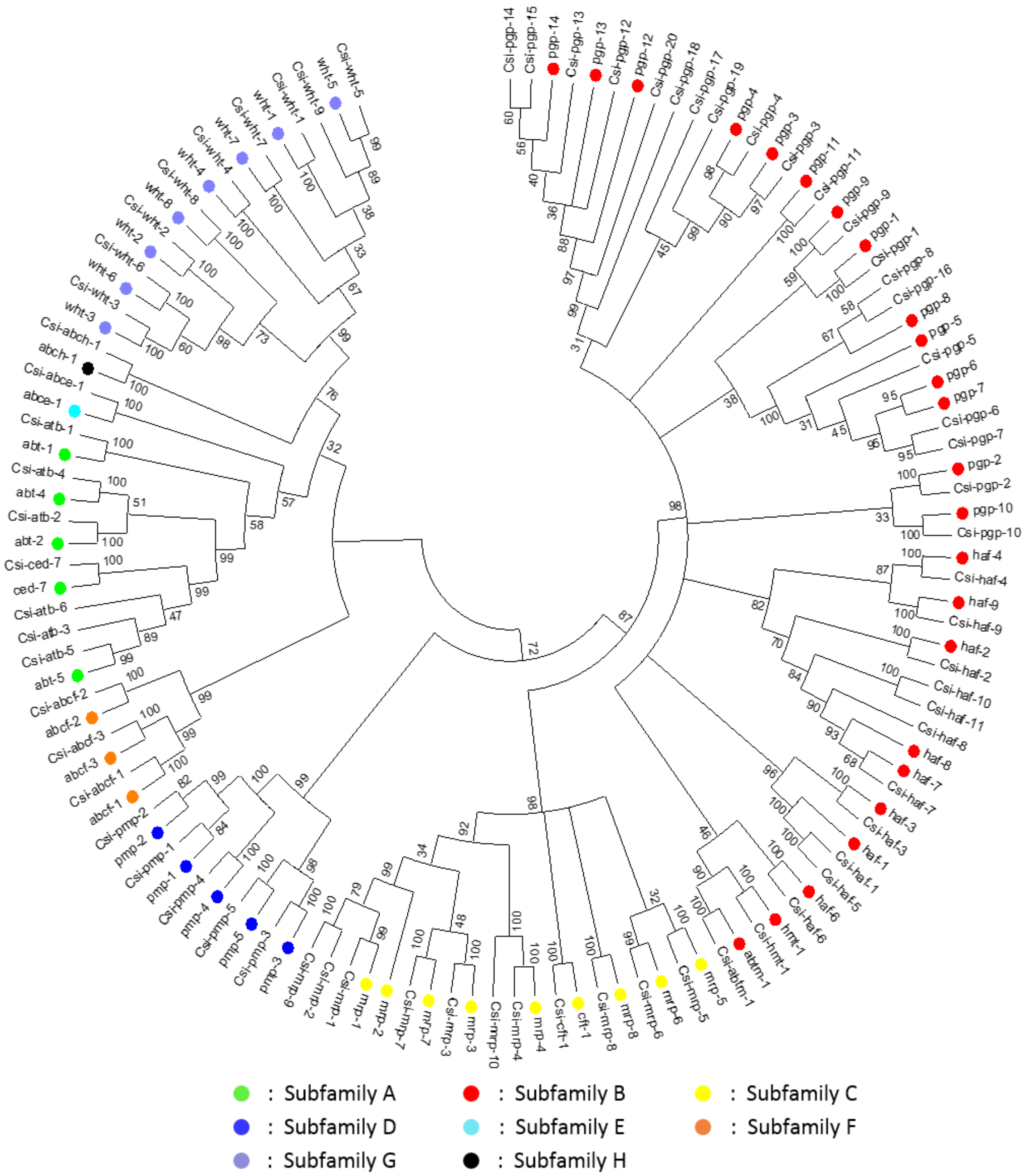


Figure 3.11: Phylogenetic analysis between *C. sinica* and *C. elegans*

Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *C. sinica* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *C. sinica* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.6. Annotation of ABC transporter genes in *C. brenneri*

C. brenneri has both male and female adults, unlike the hermaphroditic species such as *C. elegans* and *C. briggsae*. The genome of *C. brenneri* is about 40% larger than the genome of either *C. elegans* or *C. briggsae*, but is only slightly larger than *C. remanei's* genome (Fierst et al. 2015). By applying the annotation pipeline to *C. brenneri*, we obtained 106 ABC transporter gene candidates (91 candidates from InterProScan searches, 15 additional ones from BLAST searches), none of which was due to contamination. Among these 106 candidates, 56 were high-quality ABC transporter genes. All of these 56 genes also encoded appropriate TM domain (s). For the remaining 50 candidates, we tried to improve each of them and we eventually generated 16 revised gene models of high-quality (Table 3.5), three of which with only TM domain improved. Three ABC transporter candidates, CBN31112, CBN30138 and CBN31679, should be merged into one single gene which was annotated as a full ABC transporter gene encoding two high-quality ABC domains in subfamily B based on the improvement result (Figure 3.12). The new gene model had RNA-seq data supporting all of its introns. CBN31544 was annotated as a half ABC transporter genes in subfamily G, encoding a high-quality ABC domain but no TM domain. Through improvement, we recovered a longer ABC transporter gene model, which resulted from merging CBN31544 and its adjacent gene, CBN29600 (Figure 3.13). This new gene model encoded four predicted TM helices, as well as a high-quality ABC domain. In addition, it had RNA-seq data supporting nine of its introns. After improvement, 23 candidates were still defective, 16 of which might be caused by incomplete assembly or incomplete sequencing of their genomic region. As a result, there could be 12 potential full length ABC transporter genes if the *C. brenneri* genome is fully reconstructed. In a representative case, CBN28163 and CBN09215 were both annotated as a full ABC transporter gene in subfamily A and were suggested to be merged into one. However, the newly constructed gene model only encoded one high-quality ABC domain and contained a sequencing gap, which might result in the incompleteness of this ABC transporter gene (Figure 3.14). Taking together, we annotated 78 high-quality ABC transporter genes in *C. brenneri*, 73 of which had appropriate TM domains (Table 3.5).

Table 3.5: High-quality ABC transporter genes in *C. brenneri* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Cbn-abt-1*	CBN20002	CBN20002	6TM-ABC	2707	12	605	Exons were improved; No stop codon
	Cbn-abt-2	CBN02057	CBN02057	9TM-ABC-7TM-ABC	11149	33	2330	
	Cbn-abt-3	CBN29987	CBN29987	9TM-ABC-7TM-ABC	12408	28	2292	
	Cbn-abt-4	CBN21860	CBN21860	7TM-ABC-9TM-ABC	6402	14	1810	
	Cbn-abt-5	CBN20798	CBN20798	9TM-ABC-9TM-ABC	7069	18	1608	
	Cbn-abt-6	CBN29424	CBN29424	9TM-ABC-9TM-ABC	7322	19	1643	
	Cbn-abt-7	CBN08786	CBN08786	1TM-ABC-7TM-ABC	4413	17	1002	
	Cbn-ced-7	CBN05249	CBN05249	7TM-ABC-7TM-ABC	6050	14	1691	
B	*Cbn-abtm-1*	CBN05427	CBN05427	7TM-ABC	3111	8	704	
	Cbn-abtm-2	CBN11138	CBN11138	7TM-ABC	3107	7	680	
	Cbn-haf-1	CBN06375	CBN06375	6TM-ABC	2408	9	677	
	Cbn-haf-10	CBN28702	CBN28702	9TM-ABC	7833	14	738	
	Cbn-haf-11	CBN23436	CBN23436	8TM-ABC	3243	4	722	
	Cbn-haf-12	CBN13139	CBN13139	7TM-ABC	2574	7	763	
	Cbn-haf-13	CBN10479	CBN10479	7TM-ABC	2765	5	791	
	Cbn-haf-2	CBN28239	CBN28239	7TM-ABC	2572	7	763	
	Cbn-haf-3	CBN22002	CBN22002	6TM-ABC	4002	11	624	
	Cbn-haf-4	CBN11644	CBN11644	9TM-ABC	4108	11	787	No start codon
	Cbn-haf-5	CBN20964	CBN20964	9TM-ABC	7267	11	787	
	Cbn-haf-6	CBN16346	CBN16346	6TM-ABC	9992	6	667	
	Cbn-haf-7	CBN01809	CBN01809	9TM-ABC	3007	4	812	
	Cbn-haf-8	CBN06709	CBN06709	0TM-ABC	842	2	272	Exons were improved; No start codon
	Cbn-haf-9	CBN20431	CBN20431	9TM-ABC	7535	16	815	
	Cbn-hmt-1	CBN31144	CBN31144	10TM-ABC	3147	10	812	No stop codon
	Cbn-pgp-1	CBN28232	CBN28232	6TM-ABC-6TM-ABC	6542	13	1320	
	Cbn-pgp-10	CBN05995	CBN05995	6TM-ABC-7TM-ABC	6190	26	1357	
	Cbn-pgp-11	CBN30138	CBN30138	6TM-ABC-5TM-ABC	17587	16	1255	CBN31112 and CBG31679 were merged with CBN30138; No start codon
	Cbn-pgp-12	CBN29443	CBN29443	6TM-ABC-5TM-ABC	4856	13	1314	
	Cbn-pgp-13	CBN05968	CBN05968	6TM-ABC-6TM-ABC	4525	11	1279	No start codon
	Cbn-pgp-14	CBN05145	CBN05145	6TM-ABC-6TM-ABC	4520	8	1327	
	Cbn-pgp-2	CBN19106	CBN19106	6TM-ABC-6TM-ABC	4497	11	1326	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
B	*Cbn-pgp-3*	CBN22429	CBN22429	6TM-ABC-6TM-ABC	4799	16	1251	Exons were improved
	Cbn-pgp-4	CBN30170	CBN30170	6TM-ABC-5TM-ABC	5293	13	1321	
	Cbn-pgp-5	CBN32747	CBN32747	6TM-ABC-6TM-ABC	4768	18	1200	
	Cbn-pgp-6	CBN32748	CBN32748	5TM-ABC-6TM-ABC	5296	20	1221	Exons were improved; No start codon
	Cbn-pgp-7	CBN29992	CBN29992	6TM-ABC-6TM-ABC	4985	15	1278	
	Cbn-pgp-8	CBN31256	CBN31256	6TM-ABC-4TM-ABC	5175	17	1213	CBN31805 were merged with CBN31256
	Cbn-pgp-9	CBN23865	CBN23865	6TM-ABC-6TM-ABC	4256	9	1293	
C	*Cbn-mrp-1*	CBN28994	CBN28994	11TM-ABC-5TM-ABC	10692	22	1523	Exons were improved
	Cbn-mrp-10	CBN16903	CBN16903	11TM-ABC-5TM-ABC	5296	13	1498	
	Cbn-mrp-11	CBN05120	CBN05120	11TM-ABC-5TM-ABC	7293	13	1477	
	Cbn-mrp-12	CBN19130	CBN19130	4TM-ABC-6TM-ABC	5386	14	1161	Exons were improved
	Cbn-mrp-13	CBN20416	CBN20416	2TM-ABC-5TM-ABC	7079	12	1042	
	Cbn-mrp-14	CBN30570	CBN30570	11TM-ABC-6TM-ABC	8789	11	1562	
	Cbn-mrp-2	CBN26160	CBN26160	11TM-ABC-5TM-ABC	6570	12	1532	
	Cbn-mrp-3	CBN16989	CBN16989	11TM-ABC-6TM-ABC	6963	9	1575	
	Cbn-mrp-4	CBN14450	CBN14450	11TM-ABC-6TM-ABC	5120	9	1569	No stop codon
	Cbn-mrp-5	CBN15696	CBN15696	8TM-ABC-6TM-ABC	6154	15	1422	
	Cbn-mrp-6	CBN05362	CBN05362	7TM-ABC-6TM-ABC	5846	17	1400	
	Cbn-mrp-7	CBN16379	CBN16379	11TM-ABC-5TM-ABC	5520	12	1495	
	Cbn-mrp-8	CBN10730	CBN10730	11TM-ABC-5TM-ABC	16895	18	1449	
	Cbn-mrp-9	CBN25075	CBN25075	11TM-ABC-5TM-ABC	15927	18	1449	
D	*Cbn-pmp-1*	CBN32085	CBN32085	0TM-ABC	1256	3	250	No start codon
	Cbn-pmp-2	CBN31710	CBN31710	5TM-ABC	3164	9	650	
	Cbn-pmp-3	CBN16310	CBN16310	6TM-ABC	5178	7	660	
	Cbn-pmp-4	CBN06980	CBN06980	6TM-ABC	2815	12	733	
	Cbn-pmp-5	CBN04747	CBN04747	1TM-ABC	2043	5	481	Exons were improved; No start codon
	Cbn-pmp-6	CBN13604	CBN13604	5TM-ABC	2523	8	592	Exons were improved
E	*Cbn-abce-1*	CBN15483	CBN15483	ABC-ABC	2610	6	610	No stop codon
F	*Cbn-abcf-1*	CBN31270	CBN31270	ABC-ABC	2390	6	618	
	Cbn-abcf-2	CBN21766	CBN21766	ABC-ABC	3719	6	620	No stop codon
	Cbn-abcf-3	CBN31269	CBN31269	ABC-ABC	7015	5	712	
	Cbn-abcf-4	CBN21377	CBN21377	ABC-ABC	2241	5	623	No start and stop codon
	Cbn-abcf-5	CBN13072	CBN13072	ABC-ABC	2284	6	615	Exons were improved; No start codon

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
G	*Cbn-wht-1*	CBN11286	CBN11286	ABC-8TM	5370	13	661	No stop codon
	Cbn-wht-10	CBN08316	CBN08316	ABC-7TM	4754	8	690	Exons were improved; No start codon
	Cbn-wht-11	CBN01826	CBN01826	ABC-4TM	5996	9	533	CBN29225 were merged with CBN01826; TM helices were improved
	Cbn-wht-12	CBN29600	CBN29600	ABC-4TM	9492	12	613	CBN31544 were merged with CBN29600; TM helices were improved
	Cbn-wht-13	CBN23498	CBN23498	ABC-5TM	2623	11	663	
	Cbn-wht-2	CBN19674	CBN19674	ABC-6TM	2356	12	610	
	Cbn-wht-3	CBN31363	CBN31363	ABC-6TM	5247	8	580	
	Cbn-wht-4	CBN16387	CBN16387	ABC-5TM	3383	10	570	
	Cbn-wht-5	CBN14682	CBN14682	ABC-7TM	5446	8	647	
	Cbn-wht-6	CBN01087	CBN01087	ABC-6TM	10738	9	627	
	Cbn-wht-7	CBN25724	CBN25724	ABC-5TM	15009	10	665	Exons were improved
	Cbn-wht-8	CBN06720	CBN06720	ABC-6TM	2223	11	562	
	Cbn-wht-9	CBN09809	CBN09809	ABC-6TM	2355	12	610	
H	*Cbn-abch-1*	CBN11737	CBN11737	ABC-6TM	2314	12	596	TM helices were improved

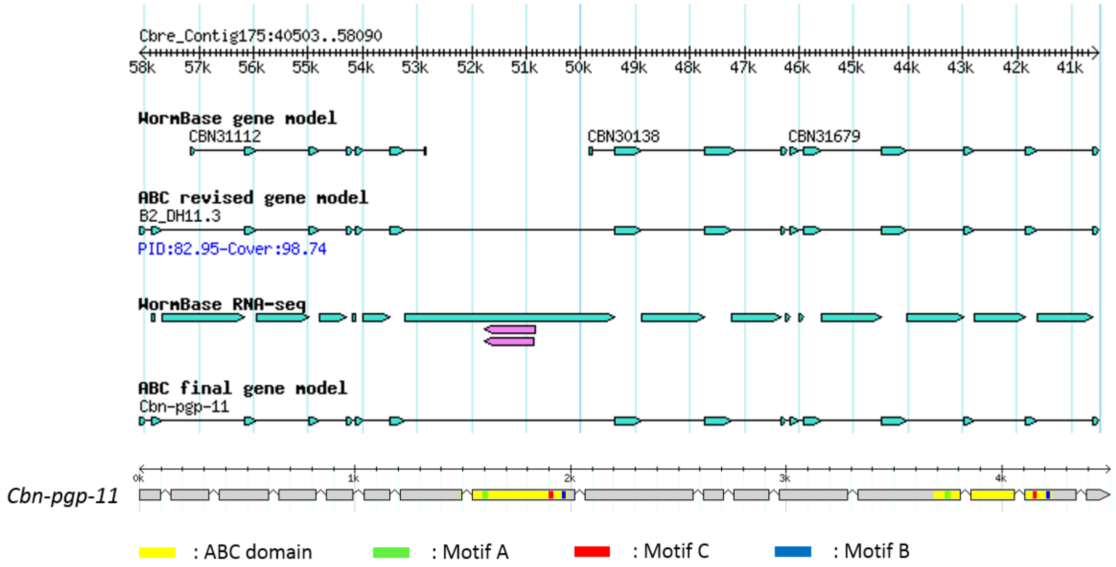
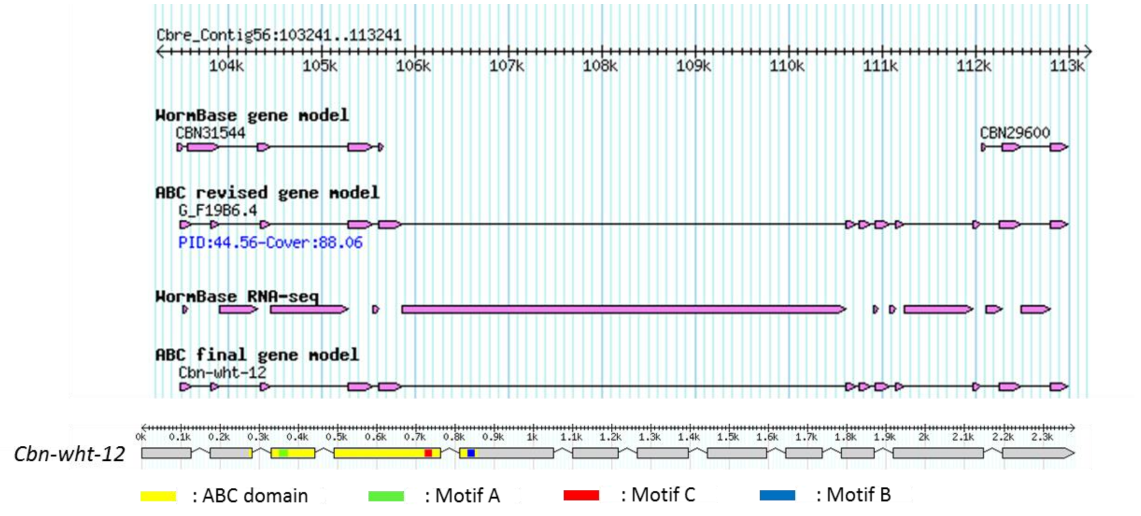


Figure 3.12: A representative case that three candidates were merged into one high-quality ABC transporter gene

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “WormBase RNA-seq” track includes introns predicted by RNA-seq data; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. CBN31112, CBN30138 and CBN31679, should be merged into one single gene which was annotated as a full ABC transporter gene in subfamily B. The new gene model had RNA-seq data supporting all of its introns.



TM domain prediction

```
CBN31544
280
NO

0
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

```
Cbn-wht-12
612
YES
IN
4
1. 368-388, 2. 453-473, 3. 497-517, 4. 582-602
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiMMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiii
iiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMM
MMiiiiiiiiii
```

Figure 3.13: A representative case that TM domain was improved after merging two adjacent genes

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "WormBase RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. CBN31544, was annotated as a half ABC transporter gene in subfamily G encoding no TM domain. Through improvement, we obtained a longer ABC transporter gene model, which resulted from merging CBN31544 and its adjacent gene CBN29600. And this revised gene model was supported by RNA-seq data and encoded proper TM domain and ABC domain.

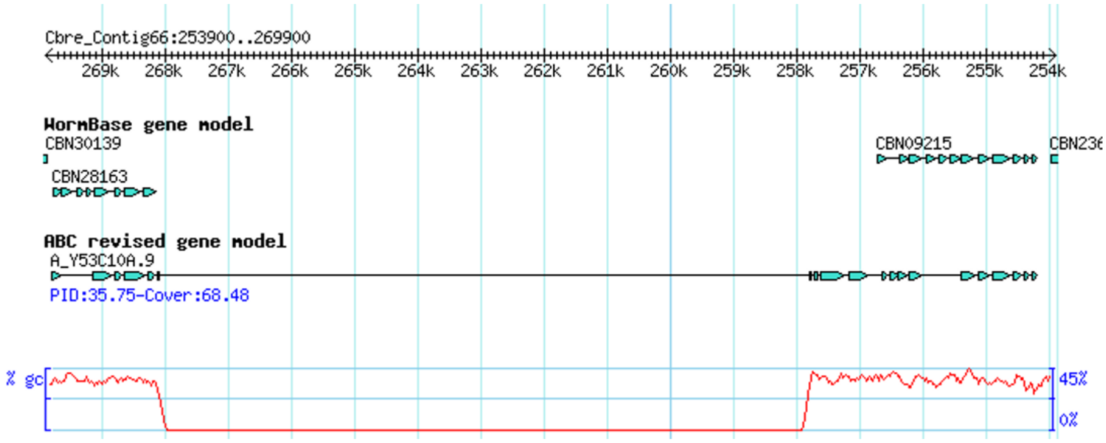


Figure 3.14: A representative case that the incompleteness of a defective candidate could be caused by technical issues

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. CBN28163 and CBN09215 were fragments of a full ABC transporter gene in subfamily A. However, the revised gene model only encoded one high-quality ABC domain and there was a sequencing gap, which resulted in the incompleteness of this ABC transporter gene.

Through phylogenetic analysis, we found only 34 out of 78 ABC transporter genes in *C. brenneri* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *C. brenneri* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.15). The large number of high-quality ABC transporter genes, small number of one-to-one orthologous relationship as well as 12 potential candidates prompted us to make a comparison between ABC transporter genes in *C. elegans*, *C. briggsae* and *C. brenneri*. Interestingly, we found 13 cases that in a single cluster, there were only one ABC transporter gene in each of *C. elegans* and *C. briggsae* genome, but more than one gene in *C. brenneri* (Table 3.6). For *C. brenneri* ABC transporter genes in the same cluster, the gene structures were very similar to each other (Figure 3.16). For example, *wht-2* had one signle ortholog in *C. briggsae* (*Cbr-wht-2*), but two orthologs, *Cbn-wht-2* and *Cbn-wht-9*, which shared almost identical protein sequences (only two base-pair difference) and gene structure in *C. brenneri* (Figure 3.17). This result suggested that these expansion cases could be resulted from the heterozygosity as well as the large genome of *C. brenneri* (Barriere et al. 2009).

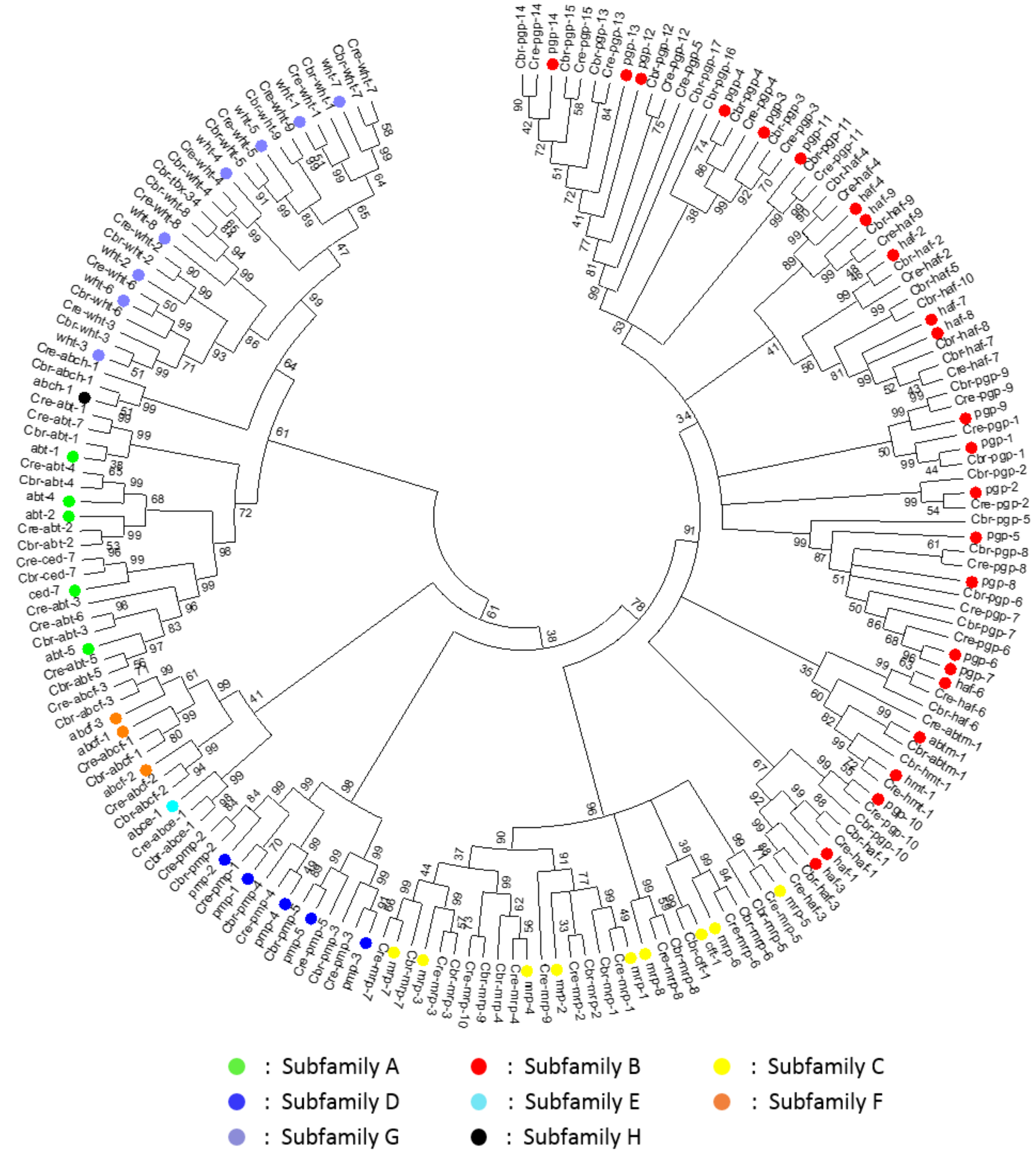


Figure 3.15: Phylogenetic analysis among *C. brenneri*, *C. briggsae* and *C. elegans*

Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *C. brenneri*, *C. briggsae* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *C. brenneri* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

Table 3.6: 13 expanded cases in *C. brenneri* compared to *C. elegans* and *C. briggsae*

C. elegans	Ortholog in *C. briggsae*	Ortholog in *C. brenneri*
wht-2	*Cbr-wht-2*	*Cbn-wht-2*
		Cbn-wht-9
wht-4	*Cbr-wht-4*	*Cbn-wht-13*
		Cbn-wht-4
wht-5	*Cbr-wht-5*	*Cbn-wht-5*
		Cbn-wht-10
abt-2	*Cbr-abt-2*	*Cbn-abt-3*
		Cbn-abt-2
abt-5	*Cbr-abt-5*	*Cbn-abt-5*
		Cbn-abt-6
abcf-1	*Cbr-abcf-1*	*Cbn-abcf-5*
		Cbn-abcf-4
		Cbn-abcf-1
pmp-2	*Cbr-pmp-2*	*Cbn-pmp-6*
		Cbn-pmp-2
mrp-7	*Cbr-mrp-7*	*Cbn-mrp-7*
		Cbn-mrp-10
mrp-8	*Cbr-mrp-8*	*Cbn-mrp-9*
		Cbn-mrp-8
abtm-1	*Cbr-abtm-1*	*Cbn-abtm-1*
		Cbn-abtm-2
haf-4	*Cbr-haf-4*	*Cbn-haf-4*
		Cbn-haf-5
haf-9	*Cbr-haf-9*	*Cbn-haf-9*
		Cbn-haf-10
haf-2	*Cbr-haf-2*	*Cbn-haf-2*
		Cbn-haf-12

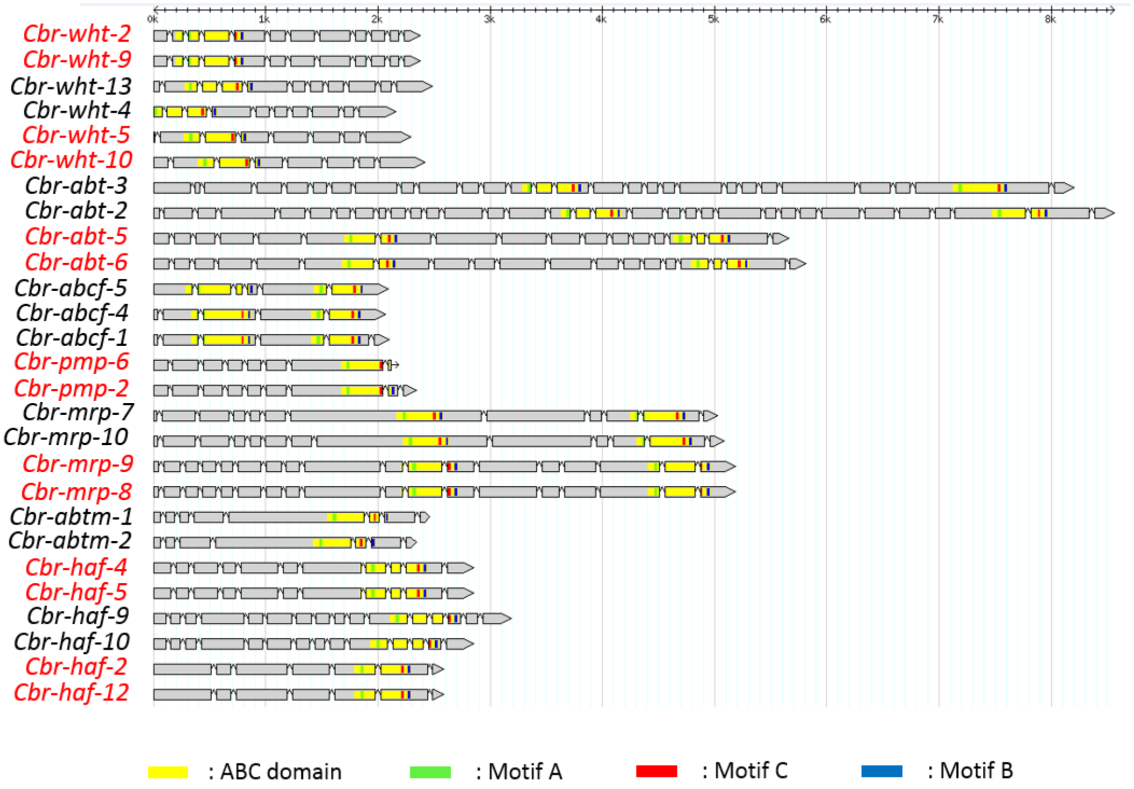
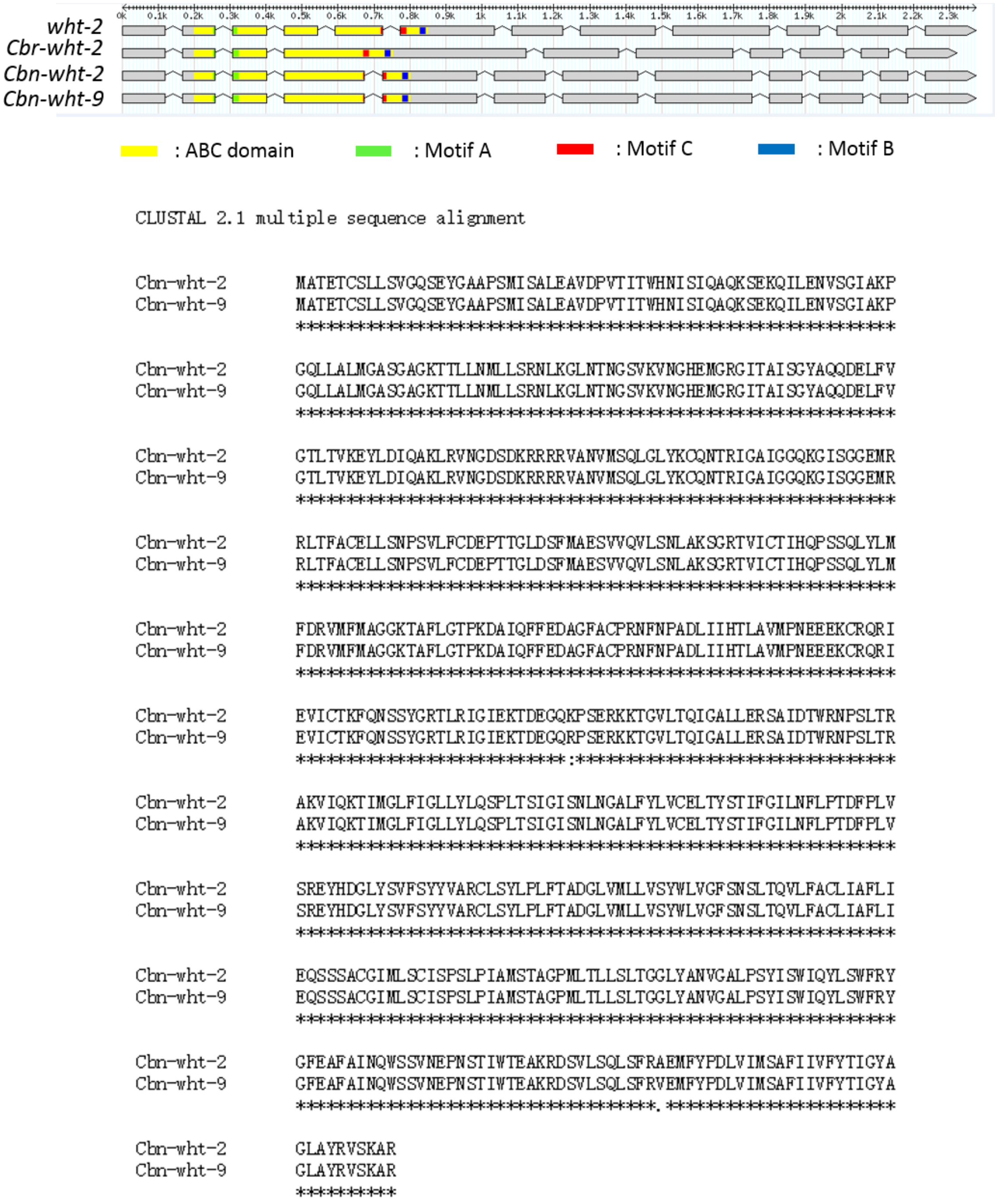


Figure 3.16: Gene structure for *C. brenneri* ABC transporter genes in the 13 expansion cases

Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B.


0k 0.1k 0.2k 0.3k 0.4k 0.5k 0.6k 0.7k 0.8k 0.9k 1k 1.1k 1.2k 1.3k 1.4k 1.5k 1.6k 1.7k 1.8k 1.9k 2k 2.1k 2.2k 2.3k
wht-2
Cbr-wht-2
Cbn-wht-2
Cbn-wht-9
🟨 : ABC domain
🟩 : Motif A
🟥 : Motif C
🟦 : Motif B

```
CLUSTAL 2.1 multiple sequence alignment


Cbn-wht-2       MATETCSLLSVGQSEYGAAPSMISALEAVDPVTITWHNISIQAQKSEKQILENVSGIAKP
Cbn-wht-9       MATETCSLLSVGQSEYGAAPSMISALEAVDPVTITWHNISIQAQKSEKQILENVSGIAKP
                ************************************************************

Cbn-wht-2       GQLLALMGASGAGKTTLLNMLLSRNLKGLNTNGSVKVNGHEMGRGITAISGYAQQDELFV
Cbn-wht-9       GQLLALMGASGAGKTTLLNMLLSRNLKGLNTNGSVKVNGHEMGRGITAISGYAQQDELFV
                ************************************************************

Cbn-wht-2       GTLTVKEYLDIQAKLRVNGDSDKRRRRVANVMSQLGLYKCQNTRIGAIGGQKGISGGEMR
Cbn-wht-9       GTLTVKEYLDIQAKLRVNGDSDKRRRRVANVMSQLGLYKCQNTRIGAIGGQKGISGGEMR
                ************************************************************

Cbn-wht-2       RLTFACELLSNPSVLFCDEPTTGLDSFMAESVVQVLSNLAKSGRTVICTIHQPSSQLYLM
Cbn-wht-9       RLTFACELLSNPSVLFCDEPTTGLDSFMAESVVQVLSNLAKSGRTVICTIHQPSSQLYLM
                ************************************************************

Cbn-wht-2       FDRVMFMAGGKTAFLGTPKDAIQFFEDAGFACPRNFNPADLIIHTLAVMPNEEEKCRQRI
Cbn-wht-9       FDRVMFMAGGKTAFLGTPKDAIQFFEDAGFACPRNFNPADLIIHTLAVMPNEEEKCRQRI
                ************************************************************

Cbn-wht-2       EVICTKFQNSSYGRTLRIGIEKTDEGQKPSERKKTGVLTQIGALLERSAIDTWRNPSLTR
Cbn-wht-9       EVICTKFQNSSYGRTLRIGIEKTDEGQRPSERKKTGVLTQIGALLERSAIDTWRNPSLTR
                ***************************:********************************

Cbn-wht-2       AKVIQKTIMGLFIGLLYLQSPLTSIGISNLNGALFYLVCELTYSTIFGILNFLPTDFPLV
Cbn-wht-9       AKVIQKTIMGLFIGLLYLQSPLTSIGISNLNGALFYLVCELTYSTIFGILNFLPTDFPLV
                ************************************************************

Cbn-wht-2       SREYHDGLYSVFSYYVARCLSYLPLFTADGLVMLLVSYWLVGFSNSLTQVLFACLIAFLI
Cbn-wht-9       SREYHDGLYSVFSYYVARCLSYLPLFTADGLVMLLVSYWLVGFSNSLTQVLFACLIAFLI
                ************************************************************

Cbn-wht-2       EQSSSACGIMLSCISPSLPIAMSTAGPMLTLLSLTGGLYANVGALPSYISWIQYLSWFRY
Cbn-wht-9       EQSSSACGIMLSCISPSLPIAMSTAGPMLTLLSLTGGLYANVGALPSYISWIQYLSWFRY
                ************************************************************

Cbn-wht-2       GFEAFAINQWSSVNEPNSTIWTEAKRDSVLSQLSFRAEMFYPDLVIMSAFIIVFYTIGYA
Cbn-wht-9       GFEAFAINQWSSVNEPNSTIWTEAKRDSVLSQLSFRVEMFYPDLVIMSAFIIVFYTIGYA
                ************************************.***********************

Cbn-wht-2       GLAYRVSKAR
Cbn-wht-9       GLAYRVSKAR
                **********
```

Figure 3.17: A representative case that shows the expansion of ABC transporter genes in *C. brenneri* could result from heterozygosity

Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The gene models of *Cbn-wht-2*, *Cbn-wht-9*, *Cbr-wht-2* and *wht-2* were quite similar and the exon structures of *Cbn-wht-2* and *Cbn-wht-9* were exactly the same. In addition, CLUSTAL alignment showed the protein sequences of *Cbn-wht-2*, *Cbn-wht-9* were almost identical, suggesting that the expansion might be caused by heterozygosity.

3.7. Annotation of ABC transporter genes in *C. remanei*

C. remanei is a free-living nematode which shares a more recent common ancestor with *C. briggsae* than with *C. elegans* (Haag et al. 2007). After applying the annotation pipeline to *C. remanei*, we obtained 84 ABC transporter gene candidates (81 candidates from InterProScan searches, three additional ones from BLAST searches), four (CRE03604, CRE25047, CRE2690 and CRE31641) of which was due to contamination from bacteria. After excluding the contamination and examining the quality of the remaining 80 candidates, 51 were high-quality ABC transporter genes. All of these 51 genes also encoded appropriate TM domain (s). Then, we tried to further improve the 29 defective candidates and eventually, we generated eight revised gene models of high-quality, two of which with only TM domain improved (Table 3.7). For example, CRE07432, was annotated as a full ABC transporter gene in subfamily B and encoded four predicted ABC domains. Through the improvement procedure, we obtained two candidates, split from CRE07432. Both of the revised gene model encoded two typical ABC domains and the revision had RNA-seq data support (Figure 3.18). CRE01587 was annotated as a full ABC transporter gene from subfamily B and encoded a slight smaller number of TM helices (nine) than expected. The revised gene model of CRE01587 encoded 10 TM helices in total, clustering in two groups. Although the number of TM helices did not change much, the new gene model had RNA-seq data supporting the newly constructed intron, suggesting the new gene model is better than the original one (Figure 3.19). Among the remaining 16 candidates that could not be further improved to be high-quality ABC transporter genes, two (CRE05353 and CRE27936) could be complete ABC transporter genes when we have a better genome quality. CRE05353, annotated as a half ABC transporter gene in subfamily B, had a short ABC domain (101 aa), which could be caused by the sequencing gap within the ABC domain (Figure 3.20) and it could be a complete ABC transporter gene when genome assembly improves. CRE27936 was annotated as a full ABC transporter gene in subfamily B and encoded only one typical ABC domai. It was located at the end of a small contig, probably leading to this truncated candidate (Figure 3.21). In total, we annotated 61 high-quality ABC transporter genes in *C. remanei*, all of which had appropriate TM domain (Table 3.7).

Table 3.7: High-quality ABC transporter genes in *C. remanei* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Cre-abt-1*	CRE30002	CRE30002	5TM-ABC-7TM-ABC	5984	22	1626	
	Cre-abt-2	CRE08294	CRE08294	7TM-ABC-8TM-ABC	14632	38	2316	
	Cre-abt-3	CRE25322	CRE25322	8TM-ABC-9TM-ABC	5868	22	1594	
	Cre-abt-4	CRE09188	CRE09188	7TM-ABC-9TM-ABC	5903	11	1817	
	Cre-abt-5	CRE28446	CRE28446	9TM-ABC-7TM-ABC	5758	20	1583	
	Cre-abt-6	CRE26165	CRE26165	8TM-ABC-7TM-ABC	5984	25	1613	
	Cre-abt-7	CRE14435	CRE14435	4TM-ABC-4TM-ABC	5289	20	1183	CRE14436 was merged with CRE14435; TM helices were improved; No start codon
	Cre-ced-7	CRE25285	CRE25285	7TM-ABC-7TM-ABC	6591	15	1759	
B	*Cre-abtm-1*	CRE18361	CRE18361	4TM-ABC	1692	1	563	
	Cre-haf-1	CRE28714	CRE28714	5TM-ABC	2936	7	674	
	Cre-haf-2	CRE26135	CRE26135	10TM-ABC	2625	7	775	
	Cre-haf-3	CRE09098	CRE09098	7TM-ABC	2543	10	681	
	Cre-haf-4	CRE03811	CRE03811	9TM-ABC	3684	10	803	
	Cre-haf-6	CRE28113	CRE28113	6TM-ABC	3219	9	668	No start and stop codon
	Cre-haf-7	CRE04975	CRE04975	9TM-ABC	4795	4	801	
	Cre-haf-9	CRE28438	CRE28438	9TM-ABC	3798	17	815	
	Cre-hmt-1	CRE04881	CRE04881	10TM-ABC	2905	10	830	
	Cre-pgp-1	CRE31378	CRE31378	6TM-ABC-6TM-ABC	9367	15	1363	
	Cre-pgp-10	CRE00562	CRE00562	6TM-ABC-5TM-ABC	5982	28	1347	
	Cre-pgp-11	CRE02037	CRE02037	6TM-ABC-5TM-ABC	7978	15	1246	
	Cre-pgp-12	CRE24118	CRE24118	6TM-ABC-5TM-ABC	4749	14	1341	
	Cre-pgp-13	CRE24117	CRE24117	6TM-ABC-6TM-ABC	4824	13	1327	
	Cre-pgp-14	CRE24072	CRE24072	6TM-ABC-6TM-ABC	4588	8	1327	
	Cre-pgp-15	CRE24071	CRE24071	4TM-ABC-6TM-ABC	4491	10	1282	
	Cre-pgp-2	CRE03982	CRE03982	6TM-ABC-5TM-ABC	12446	19	1393	
	Cre-pgp-3	CRE07303	CRE07303	6TM-ABC-6TM-ABC	4732	14	1317	
	Cre-pgp-4	CRE07304	CRE07304	6TM-ABC-5TM-ABC	4997	15	1283	
	Cre-pgp-5	CRE01587	CRE01587	5TM-ABC-5TM-ABC	4622	15	1318	TM helices were improved
	Cre-pgp-6	CRE07433	CRE07433	6TM-ABC-6TM-ABC	4811	16	1277	No stop codon
	Cre-pgp-7	CRE07432a	CRE07432	4TM-ABC-6TM-ABC	4506	17	1253	Split from CRE07432; No start codon
	Cre-pgp-8	CRE07432b	CRE07432	5TM-ABC-5TM-ABC	4421	17	1225	Split from CRE07432
	Cre-pgp-9	CRE22140	CRE22140	6TM-ABC-4TM-ABC	5806	8	1189	CRE22140 was merged with CRE22141

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
C	*Cre-mrp-1*	CRE17131	CRE17131	11TM-ABC-3TM-ABC	8897	20	1347	CRE17132 was merged with CRE17131
	Cre-mrp-10	CRE25095	CRE25095	11TM-ABC-6TM-ABC	5039	9	1562	
	Cre-mrp-2	CRE17133	CRE17133	11TM-ABC-4TM-ABC	6663	23	1508	Exons were improved
	Cre-mrp-3	CRE16789	CRE16789	10TM-ABC-6TM-ABC	5673	21	1528	
	Cre-mrp-4	CRE03284	CRE03284	11TM-ABC-6TM-ABC	5101	6	1620	
	Cre-mrp-5	CRE15405	CRE15405	8TM-ABC-6TM-ABC	5676	15	1434	
	Cre-mrp-6	CRE00343	CRE00343	7TM-ABC-6TM-ABC	6180	16	1439	
	Cre-mrp-7	CRE06044	CRE06044	10TM-ABC-5TM-ABC	13442	13	1499	
	Cre-mrp-8	CRE03108	CRE03108	11TM-ABC-5TM-ABC	9847	21	1469	
	Cre-mrp-9	CRE14222	CRE14222	6TM-ABC-4TM-ABC	4489	14	1285	
D	*Cre-pmp-1*	CRE26211	CRE26211	4TM-ABC	2899	9	663	
	Cre-pmp-2	CRE26210	CRE26210	5TM-ABC	3238	9	662	No start codon
	Cre-pmp-3	CRE04992	CRE04992	6TM-ABC	5222	7	660	No stop codon
	Cre-pmp-4	CRE28729	CRE28729	6TM-ABC	2792	10	763	
	Cre-pmp-5	CRE31152	CRE31152	6TM-ABC	2702	10	627	No stop codon
E	*Cre-abce-1*	CRE24506	CRE24506	ABC-ABC	6543	6	610	
F	*Cre-abcf-1*	CRE27470	CRE27470	ABC-ABC	2061	5	622	
	Cre-abcf-2	CRE31460	CRE31460	ABC-ABC	3552	6	621	CRE31461 was merged with CRE31460
	Cre-abcf-3	CRE03121	CRE03121	ABC-ABC	4240	4	730	No start codon
G	*Cre-wht-1*	CRE03119	CRE03119	ABC-6TM	3761	12	654	
	Cre-wht-2	CRE12815	CRE12815	ABC-6TM	2412	13	614	
	Cre-wht-3	CRE06938	CRE06938	ABC-6TM	3670	9	609	
	Cre-wht-4	CRE06601	CRE06601	ABC-5TM	4120	10	649	No start codon
	Cre-wht-5	CRE30154	CRE30154	ABC-6TM	7364	10	715	
	Cre-wht-6	CRE03182	CRE03182	ABC-6TM	2295	10	626	
	Cre-wht-7	CRE02918	CRE02918	ABC-5TM	3567	5	587	
	Cre-wht-8	CRE04679	CRE04679	ABC-5TM	4160	11	939	
	Cre-wht-9	CRE08571	CRE08571	ABC-6TM	2191	10	555	
H	*Cre-abch-1*	CRE26265	CRE26265	ABC-6TM	2310	12	599	

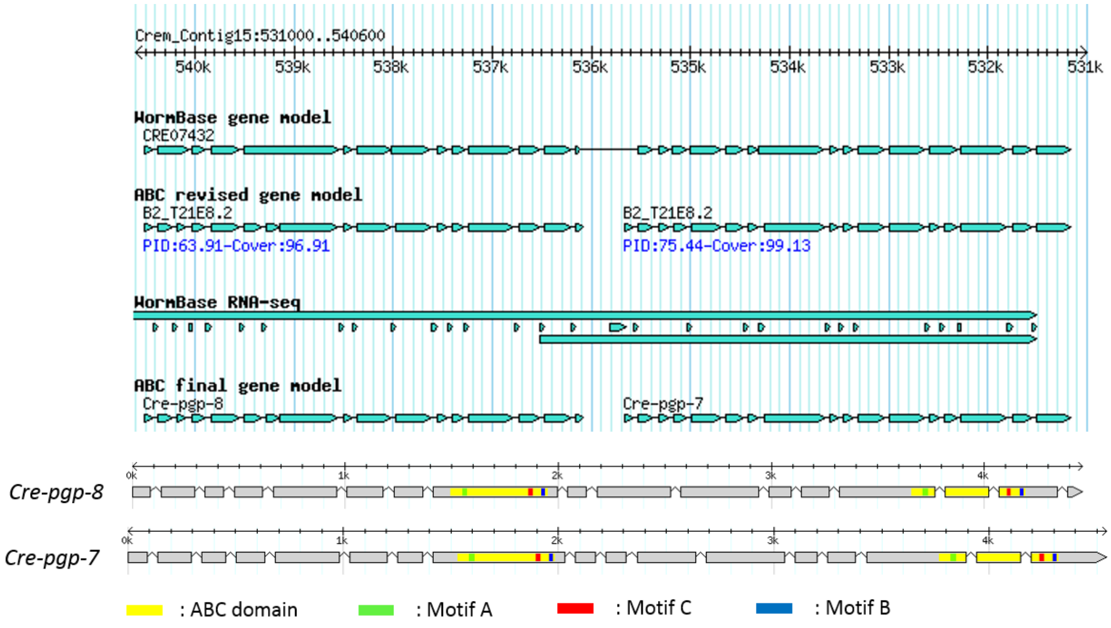
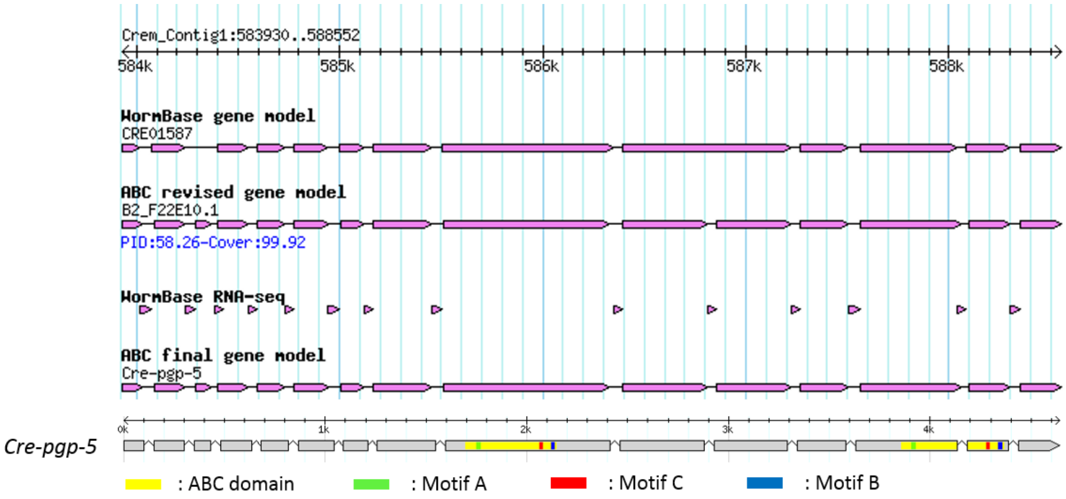
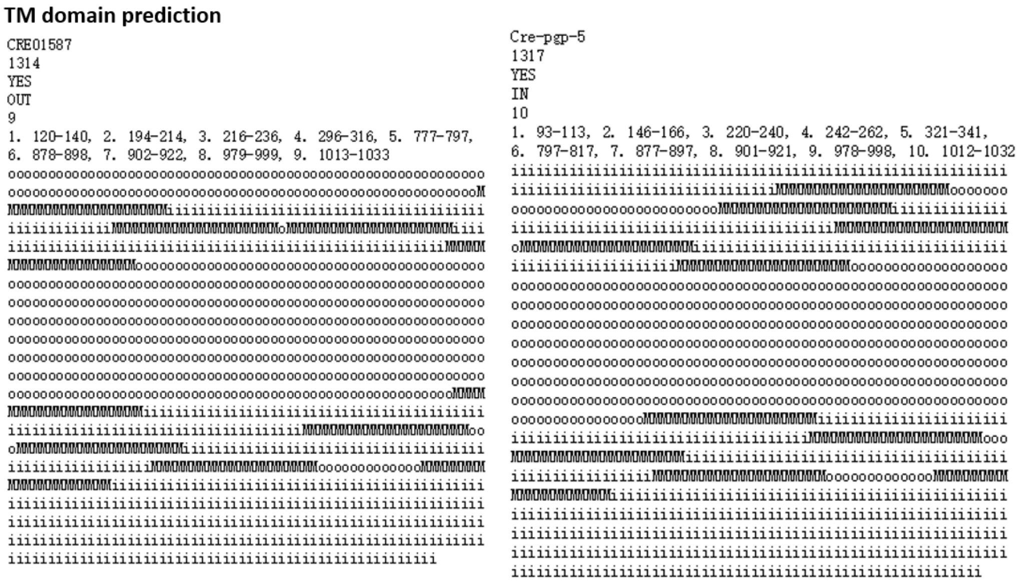


Figure 3.18: A representative case that one candidate was split into two high-quality ABC transporter genes

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “WormBase RNA-seq” track includes introns predicted by RNA-seq data; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. CRE07432 was annotated as a full ABC transporter gene in subfamily B and encoded four predicted ABC domains. Through the improvement procedure, two full ABC transporter genes were obtained by splitting CRE07432. Both of them had two high-quality ABC domains and the revision had RNA-seq data support.





```
CRE01587
1314
YES
OUT
9
1. 120-140, 2. 194-214, 3. 216-236, 4. 296-316, 5. 777-797,
6. 878-898, 7. 902-922, 8. 979-999, 9. 1013-1033
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooM
MMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMM
MMMMMMMMMMMMMMMMoooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooooooMMMM
MMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoo
oMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooMMMMMMMM
MMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

```
Cre-pgp-5
1317
YES
IN
10
1. 93-113, 2. 146-166, 3. 220-240, 4. 242-262, 5. 321-341,
6. 797-817, 7. 877-897, 8. 901-921, 9. 978-998, 10. 1012-1032
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooo
oooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMM
oMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooo
MMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooMMMMMMMMM
MMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

Figure 3.19: A representative case that the TM domain of an ABC transporter gene candidate was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "WormBase RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. CRE01587, was annotated as a full ABC transporter gene from subfamily B and encoded a slight smaller number of TM helices (nine) than expected. The revised gene model encoded 10 TM helices in total, clustering in two groups and all the introns were supported by RNA-seq data.

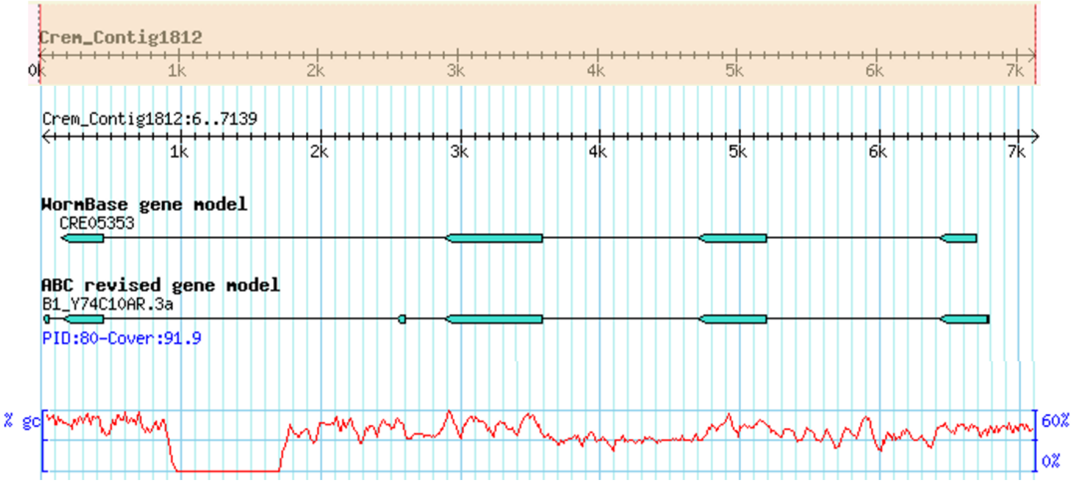


Figure 3.20: A representative case that the incompleteness of a defective candidate could be caused by technical issues

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. CRE05353 was annotated as a half ABC transporter gene in subfamily B and encoded a short ABC domain (101 aa), which could be caused by the sequencing gap within the ABC domain.

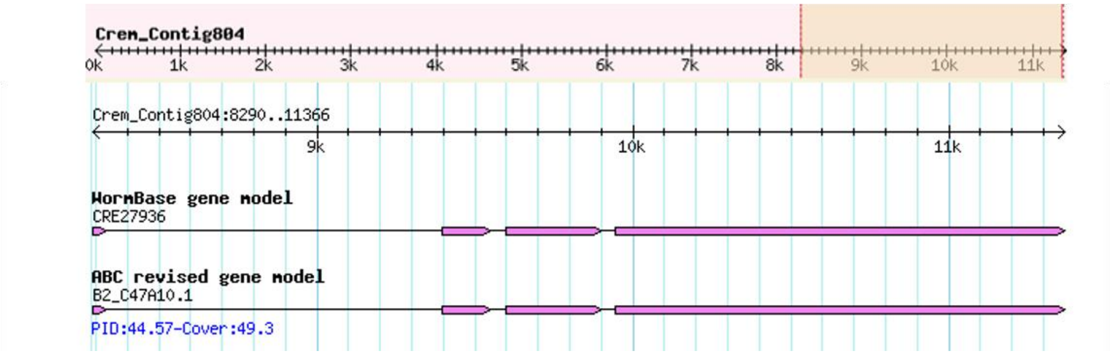


Figure 3.21: A representative case that the incompleteness of a defective candidate could be caused by technical issues

WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. CRE27936 was annotated as a full ABC transporter gene in subfamily B and encoded only one ABC domain with high-quality. It was located at the end of a small contig, probably leading to this truncated ABC transporter gene.

Through phylogenetic analysis, we found 47 out of 61 ABC transporter genes in *C. remanei* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *C. remanei* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.22). Similar to the comparison between *C. elegans* and *C. sinica*, ABCD, ABCE, ABCF and ABCH subfamily were well conserved between *C. elegans* and *C. remanei*. Small expansions were present in subfamily A (*Cre-abt-3*, *Cre-abt-6* and *Cre-abt-7*), subfamily C (*Cre-mrp-9* and *C.re-mrp-10*) and subfamily G (*Cre-wht-9*) in *C. remanei*. Although the total number of ABC transporter genes in subfamily B were identical between *C. elegans* and *C. remanei*, there were some species specific expansions, suggesting that subfamily B is more dynamic during evolution.

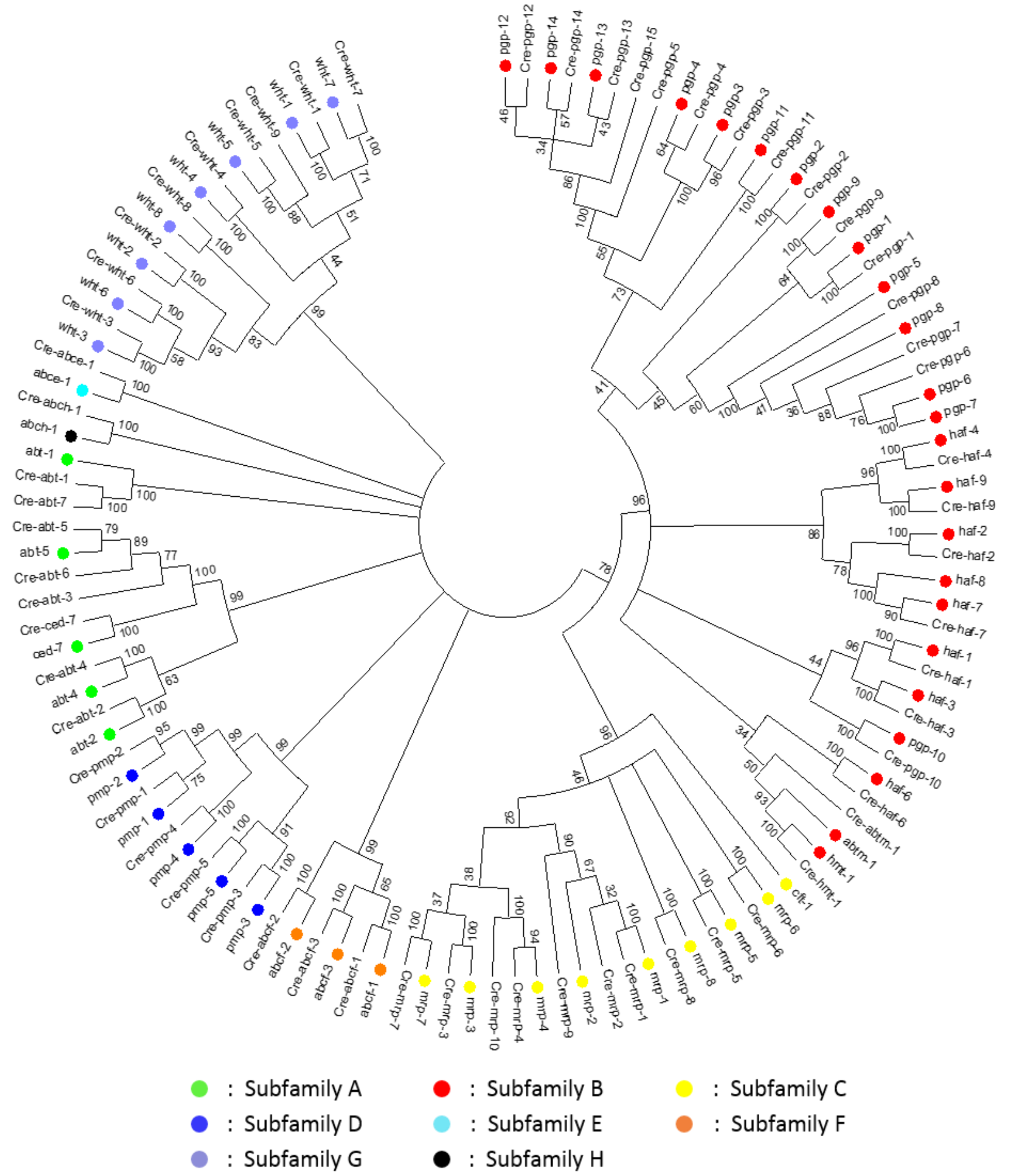


Figure 3.22: Phylogenetic analysis between *C. remanei* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *C. remanei* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *C. remanei* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.8. Annotation of ABC transporter genes in *C. japonica*

Unlike *C. briggsae, C. remanei, or C. brenneri, C. japonica* is not a member of the *Elegans* group*,* but of its sister clade called the *Japonica group* (http://www.wormbase.org). It was selected for genomic sequencing on account of its providing an available outgroup for genomic comparisons with *Elegans group* members. After applying the annotation pipeline to *C. japonica*, we identified 68 ABC transporter gene candidates from InterProScan searches, 25 additional ones from BLAST searches, totally 93 ABC transporter gene candidates. One of these candidates, CJA42478a, was due to contamination. Thus, after checking the quality of 92 remained candidates, 30 were high-quality ABC transporter genes. All of these 30 genes also encoded appropriate TM domain (s). For the 62 candidates, we tried to improve each of them and we ended up with nine revised gene models of high-quality, two of which with only TM domain improved (Table 3.8). CJA14269 and CJA18437 both were annotated as an ABC transporter gene in ABCF subfamily but each of them encoded only one ABC domain. After improvement, we generated a high-quality ABC transporter gene as the result of merging these two candidates together. The revision was supported by RNA-seq data (Figure 3.23). Another example represents TM domain improvement. Unlike most of the TM domain improvement case, the original gene model of CJA03941 encoded 20 TM helices (clustering in two groups), which was more than expected. Through improvement, we generated a new gene model which encoded typically 12 TM helices (Figure 3.24). Among the remaining 33 candidates that could not be further improved to be high-quality ABC transporter genes, nine could be complete ABC transporter genes when the genome is well sequenced and assembled. For instance, the incomplete CJA31802 which only encoded one high-quality ABC domain could be caused by the sequencing gap (Figure 3.25). In summary, we annotated 46 high-quality ABC transporter genes in *C. japonica,* 38 of which had proper TM domain (s) (Table 3.8).

Table 3.8: High-quality ABC transporter genes in *C. japonica* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Cja-abt-1*	CJA22577a	CJA22577a	4TM-ABC-2TM-ABC	2597	8	642	
	Cja-abt-2	CJA10457	CJA10457	4TM-ABC-6TM-ABC	21171	23	1486	No stop codon
	Cja-abt-4	CJA04352	CJA04352	6TM-ABC-8TM-ABC	11311	19	1808	
	Cja-abt-5	CJA03227b	CJA03227b	8TM-ABC-6TM-ABC	5898	15	1428	
B	*Cja-abtm-1*	CJA17248	CJA17248	0TM-ABC	6087	6	705	TM helices were improved
	Cja-haf-2	CJA04699	CJA04699	0TM-ABC	1437	3	384	
	Cja-haf-3	CJA10731	CJA10731	0TM-ABC	3837	7	340	No start codon
	Cja-haf-4	CJA00495b	CJA00495b	9TM-ABC	4588	9	788	No stop codon
	Cja-haf-6	CJA08656	CJA08656	6TM-ABC	5662	10	676	No start codon
	Cja-haf-7	CJA03809b	CJA03809b	4TM-ABC	4055	4	579	TM helices were improved; No start codon
	Cja-haf-9	CJA03278	CJA03278	9TM-ABC	5979	16	817	
	Cja-hmt-1	CJA18704	CJA18704	10TM-ABC	11738	13	802	
	Cja-pgp-10	CJA11203	CJA11203	6TM-ABC-7TM-ABC	6688	28	1361	
	Cja-pgp-12	CJA28064	CJA28064	6TM-ABC-5TM-ABC	4536	12	1314	
	Cja-pgp-13	CJA12756b	CJA12756b	4TM-ABC-7TM-ABC	4127	9	1237	
	Cja-pgp-14	CJA05352	CJA05352	6TM-ABC-6TM-ABC	4978	9	1323	No stop codon
	Cja-pgp-2	CJA14126	CJA14126	7TM-ABC-6TM-ABC	13389	12	1265	No stop codon
	Cja-pgp-3	CJA03941	CJA03941	6TM-ABC-6TM-ABC	4926	15	1268	TM helices were improved
	Cja-pgp-4	CJA06464	CJA06464	6TM-ABC-6TM-ABC	5792	11	1281	
	Cja-pgp-6	CJA03948	CJA03948	5TM-ABC-6TM-ABC	4216	13	1204	
	Cja-pgp-7	CJA28269b	CJA28269b	5TM-ABC-6TM-ABC	4202	13	1196	
	Cja-pgp-8	CJA42892	CJA42892	5TM-ABC-6TM-ABC	4370	10	1293	
	Cja-pgp-9	CJA43094	CJA43094	6TM-ABC-6TM-ABC	9115	11	1234	
C	*Cja-mrp-1*	CJA06868	CJA06868	11TM-ABC-5TM-ABC	11164	18	1529	
	Cja-mrp-3	CJA17614a	CJA17614a	10TM-ABC-5TM-ABC	6704	23	1461	
	Cja-mrp-4	CJA17153	CJA17153	9TM-ABC-6TM-ABC	5254	8	1572	
	Cja-mrp-5	CJA11844	CJA11844	7TM-ABC-7TM-ABC	6958	20	1388	
	Cja-mrp-6	CJA10725	CJA10725	5TM-ABC-6TM-ABC	5311	15	1247	
	Cja-mrp-8	CJA07057	CJA07057	6TM-ABC-6TM-ABC	11418	14	1311	CJA36419 was merged with CJA07057; TM helices were improved; No start codon
D	*Cja-pmp-1*	CJA11388b	CJA11388b	0TM-ABC	2138	3	366	
	Cja-pmp-2	CJA13407	CJA13407	4TM-ABC	4897	7	662	No start codon
	Cja-pmp-3	CJA12201	CJA12201	0TM-ABC	2964	4	376	
	Cja-pmp-4	CJA09316b	CJA09316b	1TM-ABC	3620	7	410	
	Cja-pmp-5	CJA04437	CJA04437	6TM-ABC	7021	13	614	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
E	*Cja-abce-1*	CJA13722	CJA13722	ABC-ABC	4471	4	610	
F	*Cja-abcf-1*	CJA02765	CJA02765	ABC-ABC	6478	6	623	No start codon
	Cja-abcf-2	CJA14269	CJA14269	ABC-ABC	5105	7	622	CJA18437a was merged with CJA14269
	Cja-abcf-3	CJA14647a	CJA14647a	ABC-ABC	5796	5	763	
G	*Cja-wht-1*	CJA16799	CJA16799	ABC-6TM	10937	12	649	No stop codon
	Cja-wht-2	CJA05083b	CJA05083b	ABC-5TM	9367	14	588	
	Cja-wht-3	CJA08136a	CJA08136a	ABC-6TM	6758	9	577	CJA32255 was merged with CJA08136a; No start codon
	Cja-wht-4	CJA14157	CJA14157	ABC-4TM	7689	10	555	Exons were improved; No start codon
	Cja-wht-5	CJA04863	CJA04863	ABC-6TM	4808	6	699	
	Cja-wht-6	CJA15564	CJA15564	ABC-4TM	7874	7	615	CJA21299a was merged with CJA15564
	Cja-wht-7	CJA15178b	CJA15178b	ABC-5TM	11552	9	583	Exons were improved; No start codon
	Cja-wht-8	CJA13911b	CJA13911b	ABC-0TM	1104	3	228	No stop codon

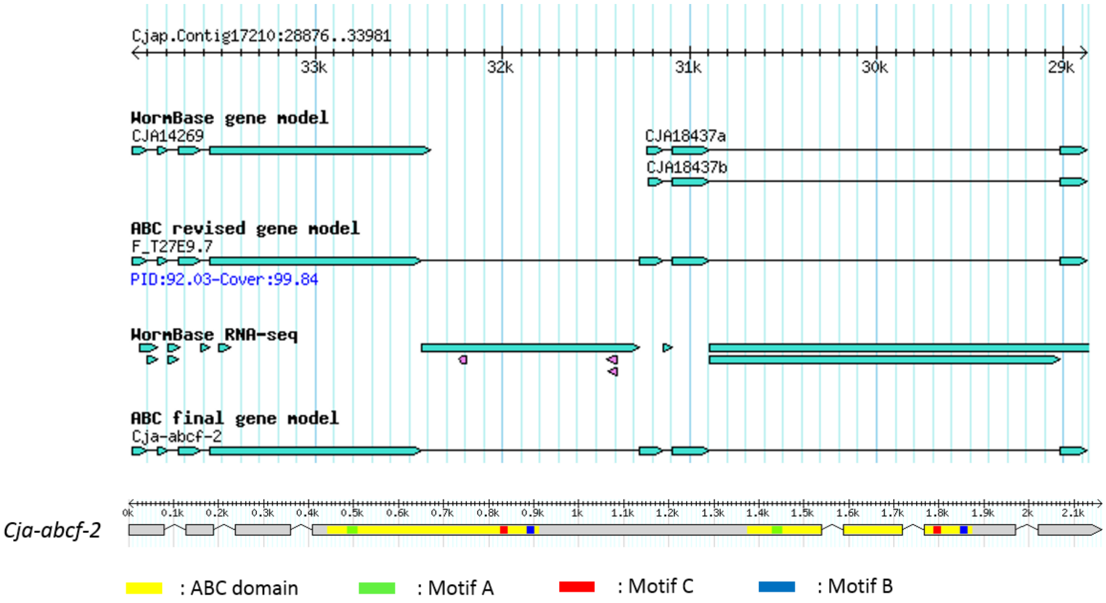
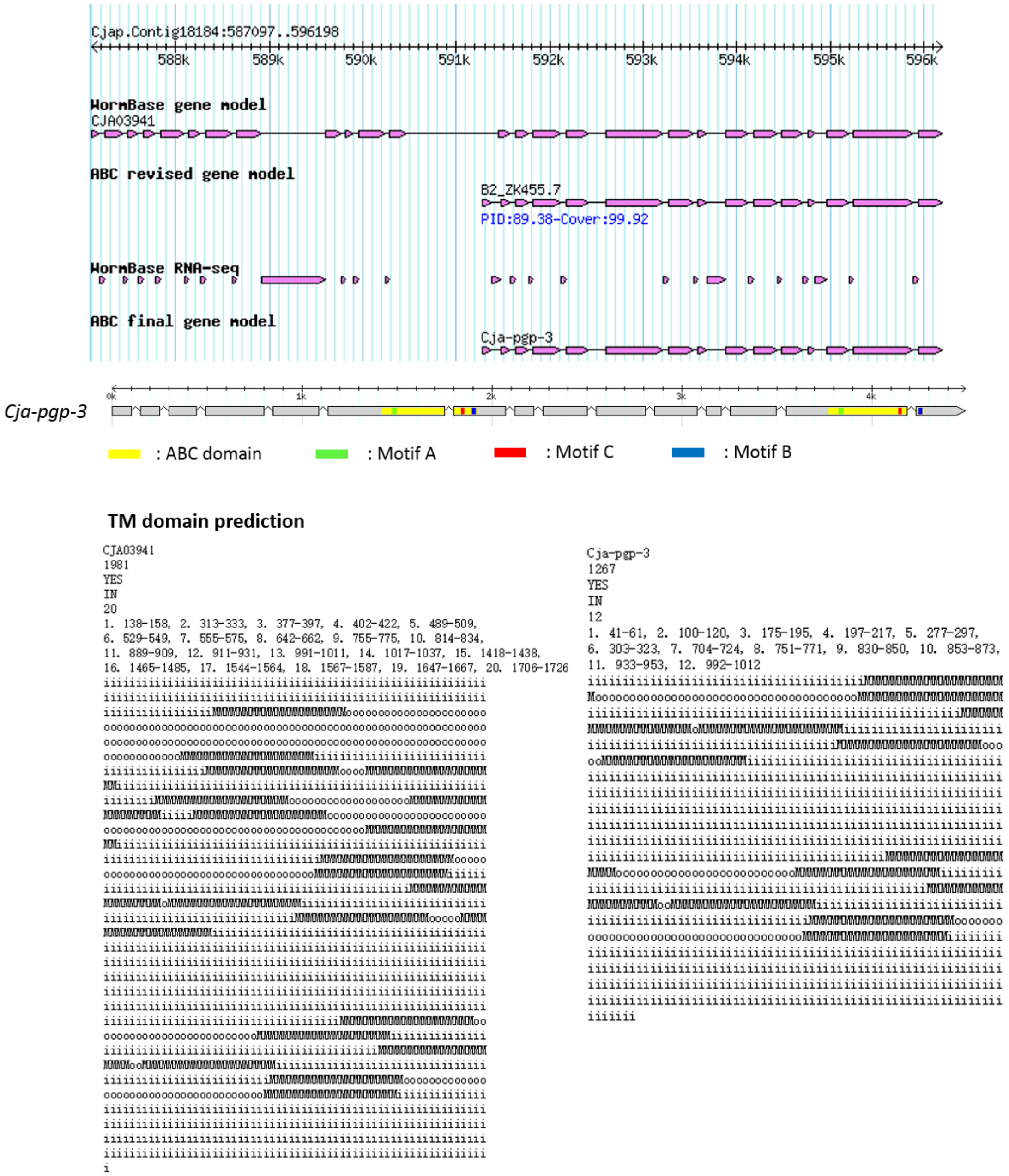


Figure 3.23: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "WormBase RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. CJA14269 and CJA18437 both were annotated as an ABC transporter gene in ABCF subfamily but each of them encoded only one predicted ABC domain. After improvement, a high-quality ABC transporter gene model with RNA-seq support was obtained as a result of merging CJA14269 and CJA18437 together.



TM domain prediction

```
CJA03941
1981
YES
IN
20
1. 138-158, 2. 313-333, 3. 377-397, 4. 402-422, 5. 489-509,
6. 529-549, 7. 555-575, 8. 642-662, 9. 755-775, 10. 814-834,
11. 889-909, 12. 911-931, 13. 991-1011, 14. 1017-1037, 15. 1418-1438,
16. 1465-1485, 17. 1544-1564, 18. 1567-1587, 19. 1647-1667, 20. 1706-1726
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooMMMMMMMMMMMMMMMMMMM
MMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooMMMMMMMMMMMM
MMMMMMMMMiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMM
MMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooo
oooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMM
MMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooMMMM
MMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoo
ooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMM
MMMMooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooo
oooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
i
```

```
Cja-pgp-3
1267
YES
IN
12
1. 41-61, 2. 100-120, 3. 175-195, 4. 197-217, 5. 277-297,
6. 303-323, 7. 704-724, 8. 751-771, 9. 830-850, 10. 853-873,
11. 933-953, 12. 992-1012
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMM
MooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMM
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMM
MMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooo
ooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMM
MMMMooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMM
MMMMMMMMMMooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooo
oooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiii
```

Figure 3.24: A representative case that the TM domain of an ABC transporter gene candidate was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "WormBase RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The original gene model of CJA03941 was annotated as a full ABC transporter gene in subfamily B and encoded 20 TM helices (clustering in two groups), which was more than expected. Through improvement, we generated a new gene model which encoded typically 12 TM helices.

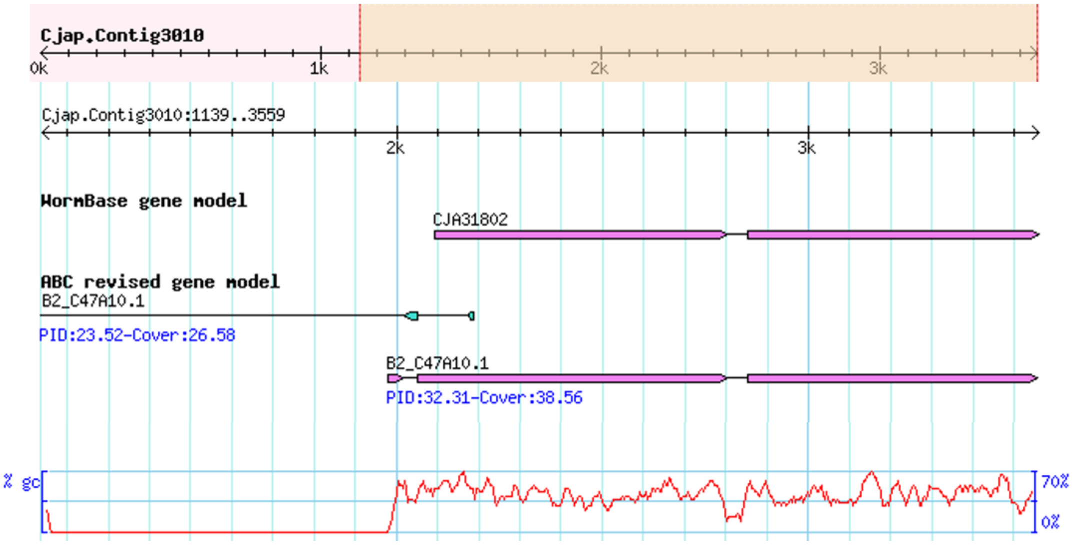


Figure 3.25: A representative case that sequencing error could result in incompleteness of an ABC transporter gene candidate

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. The incomplete CJA31802 which only encoded one high-quality ABC domain could be caused by the sequencing gap in this region.

Through phylogenetic analysis, we found 39 out of 46 ABC transporter genes in *C. japonica* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned gene names for ABC transporter genes in *C. japonica* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.26). ABC transporter genes were generally conserved between these two species. Considering that the total number of ABC transporter genes in *C. japonica* is obviously smaller than that in *C. elegans*, there might be some ABC transporter genes that were truly lost or failed to be annotated due to sequecing errors or assembly errors in *C. japonica.*

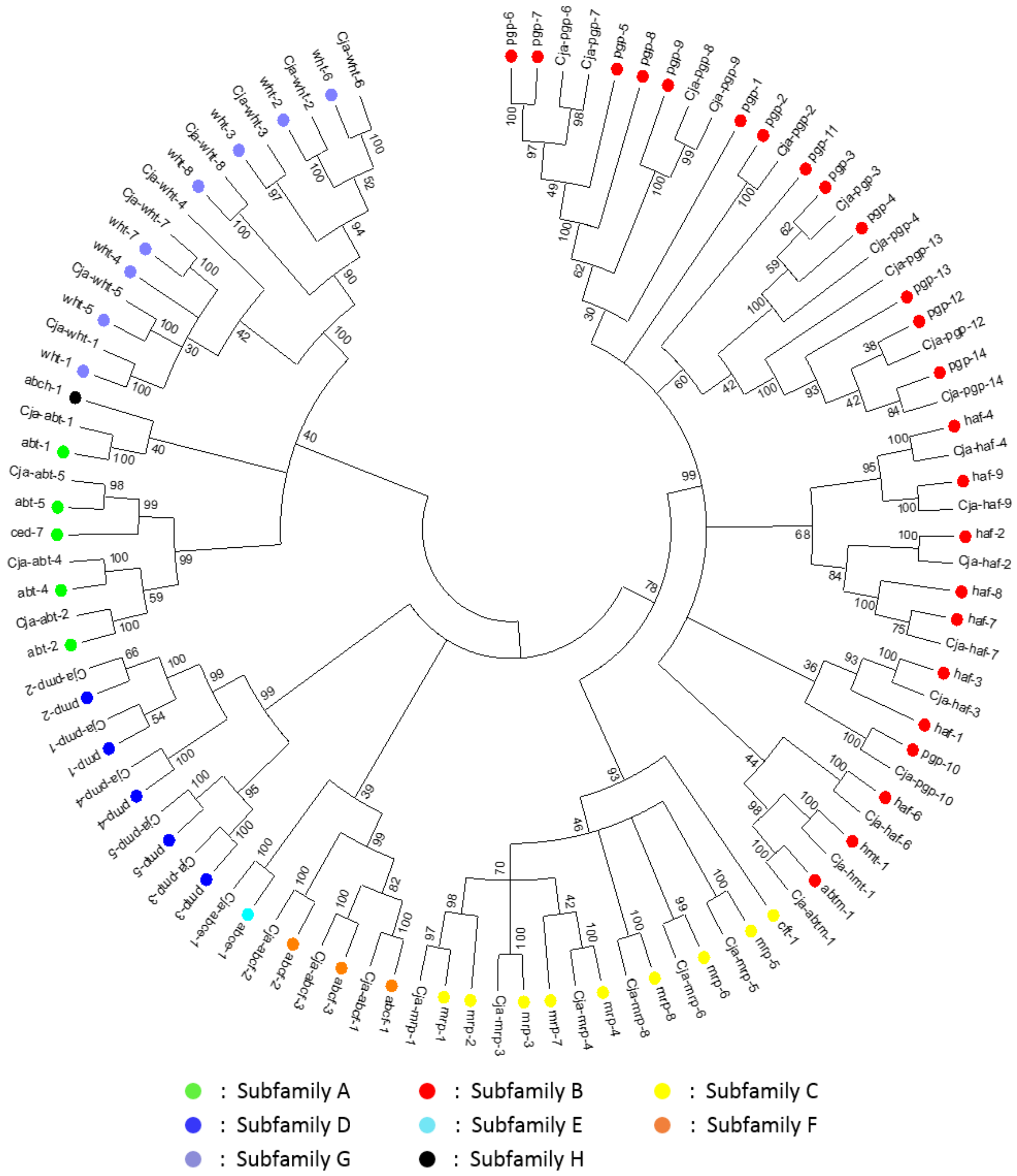


Figure 3.26: Phylogenetic analysis between *C. japonica* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *C. japonica* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *C. japonica* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.9. Annotation of ABC transporter genes in *C. angaria*

C. angaria (ex-species 3) is part of the *Drosophilae* super-group of *Caenorhabditis* species*,* with quite distinct morphology and behavior compared to *C. elegans* (Mortazavi et al. 2010)*.* After applying the annotation pipeline to *C. angaria*, we identified 279 ABC transporter gene candidates (219 candidates from InterProScan searches, 62 additional ones from BLAST searches), which was a much large number compared to those identified in other *Caenorhabditis* species. According to contamination filtering process, we found 136 candidates were due to bacteria contamination. This result is consistent with previous study, demonstrating that DNA of apparently recent bacterial origin was found in the genomic sequences of *C. angaria* (Percudani 2013)*.* After excluding the contamination and checking the quality of candidates, we found most of them were defective, only 11 were high-quality ABC transporter genes. For the 132 defective candidates, we tried to improve each of their gene models and examined the quality of newly constructed gene models. 24 improved gene models with high-quality were generated, six of which with only TM domain improved (Table 3.9). One of the improved gene models, Cang_2012_03_13_00228.g7427, obtained from BLAST searches, was annotated as a half ABC transporter gene in subfamily G but encoded no ABC domain. The revised gene model encoded one high-quality ABC domain (Figure 3.27), making it a high-quality ABC transporter gene. Another example is for merging case: Cang_2012_03_13_00071.g3374 and Cang_2012_03_13_00071.g3375, both encoded only one high-quality ABC domain and were annotated as a full ABC transporter gene in subfamily C. The revised gene model was a result of merging the above two candidates and encoded two typical ABC domains (Figure 3.28). TM domain improvement occurred in Cang_2012_03_13_00140.g5425, a full ABC transporter gene in subfamily B. After revision, the new gene model encoded typically 12 TM helices, three more compared to that of original gene model (Figure 3.29). Surprisingly, 97 candidates could be further improved to be high-quality ABC transporter genes, which reflected that the current genome was poorly assembled. 23 of these defective candidates that had sequencing errors or assembly errors could be complete ABC transporter genes when genome is fully sequenced and assembled. For example, two fragments of the ortholog of *abce-1* showed similarity to the different parts of *abce-1* (Figure 3.30)*.* However, one of them (Cang_2012_03_13_06705.g20433) was in a small contig with only one gene and the other (Cang_2012_03_13_00119.g4875) had some sequencing gaps within its genomic region (Figure 3.30). It suggests that these technique issues could be the cause of defect. Taking together, we annotated only 37 high-quality ABC transporter genes in *C. angaria,* 32 of which had proper TM domain (s) (Table 3.9).

Table 3.9: High-quality ABC transporter genes in *C. angaria* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Can-abt-2*	Cang_2012_03_13_00010.g803.t1	Cang_2012_03_13_00010.g803.t1	5TM-ABC-6TM-ABC	11554	19	1300	Exons were improved
	Can-abt-3	Cang_2012_03_13_00297.g8719.t1	Cang_2012_03_13_00297.g8719.t1	8TM-ABC-5TM-ABC	7845	19	1463	Cang_2012_03_13_00297.g8720 was merged with Cang_2012_03_13_00297.g8719
	Can-abt-4	Cang_2012_03_13_00812.g14307.t3	Cang_2012_03_13_00812.g14307.t3	4TM-ABC-7TM-ABC	5878	10	1599	
	Can-abt-5	Cang_2012_03_13_00567.g12314.t1	Cang_2012_03_13_00567.g12314.t1	8TM-ABC-6TM-ABC	7332	18	1544	Cang_2012_03_13_00567.g12315and Cang_2012_03_13_00567.g12316 were merged with Cang_2012_03_13_00567.g12314
	Can-abt-6	Cang_2012_03_13_00046.g2453.t1	Cang_2012_03_13_00046.g2453.t1	5TM-ABC-5TM-ABC	6150	11	1199	Exons were improved
B	*Can-abtm-1*	Cang_2012_03_13_00482.g11360.t1	Cang_2012_03_13_00482.g11360.t1	7TM-ABC	7462	11	604	Exons were improved; No start codon
	Can-haf-1	Cang_2012_03_13_00004.g357.t1	Cang_2012_03_13_00004.g357.t1	5TM-ABC	2154	5	651	
	Can-haf-2	Cang_2012_03_13_00115.g4774.t1	Cang_2012_03_13_00115.g4774.t1	9TM-ABC	3241	9	762	Exons were improved; No start codon
	Can-haf-3	Cang_2012_03_13_00696.g13506.t1	Cang_2012_03_13_00696.g13506.t1	4TM-ABC	3227	7	571	
	Can-haf-4	Cang_2012_03_13_08944.g21792.t1	Cang_2012_03_13_08944.g21792.t1	0TM-ABC	602	1	200	No stop codon
	Can-haf-6	Cang_2012_03_13_00493.g11502.t1	Cang_2012_03_13_00493.g11502.t1	7TM-ABC	6751	7	692	
	Can-haf-9	Cang_2012_03_13_00204.g6953.t2	Cang_2012_03_13_00204.g6953.t2	8TM-ABC	7171	10	753	Cang_2012_03_13_00204.g6954 was merged with Cang_2012_03_13_00204.g6953; TM helices were improved
	Can-hmt-1	Cang_2012_03_13_00036.g2043.t1	Cang_2012_03_13_00036.g2043.t1	11TM-ABC	7072	11	788	Exons were improved
	Can-pgp-10	Cang_2012_03_13_00086.g3903.t1	Cang_2012_03_13_00086.g3903.t1	6TM-ABC-7TM-ABC	6594	21	1370	Exons were improved; No start codon
	Can-pgp-11	Cang_2012_03_13_00140.g5425.t1	Cang_2012_03_13_00140.g5425.t1	5TM-ABC-5TM-ABC	5171	17	1258	Exons were improved; TM helices were improved; No start codon
	Can-pgp-12	Cang_2012_03_13_01165.g15866.t1	Cang_2012_03_13_01165.g15866.t1	7TM-ABC-4TM-ABC	10739	12	1089	Exons were improved
	Can-pgp-2	Cang_2012_03_13_00635.g13012.t2	Cang_2012_03_13_00635.g13012.t2	4TM-ABC-6TM-ABC	17021	18	1126	
	Can-pgp-3	Cang_2012_03_13_00056.g2875.t1	Cang_2012_03_13_00056.g2875.t1	5TM-ABC-5TM-ABC	10160	14	1220	Exons were improved; No start codon
	Can-pgp-4	Cang_2012_03_13_00158.g5875.t1	Cang_2012_03_13_00158.g5875.t1	6TM-ABC-6TM-ABC	4778	12	1265	
	Can-pgp-5	Cang_2012_03_13_00373.g9936.t2	Cang_2012_03_13_00373.g9936.t2	6TM-ABC-6TM-ABC	5005	13	1239	Exons were improved; TM helices were improved;
C	*Can-mrp-1*	Cang_2012_03_13_00071.g3374.t1	Cang_2012_03_13_00071.g3374.t1	11TM-ABC-5TM-ABC	7697	14	1530	Cang_2012_03_13_00071.g3375 was merged with Cang_2012_03_13_00071.g3374
	Can-mrp-2	Cang_2012_03_13_00021.g1308.t2	Cang_2012_03_13_00021.g1308.t2	1TM-ABC-7TM-ABC	12062	9	1154	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
C	*Can-mrp-4*	Cang_2012_03_13_01287.g16169.t2	Cang_2012_03_13_01287.g16169.t2	7TM-ABC-5TM-ABC	4832	11	1445	No start codon
	Can-mrp-5	Cang_2012_03_13_00003.g282.t1	Cang_2012_03_13_00003.g282.t1	8TM-ABC-5TM-ABC	5818	18	1287	Exons were improved
	Can-mrp-8	Cang_2012_03_13_00278.g8362.t1	Cang_2012_03_13_00278.g8362.t1	7TM-ABC-6TM-ABC	16459	24	1316	Cang_2012_03_13_00278.g8364 was merged with Cang_2012_03_13_00278.g8362; No start codon
D	*Can-pmp-1*	Cang_2012_03_13_00128.g5121.t2	Cang_2012_03_13_00128.g5121.t2	4TM-ABC	3413	8	647	Exons were improved
	Can-pmp-2	Cang_2012_03_13_00128.g5123.t1	Cang_2012_03_13_00128.g5123.t1	5TM-ABC	4055	10	673	Cang_2012_03_13_00128.g5122 was merged with Cang_2012_03_13_00128.g5123
	Can-pmp-3	Cang_2012_03_13_00477.g11302.t2	Cang_2012_03_13_00477.g11302.t2	6TM-ABC	5639	6	688	Exons were improved
	Can-pmp-4	Cang_2012_03_13_00515.g11743.t1	Cang_2012_03_13_00515.g11743.t1	6TM-ABC	3981	8	715	Exons were improved
	Can-pmp-5	Cang_2012_03_13_00533.g11943.t1	Cang_2012_03_13_00533.g11943.t1	0TM-ABC	2041	3	271	
F	*Can-abcf-1*	Cang_2012_03_13_00427.g10727.t1	Cang_2012_03_13_00427.g10727.t1	ABC-ABC	2508	8	664	
G	*Can-wht-1*	Cang_2012_03_13_00218.g7257.t1	Cang_2012_03_13_00218.g7257.t1	ABC-0TM	1859	4	329	
	Can-wht-2	Cang_2012_03_13_00229.g7438.t1	Cang_2012_03_13_00229.g7438.t1	ABC-5TM	3611	7	589	
	Can-wht-3	Cang_2012_03_13_00620.g12870.t1	Cang_2012_03_13_00620.g12870.t1	ABC-6TM	2494	5	567	Cang_2012_03_13_00620.g12869 was merged with Cang_2012_03_13_00620.g12870; TM helices were improved
	Can-wht-4	Cang_2012_03_13_00546.g12078.t2	Cang_2012_03_13_00546.g12078.t2	ABC-4TM	4243	7	569	Cang_2012_03_13_00546.g12079 was merged with Cang_2012_03_13_00546.g12078; TM helices were improved; No start codon
	Can-wht-5	Cang_2012_03_13_00546.g12080.t1	Cang_2012_03_13_00546.g12080.t1	ABC-4TM	3213	7	517	Exons were improved; TM helices were improved; No start codon
	Can-wht-6	Cang_2012_03_13_00228.g7427.t2	Cang_2012_03_13_00228.g7427.t2	ABC-6TM	5303	9	603	Exons were improved

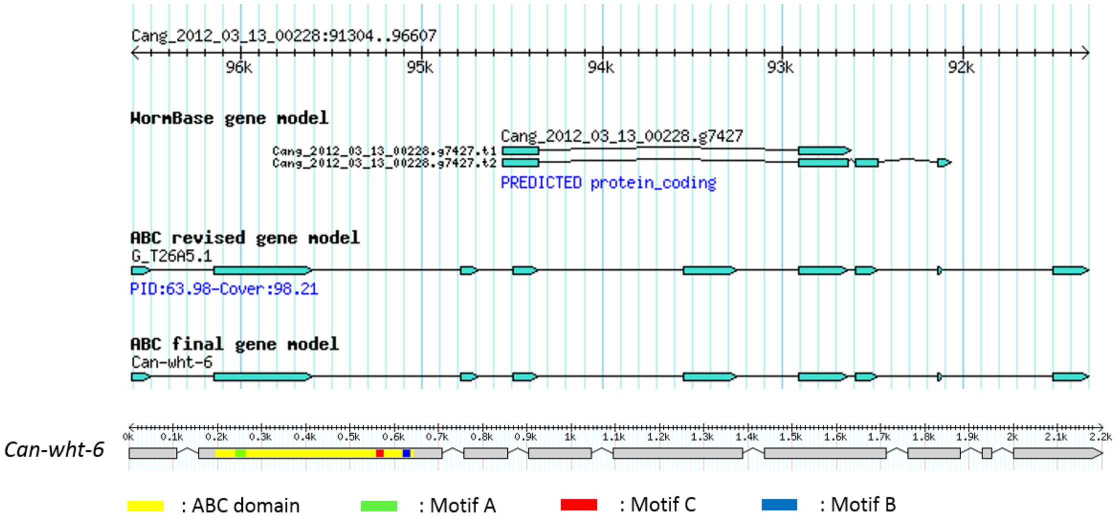


Figure 3.27: A representative case that the exons of one defective ABC transporter gene were improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Cang_2012_03_13_00228.g7427 was annotated as a half ABC transporter gene in subfamily G but encoded no ABC domain. Through genBlastG improvement, the revised gene model encoded one high-quality ABC domain.

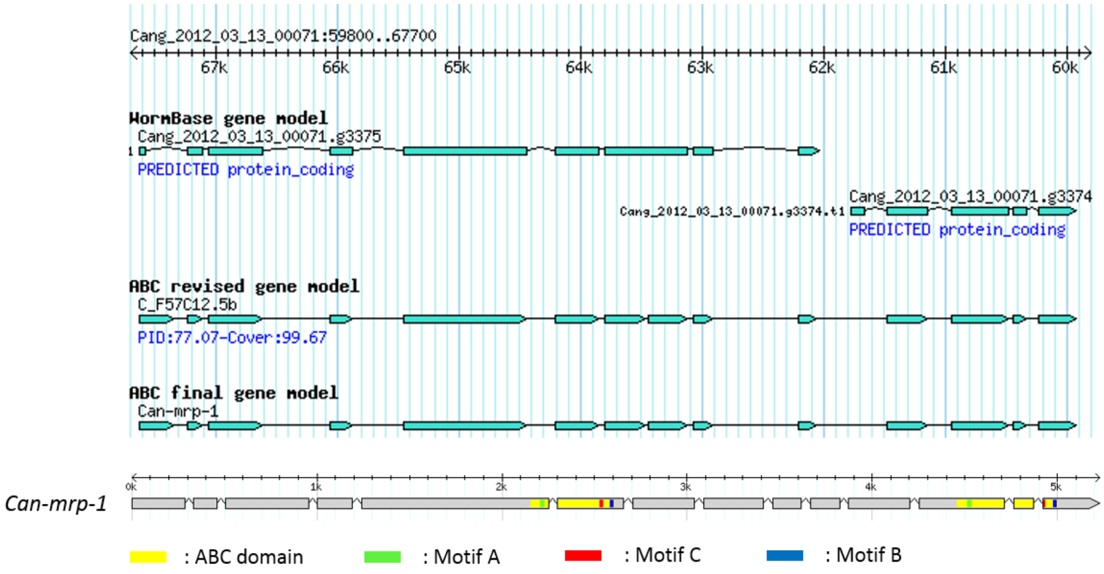
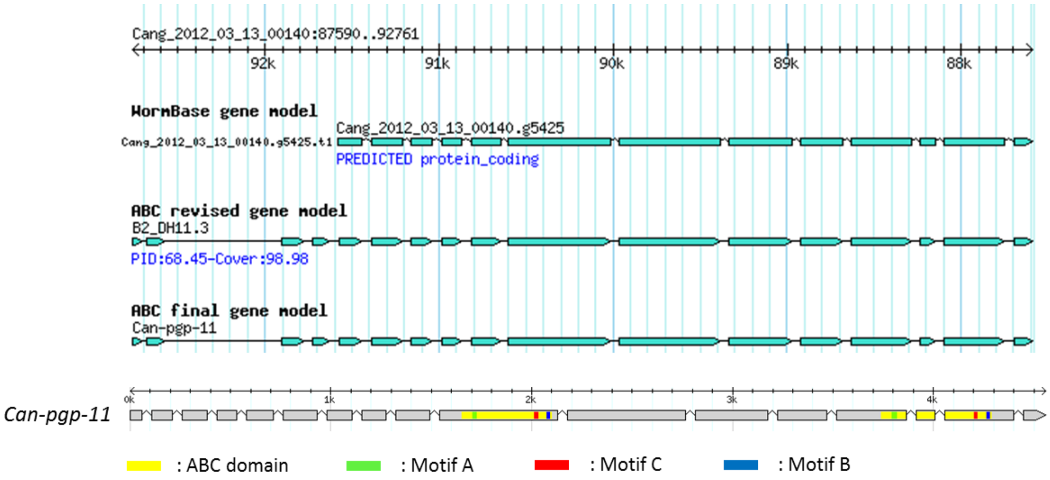


Figure 3.28: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Cang_2012_03_13_00071.g3374 and Cang_2012_03_13_00071.g3375, both encoded only one high-quality ABC domain and were annotated as a full ABC transporter gene in subfamily C. The revised gene model was a result of merging the above two candidates and encoded two high-quality ABC domains.



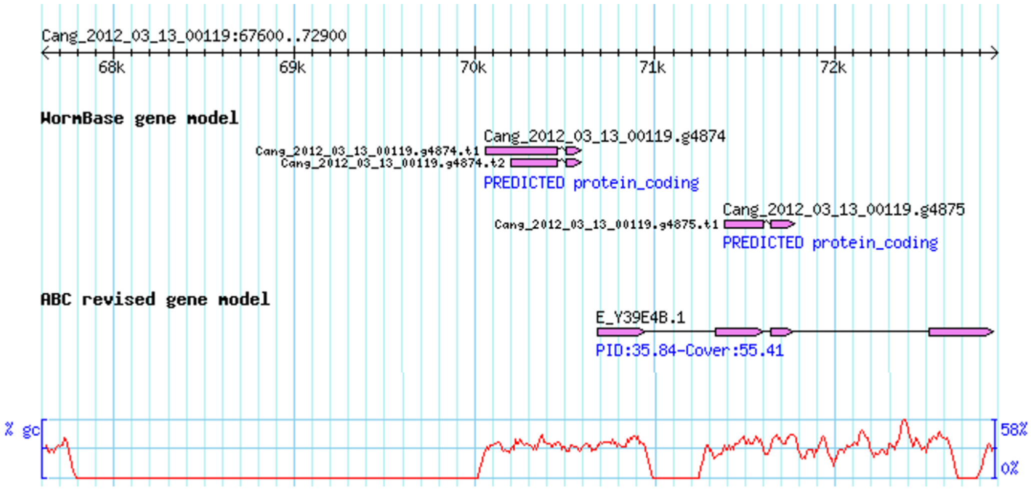
TM domain prediction

```
Cang_2012_03_13_00140.g5425.t1
1132
YES
IN
9
1. 60-80, 2. 82-102, 3. 172-192, 4. 199-219, 5. 573-593,
6. 621-641, 7. 696-716, 8. 719-739, 9. 809-829
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiM
MMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMM
MMMMMMMMMMMMooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooo
ooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooMM
MMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooo
```

```
Can-pgp-11
1257
YES
OUT
12
1. 19-39, 2. 59-79, 3. 112-132, 4. 185-205, 5. 207-227,
6. 297-317, 7. 324-344, 8. 698-718, 9. 746-766, 10. 821-841,
11. 844-864, 12. 934-954
ooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiMM
MMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooooooooMMMMMMMMM
MMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMM
MMMMMMMMMMMMMMMMMooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoo
oooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMM
MooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
```

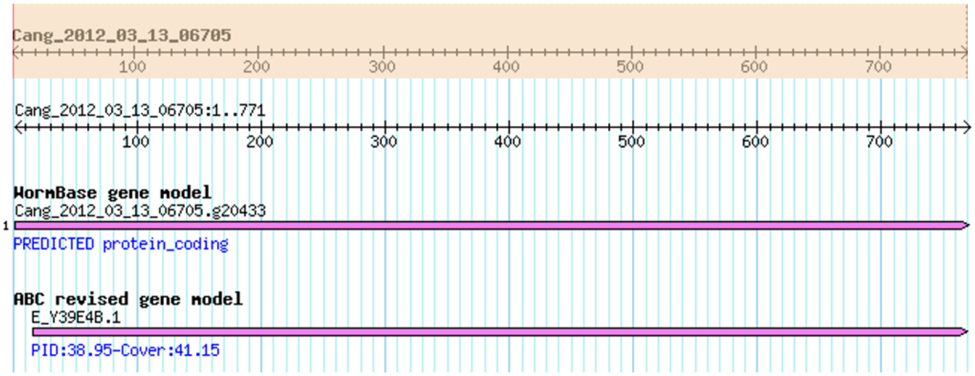
Figure 3.29: A representative case that the TM domain of an ABC transporter gene candidate was improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Cang_2012_03_13_00140.g5425 was annotated as a full ABC transporter gene in subfamily B with a slightly lower number of TM helices. After improvement, the revised gene model of encoded typically 12 TM helices, three more compared to that of original gene model.


Cang_2012_03_13_00119:67600..72900
68k
69k
70k
71k
72k
WormBase gene model
Cang_2012_03_13_00119.g4874
Cang_2012_03_13_00119.g4874.t1
Cang_2012_03_13_00119.g4874.t2
PREDICTED protein_coding
Cang_2012_03_13_00119.g4875
Cang_2012_03_13_00119.g4875.t1
PREDICTED protein_coding
ABC revised gene model
E_Y39E4B.1
PID:35.84-Cover:55.41
% gc
58%
0%

genBlastG result

```
E_Y39E4B.1|Cang_2012_03_13_00119:70670..72651|+|gene cover:265(43.4426%)|score:14.1635|rank:1
HSP_ID[4]:(72478-72651);query:(545-602); pid: 75.8621
HSP_ID[5]:(71632-71769);query:(512-557); pid: 89.1304
HSP_ID[3]:(71357-71599);query:(436-516); pid: 93.8272
HSP_ID[2]:(70670-70963);query:(335-432); pid: 91.8367
```



genBlastG result

```
E_Y39E4B.1|Cang_2012_03_13_06705:25..768|+|gene cover:248(40.6557%)|score:-67.9281|rank:2
HSP_ID[12]:(25-768);query:(12-259); pid: 94.7581
```

Figure 3.30: A representative case that technical issues could result in incompleteness of an ABC transporter gene candidate

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and C. elegans orthologs as query proteins. Cang_2012_03_13_06705.g20433 and Cang_2012_03_13_00119.g4875 showed similarity to the different parts of *abce-1* based on genBlastG result. However, Cang_2012_03_13_06705.g20433 was in a small contig with only one gene and Cang_2012_03_13_00119.g4875 had some sequencing gaps within its genomic region, suggesting that these technique issues could be the cause of defect.

Through phylogenetic analysis, we found 22 out of 37 ABC transporter genes in *C. angaria* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *C. angaria* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.31). The biggest difference between *C. angaria* and *C. elegans* is the contraction of subfamily B in *C. angaria*. For example, there is only one *C. angaria* gene in the cluster that contain *pgp-5.* However, this contraction could be validated when genomic region of the potential candidates is fully sequenced and assembled. In general, the small number of one-to-one orthologous relationship was consistent with relatively distant evolutionary relationship between *C. angaria* and *C. elegans*.

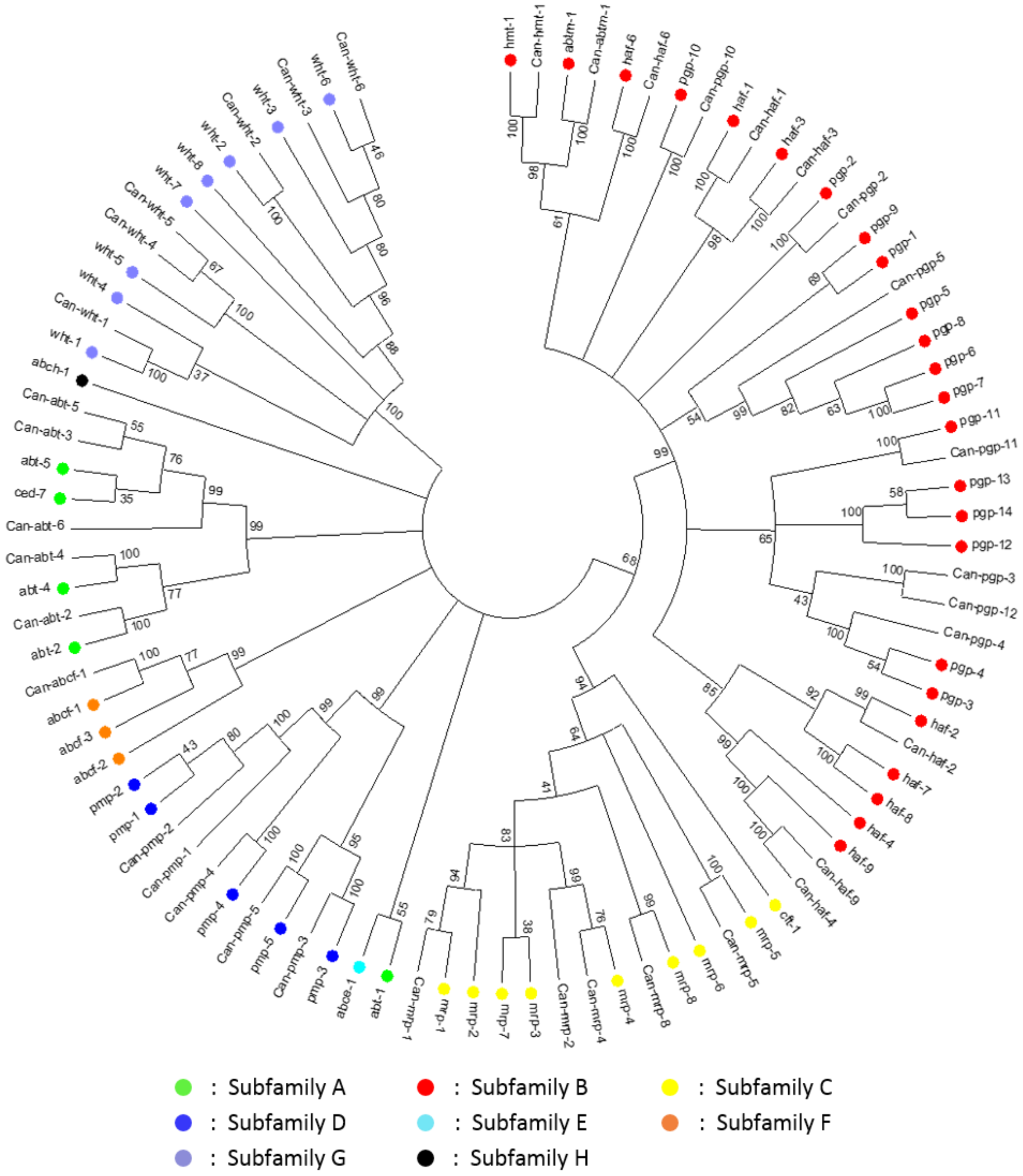


Figure 3.31: Phylogenetic analysis between *C. angaria* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *C. angaria* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *C. angaria* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.10. Annotation of ABC transporter genes in *P. pacificus*

P. pacificus is a necromenic nematode, specifically associated with several species of phytophagous beetles around the globe. Free living *P. pacificus* populations can also be found in the soil and maintained on strict bacterial diets in the laboratory (Kroetz et al. 2012). The hermaphroditic nematode *P. pacificus* is an established model system for comparative studies with *C. elegans* in developmental biology, ecology, and population genetics (Dieterich et al. 2008). After applying the annotation pipeline to *P. pacificus*, we identified 83 ABC transporter gene candidates from InterProScan searches, 22 additional ones from BLAST searches, totally 105 candidates. None of these candidates was due to contamination. After quality checking process, only 16 candidates were high-quality ABC transporter genes. All of these 16 genes also encoded appropriate TM domain (s). For the 89 defective candidates, we tried to improve them and we ended up with 35 revised gene models of high-quality, 10 of which with only TM domain improved (Table 3.10). For example, PPA28101 was annotated as a full ABC transporter gene in subfamily B but the original gene model only encoded one high-quality ABC domain. After improvement, the new gene model had extended exons, making it to be a high-quality ABC transporter gene encoding two typical ABC domains (Figure 3.32). PPA32860 and PPA29777 both were annotated as a full ABC transporter gene in subfamily B but each of them encoded only a single ABC domain. After improvement, we generated a high-quality ABC transporter gene model which was the outcome of merging these two candidates together (Figure 3.33). Another example represents TM domain improvement. PPA25170 encoded TM domain and its adjacent gene PPA25171 encoded a high-quality ABC domain. Through improvement, these two genes together formed an half ABC transporter gene in subfamily D, which encoded both TM domain and high-quality ABC domain (Figure 3.34). 31 candidates that could not be further improved to be high-quality ABC transporter genes, 10 could be complete ABC transporter genes when the genome assembly improves. For instance, although the improved gene model (merging three candidate genes: PPA07651, PPA07652 and PPA07657) had two ABC domain satisfying our criteria, there were two sequencing gaps with this region, making this new model much longer than real ABC transporter genes. Therefore, there could be more than one ABC transporter genes in this region when the genome is fully assembled (Figure 3.35). In summary, we annotated 55 high-quality ABC transporter genes in *P. pacificus,* 34 of which had proper TM domain (s) (Table 3.10).

In previous studies, initial analysis of the *P. pacificus* genome found a larger number of putative ABC transporter genes (129) (Dieterich et al. 2008). Later on, a much smaller number of putative ABC transporter gene (65) were reported by the same group (Markov et al. 2015). However, they did not explain the reason for such difference and did not provide subfamily information for these ABC transporter genes. In our analysis, we evaluated all ABC transporter gene candidates and ended up with 55 high-quality ABC transporter genes. 18 of these high-quality ABC transporter genes were obtained from merging at least two adjacent candidates, which could explain the reason why we identify a smaller number of high-quality ABC transporter genes.

Table 3.10: High-quality ABC transporter genes in *P. pacificus* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Ppa-abt-1*	PPA01242	PPA01242	4TM-ABC-4TM-ABC	8663	38	1118	
	Ppa-abt-2	PPA20763	PPA20763	7TM-ABC-7TM-ABC	15889	41	1367	PPA20762 was merged with PPA20763; TM helices were improved; No start codon
	Ppa-abt-3	PPA10514	PPA10514	7TM-ABC-8TM-ABC	9849	50	1465	PPA10515 was merged with PPA10514
	Ppa-abt-4	PPA04003	PPA04003	8TM-ABC-9TM-ABC	25749	51	1539	PPA04000 was merged with PPA04003; TM helices were improved
	Ppa-abt-5	PPA00756	PPA00756	4TM-ABC-7TM-ABC	35408	41	1314	
B	*Ppa-haf-1*	PPA14180	PPA14180	3TM-ABC	2951	16	532	
	Ppa-haf-2	PPA06384	PPA06384	8TM-ABC	4387	23	749	
	Ppa-haf-3	PPA26513	PPA26513	6TM-ABC	14140	21	606	TM helices were improved; No start codon
	Ppa-haf-4	PPA16516	PPA16516	3TM-ABC	3121	16	570	
	Ppa-haf-6	PPA06443	PPA06443	4TM-ABC	5956	13	416	
	Ppa-haf-9	PPA00989	PPA00989	9TM-ABC	7903	33	854	
	Ppa-hmt-1	PPA14422	PPA14422	10TM-ABC	25377	23	789	PPA14427 was merged with PPA14422; TM helices were improved
	Ppa-pgp-1	PPA17189	PPA17189	6TM-ABC-4TM-ABC	8817	40	1274	PPA17188 was merged with PPA17189; TM helices were improved; No start codon
	Ppa-pgp-10	PPA23730	PPA23730	6TM-ABC-6TM-ABC	9443	37	1239	PPA23731 was merged with PPA23730
	Ppa-pgp-11	PPA22128	PPA22128	0TM-ABC-2TM-ABC	5933	15	618	PPA22129 was merged with PPA22128; No start codon
	Ppa-pgp-12	PPA21538	PPA21538	6TM-ABC-5TM-ABC	7656	34	1186	
	Ppa-pgp-13	PPA28101	PPA28101	6TM-ABC-6TM-ABC	11557	40	1301	Exons were improved
	Ppa-pgp-14	PPA25211	PPA25211	5TM-ABC-4TM-ABC	6216	34	1124	PPA29777 was merged with PPA25211; No start codon
	Ppa-pgp-15	PPA19458	PPA19458	6TM-ABC-8TM-ABC	12258	41	1209	Exons were improved; No start codon
	Ppa-pgp-16	PPA03557	PPA03557	7TM-ABC-6TM-ABC	21373	39	1145	TM helices were improved; No start codon
	Ppa-pgp-17	PPA21537	PPA21537	0TM-ABC-5TM-ABC	5781	30	1017	
	Ppa-pgp-18	PPA09633	PPA09633	0TM-ABC-0TM-ABC	4156	13	557	PPA04233 was merged with PPA09633; No start codon
	Ppa-pgp-19	PPA05777	PPA05777	0TM-ABC-2TM-ABC	4598	14	722	Exons were improved; No start codon
	Ppa-pgp-2	PPA04690	PPA04690	6TM-ABC-6TM-ABC	10764	44	1171	Exons were improved; No start codon
	Ppa-pgp-20	PPA16243	PPA16243	0TM-ABC-3TM-ABC	6665	21	628	Exons were improved
	Ppa-pgp-21	PPA17954	PPA17954	6TM-ABC-4TM-ABC	5797	29	1068	Exons were improved; No start codon
	Ppa-pgp-22	PPA25898	PPA25898	6TM-ABC-2TM-ABC	7847	26	1000	PPA25900 was merged with PPA25898; TM helices were improved
	Ppa-pgp-23	PPA15485	PPA15485	6TM-ABC-6TM-ABC	10720	34	1270	
	Ppa-pgp-24	PPA24275	PPA24275	4TM-ABC-4TM-ABC	5792	23	1072	Exons were improved

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
B	*Ppa-pgp-3*	PPA01136	PPA01136	6TM-ABC-3TM-ABC	7803	31	992	
	Ppa-pgp-4	PPA01137	PPA01137	3TM-ABC-0TM-ABC	8858	16	640	PPA06272 was merged with PPA01137; No start codon
	Ppa-pgp-5	PPA01140	PPA01140	3TM-ABC-0TM-ABC	13141	21	801	Exons were improved; No start codon
	Ppa-pgp-6	PPA02842	PPA02842	2TM-ABC-6TM-ABC	10069	34	1109	
	Ppa-pgp-7	PPA24230	PPA24230	2TM-ABC-1TM-ABC	5033	14	561	PPA32738 and PPA32114 were merged with PPA24230; No start codon
	Ppa-pgp-8	PPA08573	PPA08573	1TM-ABC-6TM-ABC	6406	34	1111	
	Ppa-pgp-9	PPA07555	PPA07555	6TM-ABC-6TM-ABC	10649	43	1280	
C	*Ppa-mrp-1*	PPA20573	PPA20573	8TM-ABC-5TM-ABC	11110	36	1373	PPA20574 was merged with PPA20573; No start codon
	Ppa-mrp-2	PPA07998	PPA07998	9TM-ABC-6TM-ABC	11185	50	1453	
	Ppa-mrp-3	PPA06907	PPA06907	5TM-ABC-4TM-ABC	11761	39	1279	PPA06910 was merged with PPA06907
	Ppa-mrp-4	PPA17668	PPA17668	10TM-ABC-7TM-ABC	23570	47	1488	
	Ppa-mrp-5	PPA20781	PPA20781	11TM-ABC-2TM-ABC	7281	36	1169	
	Ppa-mrp-6	PPA24297	PPA24297	11TM-ABC-6TM-ABC	7498	41	1470	TM helices were improved
	Ppa-mrp-7	PPA25269	PPA25269	10TM-ABC-5TM-ABC	8168	47	1438	Exons were improved
	Ppa-mrp-8	PPA16626	PPA16626	3TM-ABC-3TM-ABC	20953	33	907	PPA16627 was merged with PPA16626; No start codon
D	*Ppa-pmp-2*	PPA25171	PPA25171	3TM-ABC	3515	16	600	PPA25170 was merged with PPA25171; TM helices were improved; No start codon
	Ppa-pmp-4	PPA11598	PPA11598	6TM-ABC	8819	30	725	
	Ppa-pmp-5	PPA02112	PPA02112	5TM-ABC	3030	17	590	No stop codon
E	*Ppa-abce-1*	PPA10310	PPA10310	ABC-ABC	4937	20	611	Exons were improved
F	*Ppa-abcf-1*	PPA00336	PPA00336	ABC-ABC	7575	22	732	
	Ppa-abcf-2	PPA08123	PPA08123	ABC-ABC	10138	19	627	Exons were improved
G	*Ppa-wht-1*	PPA19948	PPA19948	ABC-7TM	3860	21	633	
	Ppa-wht-2	PPA06433	PPA06433	ABC-7TM	2999	18	613	Exons were improved
	Ppa-wht-4	PPA28021	PPA28021	ABC-5TM	2993	19	510	Exons were improved; No start codon
	Ppa-wht-7	PPA08267	PPA08267	ABC-6TM	14659	17	624	PPA08263 was merged with PPA08267; TM helices were improved
H	*Ppa-abch-1*	PPA18570	PPA18570	ABC-0TM	2381	11	364	Exons were improved

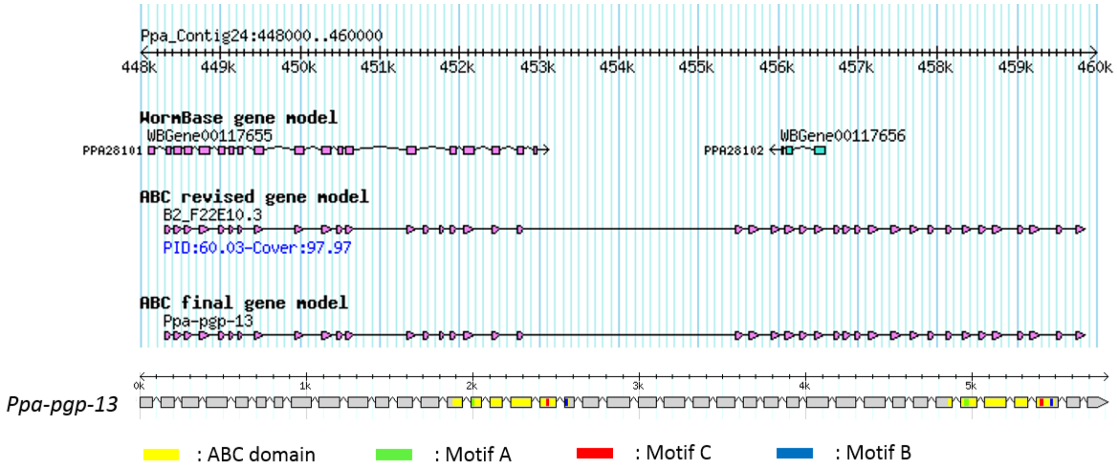


Figure 3.32: A representative case that the exons of one defective ABC transporter gene were improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. PPA28101 was annotated as a full ABC transporter gene in subfamily B but encoded only one high-quality ABC domain. After improvement, the revised gene model had extended exons, making it to be a high-quality ABC transporter gene encoding two typical ABC domains.

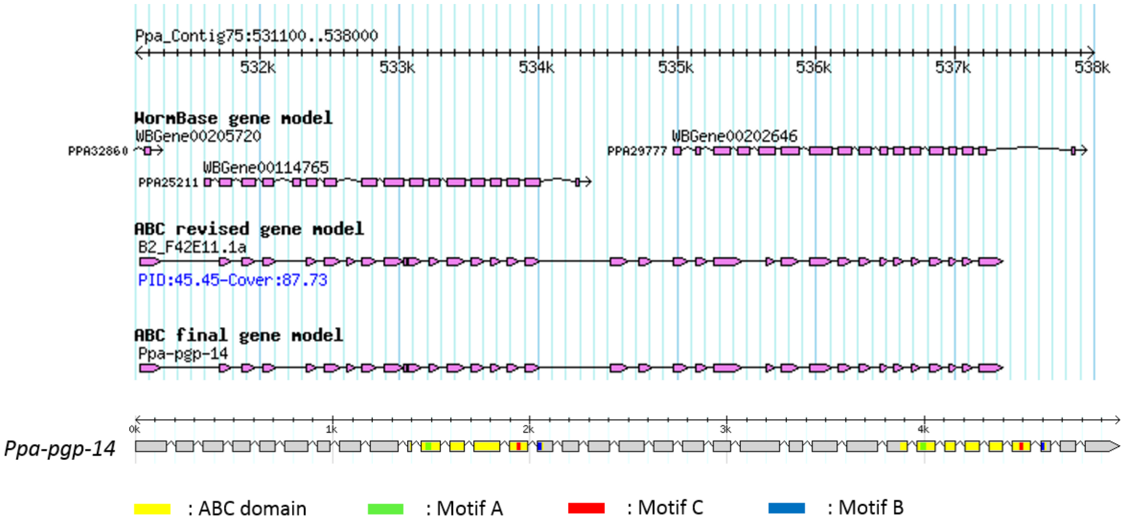
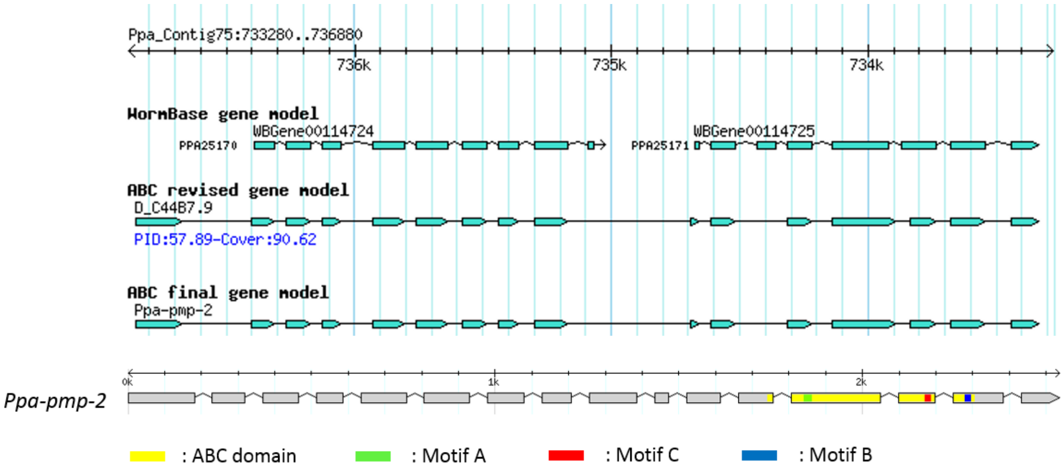
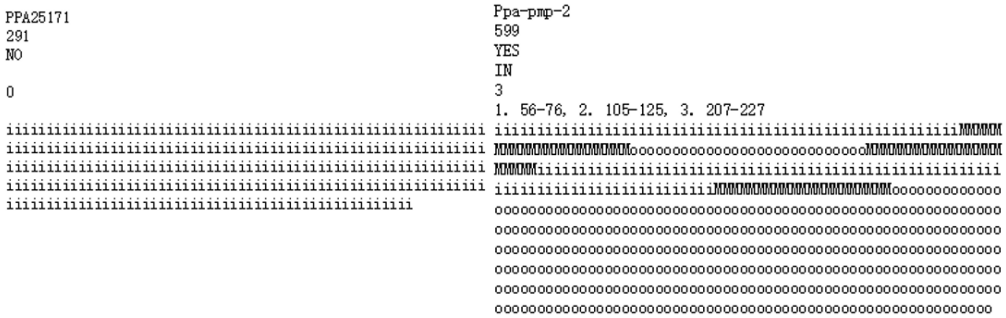


Figure 3.33: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. PPA32860 and PPA29777 both were two fragments of a full ABC transporter gene in subfamily B. The revised gene model, encoding two high-quality ABC domain, was the outcome of merging these two candidates together.



TM domain prediction



```
PPA25171
291
NO

0
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

```
Ppa-pmp-2
599
YES
IN
3
1. 56-76, 2. 105-125, 3. 207-227
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMM
MMMMMMMMMMMMMMMMooooooooooooooooooooooooooooMMMMMMMMMMMMMMMM
MMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
```

Figure 3.34: A representative case that the TM domain of an ABC transporter gene was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. PPA25170 encoded TM domain and its adjacent gene PPA25171 encoded a high-quality ABC domain. Through improvement, these two genes together formed a high-quality half ABC transporter gene in subfamily D.

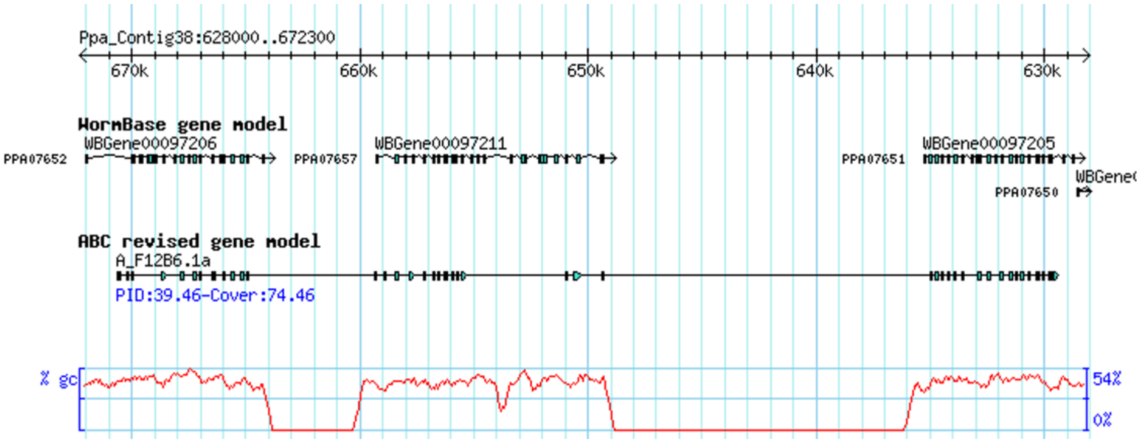


Figure 3.35: A representative case that sequencing errors could result in incompleteness of ABC transporter gene candidates

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. PPA07651, PPA07652 and PPA07657 were merged together and had two ABC domain satisfying our criteria. However, there were two sequencing gaps with this region made this new model much longer than real ABC transporter genes. Therefore, there could be more than one ABC transporter genes in this region when the genome is fully assembled.

Through phylogenetic analysis, we found only 19 out of 55 ABC transporter genes in *P. pacificus* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*, which reflected the distant evolutionary relationship between *P. pacificus* and *C. elegans.* We assigned the gene names for ABC transporter genes in *P. pacificus* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.36). These two species both showed species specific expansions, especially in subfamily A, B and C. For example, in subfamily C, *Ppa-mrp-2*, *Ppa-mrp-2*, *Ppa-mrp-2* and *Ppa-mrp-2* in *P. pacificus* were clustered together without any obvious ortholog in *C. elegans*. Similarly, *mrp-5*, *mrp-6, mrp-8* and *cft-1* became a group that did not contain any *P. pacificus* gene. Besides, we saw a contraction (4 members) of subfamily G in *P. pacificus* compared *C. elegans* (8 members), suggesting that some ABC transporter genes in subfamily G might not be essential for *P. pacificus* and thus, were lost after speciation. Interestingly, the ABC transporter gens in *P. pacificus* generally consisted of more but shorter exons, as well as longer gene model due to the longer introns compared to the ABC transporter genes in *Caenorhabditis* species. This proteome complexity may be related to the ecology of this organism (Dieterich et al. 2008).

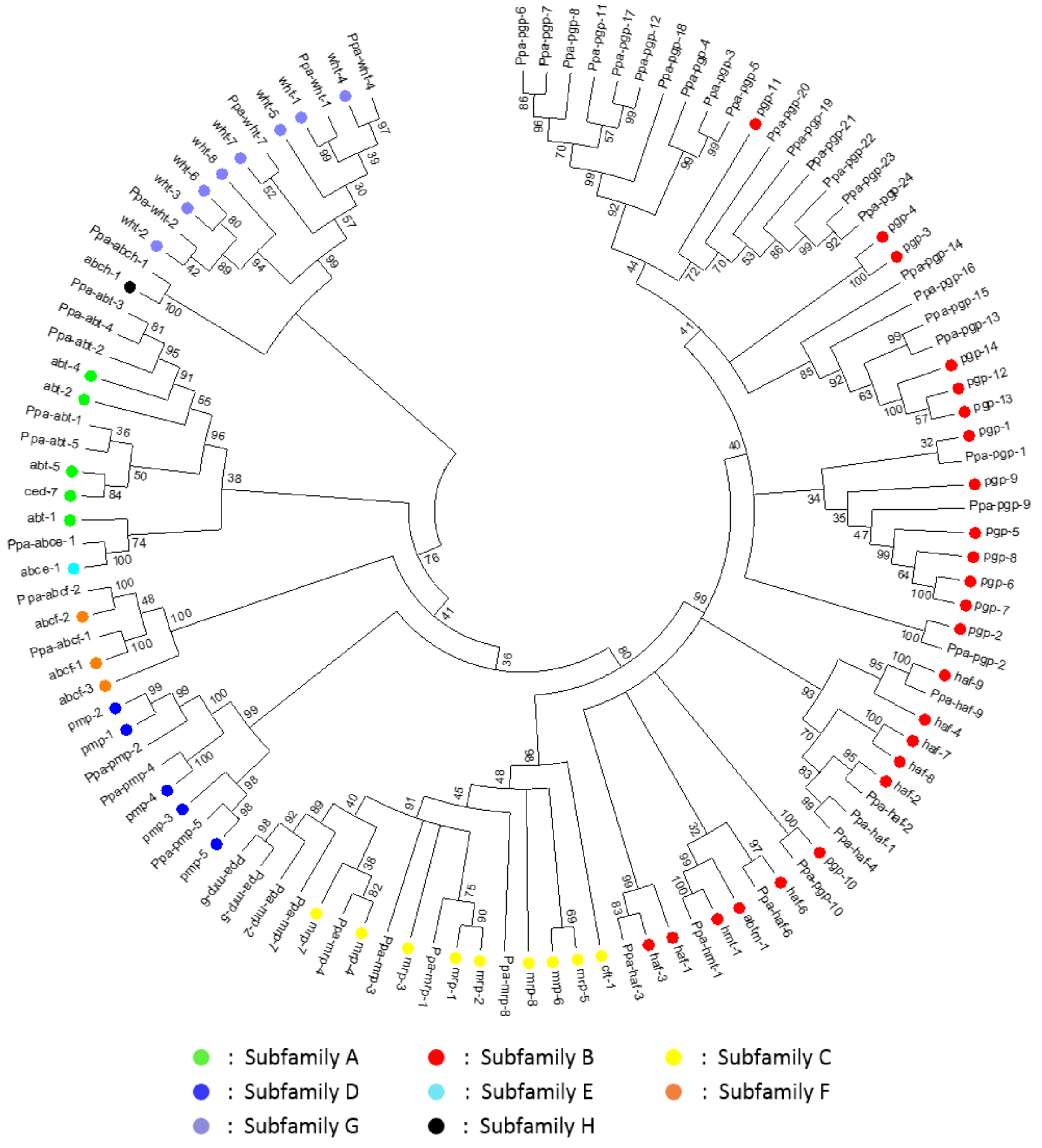


Figure 3.36: Phylogenetic analysis between *P. pacificus* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *P. pacificus* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *P. pacificus* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.11. Annotation of ABC transporter genes in *P. exspectatus*

P. exspectatus, a very closely related outcrossing sister species of *P. pacificus.* It has been isolated from stag beetles in Japan (Kanzaki et al. 2012). The close phylogenetic relationship between *P. pacificus* and *P. exspectatus* provides a powerful framework for studying genome evolution. After applying the annotation pipeline to *P. exspectatus*, we obtained 84 ABC transporter gene candidates (72 candidates from InterProScan searches, 18 additional ones from BLAST searches), none of which was due to contamination. After examining the quality of all the candidates, 27 were high-quality ABC transporter genes. All of these 27 genes also encoded appropriate TM domain (s). After trying to improve the defective candidates, we generated 28 revised gene models of high-quality, four of which with only TM domain improved (Table 3.11). Among these revised gene models, scaffold544-EXSNAP2012.9 was annotated as a half ABC transporter gene in subfamily D, did not encode any predicted ABC domain but showed some similarity to ABC transporter in *C. elegans*. Through the improvement procedure, we constructed a longer gene model which encoded a protein with a high-quality ABC domain (146 aa, 3.6E-17) (Figure 3.37). Another example is a merging case in which two candidates, scaffold505-EXSNAP2012.16 and scaffold505-EXSNAP2012.17, both were annotated as a full ABC transporter gene in subfamily B encoding a high-quality ABC domain in each. After merging, the revised gene model was a high-quality ABC transporter gene encoding a protein with two typical ABC domains (Figure 3.38). A representative case for TM domain improvement is scaffold170-EXSNAP2012.8 in subfamily G. Although this ABC transporter had a high-quality ABC domain, it did not contain any predicted TM helices. After improvement, the new gene model extended to the neighboring region and became a longer gene model with eight predicted TM helices (Figure 3.39). The remaining 16 candidates, could not be further improved. In conclusion, we annotated totally 62 high-quality ABC transporter genes in *P. exspectatus*, 47 of which had appropriate TM domain (s) (Table 3.11).

Table 3.11: High-quality ABC transporter genes in *P. exspectatus* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Pex-abt-1*	scaffold1304-EXSNAP2012.2	scaffold1304-EXSNAP2012.2	4TM-ABC-6TM-ABC	9125	47	1334	
	Pex-abt-2	scaffold732-EXSNAP2012.2	scaffold732-EXSNAP2012.2	7TM-ABC-9TM-ABC	42953	50	1442	Exons were improved; No start codon
	Pex-abt-3	scaffold196-EXSNAP2012.13	scaffold196-EXSNAP2012.13	6TM-ABC-8TM-ABC	27068	55	1581	scaffold196-EXSNAP2012.14 was merged with scaffold196-EXSNAP2012.13
	Pex-abt-4	scaffold1-EXSNAP2012.220	scaffold1-EXSNAP2012.220	6TM-ABC-8TM-ABC	10849	44	1501	scaffold1-EXSNAP2012.221 and scaffold1-EXSNAP2012.222 was merged with scaffold1-EXSNAP2012.220
	Pex-abt-5	scaffold151-EXSNAP2012.38	scaffold151-EXSNAP2012.38	3TM-ABC-7TM-ABC	13279	36	1232	
	Pex-abt-6	scaffold1443-EXSNAP2012.3	scaffold1443-EXSNAP2012.3	4TM-ABC-4TM-ABC	11143	42	1243	
	Pex-abt-7	scaffold339-EXSNAP2012.10	scaffold339-EXSNAP2012.10	5TM-ABC-8TM-ABC	13420	45	1361	
	Pex-abt-8	scaffold571-EXSNAP2012.13	scaffold571-EXSNAP2012.13	7TM-ABC-9TM-ABC	17193	57	1552	
	Pex-abt-9	scaffold81-EXSNAP2012.6	scaffold81-EXSNAP2012.6	4TM-ABC-1TM-ABC	7452	19	673	Exons were improved; No start codon
B	*Pex-haf-2*	scaffold288-EXSNAP2012.12	scaffold288-EXSNAP2012.12	8TM-ABC	4011	23	766	
	Pex-haf-4	scaffold717-EXSNAP2012.4	scaffold717-EXSNAP2012.4	4TM-ABC	2777	16	634	
	Pex-haf-5	scaffold717-EXSNAP2012.5	scaffold717-EXSNAP2012.5	0TM-ABC	2590	10	323	Exons were improved; No start codon
	Pex-haf-6	scaffold149-EXSNAP2012.3	scaffold149-EXSNAP2012.3	5TM-ABC	3846	16	624	Exons were improved; No start codon
	Pex-haf-7	scaffold1229-EXSNAP2012.7	scaffold1229-EXSNAP2012.7	8TM-ABC	4654	21	747	
	Pex-haf-8	scaffold194-EXSNAP2012.8	scaffold194-EXSNAP2012.8	3TM-ABC	2877	11	403	
	Pex-haf-9	scaffold644-EXSNAP2012.12	scaffold644-EXSNAP2012.12	9TM-ABC	7089	30	817	
	Pex-hmt-1	scaffold769-EXSNAP2012.9	scaffold769-EXSNAP2012.9	10TM-ABC	22092	23	749	scaffold769-EXSNAP2012.7 and scaffold769-EXSNAP2012.8 were merged with scaffold769-EXSNAP2012.9; TM helices were improved
	Pex-pgp-1	scaffold539-EXSNAP2012.1	scaffold539-EXSNAP2012.1	6TM-ABC-6TM-ABC	8622	51	1370	
	Pex-pgp-10	scaffold391-EXSNAP2012.12	scaffold391-EXSNAP2012.12	6TM-ABC-6TM-ABC	9355	43	1317	
	Pex-pgp-11	scaffold276-EXSNAP2012.12	scaffold276-EXSNAP2012.12	4TM-ABC-5TM-ABC	7978	38	1204	
	Pex-pgp-12	scaffold4-EXSNAP2012.2	scaffold4-EXSNAP2012.2	5TM-ABC-6TM-ABC	9318	40	1237	scaffold4-EXSNAP2012.1 was merged with scaffold4-EXSNAP2012.2; TM helices were improved; No start codon
	Pex-pgp-13	scaffold1967-EXSNAP2012.1	scaffold1967-EXSNAP2012.1	6TM-ABC-6TM-ABC	9459	44	1305	
	Pex-pgp-14	scaffold505-EXSNAP2012.16	scaffold505-EXSNAP2012.16	4TM-ABC-5TM-ABC	6402	33	1116	scaffold505-EXSNAP2012.17 was merged with scaffold505-EXSNAP2012.16; No start codon
	Pex-pgp-15	scaffold142-EXSNAP2012.4	scaffold142-EXSNAP2012.4	5TM-ABC-5TM-ABC	10243	34	1099	
	Pex-pgp-16	scaffold142-EXSNAP2012.5	scaffold142-EXSNAP2012.5	3TM-ABC-5TM-ABC	7356	32	1044	
	Pex-pgp-17	scaffold142-EXSNAP2012.6	scaffold142-EXSNAP2012.6	6TM-ABC-5TM-ABC	8137	38	1228	
	Pex-pgp-18	scaffold510-EXSNAP2012.8	scaffold510-EXSNAP2012.8	5TM-ABC-6TM-ABC	10425	41	1272	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
B	*Pex-pgp-19*	scaffold346-EXSNAP2012.13	scaffold346-EXSNAP2012.13	6TM-ABC-2TM-ABC	6408	28	982	
	Pex-pgp-2	scaffold46-EXSNAP2012.8	scaffold46-EXSNAP2012.8	6TM-ABC-6TM-ABC	10962	43	1143	Exons were improved; No start codon
	Pex-pgp-20	scaffold505-EXSNAP2012.10	scaffold505-EXSNAP2012.10	6TM-ABC-1TM-ABC	4812	24	775	
	Pex-pgp-21	scaffold631-EXSNAP2012.13	scaffold631-EXSNAP2012.13	4TM-ABC-6TM-ABC	5374	34	1195	
	Pex-pgp-22	scaffold78-EXSNAP2012.23	scaffold78-EXSNAP2012.23	0TM-ABC-5TM-ABC	4600	25	962	
	Pex-pgp-23	scaffold78-EXSNAP2012.24	scaffold78-EXSNAP2012.24	5TM-ABC-6TM-ABC	6383	35	1224	
	Pex-pgp-24	scaffold879-EXSNAP2012.4	scaffold879-EXSNAP2012.4	1TM-ABC-2TM-ABC	13396	17	804	scaffold879-EXSNAP2012.5 was merged with scaffold879-EXSNAP2012.4; No start codon
	Pex-pgp-3	scaffold383-EXSNAP2012.19	scaffold383-EXSNAP2012.19	3TM-ABC-3TM-ABC	8264	25	1054	Exons were improved; No start codon
	Pex-pgp-4	scaffold161-EXSNAP2012.36	scaffold161-EXSNAP2012.36	2TM-ABC-4TM-ABC	4799	25	926	
	Pex-pgp-5	scaffold161-EXSNAP2012.37	scaffold161-EXSNAP2012.37	6TM-ABC-5TM-ABC	6988	30	1170	Exons were improved
	Pex-pgp-6	scaffold11-EXSNAP2012.51	scaffold11-EXSNAP2012.51	6TM-ABC-6TM-ABC	7898	36	1288	
	Pex-pgp-7	scaffold197-EXSNAP2012.1	scaffold197-EXSNAP2012.1	7TM-ABC-5TM-ABC	5489	26	1112	Exons were improved; No start codon
	Pex-pgp-8	scaffold276-EXSNAP2012.11	scaffold276-EXSNAP2012.11	4TM-ABC-1TM-ABC	6955	27	995	Exons were improved; No start codon
	Pex-pgp-9	scaffold144-EXSNAP2012.40	scaffold144-EXSNAP2012.40	6TM-ABC-6TM-ABC	8743	41	1284	
C	*Pex-mrp-1*	scaffold5-EXSNAP2012.81	scaffold5-EXSNAP2012.81	9TM-ABC-5TM-ABC	12368	38	1433	scaffold5-EXSNAP2012.82 was merged with scaffold5-EXSNAP2012.81; No start codon
	Pex-mrp-2	scaffold1482-EXSNAP2012.1	scaffold1482-EXSNAP2012.1	4TM-ABC-4TM-ABC	8768	28	922	scaffold1482-EXSNAP2012.2 was merged with scaffold1482-EXSNAP2012.1; TM helices were improved; No start codon
	Pex-mrp-3	scaffold431-EXSNAP2012.15	scaffold431-EXSNAP2012.15	6TM-ABC-4TM-ABC	12653	43	1388	scaffold431-EXSNAP2012.17 was merged with scaffold431-EXSNAP2012.15; No start codon
	Pex-mrp-4	scaffold962-EXSNAP2012.4	scaffold962-EXSNAP2012.4	11TM-ABC-6TM-ABC	19786	45	1561	Exons were improved
	Pex-mrp-5	scaffold138-EXSNAP2012.1	scaffold138-EXSNAP2012.1	7TM-ABC-7TM-ABC	9857	41	1421	
	Pex-mrp-6	scaffold926-EXSNAP2012.4	scaffold926-EXSNAP2012.4	11TM-ABC-5TM-ABC	8486	44	1370	scaffold926-EXSNAP2012.5 was merged with scaffold926-EXSNAP2012.4
	Pex-mrp-7	scaffold385-EXSNAP2012.18	scaffold385-EXSNAP2012.18	6TM-ABC-6TM-ABC	9966	42	1384	scaffold385-EXSNAP2012.19 and scaffold385-EXSNAP2012.20 were merged with scaffold385-EXSNAP2012.18; No start codon
	Pex-mrp-8	scaffold425-EXSNAP2012.15	scaffold425-EXSNAP2012.15	6TM-ABC-7TM-ABC	20121	52	1356	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
D	*Pex-pmp-1*	scaffold394-EXSNAP2012.8	scaffold394-EXSNAP2012.8	4TM-ABC	3111	18	634	
	Pex-pmp-2	scaffold80-EXSNAP2012.8	scaffold80-EXSNAP2012.8	4TM-ABC	2765	15	537	Exons were improved
	Pex-pmp-3	scaffold544-EXSNAP2012.9	scaffold544-EXSNAP2012.9	6TM-ABC	13578	22	680	Exons were improved; No start codon
	Pex-pmp-4	scaffold486-EXSNAP2012.20	scaffold486-EXSNAP2012.20	6TM-ABC	7043	28	717	
	Pex-pmp-5	scaffold81-EXSNAP2012.2	scaffold81-EXSNAP2012.2	6TM-ABC	3418	18	573	Exons were improved
E	*Pex-abce-1*	scaffold12-EXSNAP2012.69	scaffold12-EXSNAP2012.69	ABC-ABC	3774	20	611	Exons were improved
F	*Pex-abcf-1*	scaffold263-EXSNAP2012.8	scaffold263-EXSNAP2012.8	ABC-ABC	6513	20	640	
	Pex-abcf-2	scaffold40-EXSNAP2012.36	scaffold40-EXSNAP2012.36	ABC-ABC	7928	19	627	Exons were improved
G	*Pex-wht-1*	scaffold777-EXSNAP2012.5	scaffold777-EXSNAP2012.5	ABC-5TM	3961	20	560	
	Pex-wht-4	scaffold267-EXSNAP2012.10	scaffold267-EXSNAP2012.10	ABC-6TM	3017	20	547	Exons were improved; No start codon
	Pex-wht-6	scaffold140-EXSNAP2012.82	scaffold140-EXSNAP2012.82	ABC-7TM	5206	18	613	Exons were improved
	Pex-wht-7	scaffold338-EXSNAP2012.9	scaffold338-EXSNAP2012.9	ABC-6TM	12222	24	756	
	Pex-wht-8	scaffold170-EXSNAP2012.8	scaffold170-EXSNAP2012.8	ABC-8TM	3359	18	607	TM helices were improved

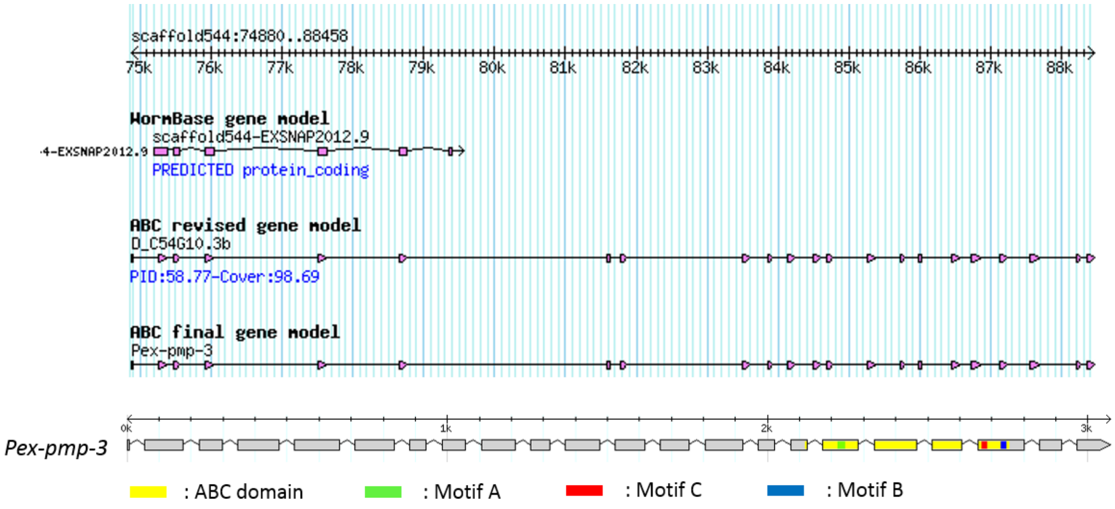


Figure 3.37: A representative case that the exons of one defective ABC transporter gene were improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The original gene model of scaffold544-EXSNAP2012.9 was annotated as a half ABC transporter gene in subfamily D, did not encoded any predicted ABC domain. The revised gene model encoded a typical ABC domain.

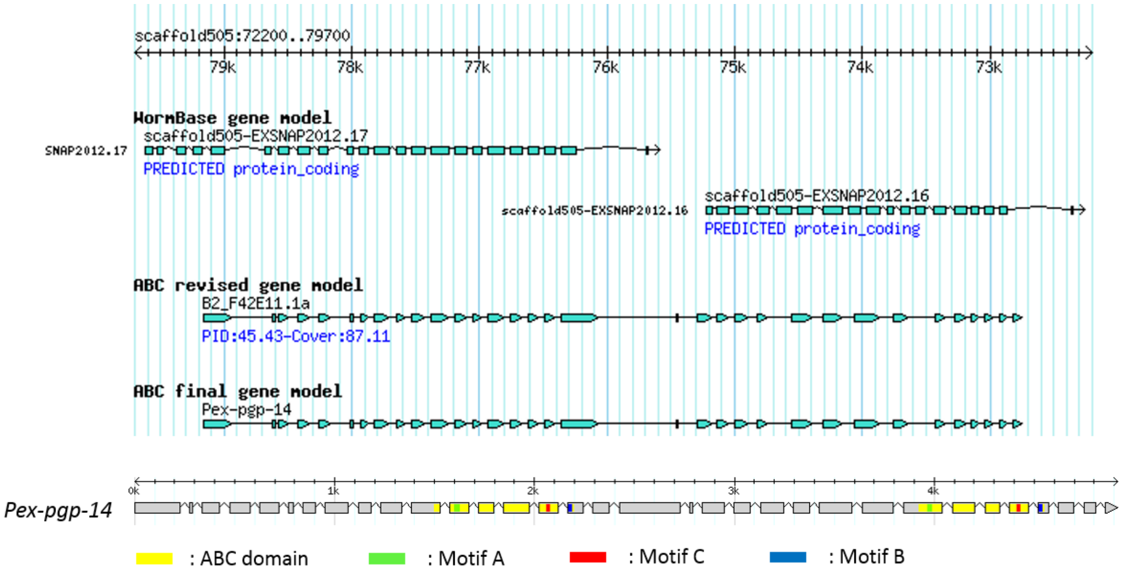
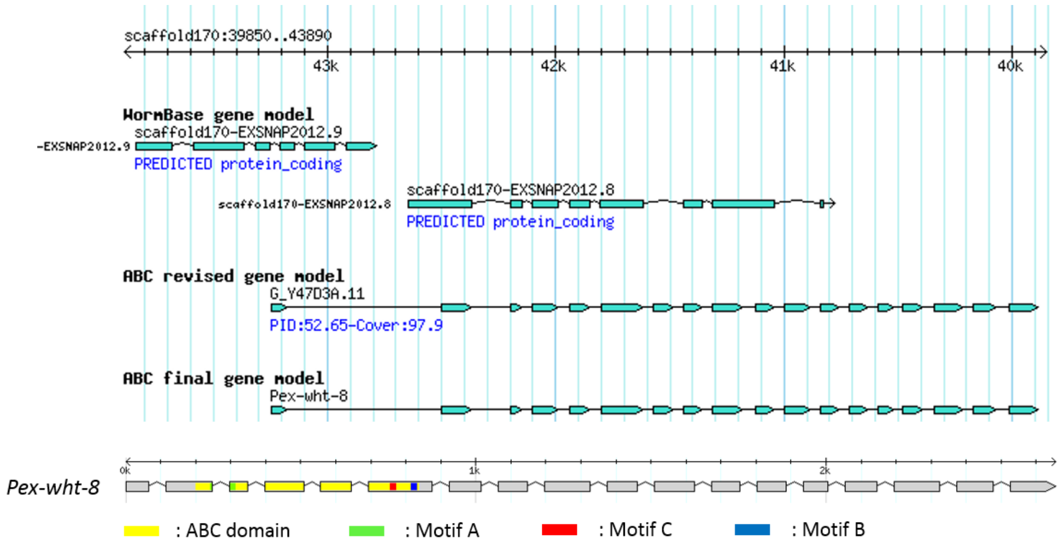
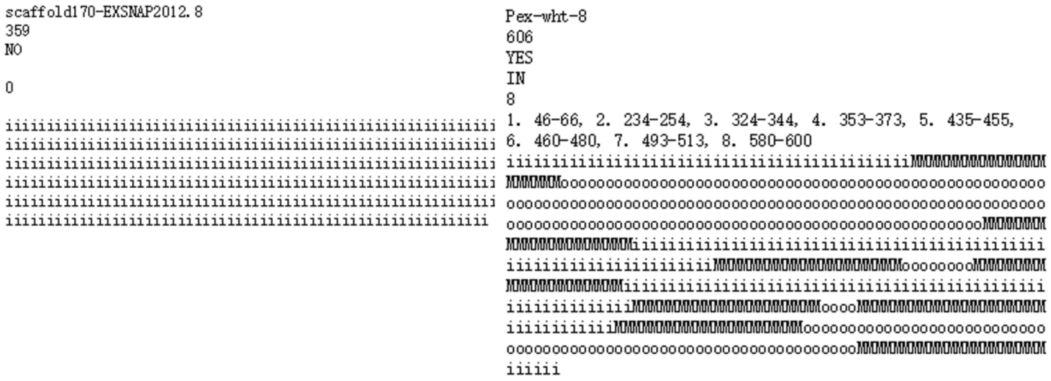


Figure 3.38: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. scaffold505-EXSNAP2012.16 and scaffold505-EXSNAP2012.17 were both annotated as a full ABC transporter gene in subfamily B, encoding a high-quality ABC domain in each. These two genes were merged into one high-quality ABC transporter gene.



TM domain prediction



```
scaffold170-EXSNAP2012.8
359
NO

0

iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

```
Pex-wht-8
606
YES
IN
8
1. 46-66, 2. 234-254, 3. 324-344, 4. 353-373, 5. 435-455,
6. 460-480, 7. 493-513, 8. 580-600
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMM
MMMMMMoooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooMMMMMMM
MMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooMMMMMMMM
MMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooMMMMMMMMMMMMMMMMMMMMM
iiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMM
iiiiii
```

Figure 3.39: A representative case that the TM domain of an ABC transporter gene was improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and C. elegans orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The original gene model of scaffold170-EXSNAP2012.8 encoded a high-quality ABC domain but no TM helices. The revised gene model extended to the neighboring region and became a longer gene model with eight predicted TM helices.

Through phylogenetic analysis, we found 18 out of 62 ABC transporter genes in *P. exspectatus* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*, which was a similar situation to the comparison between *P. pacificus* and *C. elegans*. We assigned the gene names for ABC transporter genes in *P. exspectatus* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.40). Similarly to *P. pacificus*, species specific expansions most occurred in subfamily B, some in subfaimly A and subfamily C in *P. exspectatus*. Contraction of subfamily G also happened in *P. exspectatus*, suggesting the gene loss might be present in the common ancenster of *P. pacificus* and *P. exspectatus*. In addtion, the ABC transporter genes in *P. exspectatus* also consisted of more exons in general, compared to the ABC transporter genes in *Caenorhabditis* species.

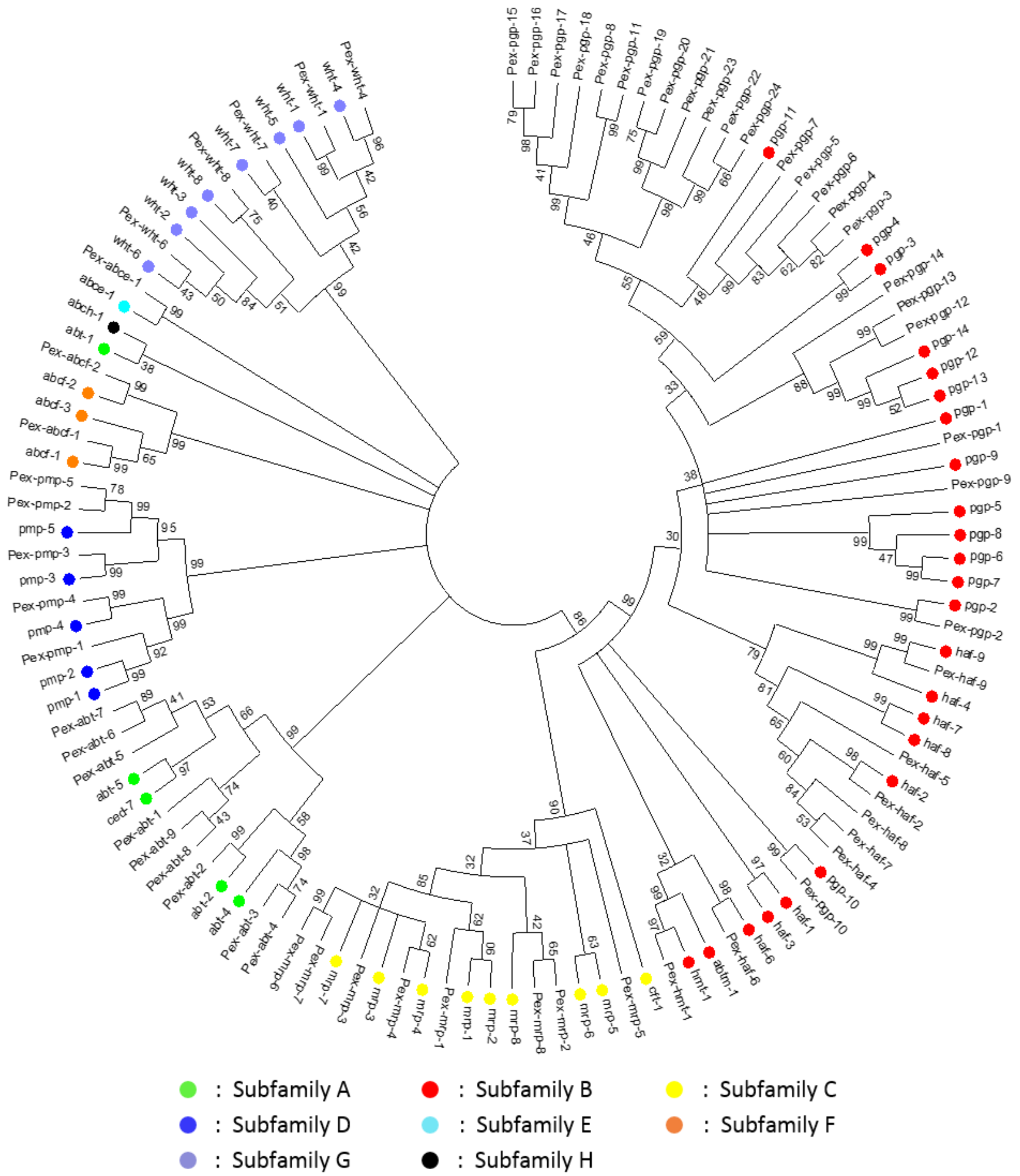


Figure 3.40: Phylogenetic analysis between *P. exspectatus* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *P. exspectatus* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *P. exspectatus* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.12. Annotation of ABC transporter genes in *H. contortus*

H. contortus is also known as Barber's pole worm. It is a highly pathogenic parasitic nematode that can infect a large number of wild and domesticated ruminant species by attaching to abomasal mucosa and feed on the blood. It is the most economically important parasite of sheep and goats worldwide (Laing et al. 2013; Schwarz et al. 2013). After InterProScan and BLAST searches, we identified 141 ABC transporter gene candidates (76 candidates from InterProScan searches, 65 additional ones from BLAST searches), two (maker-C404905-snap-gene-0.1 and snap-C421471-abinit-gene-0.1) of which were due to contamination. We excluded the contamination and checked the quality of the candidates. Among the 139 candidates, only four of them were annotated to be good quality ABC transporter genes. These four genes also encoded proper TM domain (s), suggesting that the most current genome annotation needs improvement. For the defective candidates, we tried to further improve their gene models. After examining the quality of newly constructed gene models, we successfully produced 17 high-quality gene models, two of which were TM domain improved gene models (Table 3.12). For example, augustus-scaffold10684-abinit-gene-0.0, annotated as an ABC transporter gene in subfamily E, had a defective ABC domain with a longer length (194 aa) than expected. After improvement, we got a high-quality ABC transporter gene with proper ABC domain length (134 aa) and conserved motifs (Figure 3.41). Similar example, the original gene model of maker-scaffold2142-augustus-gene-0.25, had a short predicted ABC domain (54 aa), belonging to subfamily G. The newly constructed gene model contained an improved ABC domain with a proper length (149 aa) and features (Figure 3.42). Like *C. angaria*, genome assembly quality of *H. contortus* was low, with a large number (116) of defective candidates that could not be further improved. 20 of these defective candidates could be ABC transporter genes when genome is fully sequenced and assembled. For instance, maker-C440241-snap-gene-0.3, annotated as a full ABC transporter gene in subfamily A, was located in a small contig with only one gene and this assembly error could result in the incompleteness of this candidate (Figure 3.43). In conclusion, after annotation procedure, only 22 high-quality ABC transporter genes were characterized in *H. contortus*, 18 of which had appropriate TM domain(s) (Table 3.12).

Previous study reported 46 ABC transporter genes in *H. contortus* and a significant expansion of *ced-7* was found in *H. contortus* (Laing et al. 2013) compared to that in *C. elegans.* We obtained a much smaller number of high-quality ABC transporter genes and did not see such expansion in *H. contortus.* To figure out the reason, we did a phylogenetic analysis using ABC domain sequences of 76 raw candidates. Not surprising, we observed an expansion of subfamily A in *H. contortus.* Therefore, the number of ABC transporter genes in *H. contortus* was limited by the strict annotation in our analysis*.* In other words, the small number of ABC transporter genes we obtained was mostly due to the low quality of genome assembly and might be partially due to the large number of ABC transporter pseudogenes in *H. contortus*.

Table 3.12: High-quality ABC transporter genes in *H. contortus* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Hco-abt-2*	maker-scaffold15595-augustus-gene-0.19-mRNA-1	maker-scaffold15595-augustus-gene-0.19-mRNA-1	10TM-ABC-7TM-ABC	41092	49	1841	No stop codon
B	*Hco-abtm-1*	maker-scaffold10607-snap-gene-0.34-mRNA-1	maker-scaffold10607-snap-gene-0.34-mRNA-1	4TM-ABC	6689	13	527	Exons are improved; No start codon
	Hco-haf-1	maker-C462017-snap-gene-0.7-mRNA-1	maker-C462017-snap-gene-0.7-mRNA-1	5TM-ABC	8880	18	594	Exons are improved; No start codon
	Hco-haf-4	maker-scaffold2976-snap-gene-0.18-mRNA-1	maker-scaffold2976-snap-gene-0.18-mRNA-1	9TM-ABC	27326	22	766	Exons are improved
	Hco-haf-6	maker-scaffold253-augustus-gene-0.18-mRNA-1	maker-scaffold253-augustus-gene-0.18-mRNA-1	5TM-ABC	6781	18	640	TM helices were improved; No start codon
	Hco-haf-9	snap-scaffold12028-abinit-gene-0.8-mRNA-1	snap-scaffold12028-abinit-gene-0.8-mRNA-1	4TM-ABC	12550	19	546	Exons are improved
	Hco-hmt-1	augustus-scaffold15303-abinit-gene-3.0-mRNA-1	augustus-scaffold15303-abinit-gene-3.0-mRNA-1	5TM-ABC	5022	14	566	
	Hco-pgp-11	maker-scaffold644-augustus-gene-0.18-mRNA-1	maker-scaffold644-augustus-gene-0.18-mRNA-1	0TM-ABC-5TM-ABC	13091	30	1242	No start codon
	Hco-pgp-3	maker-scaffold5316-snap-gene-0.15-mRNA-1	maker-scaffold5316-snap-gene-0.15-mRNA-1	6TM-ABC-6TM-ABC	32102	33	1316	No stop codon
C	*Hco-mrp-7*	maker-C466965-snap-gene-0.5-mRNA-1	maker-C466965-snap-gene-0.5-mRNA-1	TM-ABC-TM-ABC	21868	37	1326	Exons are improved
D	*Hco-pmp-1*	maker-scaffold5579-snap-gene-0.13-mRNA-1	maker-scaffold5579-snap-gene-0.13-mRNA-1	5TM-ABC	11098	18	606	Exons are improved; No start codon
	Hco-pmp-2	snap-C453195-abinit-gene-0.2-mRNA-1	snap-C453195-abinit-gene-0.2-mRNA-1	2TM-ABC	5116	9	382	Exons are improved; No start codon
	Hco-pmp-3	maker-scaffold1674-augustus-gene-0.20-mRNA-1	maker-scaffold1674-augustus-gene-0.20-mRNA-1	7TM-ABC	11005	18	654	Exons are improved; No start codon
	Hco-pmp-4	maker-scaffold18123-augustus-gene-0.12-mRNA-1	maker-scaffold18123-augustus-gene-0.12-mRNA-1	7TM-ABC	19633	18	716	Exons are improved
	Hco-pmp-5	maker-scaffold18585-snap-gene-0.9-mRNA-1	maker-scaffold18585-snap-gene-0.9-mRNA-1	8TM-ABC	18573	19	770	
E	*Hco-abce-1*	augustus-scaffold10684-abinit-gene-0.0-mRNA-1	augustus-scaffold10684-abinit-gene-0.0-mRNA-1	ABC-ABC	6400	13	558	Exons are improved
F	*Hco-abcf-1*	maker-scaffold8516-augustus-gene-0.15-mRNA-1	maker-scaffold8516-augustus-gene-0.15-mRNA-1	ABC-ABC	8389	15	613	Exons are improved; No start codon
G	*Hco-wht-1*	maker-scaffold2142-augustus-gene-0.25-mRNA-1	maker-scaffold2142-augustus-gene-0.25-mRNA-1	ABC-5TM	12220	19	541	Exons are improved
	Hco-wht-2	maker-C464145-snap-gene-0.5-mRNA-1	maker-C464145-snap-gene-0.5-mRNA-1	ABC-8TM	12590	15	555	TM helices were improved; No start codon
	Hco-wht-3	maker-scaffold6608-snap-gene-0.4-mRNA-1	maker-scaffold6608-snap-gene-0.4-mRNA-1	ABC-6TM	5369	14	398	Exons are improved; No stop codon
	Hco-wht-4	maker-scaffold4043-snap-gene-0.9-mRNA-1	maker-scaffold4043-snap-gene-0.9-mRNA-1	ABC-0TM	6768	11	365	Exons are improved; No stop codon
	Hco-wht-7	maker-scaffold3675-augustus-gene-0.22-mRNA-1	maker-scaffold3675-augustus-gene-0.22-mRNA-1	ABC-6TM	9612	19	648	Exons are improved

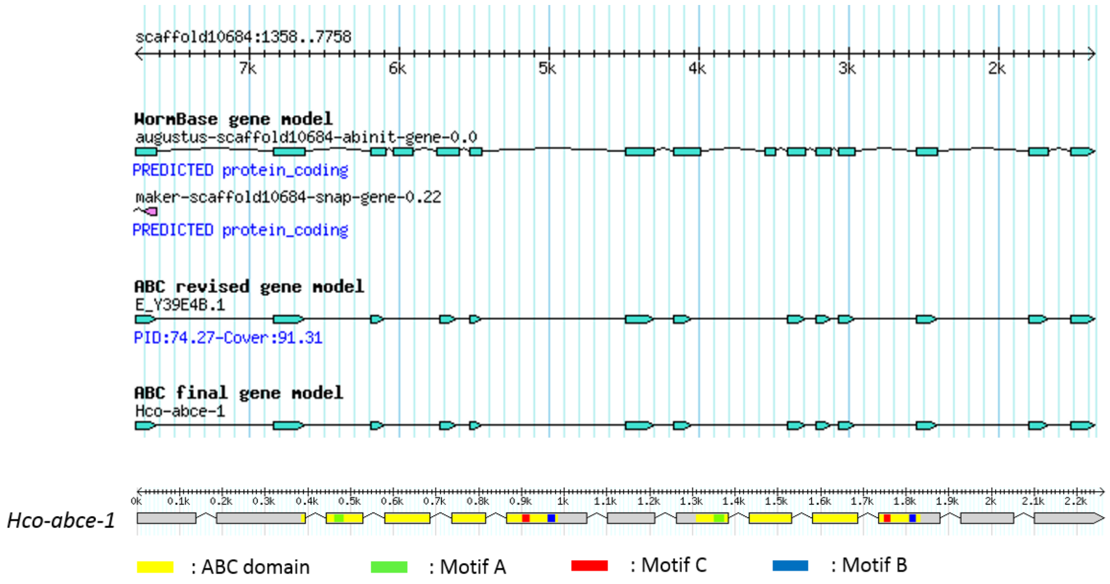


Figure 3.41: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. augustus-scaffold10684-abinit-gene-0.0, annotated as an ABC transporter gene in subfamily E, had a defective ABC domain with a longer length (194 aa) than expected. After improvement, the revised gene model had proper two ABC domains.

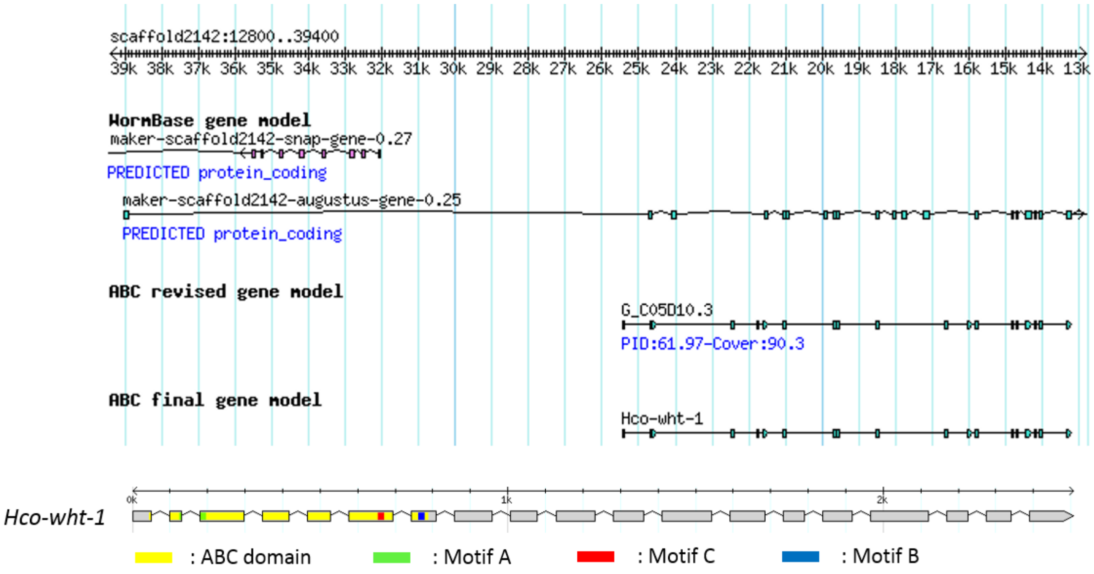

scaffold2142:12800..39400
39k 38k 37k 36k 35k 34k 33k 32k 31k 30k 29k 28k 27k 26k 25k 24k 23k 22k 21k 20k 19k 18k 17k 16k 15k 14k 13k
WormBase gene model
maker-scaffold2142-snap-gene-0.27
PREDICTED protein_coding
maker-scaffold2142-augustus-gene-0.25
PREDICTED protein_coding
ABC revised gene model
G_C05D10.3
PID:61.97-Cover:90.3
ABC final gene model
Hco-wht-1
Hco-wht-1
0k
1k
2k
: ABC domain
: Motif A
: Motif C
: Motif B

Figure 3.42: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The original gene model of maker-scaffold2142-augustus-gene-0.25, annotated as a half ABC transporter gene in subfamily G, had a short predicted ABC domain (54 aa), The revised gene model contained an improved ABC domain with a proper length (149 aa) and all three motifs.

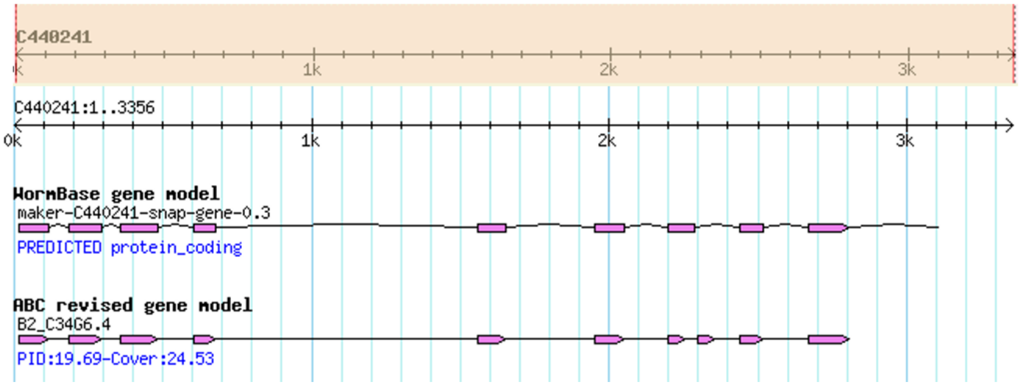


Figure 3.43: A representative case that technical issues could result in incompleteness of an ABC transporter gene candidate

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. maker-C440241-snap-gene-0.3, annotated as a full ABC transporter gene in subfamily A, was located in a small contig with only one gene and this assembly error could result in the incompleteness of this ABC transporter gene.

Through phylogenetic analysis, we found 17 out of 22 ABC transporter genes in *H. contortus* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned gene names for ABC transporter genes in *H. contortus* based on their relationship with ABC transporter genes in *C. elegans*. In *H. contortus*, there was only one high-quality ABC transporter gene in subfamily A as well as in subfamily C, compared to five and nine in *C. elegans* (Figure 3.44). Besides, subfamily B also represents a gene contraction in *H. contortus* when compared to *C. elegans*. Although there could be some potential ABC transporter genes, which we could not obtain due to the relatively low quality of the current genome, it seems that there were less ABC transporter genes in *H. contortus.* It suggests that this parasite might only keep the essential ABC transporter genes to deal with the environment in its host.

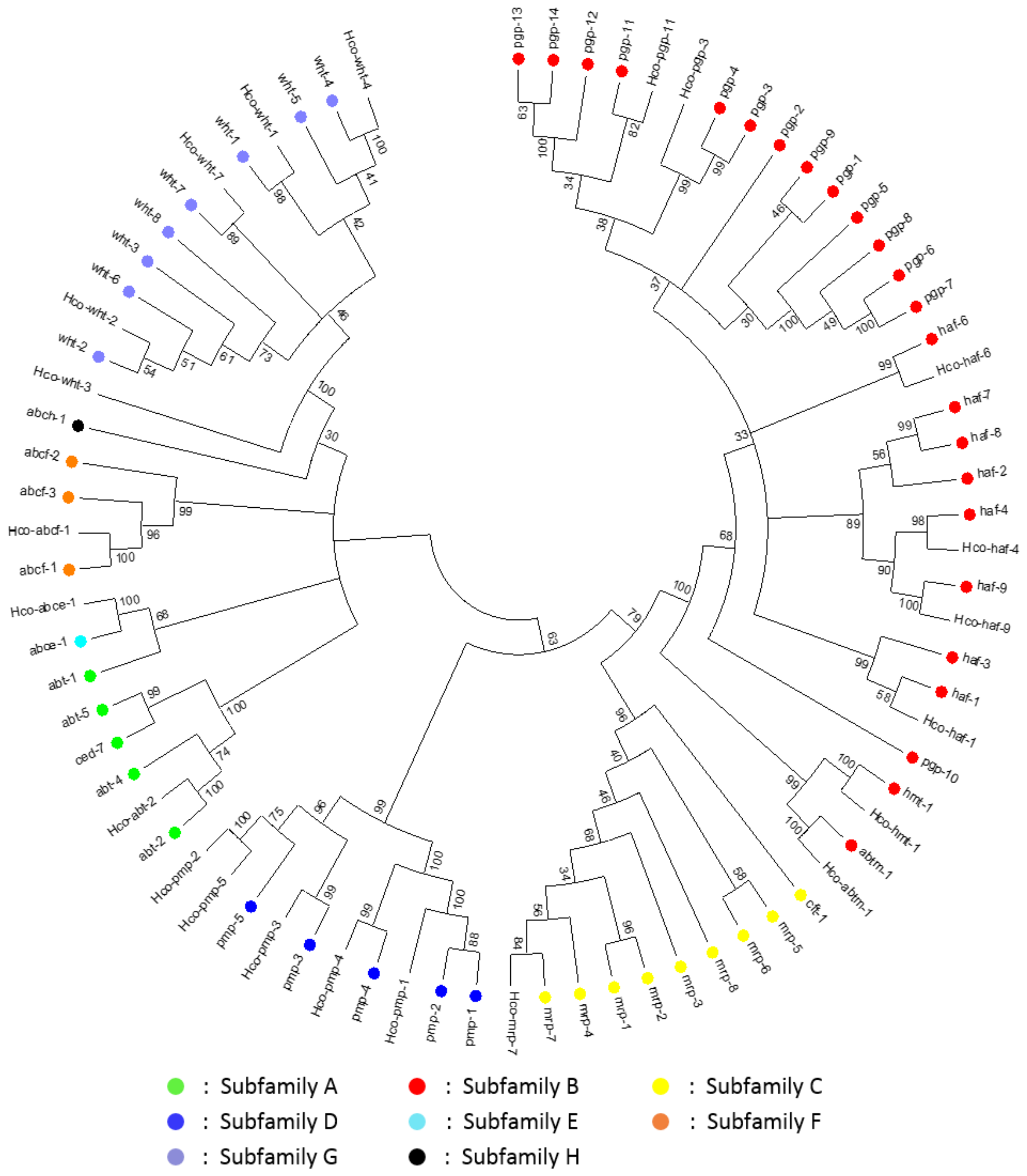


Figure 3.44: Phylogenetic analysis between *H. contortus* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *H. contortus* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *H. contortus* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.13. Annotation of ABC transporter genes in *A. ceylanicum*

A. ceylanicum is a predominant hookworm of dogs and cats and is becoming the second most common hookworm infecting humans in southeastern Asia. Unlike other hookworms*, A. ceylanicum* infects both humans and other mammals, providing a laboratory model for hookworm disease (Traub 2013; Schwarz et al. 2015). After InterProScan and BLAST searches, we identified 97 ABC transporter gene candidates (78 candidates from InterProScan searches, 21 additional ones from BLAST searches). None of these candidates were due to contamination. Thus, the quality of these 97 candidates were examined and 27 of them were high-quality ABC transporter genes. All of these 27 genes had appropriate TM domain(s). For the 70 defective candidates, we tried to further improve their gene models. After examining the quality of revised gene models, we successfully generated 15 improved gene models of high-quality, four of which with only TM domain improved (Table 3.13). Of the 15 improved gene models, 11 of them were produced by merging adjacent genes. For example, three adjacent candidates, Acey_s0031.g2380, Acey_s0031.g2381 and Acey_s0031.g2382 were basically merged together and the revised gene model was annotated and examined to be a high-quality ABC transporter gene encoding a protein with two typical ABC domains in subfamily F (Figure 3.45). Another example is that the TM domain in a half transporter gene in subfamily G was improved by merging two adjacent genes (Acey_s0601.g500 and Acey_s0601.g502). This new gene model contained a cluster of 5 TM helices compared to 3 TM helices in the original gene model (Figure 3.46). 32 candidates could not be improved. 10 of them might be potential ABC transporter genes when the genome is well sequenced and assembled. For example, Acey_s0024.g1008 was annotated as a full ABC transporter gene in subfamily B. However, because of the sequencing gap within this gene, we failed to get a high-quality ABC transporter gene (Figure 3.47). Therefore, it is possible that a high-quality ABC transporter gene is in this region when the genome assembly is improved. In summary, we annotated 50 high-quality ABC transporter genes in *A. ceylanicum,* 39 of which had appropriate TM domain(s) (Table 3.13).

Table 3.13: High-quality ABC transporter genes in *A. ceylanicum* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Ace-abt-1*	Acey_s0007.g3319.t4	Acey_s0007.g3319.t4	8TM-ABC-8TM-ABC	36125	50	1957	
	Ace-abt-2	Acey_s0119.g792.t4	Acey_s0119.g792.t4	8TM-ABC-7TM-ABC	27579	56	2243	
	Ace-abt-3	Acey_s0789.g2363.t2	Acey_s0789.g2363.t2	8TM-ABC-8TM-ABC	24904	43	1817	
	Ace-abt-4	Acey_s0546.g3254	Acey_s0546.g3254.t1	4TM-ABC-8TM-ABC	16778	30	1255	Acey_s0546.g3255, Acey_s0546.g3256 and Acey_s0546.g3258 were merged with Acey_s0546.g3254; No start codon
	Ace-abt-5	Acey_s0218.g2421.t3	Acey_s0218.g2421.t3	3TM-ABC-3TM-ABC	41035	16	767	Acey_s0218.g2423 was merged with Acey_s0218.g2421
	Ace-abt-6	Acey_s0625.g802.t2	Acey_s0625.g802.t2	4TM-ABC-1TM-ABC	22651	15	635	Acey_s0625.g804 was merged with Acey_s0625.g802; No start codon
	Ace-abt-7	Acey_s0625.g796.t10	Acey_s0625.g796.t10	4TM-ABC-9TM-ABC	28344	40	1538	
B	*Ace-abtm-1*	Acey_s0293.g1618.t1	Acey_s0293.g1618.t1	6TM-ABC	7265	19	701	
	Ace-haf-1	Acey_s0264.g613.t4	Acey_s0264.g613.t4	9TM-ABC	7465	19	810	
	Ace-haf-2	Acey_s0071.g611.t1	Acey_s0071.g611.t1	9TM-ABC	6708	17	778	
	Ace-haf-3	Acey_s0020.g59.t2	Acey_s0020.g59.t2	3TM-ABC	5112	21	751	
	Ace-haf-4	Acey_s0948.g3172.t3	Acey_s0948.g3172.t3	1TM-ABC	4412	12	423	
	Ace-haf-6	Acey_s0464.g1934.t3	Acey_s0464.g1934.t3	5TM-ABC	7805	17	655	
	Ace-haf-7	Acey_s0071.g607.t1	Acey_s0071.g607.t1	7TM-ABC	5780	16	1187	
	Ace-haf-8	Acey_s0071.g608.t2	Acey_s0071.g608.t2	9TM-ABC	5450	17	740	
	Ace-haf-9	Acey_s0180.g807.t1	Acey_s0180.g807.t1	9TM-ABC	10053	24	804	
	Ace-hmt-1	Acey_s0003.g1344.t3	Acey_s0003.g1344.t3	11TM-ABC	7466	18	821	
	Ace-pgp-1	Acey_s0006.g3036.t1	Acey_s0006.g3036.t1	6TM-ABC-5TM-ABC	22567	35	1312	
	Ace-pgp-10	Acey_s0095.g2839.t2	Acey_s0095.g2839.t2	5TM-ABC-6TM-ABC	38069	36	1346	Acey_s0095.g2845 was merged with Acey_s0095.g2839; TM helices were improved
	Ace-pgp-11	Acey_s0007.g3528.t2	Acey_s0007.g3528.t2	6TM-ABC-7TM-ABC	18880	31	1215	
	Ace-pgp-12	Acey_s0640.g1003.t3	Acey_s0640.g1003.t3	6TM-ABC-6TM-ABC	12100	32	1336	
	Ace-pgp-13	Acey_s0082.g1571.t1	Acey_s0082.g1571.t1	6TM-ABC-6TM-ABC	15421	35	1375	
	Ace-pgp-14	Acey_s0029.g1873.t4	Acey_s0029.g1873.t4	5TM-ABC-6TM-ABC	14350	28	1179	Exons were improved; No start codon
	Ace-pgp-2	Acey_s0045.g1172.t2	Acey_s0045.g1172.t2	5TM-ABC-6TM-ABC	11928	30	1162	
	Ace-pgp-3	Acey_s0202.g1767.t3	Acey_s0202.g1767.t3	6TM-ABC-5TM-ABC	14820	33	1319	
	Ace-pgp-9	Acey_s0007.g3527.t3	Acey_s0007.g3527.t3	4TM-ABC-5TM-ABC	19457	29	1158	

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
C	*Ace-cft-1*	Acey_s0006.g2897.t1	Acey_s0006.g2897.t1	6TM-ABC-6TM-ABC	23773	34	1255	
	Ace-mrp-1	Acey_s0194.g1443.t3	Acey_s0194.g1443.t3	1TM-ABC-6TM-ABC	8688	31	1440	
	Ace-mrp-3	Acey_s0057.g2801.t2	Acey_s0057.g2801.t2	10TM-ABC-7TM-ABC	13801	40	1545	
	Ace-mrp-5	Acey_s0515.g2785.t2	Acey_s0515.g2785.t2	8TM-ABC-6TM-ABC	21646	36	1462	
	Ace-mrp-6	Acey_s0263.g589.t2	Acey_s0263.g589.t2	7TM-ABC-6TM-ABC	9950	39	1403	
	Ace-mrp-7	Acey_s0728.g1883.t1	Acey_s0728.g1883.t1	10TM-ABC-6TM-ABC	17258	42	1446	
	Ace-mrp-8	Acey_s0728.g1885.t3	Acey_s0728.g1885.t3	8TM-ABC-7TM-ABC	24435	41	1457	TM helices were improved; No start codon
D	*Ace-pmp-1*	Acey_s0044.g1014.t1	Acey_s0044.g1014.t1	3TM-ABC	5600	18	661	Acey_s0044.g1014 were merged with Acey_s0044.g1016
	Ace-pmp-2	Acey_s0424.g1223.t1	Acey_s0424.g1223.t1	2TM-ABC	9623	13	627	Acey_s0424.g1224 wasmerged with Acey_s0424.g1223
	Ace-pmp-3	Acey_s0163.g3465.t1	Acey_s0163.g3465.t1	6TM-ABC	10476	16	660	
	Ace-pmp-4	Acey_s0538.g3136.t1	Acey_s0538.g3136.t1	3TM-ABC	28369	12	542	Acey_s0538.g3137 was merged with Acey_s0538.g3136; No start codon
	Ace-pmp-5	Acey_s0024.g1003.t1	Acey_s0024.g1003.t1	7TM-ABC	4299	15	619	
	Ace-pmp-6	Acey_s0148.g2622.t1	Acey_s0148.g2622.t1	4TM-ABC	6573	13	500	TM helices were improved; No start codon
E	*Ace-abce-1*	Acey_s0005.g2684.t1	Acey_s0005.g2684.t1	ABC-ABC	5907	14	610	
F	*Ace-abcf-1*	Acey_s0018.g3484.t1	Acey_s0018.g3484.t1	ABC-ABC	4547	15	639	
	Ace-abcf-2	Acey_s0226.g2769.t5	Acey_s0226.g2769.t5	ABC-ABC	10400	15	620	Exons were improved
	Ace-abcf-3	Acey_s0031.g2381.t2	Acey_s0031.g2381.t2	ABC-ABC	5649	20	711	Acey_s0031.g2382 and Acey_s0031.g2380 were merged with Acey_s0031.g2381
G	*Ace-wht-1*	Acey_s0062.g3393.t2	Acey_s0062.g3393.t2	ABC-5TM	11868	18	546	Acey_s0062.g33935 was merged with Acey_s0062.g3393; No start codon
	Ace-wht-4	Acey_s0601.g502.t3	Acey_s0601.g502.t3	ABC-5TM	10918	15	566	Acey_s0601.g500 was merged with Acey_s0601.g502; TM helices were improved
	Ace-wht-5	Acey_s0016.g2897.t3	Acey_s0016.g2897.t3	ABC-5TM	9649	26	713	
	Ace-wht-6	Acey_s0471.g2046.t1	Acey_s0471.g2046.t1	ABC-7TM	8057	16	614	
	Ace-wht-7	Acey_s0159.g3310.t2	Acey_s0159.g3310.t2	ABC-5TM	11624	20	717	
	Ace-wht-8	Acey_s0285.g1362.t2	Acey_s0285.g1362.t2	ABC-4TM	19291	15	612	Acey_s0285.g1361.t2 and Acey_s0285.g1360 were merged with Acey_s0285.g1362
H	*Ace-abch-1*	Acey_s0441.g1512.t1	Acey_s0441.g1512.t1	ABC-0TM	3853	7	381	

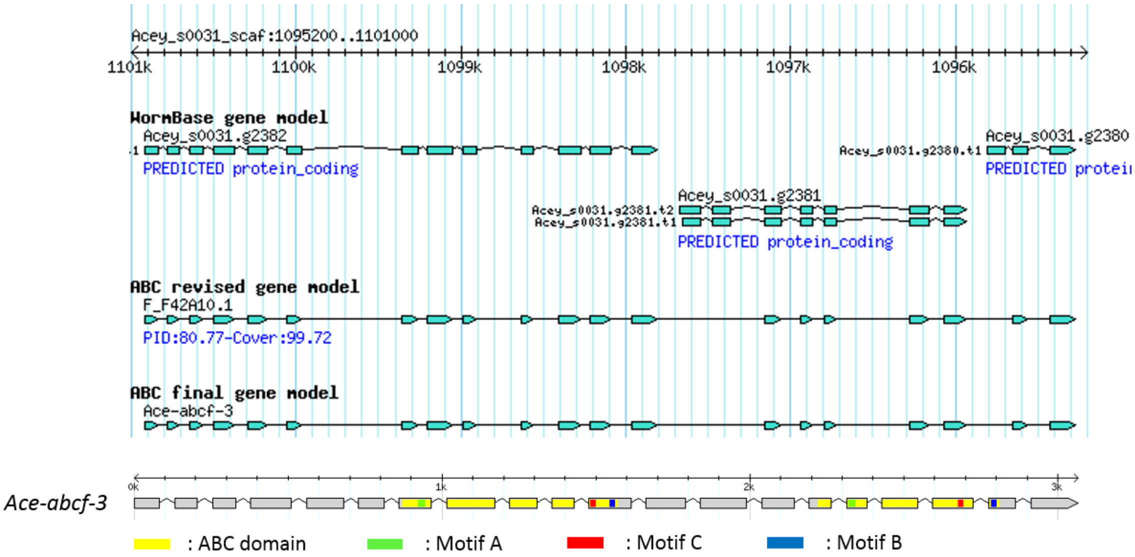
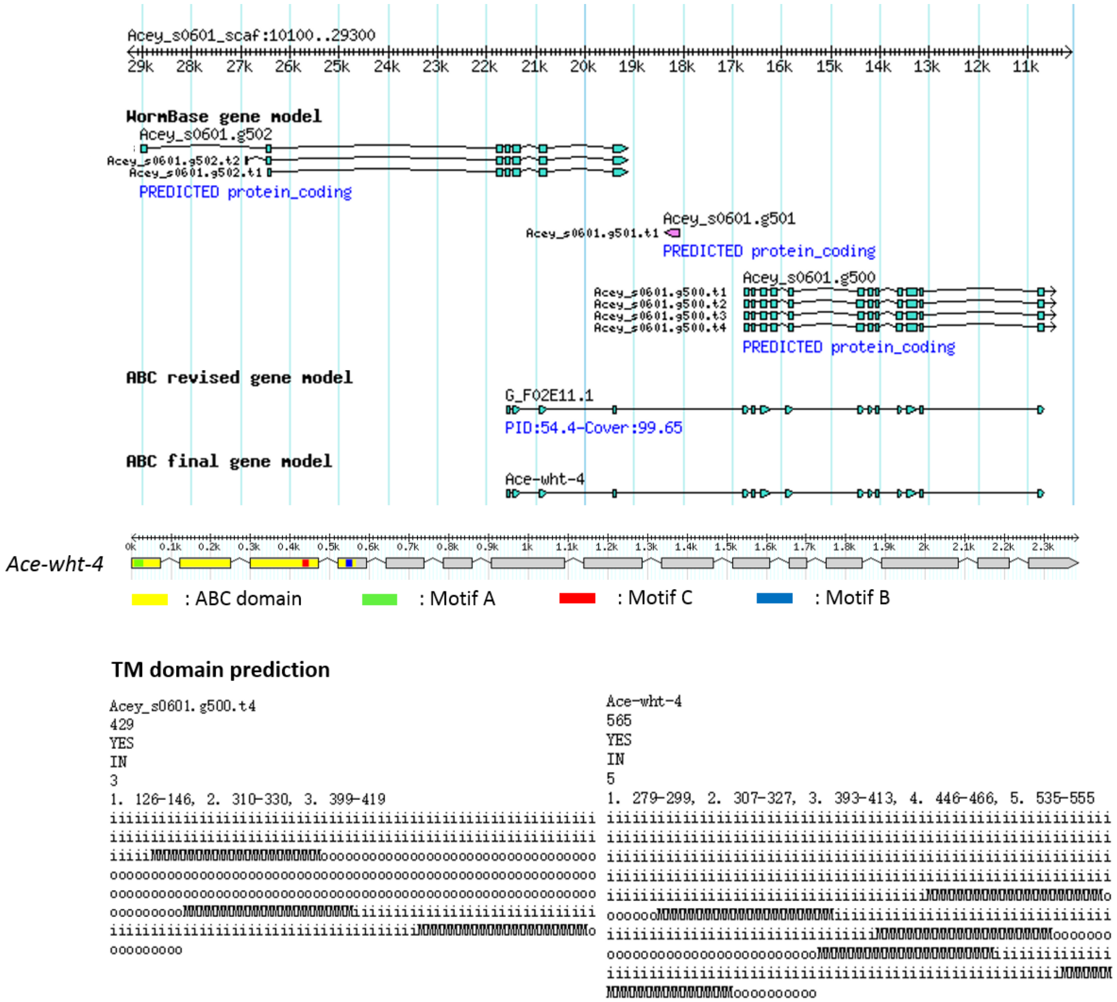


Figure 3.45: A representative case that three adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Two ABC transporter gene candidates, Acey_s0031.g2380, Acey_s0031.g2381 and Acey_s0031.g2382, were merged together and the revised gene model was annotated and examined to be a high-quality ABC transporter gene in subfamily F.



TM domain prediction

```
Acey_s0601.g500.t4
429
YES
IN
3
1. 126-146, 2. 310-330, 3. 399-419
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiMMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMo
ooooooooo
```

```
Ace-wht-4
565
YES
IN
5
1. 279-299, 2. 307-327, 3. 393-413, 4. 446-466, 5. 535-555
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMo
ooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooo
oooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMM
MMMMMMMMMMMMMMMoooooooooo
```

Figure 3.46: A representative case that the TM domain of an ABC transporter gene was improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Acey_s0601.g500 and Acey_s0601.g502 were merged to get a high-quality ABC transporter gene containing an improved TM domain.

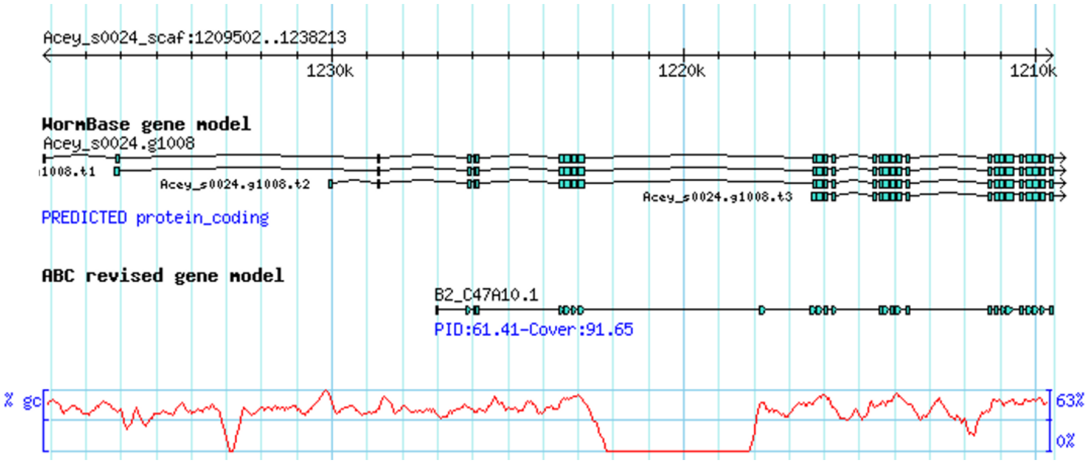


Figure 3.47: A representative case that sequencing error could result in incompleteness of an ABC transporter gene candidate

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. Acey_s0024.g1008 was annotated as a full ABC transporter gene in subfamily B. However, because of the sequencing gap within this gene, we failed to get a high-quality ABC transporter gene.

Through phylogenetic analysis, we found 27 out of 50 ABC transporter genes in *A. ceylanicum* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *A. ceylanicum* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.48). Although the total number of ABC transporter genes in *A. ceylanicum* was close to that in *C. elegans*, only about half of ABC transporter genes share orthologous relationship with those in *C. elegans*, suggesting these two species were evolutionarily distant nematodes. More specifically, species specific gene existed in all subfamilies except subfamily E, F and H. For example, in subfamily A, *Ace-abt-3* together with *Ace-abt-4* share orthologous relationship with *abt-4*. While in subfamily C, *mrp-1*, *mrp-2, mrp-4* and *mrp-8* do not have clear orthologs in *A. ceylanicum*, resulting in a gene expansion in *C. elegans.*

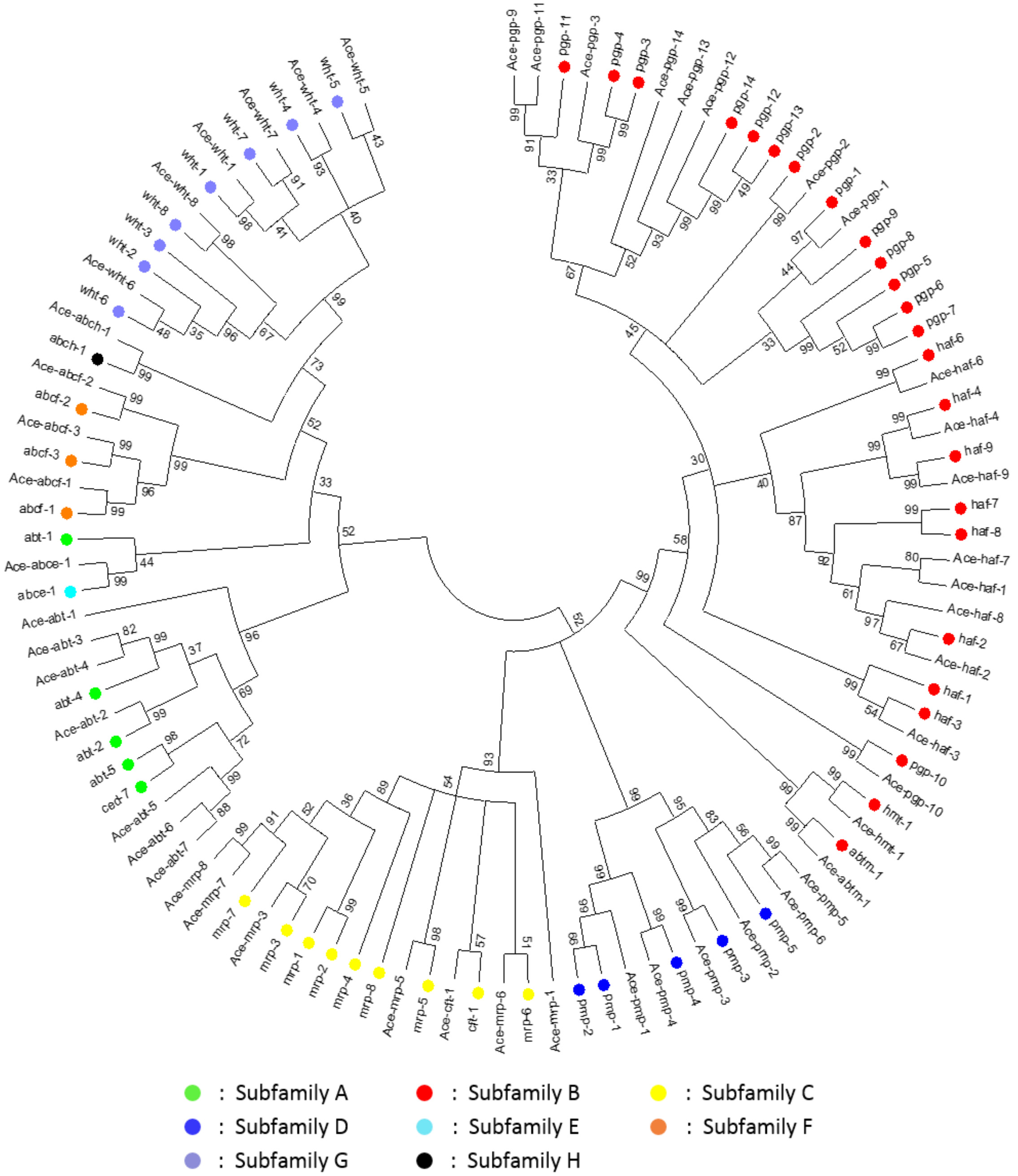


Figure 3.48: Phylogenetic analysis between *A. ceylanicum* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *A. ceylanicum* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *A. ceylanicum* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.14. Annotation of ABC transporter genes in *N. americanus*

The hookworm *N. americanus* is a soil-transmitted helminth (Tang et al. 2014), which can infect human by attaching themselves to the intestinal wall, leading to blood loss, iron-deficiency anemia, and other anemia associated-symptoms and signs (Hyun et al. 2010). We identified 108 ABC transporter gene candidates (70 candidates from InterProScan searches, 38 additional ones from BLAST searches), eight of which were due to contamination. We excluded the contamination and examined the quality of 100 candidates. Most of them were defective, only 10 were high-quality ABC transporter genes. For the 90 defective candidates, we tried to revise each of their gene models. After examining the quality of newly constructed gene models, we generated 27 improved models of high-quality, 10 of which with only TM domain improved (Table 3.14). Among these defective genes, NECAME_08294, annotated as a half ABC transporter gene in subfamily D, encoding a defective ABC domain (110 aa; 3.80E-06). After improvement, we obtained a high-quality ABC transporter gene encoding a qualified ABC domain (143 aa; 3.0E-17) (Figure 3.49). Another example, three adjacent ABC transporter gene candidates (NECAME_09146, NECAME_09147 and NECAME_09150) were merged into one high-quality ABC transporter gene. The revised gene was annotated as a full ABC transporter in subfamily A, encoding two high-quality ABC domains (Figure 3.50). NECAME_03323 had four typical ABC domains and was annotated as a full ABC transporter gene in subfamily B. After improvement, two full ABC transporter genes were obtained by splitting the original model of NECAME_03323 (Figure 3.51). NECAME_11679 was annotated as a full ABC transporter gene in subfamily B and encoded two high-quality ABC domains but only eight TM helices clustering in two TM domains. After improvement, NECAME_11679 was merged with its adjacent gene, NECAME_11678, leading to an improved TM domain with 10 TM helices (Figure 3.52). Among 34 defective candidates that could not be further improved, 19 might be ABC transporter genes when genome assembly is improved. For example, NECAME_16796, annotated as a full ABC transporter gene in subfamily C, only encoded one predicted ABC domain. We found a sequencing gap within this gene, which might be the reason of missing another ABC domain (Figure 3.53). Taking together, we annotated 39 high-quality ABC transporter genes in *N. americanus*, 33 of which had appropriate TM domain(s) (Table 3.14).

Table 3.14: High-quality ABC transporter genes in *N. americanus* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Nam-abt-1*	NECAME_04425	NECAME_04425	7TM-ABC-2TM-ABC	24379	14	696	Exons were improved; No start codon
	Nam-abt-2	NECAME_02178	NECAME_02178	9TM-ABC-6TM-ABC	27228	47	1789	NECAME_02176, NECAME_02177, NECAME_02179 and NECAME_021780 were merged with NECAME_02178; No start codon
	Nam-abt-3	NECAME_10873	NECAME_10873	3TM-ABC-3TM-ABC	37063	17	1006	NECAME_10872 was merged with NECAME_10873
	Nam-abt-4	NECAME_11754	NECAME_11754	7TM-ABC-10TM-ABC	21296	38	1615	NECAME_11750, NECAME_11751 and NECAME_11753 were merged with NECAME_11754
	Nam-abt-5	NECAME_09147	NECAME_09147	5TM-ABC-8TM-ABC	16097	35	1612	NECAME_09146 and NECAME_09150 were merged with NECAME_09147
B	*Nam-abtm-1*	NECAME_13687	NECAME_13687	6TM-ABC	5176	17	637	
	Nam-haf-2	NECAME_10273a	NECAME_10273	9TM-ABC	6565	17	772	Split from NECAME_10273
	Nam-haf-4	NECAME_17011	NECAME_17011	4TM-ABC	5346	19	686	TM helices were improved; No start codon
	Nam-haf-7	NECAME_10273b	NECAME_10273	8TM-ABC	6499	17	723	Split from NECAME_10273
	Nam-haf-8	NECAME_10275	NECAME_10275	7TM-ABC	12356	15	714	NECAME_10274 was merged with NECAME_10275; TM helices were improved
	Nam-haf-9	NECAME_11497	NECAME_11497	9TM-ABC	12678	22	731	
	Nam-hmt-1	NECAME_11896	NECAME_11896	9TM-ABC	7268	18	717	
	Nam-pgp-1	NECAME_08952	NECAME_08952	6TM-ABC-6TM-ABC	13936	31	1195	NECAME_08953 and NECAME_08954 were merged with NECAME_08952; No start codon
	Nam-pgp-10	NECAME_03322	NECAME_03322	4TM-ABC-7TM-ABC	17132	32	1198	
	Nam-pgp-11	NECAME_08861	NECAME_08861	6TM-ABC-8TM-ABC	12109	27	1111	NECAME_08860 was merged with NECAME_08861; TM helices were improved; No start codon
	Nam-pgp-12	NECAME_00050	NECAME_00050	6TM-ABC-6TM-ABC	13951	37	1467	
	Nam-pgp-13	NECAME_06920	NECAME_06920	4TM-ABC-5TM-ABC	19716	30	1218	NECAME_06921 was merged with NECAME_06920; TM helices were improved; No start codon
	Nam-pgp-14	NECAME_11679	NECAME_11679	4TM-ABC-6TM-ABC	12998	32	1236	NECAME_11678 was merged with NECAME_11679; TM helices were improved; No start codon
	Nam-pgp-2	NECAME_07485	NECAME_07485	6TM-ABC-4TM-ABC	13943	32	1227	
	Nam-pgp-3	NECAME_00384	NECAME_00384	4TM-ABC-1TM-ABC	20089	26	991	NECAME_00383 was merged with NECAME_00384; TM helices were improved; No start codon

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
B	*Nam-pgp-4*	NECAME_00386	NECAME_00386	2TM-ABC-0TM-ABC	24844	20	827	NECAME_00385, NECAME_00387 and NECAME_00388 were merged with NECAME_00386; No start codon
	Nam-pgp-5	NECAME_03323a	NECAME_03323	5TM-ABC-6TM-ABC	18867	34	1353	Split from NECAME_03323
	Nam-pgp-6	NECAME_03323b	NECAME_03323	6TM-ABC-4TM-ABC	14978	31	1199	Split from NECAME_03323
	Nam-pgp-9	NECAME_17060	NECAME_17060	6TM-ABC-3TM-ABC	13314	26	993	Exons were improved; No start codon
C	*Nam-mrp-1*	NECAME_09054	NECAME_09054	6TM-ABC-6TM-ABC	12071	29	1243	
	Nam-mrp-5	NECAME_01555	NECAME_01555	8TM-ABC-3TM-ABC	13579	32	1222	No start codon
	Nam-mrp-6	NECAME_01492	NECAME_01492	6TM-ABC-6TM-ABC	8424	31	1156	NECAME_01491 and NECAME_01493 were improved with NECAME_01492; No start codon
D	*Nam-pmp-1*	NECAME_08294	NECAME_08294	4TM-ABC	6171	18	663	Exons were improved
	Nam-pmp-2	NECAME_09801	NECAME_09801	7TM-ABC	8933	15	586	NECAME_09800 was merged with NECAME_09801; TM helices were improved
	Nam-pmp-3	NECAME_09308	NECAME_09308	5TM-ABC	9190	16	633	
	Nam-pmp-5	NECAME_00021	NECAME_00021	5TM-ABC	10798	13	549	NECAME_00020 was merged with NECAME_00021; TM helices were improved
	Nam-abce-1	NECAME_10477	NECAME_10477	ABC-ABC	4802	14	610	
F	*Nam-abcf-1*	NECAME_02046	NECAME_02046	ABC-ABC	5712	15	618	TM helices were improved
	Nam-abcf-2	NECAME_05991	NECAME_05991	ABC-ABC	10060	15	613	Exons were improved
	Nam-abcf-3	NECAME_15694	NECAME_15694	ABC-ABC	7477	20	710	
G	*Nam-wht-1*	NECAME_14853	NECAME_14853	6ABC-TM	9257	18	530	Exons were improved; No start codon
	Nam-wht-2	NECAME_11245	NECAME_11245	6ABC-TM	13639	16	648	
	Nam-wht-7	NECAME_02874	NECAME_02874	5ABC-TM	5640	17	596	
	Nam-wht-8	NECAME_05557	NECAME_05557	4ABC-TM	14395	13	484	NECAME_05556 was merged with NECAME_05557; TM helices were improved

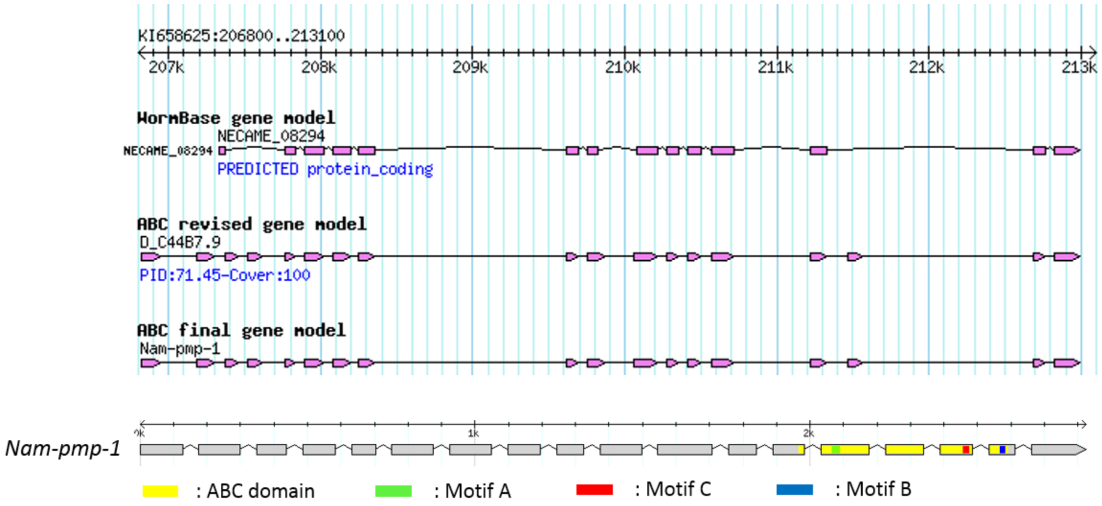


Figure 3.49: A representative case that the exons of one defective ABC transporter gene were improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. NECAME_08294 was annotated as a half ABC transporter gene in subfamily D, encoding a defective ABC domain (110 aa; 3.80E-06). The revised gene model had a typical ABC domain (143 aa; 3.0E-17).

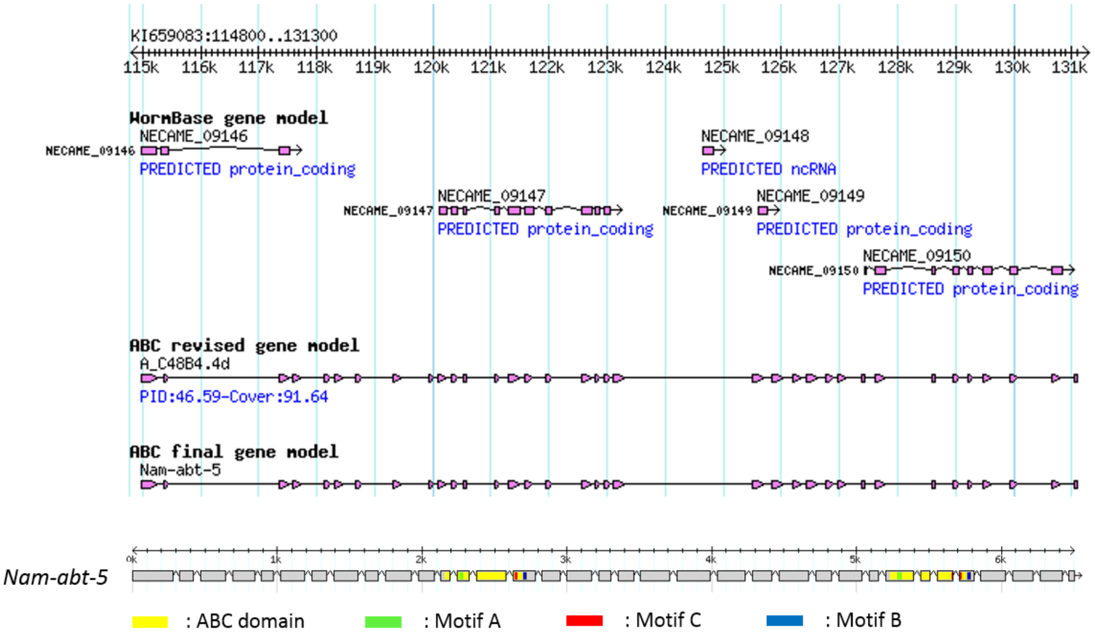


Figure 3.50: A representative case that three adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. NECAME_09146, NECAME_09147 and NECAME_09150 were merged into one ABC transporter gene which was annotated as a full ABC transporter in subfamily A, encoding two high-quality ABC domains.

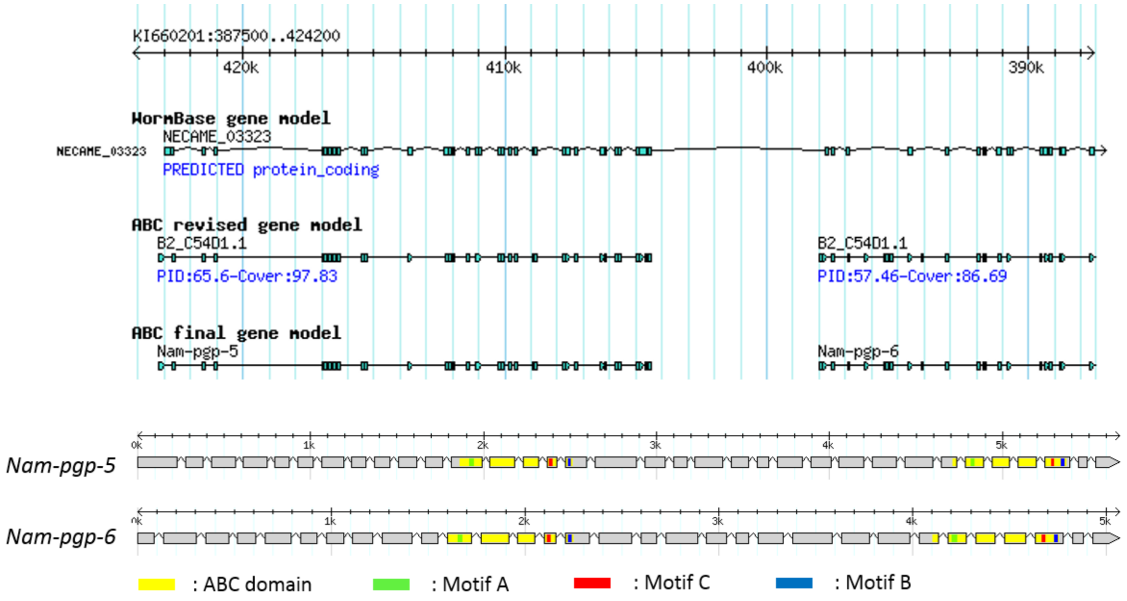
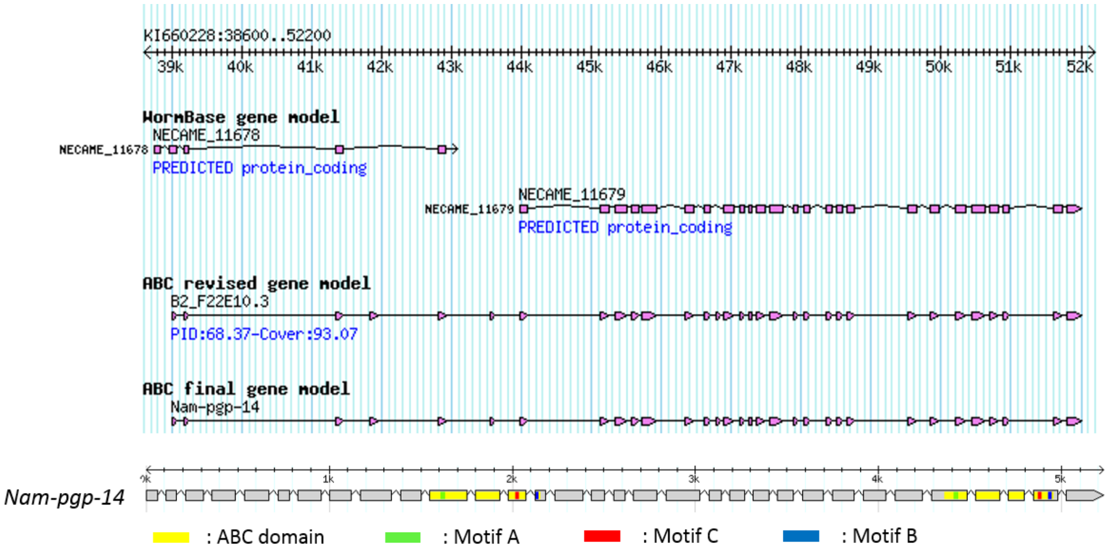
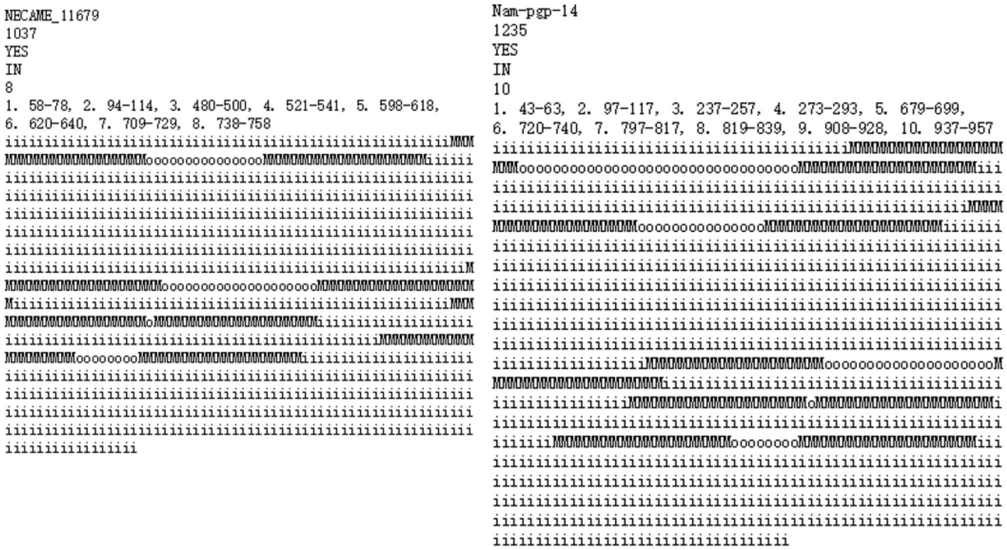


Figure 3.51: A representative case that one candidate was split into two high-quality ABC transporter genes

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. NECAME_03323 had four typical ABC domains and was annotated as a full ABC transporter gene in subfamily B. After improvement, we obtained two full ABC transporter genes as a result of splitting the original model of NECAME_03323.



TM domain prediction



```
NECAME_11679
1037
YES
IN
8
1. 58-78, 2. 94-114, 3. 480-500, 4. 521-541, 5. 598-618,
6. 620-640, 7. 709-729, 8. 738-758
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMM
MMMMMMMMMMMMMMMMMMoooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiM
MMMMMMMMMMMMMMMMMMMMooooooooooooooooooooMMMMMMMMMMMMMMMMMMMM
MiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMM
MMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMM
MMMMMMMMMooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiii
```

```
Nam-pgp-14
1235
YES
IN
10
1. 43-63, 2. 97-117, 3. 237-257, 4. 273-293, 5. 679-699,
6. 720-740, 7. 797-817, 8. 819-839, 9. 908-928, 10. 937-957
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMM
MMMoooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMM
MMMMMMMMMMMMMMMMMoooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooM
MMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMi
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooMMMMMMMMMMMMMMMMMMMMMiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

Figure 3.52: A representative case that the TM domain of an ABC transporter gene was improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. NECAME_11679 was annotated as a full ABC transporter gene in subfamily B encoding two high-quality ABC domains but only eight TM helices. After improvement, NECAME_11679 was merged with its adjacent gene, NECAME_11678, leading to an improved TM domain that had 10 TM helices clustering into two TM domains.

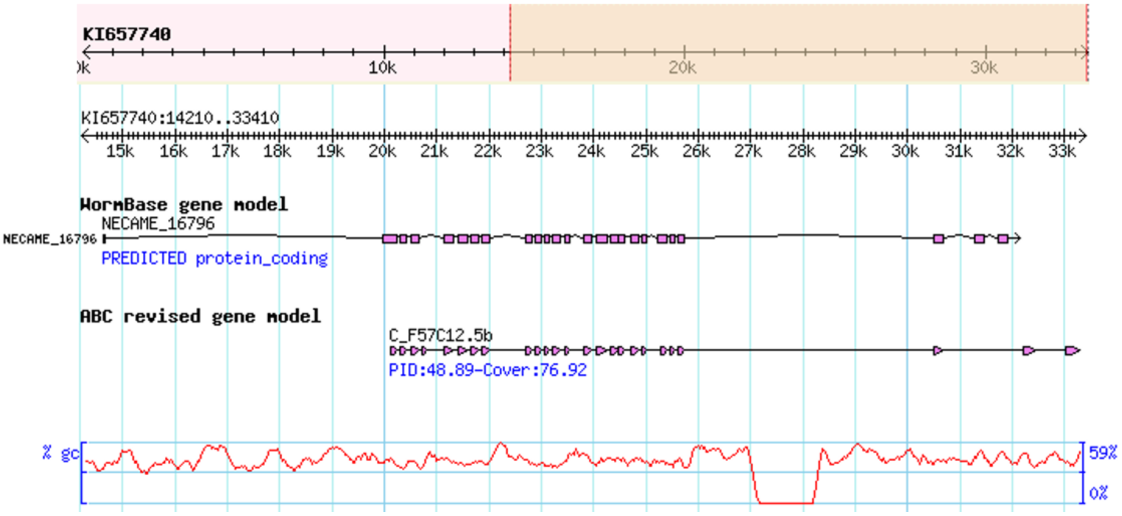


Figure 3.53: A representative case that technical issues could result in incompleteness of an ABC transporter gene candidate

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. NECAME_16796, annotated as a full ABC transporter gene in subfamily C, but only had one predicted ABC domain. The sequencing gap within the genomic region of NECAME_16796 might be the reason of missing another ABC domain in this candidate

Through phylogenetic analysis, we found 22 out of 39 ABC transporter genes in *N. americanus* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned gene names for ABC transporter genes in *N. americanus* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.54). Unlike all the species mentioned above in *N. americanus*, there is no high-quality ABC transporter gene or even putative candidate in subfamily H, suggesting this gene might be lost in *N. americanus*. Except for subfamily E and subfamily F, the number of genes in other subfamilies in *N. americanus* is less than that in *C. elegans.* Therefore, we can see more gene expansions in *C. elegans.*

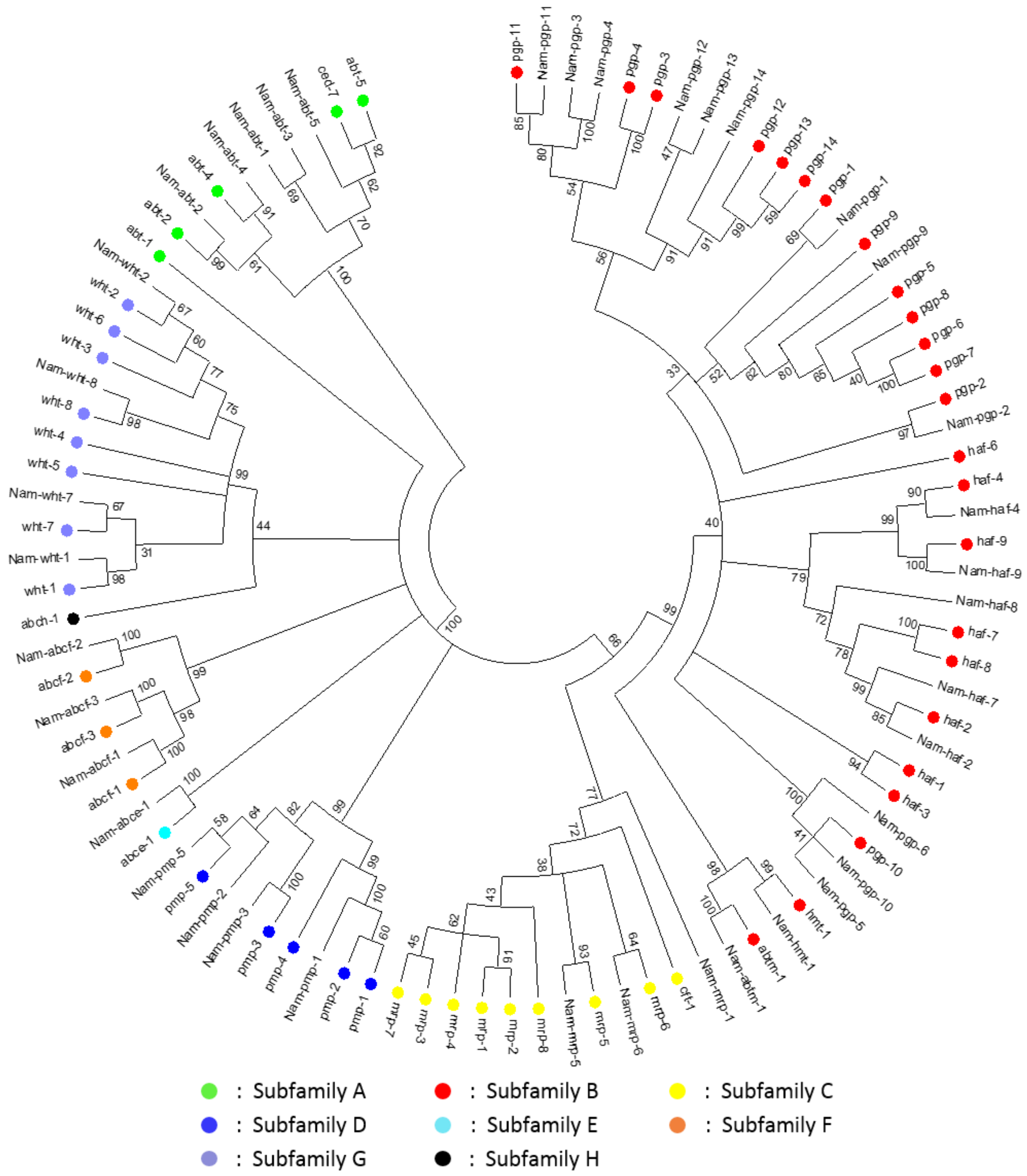


Figure 3.54: Phylogenetic analysis between *N. americanus* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *N. americanus* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *N. americanus* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.15. Annotation of ABC transporter genes in *H. bacteriophora*

H. bacteriophora is an entomopathogenic nematode that has evolved a mutualism with *P. luminescens* bacteria to function as highly virulent insect pathogens (Bai et al. 2013). It was first described in 1975 as a new genus species, and family of *(Heterorhabditidae)* of *Rhabditida.* Much of the previous research concerning *H. bacteriophora* has dealt with applied aspects related to biological control of insect. However, *H. bacteriophora* is an excellent model to investigate fundamental processes such as parasitism and mutualism in addition to its comparative value to *C. elegans* (Ciche 2007). After applying the annotation pipeline to *H. bacteriophora*, we got 86 ABC transporter gene candidates (39 candidates from InterProScan searches, 47 additional ones from BLAST searches), none of which was due to contamination. Among these 86 candidates, only three candidates were high-quality ABC transporter genes. All these three genes also encoded proper TM domain (s). Our improvement procedure generated 38 revised gene models of high-quality, two of which with only TM domain improved (Table 3.15). A representative case for exon improvement of defective gene is Hba_16500. The original gene model of Hba_16500 was annotated as a half ABC transporter gene in subfamily B but encoded a short ABC domain (177 aa). The revised gene model encoded a high-quality ABC domain with a length of 151 aa (Figure 3.55). Hba_11221 and Hba_11222 were both obtained from BLAST searches and did not encode any predicted ABC domain. The revised gene model encoded two typical ABC domains (Figure 3.56). Hba_20849 was annotated as a half ABC transporter gene in subfamily G and encoded one typical ABC domain but two TM domains. After improvement, we got two high-quality ABC transporter genes encoding one ABC domain in each (Figure 3.57). A full ABC transporter gene in subfamily B, Hba_13100, had 18 TM helices, clustering into three groups. The new gene model of Hba_13100 showed a decreased number of TM helices, (12 TM helices) forming two TM domains (Figure 3.58). For the remained 13 candidates, they were not able to be improved. In total, we annotated 41 high-quality ABC transporter genes in *H. bacteriophora*, 36 of which had appropriate TM domain(s) (Table 3.15).

Table 3.15: High-quality ABC transporter genes in *H. bacteriophora* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Hba-abt-2*	Hba_09980	Hba_09980	8TM-ABC-7TM-ABC	21636	46	1837	Hba_09983 and Hba_09987 were merged with Hba_09980
	Hba-abt-4	Hba_11221	Hba_11221	7TM-ABC-7TM-ABC	13503	42	1687	Hba_11222 was merged with Hba_11221
B	*Hba-abtm-1*	Hba_07123	Hba_07123	7TM-ABC	4024	18	693	Exons were improved; No start codon
	Hba-haf-2	Hba_08428	Hba_08428	9TM-ABC	4197	17	769	Exons were improved
	Hba-haf-3	Hba_16500	Hba_16500	6TM-ABC	3680	18	626	Exons were improved
	Hba-haf-4	Hba_08425	Hba_08425	9TM-ABC	6628	20	788	Hba_08426 was merged with Hba_08425
	Hba-haf-9	Hba_11433	Hba_11433	10TM-ABC	7521	25	808	Hba_11434 was merged with Hba_11433
	Hba-hmt-1	Hba_12580	Hba_12580	10TM-ABC	6489	19	758	Hba_12582 was merged with Hba_12580; TM helices were improved;
	Hba-pgp-1	Hba_11458	Hba_11458	7TM-ABC-6TM-ABC	7353	31	1226	Hba_11458 was merged with Hba_11458
	Hba-pgp-10	Hba_18332	Hba_18332	6TM-ABC-5TM-ABC	12911	35	1270	Hba_18333 was merged with Hba_18332
	Hba-pgp-11	Hba_14953	Hba_14953	6TM-ABC-4TM-ABC	9516	26	1110	Hba_14954 was merged with Hba_14953; No start codon
	Hba-pgp-12	Hba_00586	Hba_00586	2TM-ABC-1TM-ABC	9317	21	768	Hba_00585 was merged with Hba_00586; No start codon
	Hba-pgp-13	Hba_13100	Hba_13100	6TM-ABC-6TM-ABC	6417	30	1188	Hba_13101 was merged with Hba_13100; TM helices were improved; No start codon
	Hba-pgp-14	Hba_05467	Hba_05467	6TM-ABC-6TM-ABC	7035	30	1310	Hba_05468 was merged with Hba_05467
	Hba-pgp-2	Hba_10257	Hba_10257	7TM-ABC-6TM-ABC	12841	32	1175	Hba_10259, Hba_10260, Hba_10261 and Hba_10267 were merged with Hba_10257; No start and stop codon
	Hba-pgp-3	Hba_11174	Hba_11174	6TM-ABC-5TM-ABC	8389	31	1218	Exons were improved
	Hba-pgp-4	Hba_03102	Hba_03102	4TM-ABC-4TM-ABC	10629	25	1042	Hba_03105 was merged with Hba_03102; No start codon
	Hba-pgp-9	Hba_14181	Hba_14181	6TM-ABC-6TM-ABC	8492	30	1229	Hba_14182 was merged with Hba_14181
C	*Hba-mrp-1*	Hba_19898	Hba_19898	11TM-ABC-5TM-ABC	9370	34	1490	Hba_19899 was merged with Hba_19898
	Hba-mrp-2	Hba_14631	Hba_14631	6TM-ABC-6TM-ABC	13598	28	1231	Hba_14629 was merged with Hba_14631
	Hba-mrp-3	Hba_15253	Hba_15253	9TM-ABC-3TM-ABC	14442	29	1266	Exons were improved; No start codon
	Hba-mrp-4	Hba_01282	Hba_01282	11TM-ABC-6TM-ABC	10280	36	1547	Hba_01283 was merged with Hba_01282
	Hba-mrp-5	Hba_07200	Hba_07200	8TM-ABC-9TM-ABC	10159	33	1327	Hba_07203 and Hba_07201 were merged with Hba_07200
	Hba-mrp-7	Hba_19586	Hba_19586	10TM-ABC-4TM-ABC	9831	39	1463	Exons were improved
D	*Hba-pmp-1*	Hba_19536	Hba_19536	4TM-ABC	4305	18	661	Hba_19537, Hba_19538 and Hba_19539 were merged with Hba_19536
	Hba-pmp-2	Hba_13920	Hba_13920	7TM-ABC	7707	16	614	Exons were improved
	Hba-pmp-3	Hba_19272	Hba_19272	6TM-ABC	3876	15	624	
	Hba-pmp-4	Hba_14447	Hba_14447	5TM-ABC	4348	18	708	Hba_14448 was merged with Hba_14447
	Hba-pmp-5	Hba_04521	Hba_04521	0TM-ABC	2147	10	404	Exons were improved
	Hba-pmp-6	Hba_05200	Hba_05200	3TM-ABC	3054	11	414	Hba_05201 was merged with Hba_05200; No start codon

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
E	*Hba-abce-1*	Hba_16821	Hba_16821	ABC-ABC	3759	14	612	Exons were improved
F	*Hba-abcf-1*	Hba_07491	Hba_07491	ABC-ABC	4282	16	635	Exons were improved
	Hba-abcf-2	Hba_20712	Hba_20712	ABC-ABC	7498	14	620	Hba_20713 was merged with Hba_20712
	Hba-abcf-3	Hba_00491	Hba_00491	ABC-ABC	4416	19	699	Hba_00490 and Hba_00492 was merged with Hba_00491
G	*Hba-wht-1*	Hba_00009	Hba_00009	ABC-4TM	4197	16	592	Hba_00010 was merged with Hba_00009; No start and stop codon
	Hba-wht-2	Hba_17094	Hba_17094	ABC-8TM	5951	19	616	
	Hba-wht-3	Hba_20837	Hba_20837	ABC-7TM	5016	16	817	
	Hba-wht-4	Hba_17135	Hba_17135	ABC-0TM	5229	6	259	Exons were improved
	Hba-wht-5	Hba_10572	Hba_10572	ABC-4TM	4181	16	587	Hba_10573 was merged with Hba_10572; No start and stop codon
	Hba-wht-6	Hba_20849a	Hba_20849	ABC-5TM	4221	16	606	Split from Hba_20849; No start codon
	Hba-wht-7	Hba_20849b	Hba_20849	ABC-8TM	5642	16	657	Split from Hba_20849; No start codon

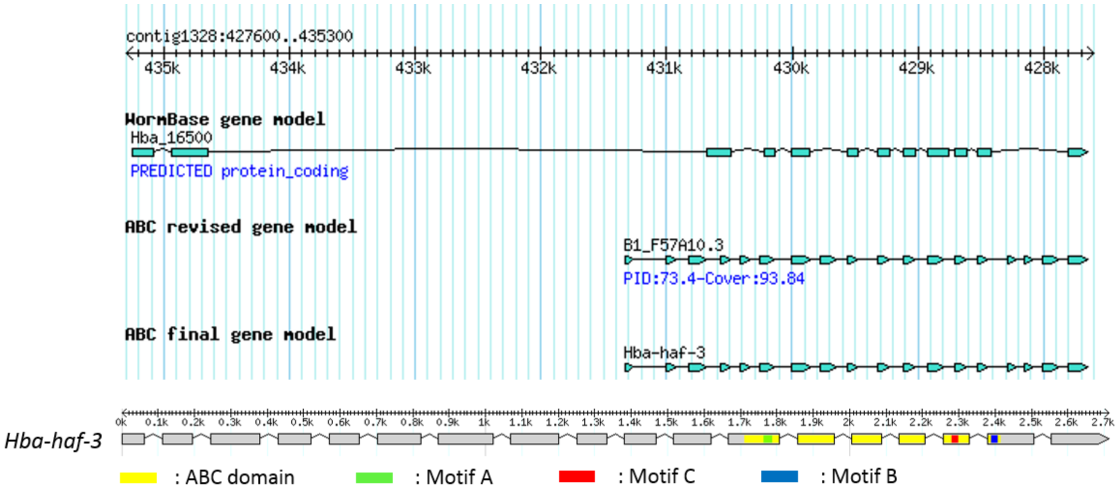


Figure 3.55: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Hba_16500 was annotated as a half ABC transporter gene in subfamily B encoding a short ABC domain (107 aa). The revised gene model contained a high-quality ABC domain with a length of 151 aa.

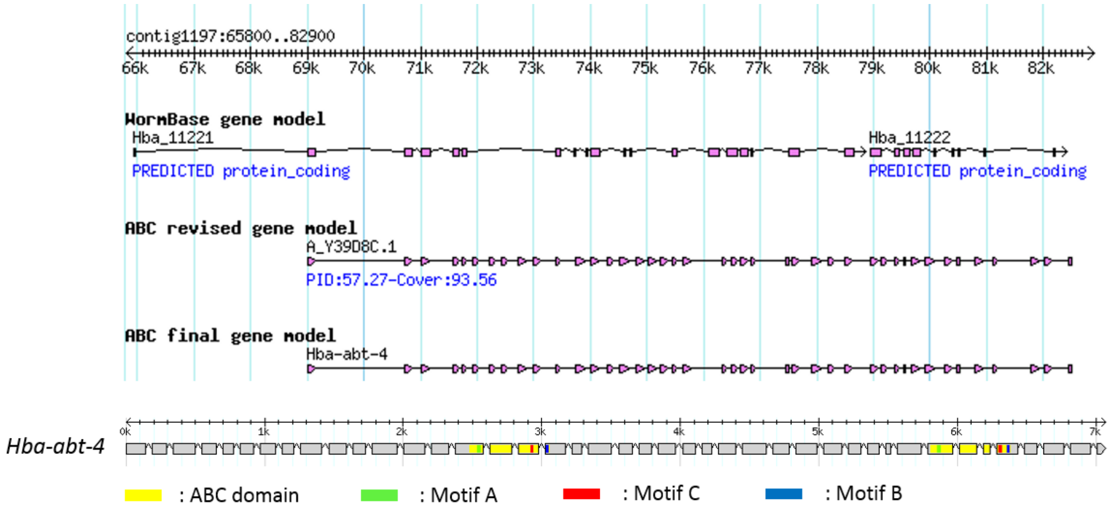


Figure 3.56: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Hba_11221 and Hba_11222 were both obtained from BLAST searches and did not encoded any predicted ABC domain. Through our revision, more exons were annotated in this genomic region, leading to one high-quality ABC transporter with two typical ABC domains.

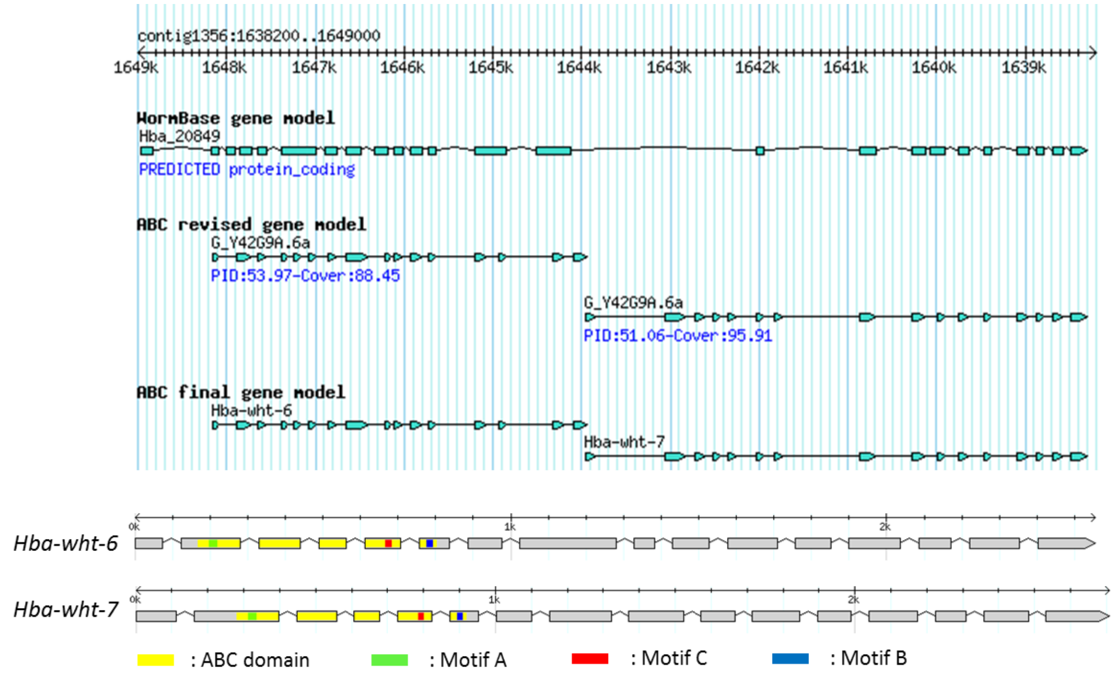
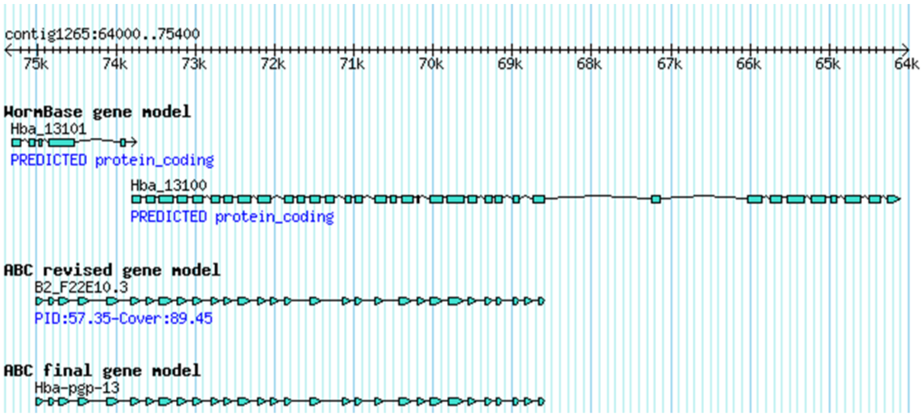
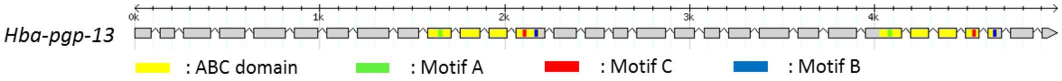


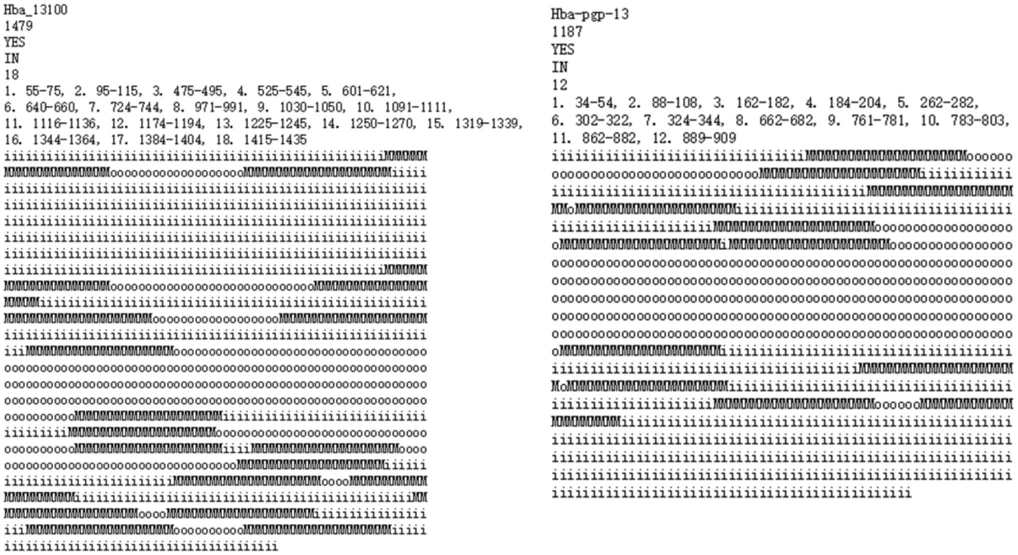
Figure 3.57: A representative case that one candidate was split into two high-quality ABC transporter genes

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Hba_20849 was annotated as a half ABC transporter gene in subfamily G and encoded one high-quality ABC domain but two TM domains. After improvement, we got two high-quality ABC transporter genes with one high-quality ABC domain in each.





TM domain prediction



```
Hba_13100
1479
YES
IN
18
1. 55-75, 2. 95-115, 3. 475-495, 4. 525-545, 5. 601-621,
6. 640-660, 7. 724-744, 8. 971-991, 9. 1030-1050, 10. 1091-1111,
11. 1116-1136, 12. 1174-1194, 13. 1225-1245, 14. 1250-1270, 15. 1319-1339,
16. 1344-1364, 17. 1384-1404, 18. 1415-1435
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMM
MMMMMMMMMMMMMMMoooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMM
MMMMMMMMMMMMMMMoooooooooooooooooooooooooooooMMMMMMMMMMMMMMMM
MMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
MMMMMMMMMMMMMMMMMMMMMooooooooooooooooooMMMMMMMMMMMMMMMMMMMMM
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooooo
ooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiMMMMMMMMMMMMMMMMMMMMMoooo
oooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooMMMMMMMMMMM
MMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMM
MMMMMMMMMMMMMMMMMMMooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiii
iiiMMMMMMMMMMMMMMMMMMMMMooooooooooMMMMMMMMMMMMMMMMMMMMMiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

```
Hba-pgp-13
1187
YES
IN
12
1. 34-54, 2. 88-108, 3. 162-182, 4. 184-204, 5. 262-282,
6. 302-322, 7. 324-344, 8. 662-682, 9. 761-781, 10. 783-803,
11. 862-882, 12. 889-909
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooo
oooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMM
MMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooo
oMMMMMMMMMMMMMMMMMMMMMiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMM
MoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooMMMMMMMMMMMM
MMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

Figure 3.58: A representative case that the TM domain of an ABC transporter gene was improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Hba_13100 was annotated as a full ABC transporter gene in subfamily B and had 18 TM helices, clustering into three groups. The new gene model of Hba_13100 showed a decreased number of TM helices (12), forming two TM domains.

Through phylogenetic analysis, we found 23 out of 41 ABC transporter genes in *H. bacteriophora* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. ABC transporter genes in *H. bacteriophora* were assinged gene names based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.59). Compared to *C. elegans*, there were obvious gene contractions of subfamily A and subfamily B in *H. bacteriophora*, with only two members in subfamily A and 16 members in subfamily B compared to five and 24 in *C. elegans*. It reflects that some ABC transporter genes in these two subfamily might not be essential for *H. bacteriophora* to confront host evironment or be compensated by ABC trnasporter genes in its symbiotic bacteria.

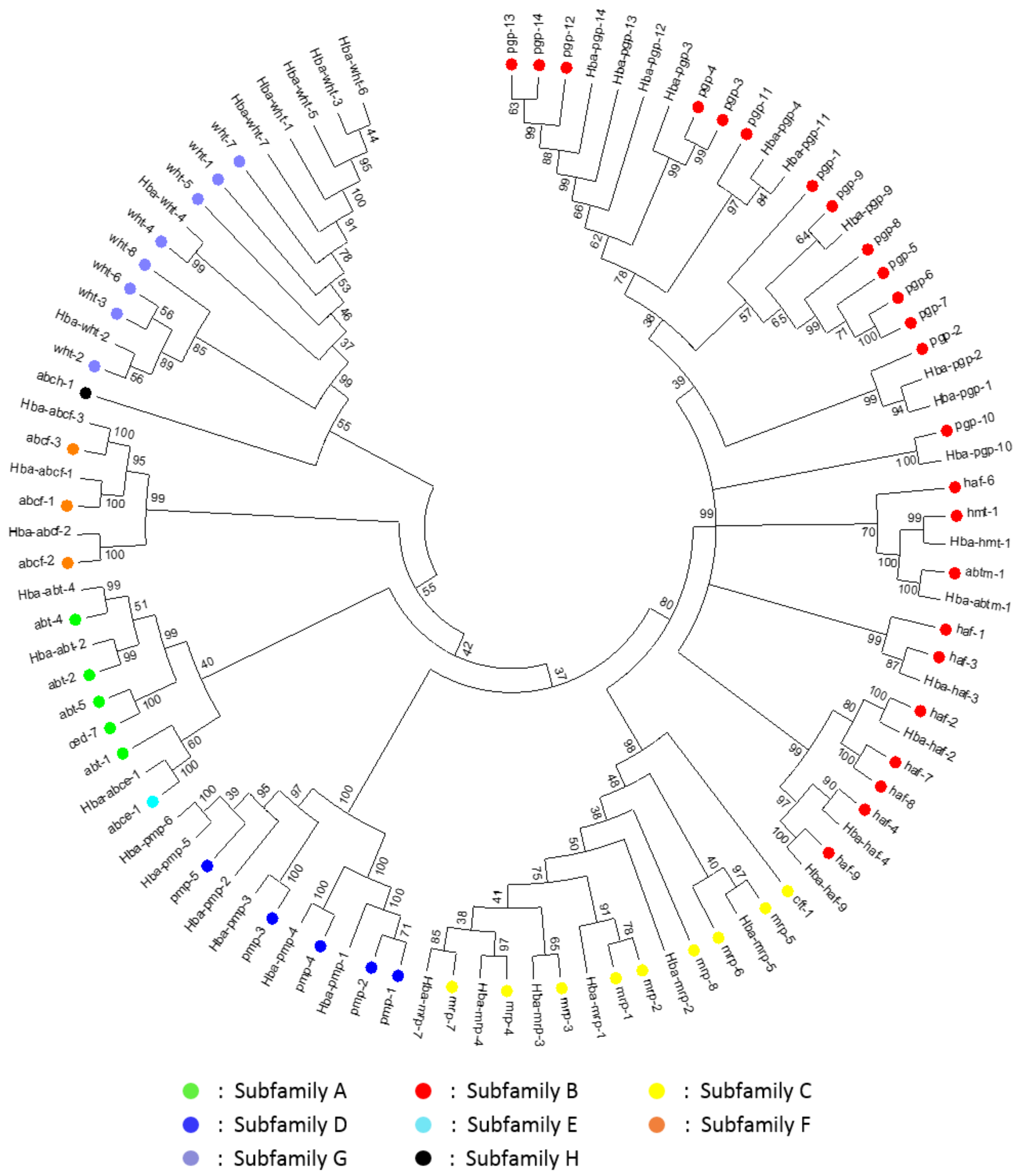


Figure 3.59: Phylogenetic analysis between *H. bacteriophora* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *H. bacteriophora* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *H. bacteriophora* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.16. Annotation of ABC transporter genes in *P. redivivus*

P. redivivus, which is also known as the “microworm” is a small, free-living nematode found in soil and it has been used as a model system to understand the evolution of developmental and behavioral processes given its phylogenetic distance to *C. elegans* (Sternberg and Horvitz 1981). After applying the annotation pipeline to *P. redivivus*, we obtained 105 ABC transporter gene candidates with 91 ABC transporter gene candidates from InterProScan searches and 14 additional ones from BLAST searches. None of these candidates was due to contamination. 44 of 105 candidates were high-quality ABC transporter genes. All of these 44 genes encoded appropriate TM domain (s). After improvement, we generated 11 revised gene models of high-quality, five of which with only TM domain improved (Table 3.16). Pan_g2208 was annotated as a full ABC transporter gene in subfamily B. However, one of its ABC domains was much longer (181 aa) than expected. After improvement, the revised gene model encoded two typical ABC domain (Figure 3.60). Another example shows TM domain improvement. Pan_g13159 merged with its neighboring gene Pan_g13157, leading to an increase number of TM helices, from 10 to 12. Thus, the new gene model had typical number of TM helices forming two TM domains (Figure 3.61). For the remained candidates that were still defective, five of them contained some technique errors in their genomic region. Therefore, they could be complete ABC transporter genes when these region are fully sequenced and assembled. For example, Pan_g12794 was annotated as a full ABC transporter gene in subfamily B but only encoded one typical ABC domain. We found a sequencing gap within this gene, which might cause the incompleteness of this ABC transporter candidate (Figure 3.62). In summary, we annotated 59 high-quality ABC transporter genes in *P. redivivus*, 56 of which had appropriate TM domains (s) (Table 3.16).

Although a large number (94) of ABC transporter gene candidates identified in previous study through InterProScan search (Srinivasan et al. 2013), the quality of these candidates was not evaluated. Moreover, the ABC transporter genes they obtained were mostly likely the ones (91) that we identified from InterProScan search. Considering the improvement and evaluation we made, it is reasonable that 59 high-quality ABC transporter genes were characterized in *P. redivivus*.

Table 3.16: High-quality ABC transporter genes in *P. redivivus* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Pre-abt-1*	g11984.t1	g11984.t1	7TM-ABC-14TM-ABC	5673	14	1654	
	Pre-abt-2	g18975.t1	g18975.t1	7TM-ABC-9TM-ABC	7477	10	2276	
	Pre-abt-3	g1560.t1	g1560.t1	8TM-ABC-10TM-ABC	5197	12	1533	
	Pre-abt-4	g11354.t1	g11354.t1	7TM-ABC-7TM-ABC	5864	11	1787	
	Pre-abt-5	g9825.t1	g9825.t1	6TM-ABC-7TM-ABC	6762	15	1516	g9826 was merged with g9825
B	*Pre-abtm-1*	g16493.t1	g16493.t1	7TM-ABC	2391	6	705	
	Pre-haf-1	g18432.t1	g18432.t1	7TM-ABC	2578	6	770	TM helices were improvement; No stop codon
	Pre-haf-2	g40.t1	g40.t1	9TM-ABC	2589	5	788	
	Pre-haf-3	g3308.t1	g3308.t1	4TM-ABC	2579	4	522	
	Pre-haf-4	g39.t1	g39.t1	9TM-ABC	2709	6	819	
	Pre-haf-6	g20293.t1	g20293.t1	6TM-ABC	2322	6	684	
	Pre-haf-9	g10515.t1	g10515.t1	10TM-ABC	2751	4	830	
	Pre-hmt-1	g21191.t1	g21191.t1	10TM-ABC	2642	7	719	TM helices were improvement; No start codon
	Pre-hmt-2	g8087.t1	g8087.t1	4TM-ABC	2098	5	561	
	Pre-hmt-3	g12612.t1	g12612.t1	11TM-ABC	2646	6	793	
	Pre-hmt-4	g9675.t1	g9675.t1	8TM-ABC	2645	6	760	
	Pre-pgp-1	g11074.t1	g11074.t1	6TM-ABC-5TM-ABC	4568	7	1253	
	Pre-pgp-10	g13159.t1	g13159.t1	6TM-ABC-6TM-ABC	6421	15	1401	g13157 was merged with g13159; TM helices were improvement;
	Pre-pgp-11	g14521.t1	g14521.t1	6TM-ABC-5TM-ABC	7584	15	1540	
	Pre-pgp-12	g4627.t1	g4627.t1	6TM-ABC-6TM-ABC	4416	7	1354	
	Pre-pgp-13	g18108.t1	g18108.t1	4TM-ABC-4TM-ABC	3788	8	1049	Exons were improved
	Pre-pgp-14	g9667.t1	g9667.t1	6TM-ABC-6TM-ABC	4542	13	1258	
	Pre-pgp-15	g10895.t1	g10895.t1	6TM-ABC-5TM-ABC	3989	8	1197	
	Pre-pgp-16	g3036.t1	g3036.t1	6TM-ABC-5TM-ABC	4154	8	1180	
	Pre-pgp-17	g22296.t1	g22296.t1	6TM-ABC-6TM-ABC	5811	10	1195	
	Pre-pgp-18	g8824.t1	g8824.t1	6TM-ABC-6TM-ABC	3974	9	1193	
	Pre-pgp-19	g19718.t1	g19718.t1	2TM-ABC-4TM-ABC	5501	7	995	g19719 was merged with g19718
	Pre-pgp-2	g22409.t1	g22409.t1	6TM-ABC-6TM-ABC	4630	10	1318	
	Pre-pgp-3	g19721.t1	g19721.t1	6TM-ABC-6TM-ABC	7851	11	1221	g19722 was merged with g19721; No start codon

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
B	*Pre-pgp-4*	g2208.t1	g2208.t1	6TM-ABC-6TM-ABC	4791	9	1193	Exons were improved; No start codon
	Pre-pgp-5	g17132.t1	g17132.t1	5TM-ABC-7TM-ABC	3989	7	1193	TM helices were improvement; No start and stop codon
	Pre-pgp-6	g12160.t1	g12160.t1	6TM-ABC-5TM-ABC	4210	7	1252	
	Pre-pgp-7	g17150.t2	g17150.t2	4TM-ABC-6TM-ABC	4208	8	1203	
	Pre-pgp-8	g18526.t1	g18526.t1	6TM-ABC-6TM-ABC	4201	10	1250	
	Pre-pgp-9	g603.t1	g603.t1	6TM-ABC-6TM-ABC	4671	9	1290	
C	*Pre-mrp-1*	g585.t3	g585.t3	10TM-ABC-7TM-ABC	6592	13	1550	
	Pre-mrp-3	g2345.t1	g2345.t1	10TM-ABC-5TM-ABC	11329	16	1469	g2347 was merged with g2345
	Pre-mrp-4	g18857.t1	g18857.t1	10TM-ABC-5TM-ABC	6075	10	1596	
	Pre-mrp-5	g14752.t1	g14752.t1	7TM-ABC-6TM-ABC	4663	4	1506	
	Pre-mrp-6	g2536.t1	g2536.t1	7TM-ABC-6TM-ABC	4363	6	1361	
	Pre-mrp-7	g7823.t1	g7823.t1	11TM-ABC-6TM-ABC	5891	9	1490	
	Pre-mrp-8	g17310.t1	g17310.t1	11TM-ABC-6TM-ABC	5909	10	1550	
D	*Pre-pmp-1*	g18747.t1	g18747.t1	5TM-ABC	2383	6	659	
	Pre-pmp-2	g7109.t1	g7109.t1	5TM-ABC	2163	9	563	g7108 was merged with g7109; TM helices were improvement;
	Pre-pmp-3	g19415.t1	g19415.t1	6TM-ABC	2703	7	763	
	Pre-pmp-4	g20447.t1	g20447.t1	6TM-ABC	2466	8	706	
	Pre-pmp-5	g7104.t1	g7104.t1	5TM-ABC	2223	9	616	
	Pre-pmp-6	g17130.t1	g17130.t1	4TM-ABC	2670	10	633	
	Pre-pmp-7	g24041.t1	g24041.t1	4TM-ABC	2457	11	612	
E	*Pre-abce-1*	g3531.t1	g3531.t1	ABC-ABC	2186	6	615	
F	*Pre-abcf-1*	g10273.t2	g10273.t2	ABC-ABC	2335	6	633	
	Pre-abcf-2	g13187.t1	g13187.t1	ABC-ABC	2080	5	620	
	Pre-abcf-3	g17665.t1	g17665.t1	ABC-ABC	2509	7	735	
G	*Pre-wht-1*	g7599.t1	g7599.t1	ABC-6TM	2480	6	667	
	Pre-wht-2	g412.t1	g412.t1	ABC-5TM	2249	7	656	
	Pre-wht-4	g2514.t1	g2514.t1	ABC-0TM	886	3	250	
	Pre-wht-7	g18408.t1	g18408.t1	ABC-4TM	2472	11	639	No stop codon
	Pre-wht-8	g15295.t1	g15295.t1	ABC-6TM	2161	5	610	
H	*Pre-abch-1*	g3293.t1	g3293.t1	ABC-7TM	2380	8	627	

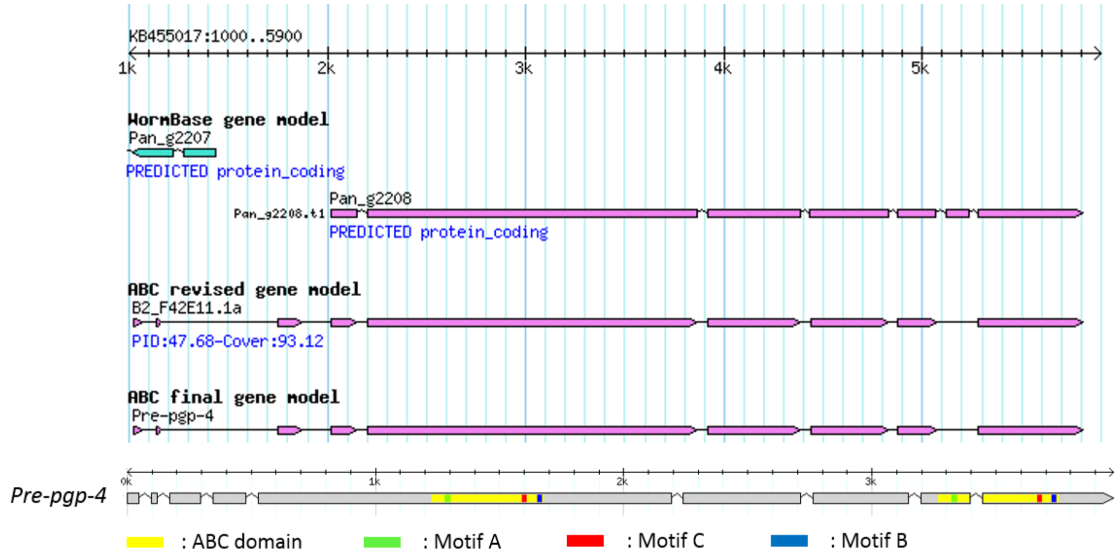
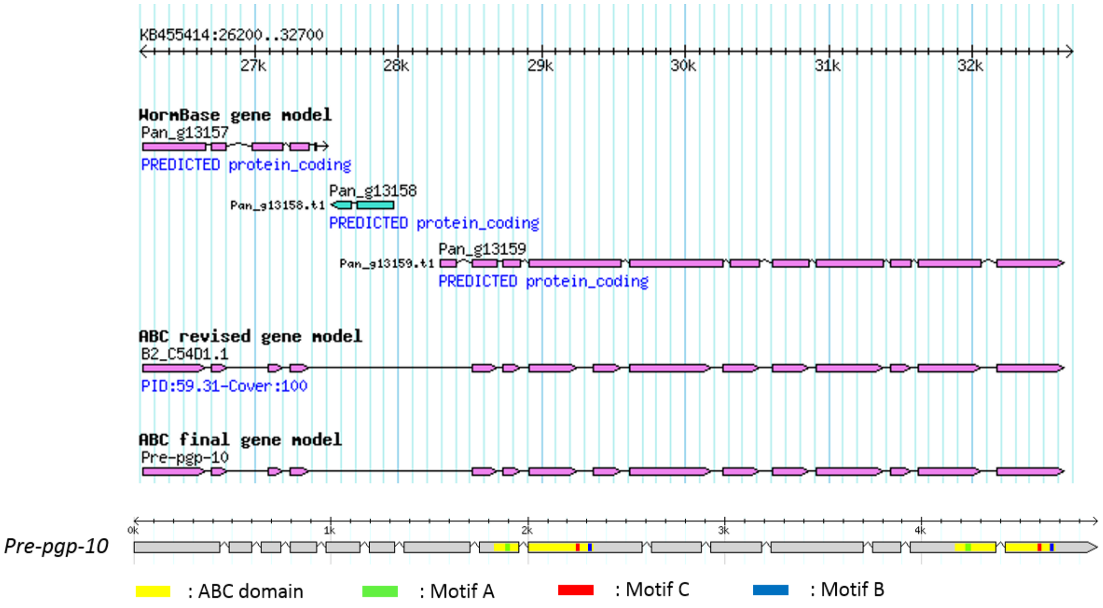
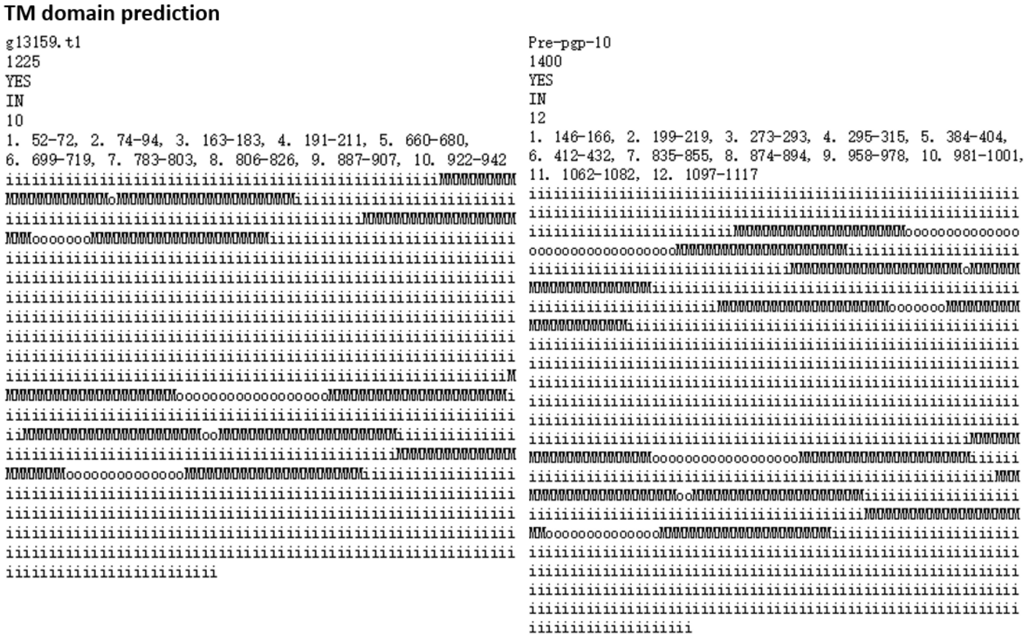


Figure 3.60: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Pan_g2208 was annotated as a full ABC transporter gene in subfamily B enconding one of its ABC domain was longer (181 aa) than expected. The defective ABC domain was improved (145 aa) in the revised gene model.





TM domain prediction

```
g13159.t1
1225
YES
IN
10
1. 52-72, 2. 74-94, 3. 163-183, 4. 191-211, 5. 660-680,
6. 699-719, 7. 783-803, 8. 806-826, 9. 887-907, 10. 922-942
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMM
MMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMM
MMMoooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiM
MMMMMMMMMMMMMMMMMMMMooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMi
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiMMMMMMMMMMMMMMMMMMMMMooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMM
MMMMMMMooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiii
```

```
Pre-pgp-10
1400
YES
IN
12
1. 146-166, 2. 199-219, 3. 273-293, 4. 295-315, 5. 384-404,
6. 412-432, 7. 835-855, 8. 874-894, 9. 958-978, 10. 981-1001,
11. 1062-1082, 12. 1097-1117
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooo
ooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMM
MMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooMMMMMMMMM
MMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMM
MMMMMMMMMMMMMMMooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMM
MMMMMMMMMMMMMMMMMMooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMM
MMooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiii
```

Figure 3.61: A representative case that the TM domain of an ABC transporter gene was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Pan_g13159 was annotated as a full ABC transporter gene in subfamily B. As a result of merging Pan_g13159 and Pan_g13157, the revised gene model had an increase number of TM helices (12), clustering in two TM domains.

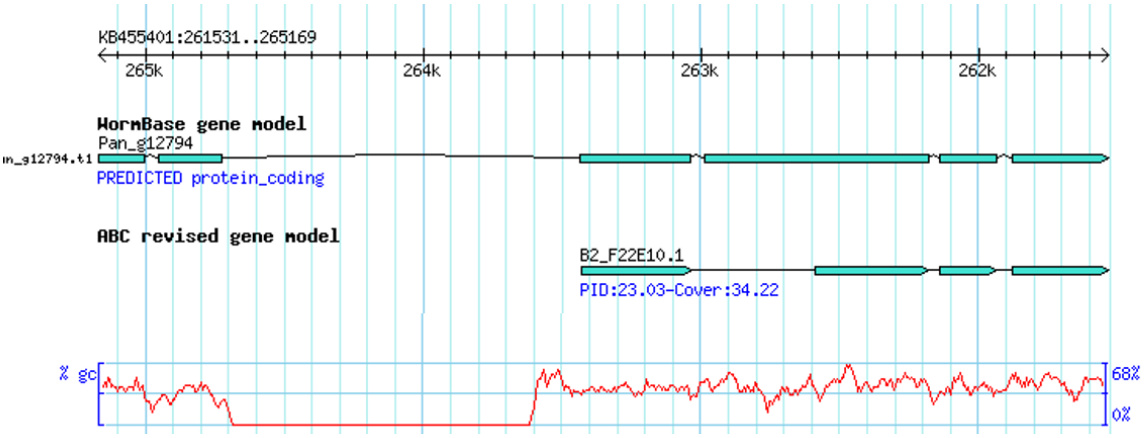


Figure 3.62: A representative case that sequencing error could result in incompleteness of an ABC transporter gene candidate

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. Pan_g12794 was annotated as a full ABC transporter gene in subfamily B but only had one satisfying ABC domain. By checking its genomic region, we found a sequencing gap within this gene, which might cause the incompleteness of this ABC transporter candidate.

Through phylogenetic analysis, we found 23 out of 59 ABC transporter genes in *P. redivivus* showed one-to-one orthologous relationship with the ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *P. redivivus* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.63). The totla number of ABC transporter genes in *P. redivivus* is very close to that in *C. elegans*. But less than half of these genes shared orthologous relationship with those in *C. elegans*, reflecting its phylogenetic distance to *C. elegans*. Subfamily B was more diverse between these two species. Most interestingly, *hmt-1,* a half ABC transporter gene in subfamily B, which is required for heavy metal detoxification in *C. elegans* (Schwartz et al. 2010), had four orthologs (*Pre-hmt-1*, *Pre-hmt-2*, *Pre-hmt-3* and *Pre-hmt-4*) in *P. redivivus*. This expansion in this particular gene may explain the high level of copper tolerance reported in *P. redivivus*, which has been shown to have higher tolerance to copper than *C. elegans* or *P. pacificus* (Boyd and Williams 2003).

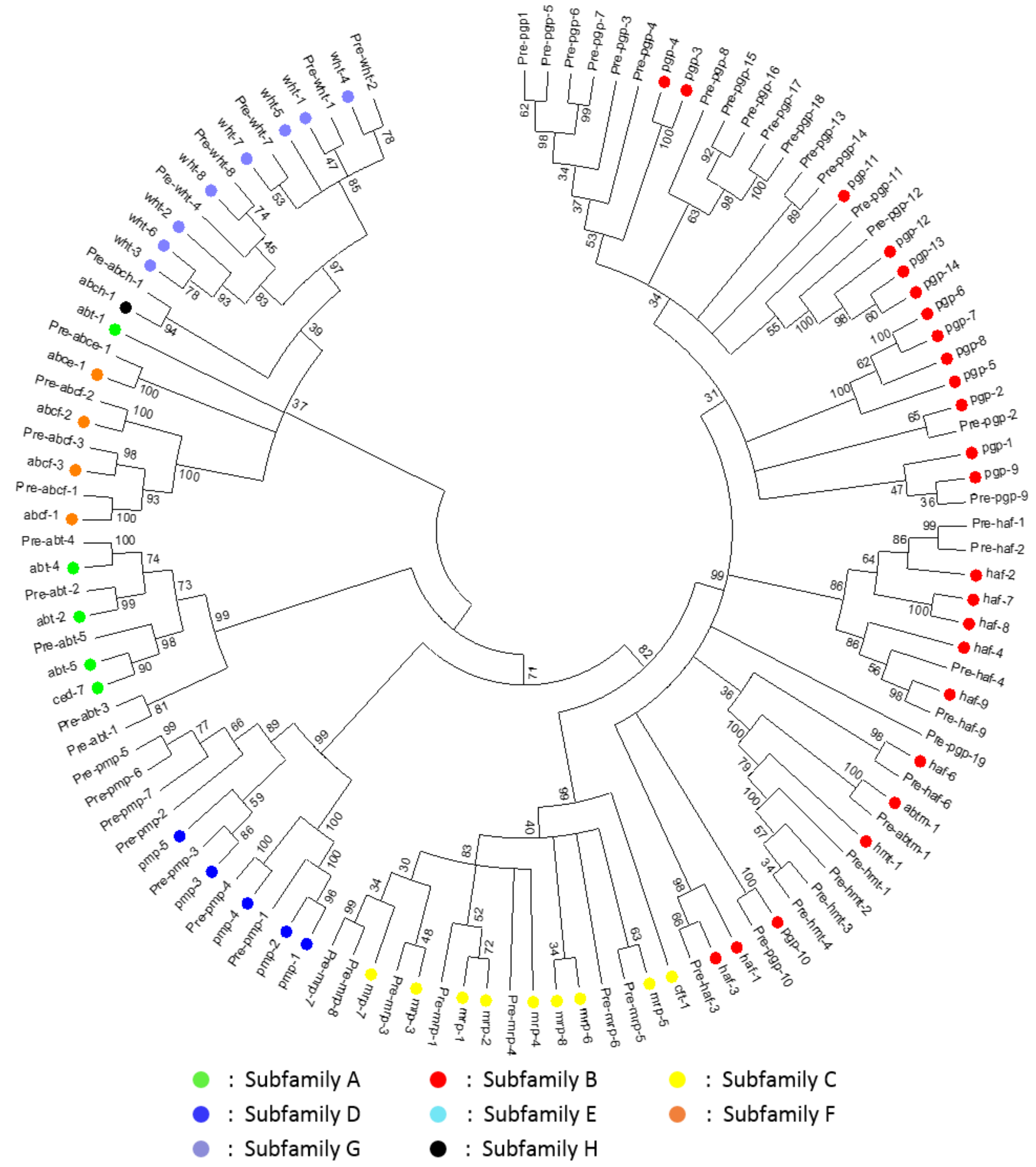


Figure 3.63: Phylogenetic analysis between *P. redivivus* and *C. elegans*

Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *P. redivivus* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *P. redivivus* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.17. Annotation of ABC transporter genes in *M. hapla*

M. hapla known as the Northern Root-Knot Nematode because it is a major pathogen of plants in cooler environments throughout the world (Wang et al. 2010). It represents not only the smallest nematode genome (54MB) but also the smallest metazoan genome. It defines a platform for elucidating mechanisms of parasitism (Opperman et al. 2008). After applying the annotation pipeline to *M. hapla*, we identified 29 ABC transporter gene candidates (26 candidates from InterProScan searches, three additional ones from BLAST searches). It is a much smaller number compared to that of the other nematode studied in this project, probably due to its small genome size. None of these candidates was due to contamination. 10 of 29 candidates were characterized to be high-quality ABC transporter genes. All of these 10 genes encoded appropriate TM domain (s). For the 19 defective candidates, we tried to further improve them. We generated five revised gene models of high-quality, two of which with only TM domain improved (Table 3.17). A representative case for exon improvement is MhA1_Contig271.frz3.gene24. It was annotated as an ABC transporter gene in subfamily F but had one of its ABC domains defective (119 aa). The revised gene model contained a slightly improved ABC domain (124 aa) (Figure 3.64). Another representative case for TM domain improvement showed in Figure 3.65. MhA1_Contig199.frz3.gene12 was annotated as a half ABC transporter gene in subfamily B. However, it lost a part of TM domain (only three TM helices). After improvement, we generated a new gene model by merging MhA1_Contig199.frz3.gene12 and MhA1_Contig199.frz3.gene14. The revised gene model encoded an improved TM domain with 10 TM helices (Figure 3.65). Three candidates that could not be improved. In summary, we annotated 24 high-quality ABC transporter genes in *M. hapla*, all of which had appropriate TM domain (s) (Table 3.17).

Table 3.17: High-quality ABC transporter genes in *M. hapla* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Mha-abt-4*	MhA1_Contig883.frz3.gene3	MhA1_Contig883.frz3.gene3	7TM-ABC-7TM-ABC	6182	24	1478	
B	*Mha-abtm-1*	MhA1_Contig898.frz3.gene6	MhA1_Contig898.frz3.gene6	6TM-ABC	2786	14	707	
	Mha-haf-1	MhA1_Contig933.frz3.gene12	MhA1_Contig933.frz3.gene12	5TM-ABC	3777	14	716	
	Mha-haf-2	MhA1_Contig1098.frz3.gene7	MhA1_Contig1098.frz3.gene7	8TM-ABC	3333	12	745	
	Mha-haf-6	MhA1_Contig1288.frz3.fgene2	MhA1_Contig1288.frz3.fgene2	5TM-ABC	3423	13	712	
	Mha-haf-9	MhA1_Contig199.frz3.gene12	MhA1_Contig199.frz3.gene12	10TM-ABC	5556	20	817	MhA1_Contig199.frz3.gene14 was merge with MhA1_Contig199.frz3.gene12; TM helices were improved; No start codon
	Mha-hmt-1	MhA1_Contig213.frz3.fgene3	MhA1_Contig213.frz3.fgene3	9TM-ABC	3052	17	702	TM helices were improved; No start codon
	Mha-pgp-10	MhA1_Contig0.frz3.gene79	MhA1_Contig0.frz3.gene79	TM-ABC-TM-ABC	10688	35	1205	MhA1_Contig0.frz3.gene82 was merged with MhA1_Contig0.frz3.gene79; No start codon
	Mha-pgp-2	MhA1_Contig76.frz3.fgene1	MhA1_Contig76.frz3.fgene1	TM-ABC-TM-ABC	9682	26	1286	
	Mha-pgp-3	MhA1_Contig622.frz3.fgene1	MhA1_Contig622.frz3.fgene1	TM-ABC-TM-ABC	7382	22	1409	
C	*Mha-mrp-1*	MhA1_Contig1584.frz3.fgene3	MhA1_Contig1584.frz3.fgene3	TM-ABC-TM-ABC	7003	32	1550	
	Mha-mrp-2	MhA1_Contig261.frz3.fgene3	MhA1_Contig261.frz3.fgene3	TM-ABC-TM-ABC	9363	31	1574	
	Mha-mrp-3	MhA1_Contig1566.frz3.gene5	MhA1_Contig1566.frz3.gene5	TM-ABC-TM-ABC	6610	26	1531	
	Mha-mrp-5	MhA1_Contig1743.frz3.fgene2	MhA1_Contig1743.frz3.fgene2	TM-ABC-TM-ABC	6698	24	1442	
	Mha-mrp-7	MhA1_Contig2079.frz3.fgene1	MhA1_Contig2079.frz3.fgene1	TM-ABC-TM-ABC	6095	21	1409	
	Mha-mrp-8	MhA1_Contig88.frz3.gene103	MhA1_Contig88.frz3.gene103	TM-ABC-TM-ABC	7081	33	1388	
D	*Mha-pmp-2*	MhA1_Contig1698.frz3.fgene1	MhA1_Contig1698.frz3.fgene1	5TM-ABC	3027	16	691	
	Mha-pmp-3	MhA1_Contig912.frz3.gene9	MhA1_Contig912.frz3.gene9	5TM-ABC	2826	11	611	Exons were improved
E	*Mha-abce-1*	MhA1_Contig915.frz3.gene13	MhA1_Contig915.frz3.gene13	ABC-ABC	2634	12	566	
F	*Mha-abcf-1*	MhA1_Contig29.frz3.gene9	MhA1_Contig29.frz3.gene9	ABC-ABC	3316	12	596	
	Mha-abcf-2	MhA1_Contig271.frz3.gene24	MhA1_Contig271.frz3.gene24	ABC-ABC	2494	14	600	Exons were improved
	Mha-abcf-3	MhA1_Contig1737.frz3.gene5	MhA1_Contig1737.frz3.gene5	ABC-ABC	3149	16	689	
G	*Mha-wht-1*	MhA1_Contig2134.frz3.gene2	MhA1_Contig2134.frz3.gene2	ABC-2TM	2510	9	533	
	Mha-wht-2	MhA1_Contig107.frz3.gene5	MhA1_Contig107.frz3.gene5	ABC-7TM	3039	11	637	

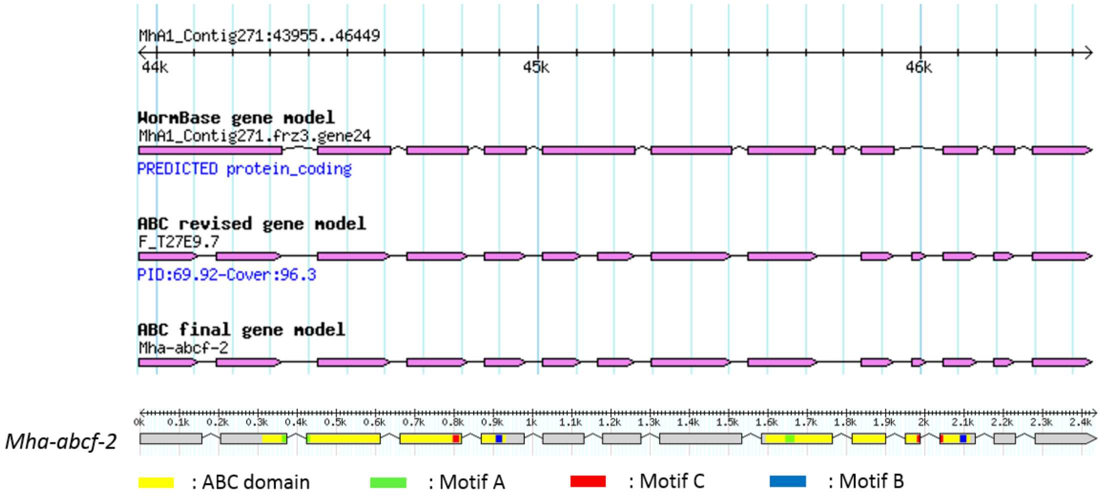
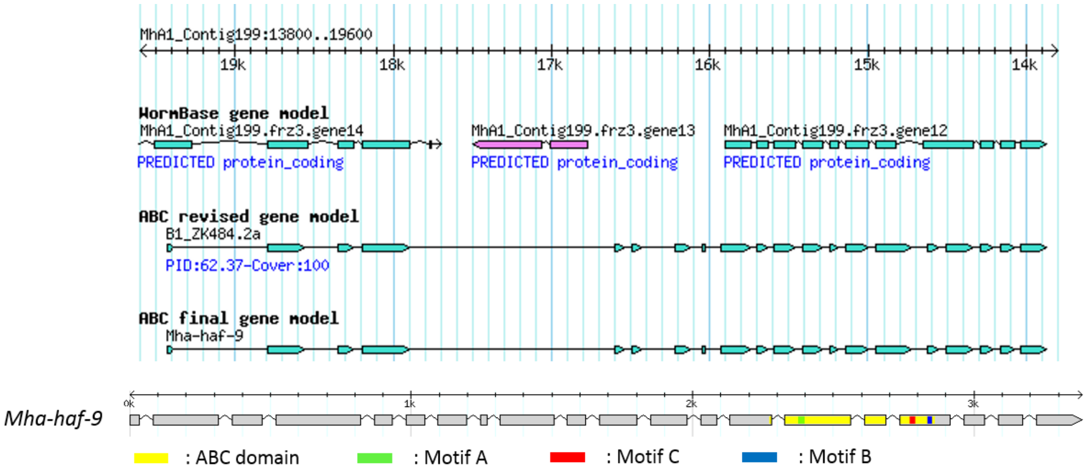


Figure 3.64: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. MhA1_Contig271.frz3.gene24 was annotated as an ABC transporter gene in subfamily F but had one of its ABC domains defective (119 aa). The revised gene model contained two ABC domains, one of which was slightly improved (124 aa).



TM domain prediction

```
MhA1_Contig199.frz3.gene12
488
YES
IN
3
1. 11-31, 2. 33-53, 3. 115-135
iiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMM
MMMMMMMMMMMMMMMooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooo
```

```
Mha-haf-9
816
YES
IN
10
1. 13-33, 2. 61-81, 3. 89-109, 4. 123-143, 5. 197-217,
6. 238-258, 7. 287-307, 8. 326-346, 9. 348-368, 10. 430-450
iiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooo
MMMMMMMMMMMMMMMMMMMMMiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooo
ooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooMMM
MMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMM
MMMMMMMooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiMMMMMMMMMMMMM
MMMMMMMMoooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

Figure 3.65: A representative case that the TM domain of an ABC transporter gene was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. MhA1_Contig199.frz3.gene12 was annotated as a half ABC transporter gene in subfamily B. However, it lost a part of TM domain with only three TM helices. The revised gene model as a result of merging MhA1_Contig199.frz3.gene12 and MhA1_Contig199.frz3.gene14 into one, had an improved TM domain with 10 TM helices.

Through phylogenetic analysis, we found 19 out of 24 ABC transporter genes in *M. hapla* showed one-to-one orthologous relationship with the ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *M. hapla* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.66). Except for subfamily E and subfamily F, other subfamilies had different levels of gene contractions in *M. hapla* when compared to those in *C. elegans*. It is consistent with significantly fewer genes encoded by *M. hapla* than those encoded by the free-living nematode *C. elegans* (Opperman et al. 2008)*.*

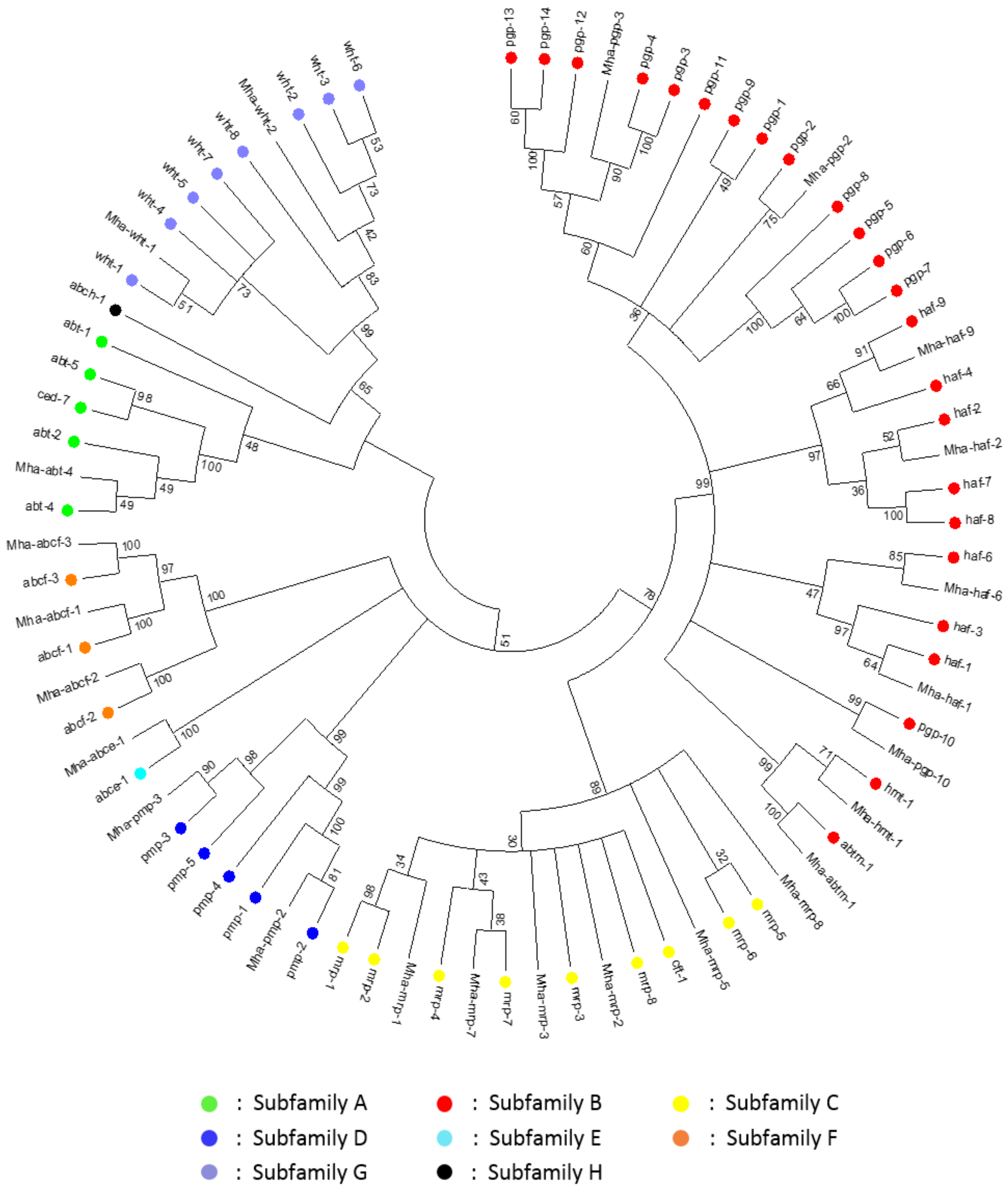


Figure 3.66: Phylogenetic analysis between *M. hapla* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *M. hapla* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *M. hapla* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.18. Annotation of ABC transporter genes in *M. incognita*

The Southern root-knot nematode *M. incognita*, a close nematode species to *M. hapla*, is able to infect the roots of almost all cultivated plants (Trudgill and Blok 2001), making this nematode a key model system for the understanding of metazoan adaptations to plant parasitism (Caillaud et al. 2008). After applying the annotation pipeline to *M. incognita*, we obtained totally 36 ABC transporter gene candidates (30 candidates from InterProScan searches, six additional ones from BLAST searches), similar to that in *M. hapla*. None of these candidates was due to contamination. 10 of 36 candidates were characterized to be high-quality ABC transporter genes. All these 10 gene encoded appropriate TM domain (s). For the 26 defective candidates, we tried to further improve them. We generated 13 revised gene models of high-quality, five of which with only TM domain improved (Table 3.18). Among those improved genes, Min01947 was annotated as an ABC transporter gene in subfamily E but one of its predicted ABC domains was defective only with a length of 37 aa. Its adjacent gene, Min01946 was also annotated as a member in subfamily E encoding a defective ABC domain (33 aa). After improvement, the defective ABC domain was revised to be a high-quality one with a length of 140 aa (Figure 3.67). The TM domain of a half ABC transporter gene in subfamily B was improved by merging Minc12057 (containing a high-quality ABC domain) and Minc12058 (containing proper TM domain), shown in Figure 3.68. Five candidates without technique errors could not be improved. Taking together, we annotated 24 high-quality ABC transporter genes in *M. incognita*, all of which had appropriate TM domain (s) (Table 3.18).

Table 3.18: High-quality ABC transporter genes in *M. incognita* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Min-abt-5*	Minc16660	Minc16660	7TM-ABC-6TM-ABC	8276	23	1223	Minc16661 was merged with Minc16660; No start codon
B	*Min-abtm-1*	Minc12058	Minc12058	6TM-ABC	2498	14	614	TM helices were improved
	Min-abtm-2	Minc13572	Minc13572	7TM-ABC	4145	18	770	
	Min-haf-1	Minc07785	Minc07785	5TM-ABC	5950	13	671	No stop codon
	Min-haf-2	Minc15235	Minc15235	8TM-ABC	3130	14	799	
	Min-haf-3	Minc08936	Minc08936	5TM-ABC	5668	15	648	Minc08937 was merged with Minc08936; TM helices were improved
	Min-haf-4	Minc02288	Minc02288	5TM-ABC	3519	12	603	Exons were improved; No start codon
	Min-haf-9	Minc04968	Minc04968	9TM-ABC	4754	21	897	No stop codon
	Min-pgp-10	Minc01351	Minc01351	6TM-ABC-4TM-ABC	10412	35	1195	
	Min-pgp-2	Minc04234	Minc04234	6TM-ABC-6TM-ABC	8718	28	1416	No stop codon
	Min-pgp-3	Minc14983	Minc14983	6TM-ABC-6TM-ABC	8847	28	1267	TM helices were improved
	Min-pgp-4	Minc13430	Minc13430	6TM-ABC-8TM-ABC	7712	22	1213	Minc13431 was merged with Minc13430
C	*Min-mrp-1*	Minc02886	Minc02886	6TM-ABC-5TM-ABC	9575	25	1138	Minc02887 was merged with Minc02886; No start codon
	Min-mrp-3	Minc07240	Minc07240	13TM-ABC-6TM-ABC	8742	32	1750	
	Min-mrp-7	Minc15796	Minc15796	4TM-ABC-6TM-ABC	6997	19	1137	TM helices were improved; No start codon
D	*Min-pmp-1*	Minc00863	Minc00863	5TM-ABC	3989	17	640	TM helices were improved; No stop codon
	Min-pmp-2	Minc09635	Minc09635	5TM-ABC	3938	15	627	No stop codon
	Min-pmp-3	Minc07030	Minc07030	5TM-ABC	2913	12	633	Exons were improved; No start codon
E	*Min-abce-1*	Minc01946	Minc01946	ABC-ABC	6049	14	593	Minc01947 was merged with Minc01946
	Min-abce-2	Minc03998	Minc03998	ABC-ABC	6340	13	618	No stop codon
F	*Min-abcf-1*	Minc16222	Minc16222	ABC-ABC	6028	12	645	No stop codon
	Min-abcf-2	Minc05867	Minc05867	ABC-ABC	2954	13	625	No start and stop codon
	Min-abcf-3	Minc17133	Minc17133	ABC-ABC	3139	17	700	Exons were improved;
	Min-abcf-4	Minc03577	Minc03577	ABC-ABC	2709	14	607	Minc03578 was merged with Minc03577

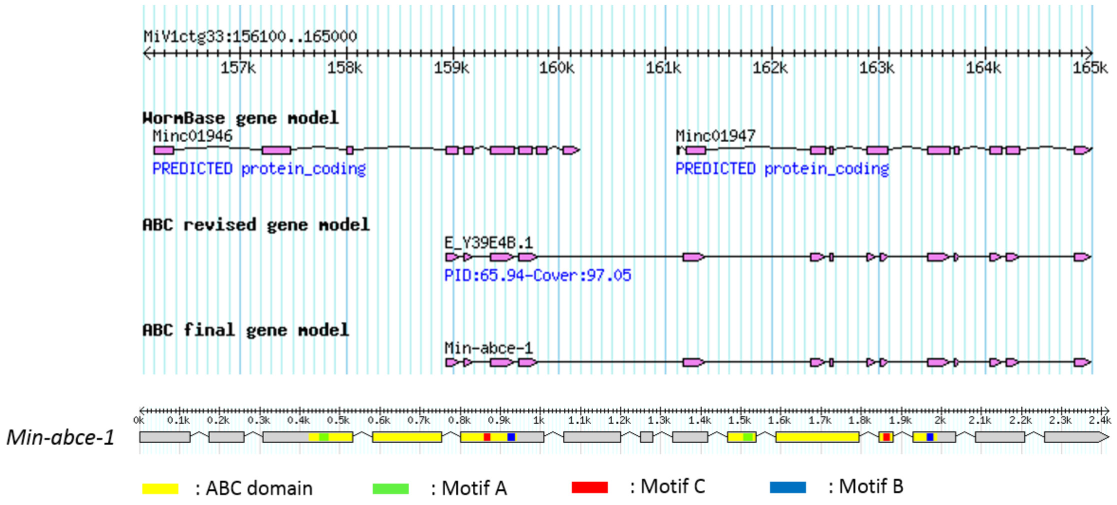
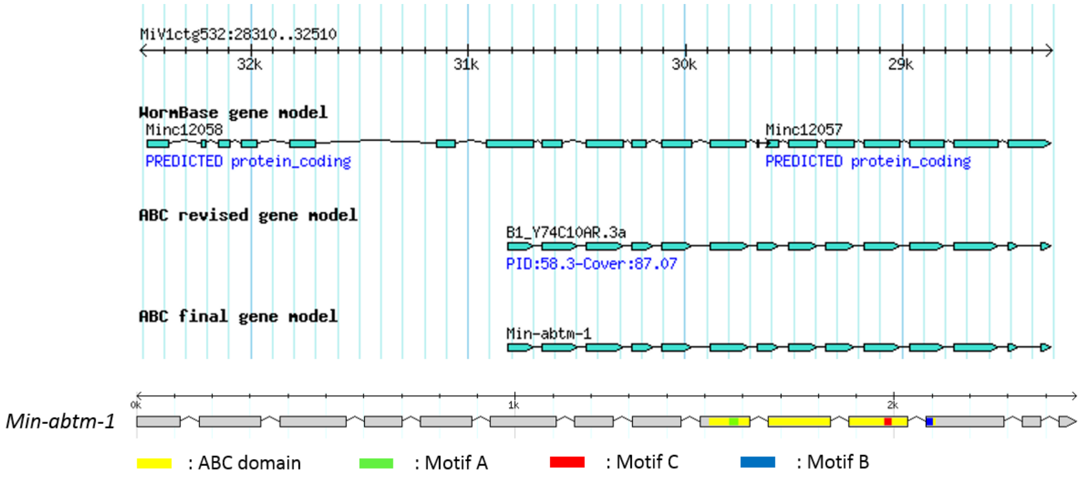
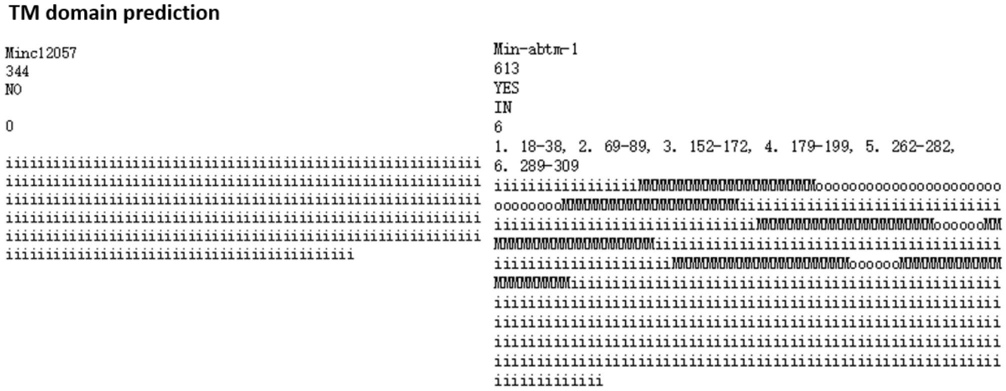


Figure 3.67: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Min01947 was annotated as an ABC transporter gene in subfamily E but one of its predicted ABC domain was defective only with a length of 37 aa. The revised gene model was a result of merging Min01947 and Minc09146 and encoded two typical ABC domains.




TM domain prediction

```
Minc12057
344
NO

0

iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

```
Min-abtm-1
613
YES
IN
6
1. 18-38, 2. 69-89, 3. 152-172, 4. 179-199, 5. 262-282,
6. 289-309
iiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooo
ooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooMM
MMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooMMMMMMMMMMMM
MMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiii
```

Figure 3.68: A representative case that the TM domain of an ABC transporter gene was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The TM domain of a half ABC transporter gene annotated in subfamily B was improved by merging Minc12057 (containing a high-quality ABC domain) and Minc12058 (containing proper TM domain).

Through phylogenetic analysis, we found 13 out of 24 ABC transporter genes in *M. incognita* showed one-to-one orthologous relationship with the ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *M. incognita* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.69). There was no annotated ABC transporter gene both in subfamily G and H in *M. incognita*. Gene contractions happened in subfamily A, B, C and D in *M. incognita* compared to those of *C. elegans.* Gene expansion occurred in subfamily E and subfamily F in *M. incognita*, which are discussed more in the next chapter (Chapter 4).

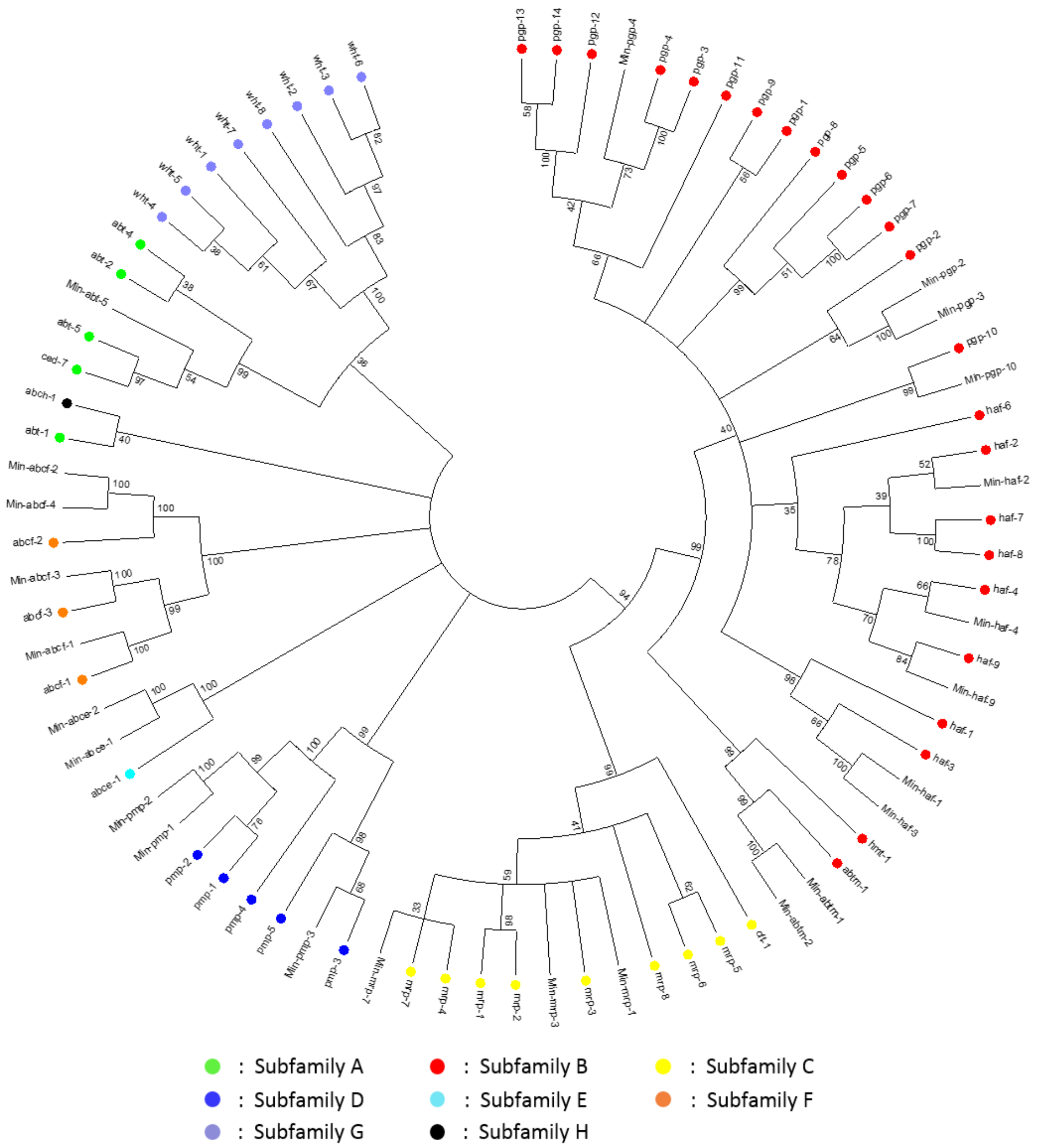


Figure 3.69: Phylogenetic analysis between *M. incognita* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *M. incognita* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *M. incognita* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.19. Annotation of ABC transporter genes in *A. suum*

A. suum, is also known as the large pig roundworm or the large white worm. It is a species of parasitic roundworms that infects pigs and wild boars worldwide (Jex et al. 2011). After applying the annotation pipeline to *A. suum*, we identified in total 78 ABC transporter gene candidates with 52 candidates from InterProScan searches and 26 additional ones from BLAST searches. One of these candidate, GS_12616 was due to contamination which were excluded from our further analysis. Among 77 candidates, 17 were high-quality ABC transporter genes. All of these 17 genes also encoded appropriate TM domain (s). For the 60 defective candidates, we tried to further improve each of them. 20 revised gene model were high-quality, five of which with only TM domain improved (Table 3.19). GS_10626 was annotated as a half ABC transporter gene in subfamily G. But it did not encoded ABC domain. After improvement, a revised gene model was obtained with a high-quality ABC domain encoded by the newly generated exons (Figure 3.70). Another representative case is for merging three candidate genes (GS_08719, GS_12341 and GS_19586) into a high-quality ABC transporter gene. This revised gene model encoded two high-quality ABC domains (Figure 3.71). TM domain could also be improved by merging adjacent candidates. For example, GS_16376, was merged with GS_22937 and GS_05523 to form a half ABC transporter gene in subfamily B (Figure 3.72). After improvement, 17 candidates could not be further improved to be high-quality ABC transporter genes. One of them with a sequencing gap could be a complete ABC transporter gene when genome assembly is improved (Figure 3.73). In summary, we annotated 38 high-quality ABC transporter genes in *A. suum,* 35 of which had proper TM domain (s) (Table 3.19).

Table 3.19: High-quality ABC transporter genes in *A. suum* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Asu-abt-1*	GS_17771	GS_17771	8TM-ABC-4TM-ABC	34299	20	1060	GS_09999 was merged with GS_17771; No start codon
	Asu-abt-2	GS_05132	GS_05132	7TM-ABC-10TM-ABC	36450	32	1471	No start and stop codon
	Asu-abt-4	GS_10190	GS_10190	7TM-ABC-8TM-ABC	30381	36	1745	No stop codon
B	*Asu-abtm-1*	GS_01865	GS_01865	6TM-ABC	13317	15	788	
	Asu-haf-2	GS_00792	GS_00792	10TM-ABC	9748	12	790	
	Asu-haf-3	GS_09342	GS_09342	5TM-ABC	12000	13	578	
	Asu-haf-4	GS_08782	GS_08782	9TM-ABC	10933	16	772	
	Asu-haf-6	GS_05427	GS_05427	6TM-ABC	14736	12	642	GS_24168, L3E_00465 and GS_05427 were merged with GS_12936
	Asu-haf-9	GS_18912	GS_18912	9TM-ABC	20576	17	822	
	Asu-hmt-1	GS_16376	GS_16376	9TM-ABC	16185	15	783	GS_22937 was merged with GS_16376; TM helices were improved
	Asu-pgp-1	GS_07518	GS_07518	6TM-ABC-6TM-ABC	17875	23	1169	
	Asu-pgp-11	GS_08285	GS_08285	5TM-ABC-6TM-ABC	21663	24	1237	GS_22608 was merged with GS_08285
	Asu-pgp-12	GS_20427	GS_20427	6TM-ABC-6TM-ABC	24883	22	1209	No start and stop codon
	Asu-pgp-13	GS_00985	GS_00985	6TM-ABC-4TM-ABC	18212	24	1248	GS_08822 was merged with GS_00985; No start codon
	Asu-pgp-14	GS_22685	GS_22685	6TM-ABC-6TM-ABC	23415	24	1280	No start and stop codon
	Asu-pgp-2	GS_12341	GS_12341	7TM-ABC-8TM-ABC	24750	26	1266	GS_19586 and GS_08719 were merged with GS_12341; No start codon
	Asu-pgp-3	GS_01681	GS_01681	6TM-ABC-6TM-ABC	25926	27	1227	GS_08942 and GS_19694 were merged with GS_01681; No start codon
	Asu-pgp-4	GS_17968	GS_17968	4TM-ABC-2TM-ABC	19755	24	1044	Exons were improved; No start codon
	Asu-pgp-5	GS_16411	GS_16411	6TM-ABC-5TM-ABC	20756	23	1169	Exons were improved; No start codon
	Asu-pgp-6	GS_21361	GS_21361	6TM-ABC-6TM-ABC	17669	25	1237	No start and stop codon
C	*Asu-mrp-1*	GS_08473	GS_08473	2TM-ABC-5TM-ABC	22172	21	1090	
	Asu-mrp-3	GS_06310	GS_06310	10TM-ABC-6TM-ABC	36761	30	1469	GS_24095 was merged with GS_06310
	Asu-mrp-4	GS_20097	GS_20097	10TM-ABC-5TM-ABC	25399	34	1544	Exons were improved
	Asu-mrp-5	GS_12380	GS_12380	8TM-ABC-8TM-ABC	23177	25	1256	
	Asu-mrp-6	GS_07037	GS_07037	5TM-ABC-7TM-ABC	32026	29	1236	No start and stop codon
	Asu-mrp-7	GS_03610	GS_03610	10TM-ABC-5TM-ABC	43503	30	1423	GS_03610 was merged with GS_08708

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
D	*Asu-pmp-1*	GS_11334	GS_11334	4TM-ABC	21569	14	652	TM helices were improved
	Asu-pmp-2	GS_00403	GS_00403	4TM-ABC	16260	15	594	
	Asu-pmp-3	GS_16618	GS_16618	6TM-ABC	12582	17	717	
	Asu-pmp-5	GS_09393	GS_09393	5TM-ABC	9892	15	625	
E	*Asu-abce-1*	GS_14232	GS_14232	ABC-ABC	14670	15	580	GS_21136 was merged with GS_14232; No start codon
F	*Asu-abcf-1*	GS_01526	GS_01526	ABC-ABC	13833	16	616	GS_01524 was merged with GS_01526; No start codon
	Asu-abcf-2	GS_14282	GS_14282	ABC-ABC	8014	13	538	
	Asu-abcf-3	GS_17626	GS_17626	ABC-ABC	11596	17	712	
G	*Asu-wht-1*	GS_10626	GS_10626	ABC-6TM	11747	17	562	Exons were improved
	Asu-wht-2	GS_05613	GS_05613	ABC-6TM	9592	13	576	GS_12024 and GS_11174 were merged with GS_05613; TM helices were improved
	Asu-wht-7	GS_05172	GS_05172	ABC-8TM	11025	17	621	GS_05136 and GS_19305 were merged with GS_05172; TM helices were improved; No start codon
H	*Asu-abch-1*	GS_05146	GS_05146	ABC-0TM	7555	9	351	

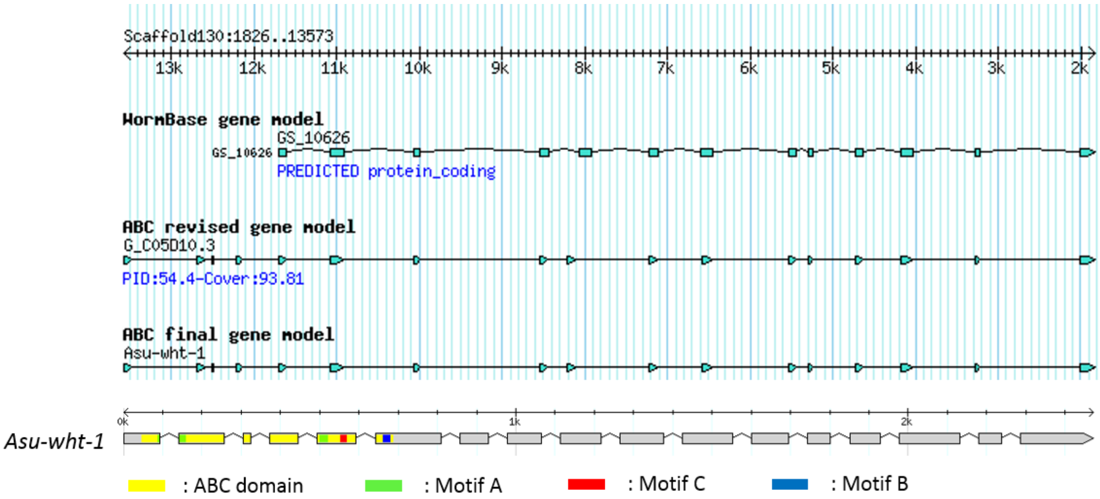


Figure 3.70: A representative case that the exons of one defective ABC transporter gene were improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. GS_10626 was annotated as a half ABC transporter gene in subfamily G but without any predicted ABC domain. The revised gene model had a high-quality ABC domain encoded by the newly generated exons.

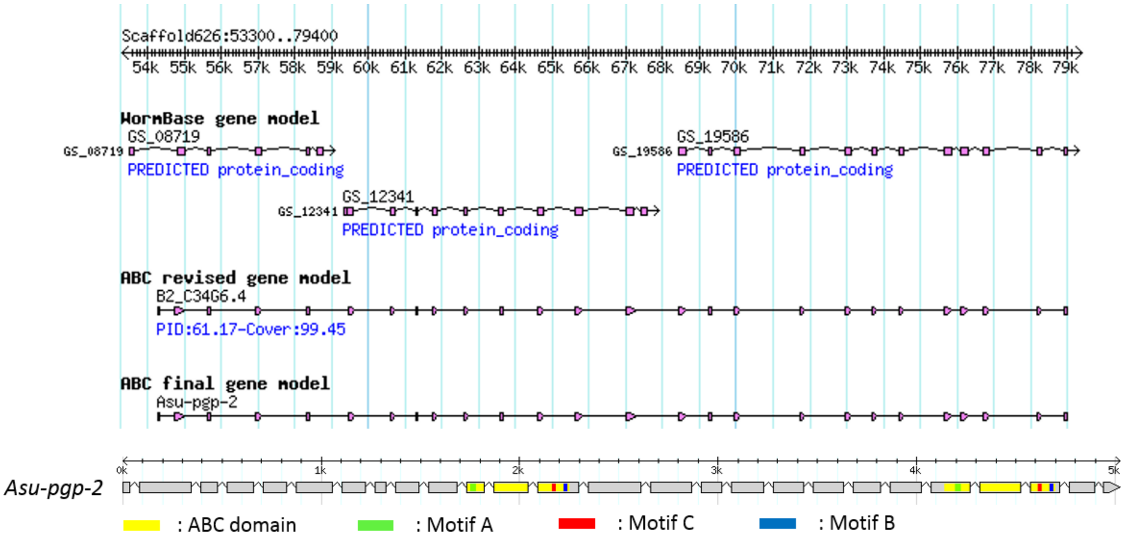
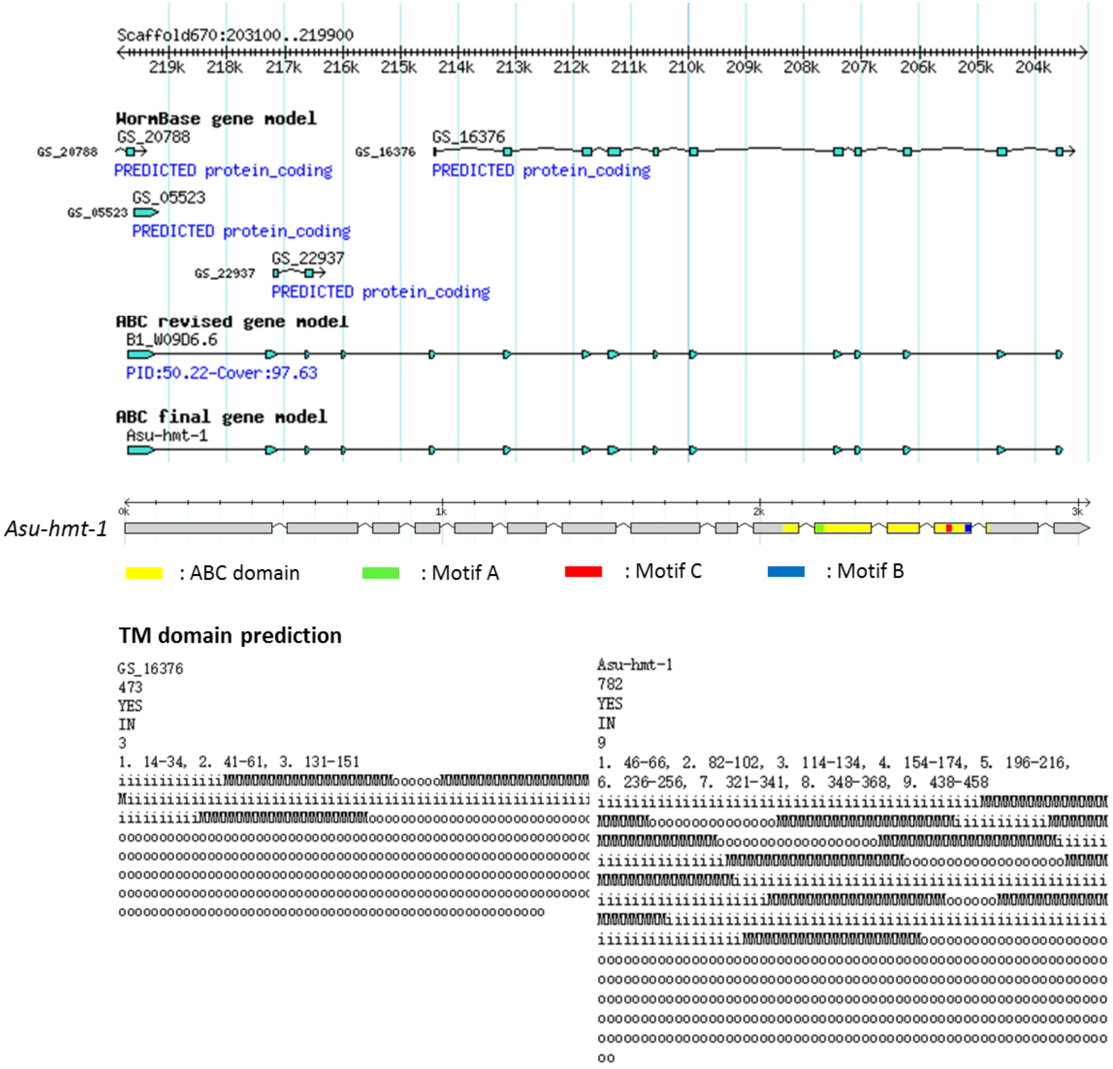


Figure 3.71: A representative case that three adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. GS_08719, GS_12341 and GS_19586 were merged into one single ABC transporter gene which encoded two high-quality ABC domains, making the revised gene a high-quality ABC transporter gene in subfamily B.



```
GS_16376
473
YES
IN
3
1. 14-34, 2. 41-61, 3. 131-151
iiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooMMMMMMMMMMMMMMMMMMMM
Miiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooo
```

```
Asu-hmt-1
782
YES
IN
9
1. 46-66, 2. 82-102, 3. 114-134, 4. 154-174, 5. 196-216,
6. 236-256, 7. 321-341, 8. 348-368, 9. 438-458
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMM
MMMMMMoooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiMMMMMMM
MMMMMMMMMMMMMMoooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiii
iiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooMMMMM
MMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooMMMMMMMMMMMMM
MMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oo
```

Figure 3.72: A representative case that the TM domain of an ABC transporter gene was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. GS_16376, was merged with GS_22937 and GS_05523 to form a half ABC transporter gene in subfamily B. The revised gene model encoded improved TM domain, nine TM helices compared to three in the original gene model.

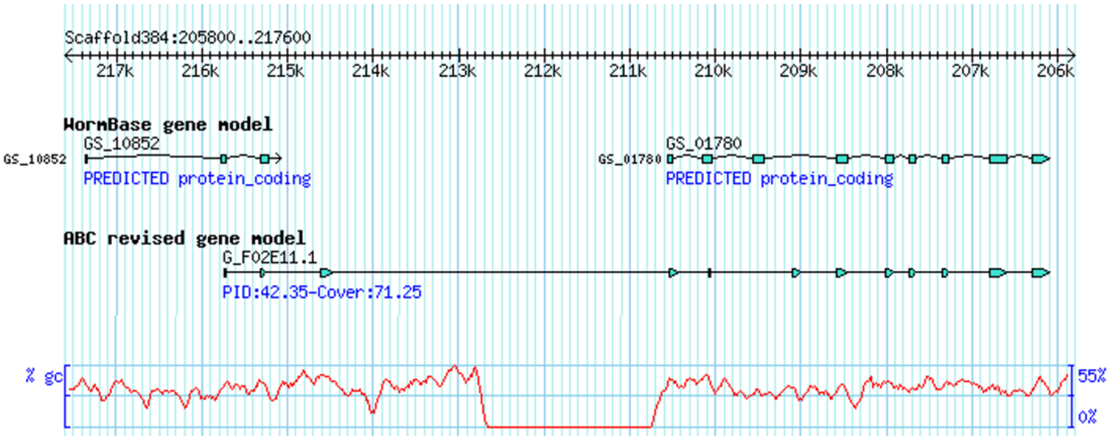


Figure 3.73: A representative case that sequencing error could result in incompleteness of an ABC transporter gene candidate

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. GS_01780 was annotated as a half transporter gene in subfamily G but had a defective ABC domain (70 aa) due to sequencing gap.

Through phylogenetic analysis, we found 23 out of 38 ABC transporter genes in *A. suum* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *A. suum* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.74). Gene contration in all the subfamilies except for subfamily E and subfamily F in *A. suum* compared to those in *C. elegans*. ABC transporter genes in *A. suum* were generally much longer than those in *C. elegans*, relating primarily to expansions of intronic regions. This observation is consistent with genome assembly and annotation results which showed a 273 Mb genome but a total number of 18500 protein-coding genes in *A. suum* (Jex et al. 2011).

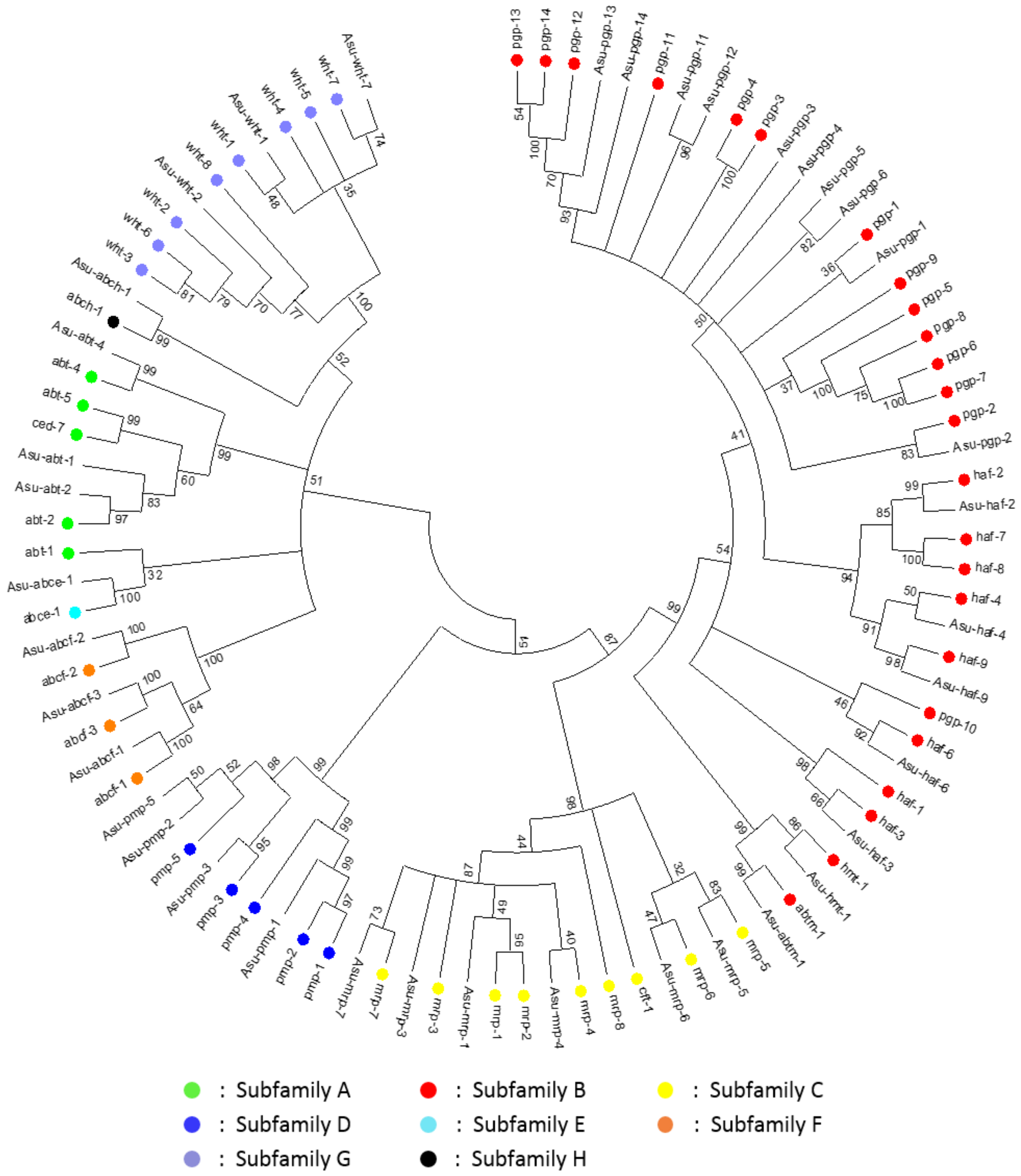


Figure 3.74: Phylogenetic analysis between *A. suum* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *A. suum* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *A. suum* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.20. Annotation of ABC transporter genes in *L. loa*

L. loa, or the African eyeworm, is a filarial worm that causes severe eye disorders in humans. Unlike most filariae, *L.loa* does not contain the obligate intracellular *Wolbachia* endosymbiont (Desjardins et al. 2013). The worm larvae are transmitted to humans through Chrysops fly bites (Padgett and Jacobsen 2008). *L. loa* affects an estimated 13 million people and causes chronic infection usually characterized by localized angioedema (Calabar swelling) and/or subconjunctival migration of adult worms across the eye (Abraham et al. 2001; Desjardins et al. 2013). After applying the annotation pipeline to *L .loa*, we obtained 32 ABC transporter gene candidates (26 candidates from InterProScan searches, six additional ones from BLAST searches). Of all the 32 candidates, none was due to contamination. After examining the quality of all candidates, we identified 12 high-quality ABC transporter genes. All of these 12 gene also encoded appropriate TM domain (s). For the defective candidates, we tried to improve their gene models. We generated seven revised gene models of high-quality, three of which with only TM domain improved (Table 3.20). For example, LOAG_18368 and LOAG_09104, both of which had one high-quality ABC domain, were merged together into a single high-quality ABC transporter gene. The revised gene was characterized as a full ABC transporter gene in subfamily C (Figure 3.75). Another representative case is for TM domain improvement, LOAG_07083 was annotated as a half ABC transporter gene in subfamily D but it did not encode any TM domain. After improvement, LOAG_07084 was merged with LOAG_07083, forming a high-quality ABC transporter gene with six TM helices in TM domain (Figure 3.76). For the remaining 6 candidates, five of them could not be further improved to be high-quality ABC transporter genes. One defective candidate which was located in the end of the contig could be a complete ABC transporter gene when the genome assembly is improved (Figure 3.77). Taking together, we annotated 20 high-quality ABC transporter genes in *L .loa*, all of which had appropriate TM domain (s) (Table 3.20).

Table 3.20: High-quality ABC transporter genes in *L. loa* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Llo-abt-2*	EJD74108.1	EJD74108.1	9TM-ABC-7TM-ABC	18916	36	1645	EFO26764.2 was merged with EJD74108.1
	Llo-abt-4	EFO28467.1	EFO28467.1	7TM-ABC-11TM-ABC	20299	32	1530	EFO28468.2, EFO28469.2 and EFO28471.1 were merged with EFO28467.1; TM helices were improved
	Llo-abt-5	EJD76542.1	EJD76542.1	8TM-ABC-9TM-ABC	15086	27	1274	EJD76543.1 was merged with EJD76542.1; TM helices were improved; No start codon
B	*Llo-abtm-1*	EFO16228.2	EFO16228.2	5TM-ABC	5644	14	754	
	Llo-haf-1	EJD76621.1	EJD76621.1	6TM-ABC	6340	15	677	
	Llo-haf-2	EJD75435.1	EJD75435.1	9TM-ABC	5022	11	757	
	Llo-pgp-10	EFO28095.2	EFO28095.2	7TM-ABC-6TM-ABC	13387	33	1407	
	Llo-pgp-11	EJD76051.1	EJD76051.1	4TM-ABC-7TM-ABC	10991	25	1188	EJD76051.1 was merged with EJD76051.1; No start codon
	Llo-pgp-12	EFO24761.1	EFO24761.1	6TM-ABC-6TM-ABC	10714	25	1280	
C	*Llo-mrp-1*	EFO25754.2	EFO25754.2	11TM-ABC-5TM-ABC	16369	33	1565	
	Llo-mrp-5	EFO28128.2	EFO28128.2	9TM-ABC-6TM-ABC	11609	30	1473	
	Llo-mrp-7	EJD74305.1	EJD74305.1	10TM-ABC-5TM-ABC	13730	30	1375	EFO19390.1 was merged with EJD74305.1
D	*Llo-pmp-3*	EFO19658.2	EFO19658.2	6TM-ABC	6323	16	670	
	Llo-pmp-4	EFO16988.2	EFO16988.2	5TM-ABC	6071	16	703	
	Llo-pmp-5	EFO21404.1	EFO21404.1	6TM-ABC	8145	13	559	EFO21405.1 was merged with EFO21404.1; TM helices were improved
E	*Llo-abce-1*	EFO24283.1	EFO24283.1	ABC-ABC	5590	13	619	
F	*Llo-abcf-1*	EFO24632.2	EFO24632.2	ABC-ABC	4779	16	642	
	Llo-abcf-2	EFO20731.1	EFO20731.1	ABC-ABC	5145	16	629	
	Llo-abcf-3	EJD74783.1	EJD74783.1	ABC-ABC	4577	17	702	EJD74784.1 was merged with EJD74783.1
G	*Llo-wht-4*	EJD75463.1	EJD75463.1	ABC-4TM	5992	13	480	

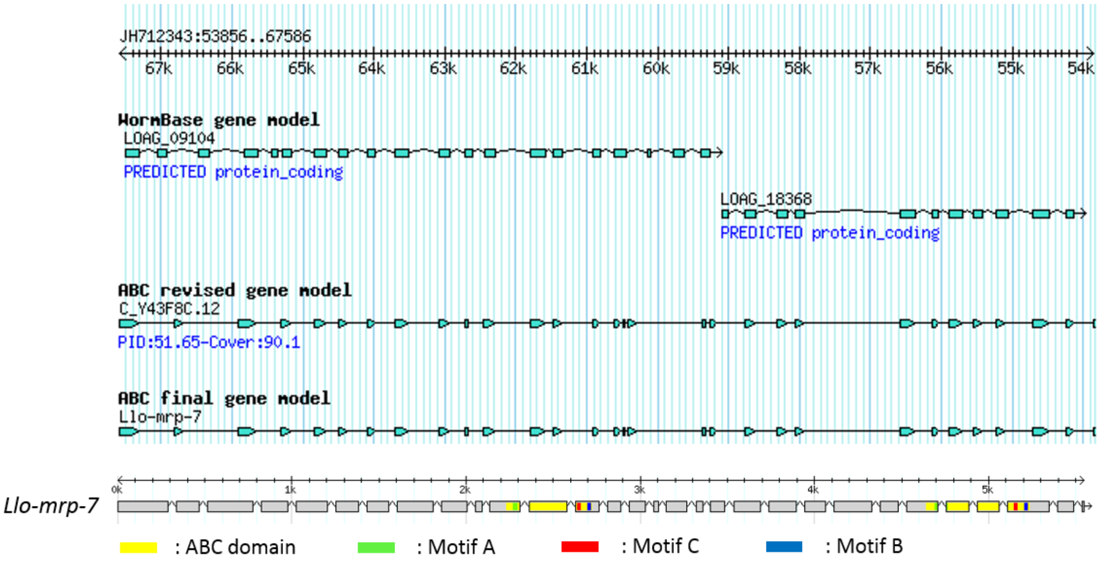
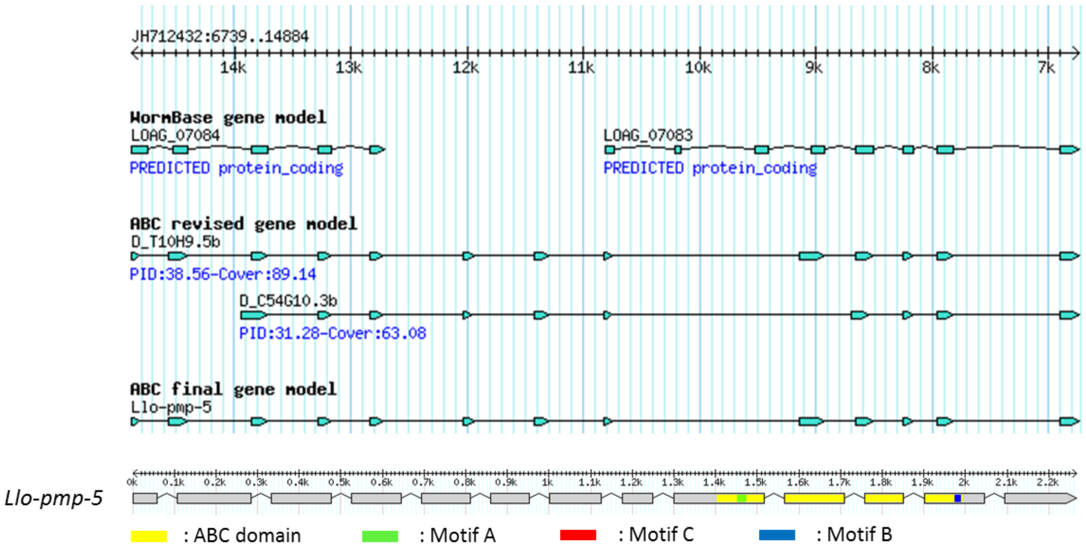
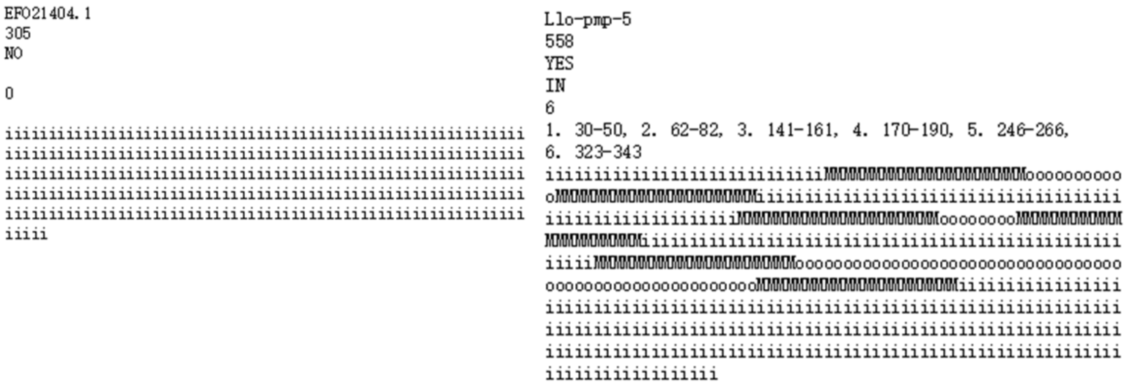


Figure 3.75: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. LOAG_18368 and LOAG_09104, both of which had one high-quality ABC domain, were merged together into a full ABC transporter gene in subfamily C.



TM domain prediction



```
EF021404.1
305
NO

0

iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiii
```

```
Llo-pmp-5
558
YES
IN
6
1. 30-50, 2. 62-82, 3. 141-161, 4. 170-190, 5. 246-266,
6. 323-343
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooo
oMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooMMMMMMMMMMM
MMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiii
```

Figure 3.76: A representative case that the TM domain of an ABC transporter gene was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. LOAG_07083 was annotated as a half ABC transporter gene in subfamily D but it did not contain any TM domain. The revised gene model was a result of merging LOAG_07084 and LOAG_07083 and was annotated as a high-quality ABC transporter gene with six TM helices in TM domain.

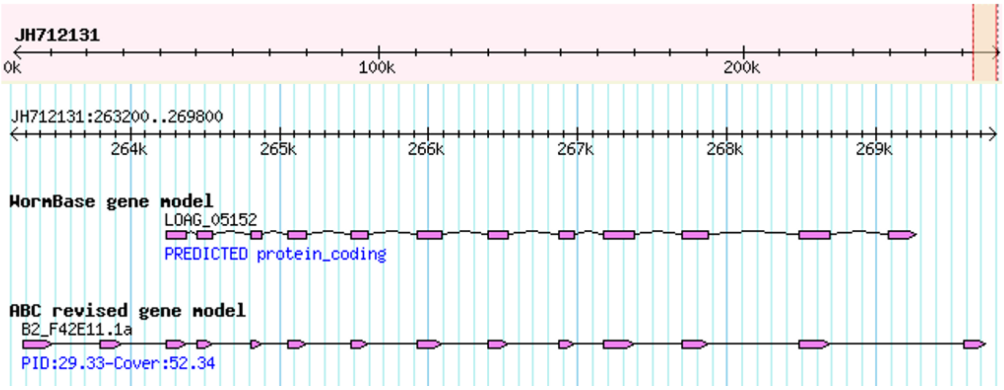


Figure 3.77: A representative case that technical issues could result in incompleteness of an ABC transporter gene candidate

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. LOAG_05152 which was located in the end of the contig encode only one ABC domain. The genome assembly error could lead to this truncated ABC transporter gene.

Through phylogenetic analysis, we found 16 out of 20 ABC transporter genes in *L. loa* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *L .loa* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.78). ABC transporter genes in *L. loa* showed gene contractions in all the subfamilies except for subfamily E, F, a similar gene contraction to those in the two plant parasites (*M. incognita* and *M. hapla*), suggesting a common parasitic metabolism they shared.

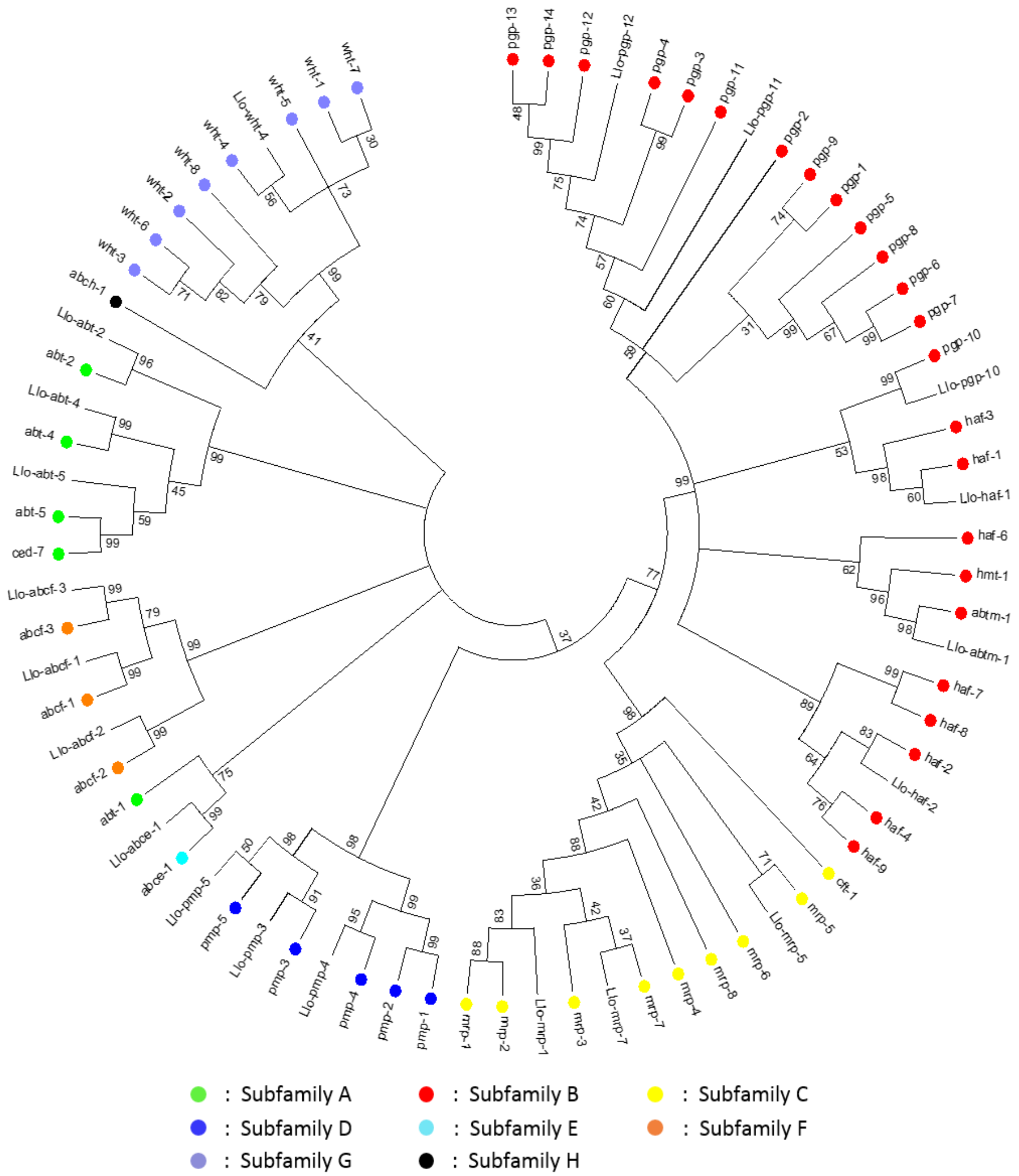


Figure 3.78: Phylogenetic analysis between *L. loa* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *L. loa* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *L. loa* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.21. Annotation of ABC transporter genes in *B. malayi*

B. malayi is a nematode (roundworm) that cause Lymphatic filariasis and river blindness, threatening hundreds of millions of people in the developing world (Ghedin et al. 2007; Knopp et al. 2012). *B. malayi* is transmitted by mosquitoes and is restricted to South and South East Asia and was chosen for whole genome sequencing because it is the only major human filarial pathogen that can be maintained in small laboratory animals (Ghedin et al. 2007; Erickson et al. 2009). After applying the annotation pipeline to *B. malayi*, we identified 33 ABC transporter gene candidates with 23 from InterProScan searches and 10 additional ones from BLAST searches. After contamination filtering process, none of these 33 candidates was due to contamination. Then, we checked the quality of these candidates and found 15 of them were high-quality ABC transporter genes. All these 15 gene encoded appropriate TM domain. After trying to further improve the 18 defective candidates, we generated only two revised gene models of high-quality (Table 3.21). One of the improved gene models was Bm3496, the original model of which was annotated as a full ABC transporter gene in subfamily C but encoded a short protein (873 aa) with only one good ABC domain. After improvement, the revised gene model encoded a longer protein (1350 aa) with two high-quality ABC domains (Figure 3.79). The new gene model was supported by RNA-seq data. Another improved candidate, Bm3156 was annotated as a half transporter gene in subfamily G but encoded a defective ABC domain (57aa 1.10E-06). After improvement, the revised gene model extended to include the neighboring region, making this gene much longer than before. Most importantly, the new gene model of Bm3156 had its 13 introns supported by RNA-seq data and the predicted ABC domain was examined to be high-quality (Figure 3.80). For the remaining 15 candidates, they could not be further improved to be high-quality ABC transporter genes and most of them were random hits from BLAST searches. In summary, we annotated 19 high-quality ABC transporter genes in *B. malayi*, all of which had appropriate TM domain (s) (Table 3.21).

Although 33 putative ABC transporter genes are found in previous study (Liu et al. 2011), the author demonstrates that *B. malayi* draft genome contains many gaps that have resulted in incomplete sequence information for some of the ABC transporter genes. For instance, they identify six ABC transporter genes with two ABC domains but no TM domain and seven ABC transporter genes even without ABC domain only with a single TM domain. Therefore, it is not surprising that we found a smaller number of high-quality ABC transporter genes than that of previous study.

Table 3.21: High-quality ABC transporter genes in *B. malayi* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Bma-abt-2*	Bm3319	Bm3319	8TM-ABC-7TM-ABC	17584	47	2225	
	Bma-abt-4	Bm2828d	Bm2828d	6TM-ABC-10TM-ABC	13578	30	1601	No stop codon
B	*Bma-abtm-1*	Bm7128	Bm7128	6TM-ABC	4660	14	706	No start codon
	Bma-haf-1	Bm4506	Bm4506	6TM-ABC	7252	18	749	
	Bma-haf-2	Bm4059	Bm4059	9TM-ABC	4754	11	757	
	Bma-pgp-10	Bm2924c	Bm2924c	6TM-ABC-6TM-ABC	11955	30	1308	
	Bma-pgp-12	Bm3524	Bm3524	6TM-ABC-6TM-ABC	9318	25	1278	
	Bma-pgp-3	Bm7476	Bm7476	5TM-ABC-6TM-ABC	11414	26	1202	
	Bma-pgp-4	Bm2594a	Bm2594a	6TM-ABC-6TM-ABC	12683	28	1303	
C	*Bma-mrp-1*	Bm4528	Bm4528	11TM-ABC-6TM-ABC	15672	33	1564	
	Bma-mrp-5	Bm3373a	Bm3373a	9TM-ABC-6TM-ABC	12019	30	1473	
	Bma-mrp-7	Bm3496	Bm3496	9TM-ABC-5TM-ABC	11631	29	1351	Exons were improved; No start codon
D	*Bma-pmp-3*	Bm2945	Bm2945	6TM-ABC	5603	15	612	
E	*Bma-abce-1*	Bm6477	Bm6477	ABC-ABC	5029	13	610	
F	*Bma-abcf-1*	Bm3436	Bm3436	ABC-ABC	4788	16	639	No start codon
	Bma-abcf-2	Bm13785a	Bm13785a	ABC-ABC	4633	16	634	No start codon
	Bma-abcf-3	Bm13655	Bm13655	ABC-ABC	4472	17	710	
G	*Bma-wht-4*	Bm3156	Bm3156	ABC-6TM	6587	15	550	Exons were improved
	Bma-wht-8	Bm6595	Bm6595	ABC-5TM	5361	13	592	

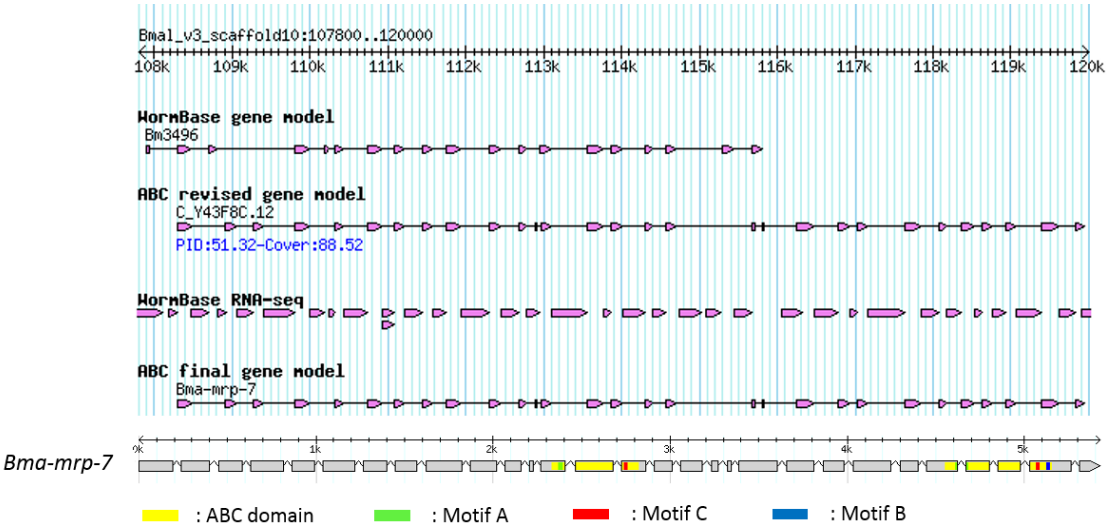


Figure 3.79: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "WormBase RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The original model of Bm3496 was annotated as a full ABC transporter genes in subfamily C but encoded a short protein (873 aa) with only one good ABC domain. The revised gene model encoded a longer protein (1350 aa) and also encoded two high-quality ABC domains, supported by RNA-seq data.

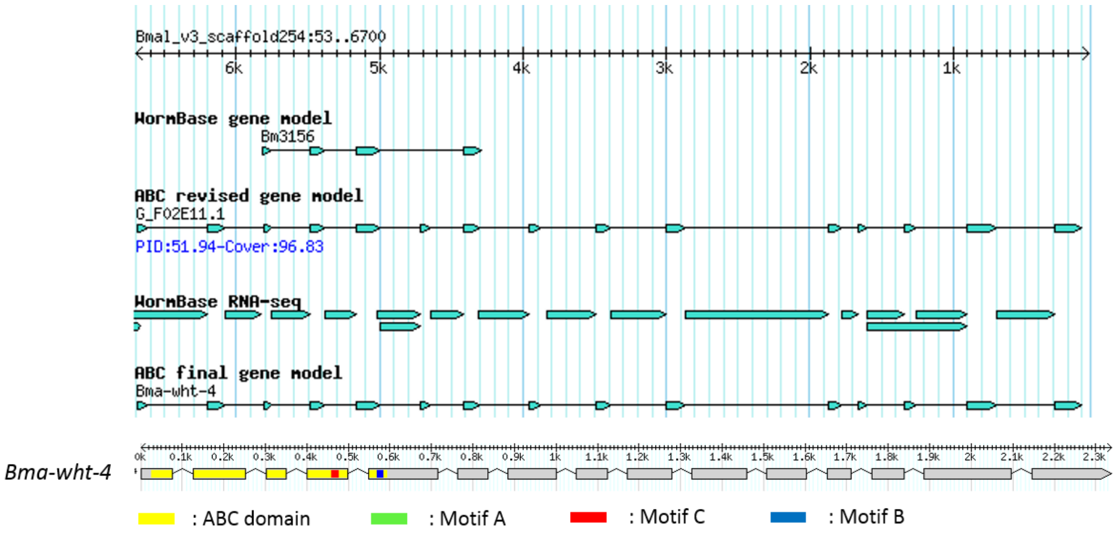


Figure 3.80: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "WormBase RNA-seq" track includes introns predicted by RNA-seq data; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Bm3156, was annotated as a half transporter gene in subfamily G but had a defective ABC domain (57aa 1.10E-06). The revised gene model had a high-quality ABC domain with all its introns supported by RNA-seq data.

Through phylogenetic analysis, we found 15 out of 19 ABC transporter genes in *B. malayi* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *B. malayi* based on their relationship with ABC transporter genes in *C. elegans* (Table). Generally, it showed obvious gene contraction of ABC transporter genes in *B. malayi* compared to those of *C. elegans* (Figure 3.81).

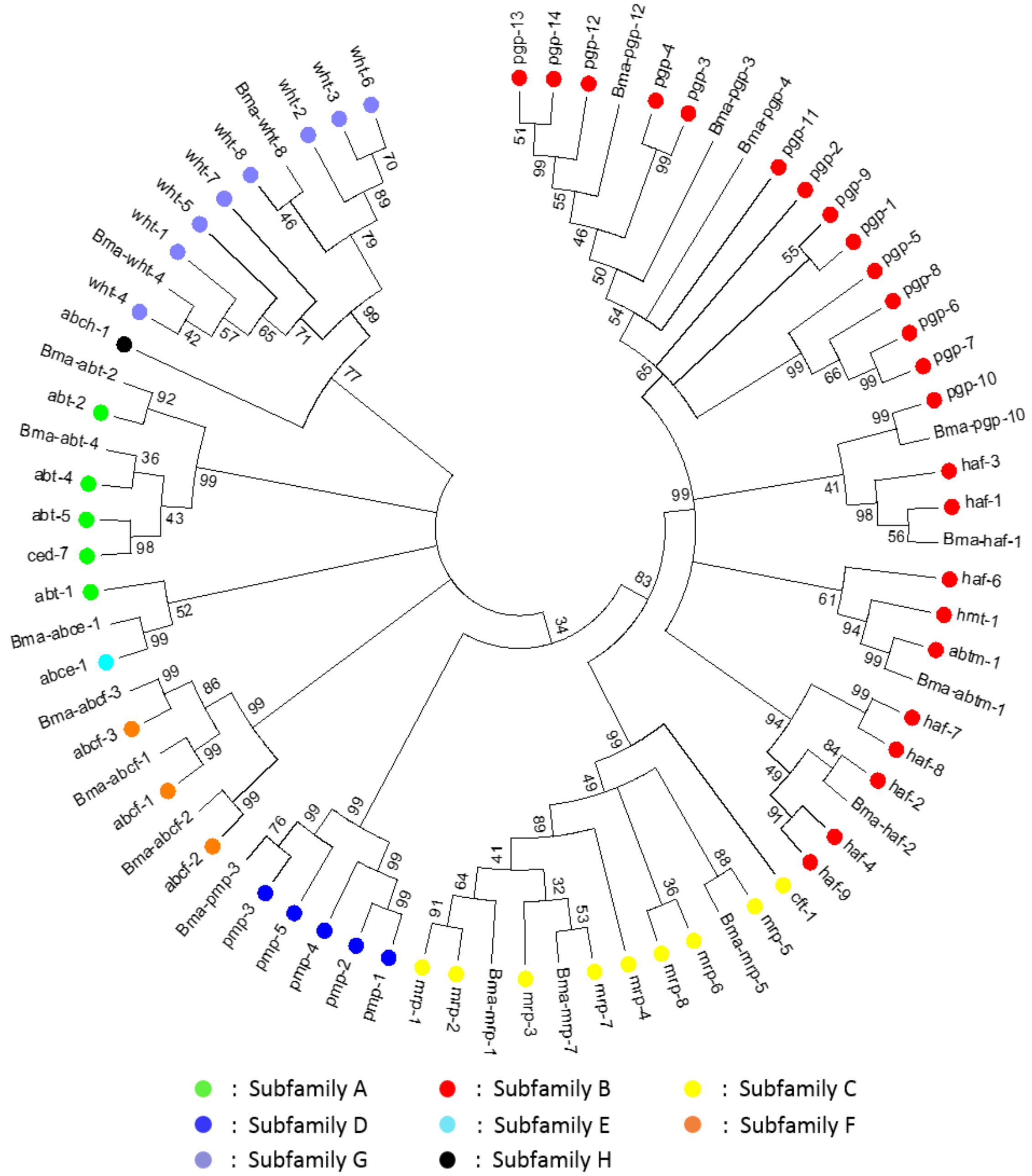


Figure 3.81: Phylogenetic analysis between *B. malayi* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *B. malayi* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *B. malayi* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.22. Annotation of ABC transporter genes in *O. volvulus*

Human onchocerciasis, also known as river blindness, is caused by the filarial parasite *O. volvulus* (Unnasch and Williams 2000). The infective larvae of *O. volvulus* enter the body through the wound made by the bite of its intermediate host, black fly (Abraham et al. 2001; Saint Andre et al. 2002). After applying the annotation pipeline to *O. volvulus*, we identified in total 24 ABC transporter gene candidates with 22 candidates from InterProScan searches and two additional ones from BLAST searches. One of these candidates, OVOC12992 was due to bacteria contamination and was excluded from our further analysis. Among 23 candidates, 15 were high-quality ABC transporter genes. All of these 15 genes also encoded appropriate TM domain (s). For the eight defective candidates, we tried to further improve. After examining the quality of revised gene models, only two were improved to be high-quality (Table 3.22). OVOC1131 was annotated as a full ABC transporter genes in subfamily A. Before improvement, OVOC1131 encoded two ABC domains, one of which was defective with a length of 98 aa. After improvement, defective ABC domain was improved to be a high-quality one in the revised gene model (136 aa). Thus two high-quality ABC domain made this revised gene model a high-quality ABC transporter gene (Figure 3.82). The second improved gene model resulted from merging two genes (OVOC7820 and OVOC7820) (Figure 3.83). And the revised gene model was annotated as a high-quality ABC transporter genes in subfamily A. Only one candidate from BLAST searches could not be improved to be ABC transporter gene and was believed to be random hit. In summary, we annotated 21 high-quality ABC transporter genes in *O. volvulus*, all of which had proper TM domain (s) (Table 3.22).

Table 3.22: High-quality ABC transporter genes in *O. volvulus* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Ovo-abt-2*	OVOC1131	OVOC1131	11TM-ABC-7TM-ABC	20751	35	1475	Exons were improved; No start codon
	Ovo-abt-4	OVOC7820	OVOC7820	8TM-ABC-7TM-ABC	25046	29	1419	OVOC7823 was merge with OVOC7820
	Ovo-abt-5	OVOC2635	OVOC2635	9TM-ABC-9TM-ABC	18445	31	1637	
B	*Ovo-abtm-1*	OVOC64	OVOC64	6TM-ABC	4940	14	747	
	Ovo-haf-2	OVOC203	OVOC203	9TM-ABC	4874	11	755	
	Ovo-haf-3	OVOC7050	OVOC7050	6TM-ABC	4720	15	679	
	Ovo-pgp-10	OVOC2790	OVOC2790	6TM-ABC-6TM-ABC	14112	30	1323	
	Ovo-pgp-3	OVOC10486	OVOC10486	6TM-ABC-6TM-ABC	10659	25	1280	
	Ovo-pgp-4	OVOC10280	OVOC10280	6TM-ABC-5TM-ABC	14011	27	1253	
C	*Ovo-mrp-1*	OVOC5425	OVOC5425	11TM-ABC-8TM-ABC	16419	34	1617	
	Ovo-mrp-3	OVOC10578	OVOC10578	11TM-ABC-5TM-ABC	16801	33	1526	
	Ovo-mrp-5	OVOC2622	OVOC2622	8TM-ABC-6TM-ABC	12533	30	1476	
D	*Ovo-pmp-3*	OVOC6105	OVOC6105	6TM-ABC	6705	16	671	
	Ovo-pmp-4	OVOC5439	OVOC5439	6TM-ABC	5958	16	703	
	Ovo-pmp-5	OVOC3292	OVOC3292	5TM-ABC	10420	15	622	
E	*Ovo-abce-1*	OVOC9163	OVOC9163	ABC-ABC	5247	13	610	
F	*Ovo-abcf-1*	OVOC2878	OVOC2878	ABC-ABC	4736	16	636	
	Ovo-abcf-2	OVOC7707	OVOC7707	ABC-ABC	5128	16	629	No stop codon
	Ovo-abcf-3	OVOC7667	OVOC7667	ABC-ABC	4725	17	710	
G	*Ovo-wht-1*	OVOC5828	OVOC5828	ABC-6TM	6945	19	652	
	Ovo-wht-8	OVOC8061	OVOC8061	ABC-4TM	7567	13	610	

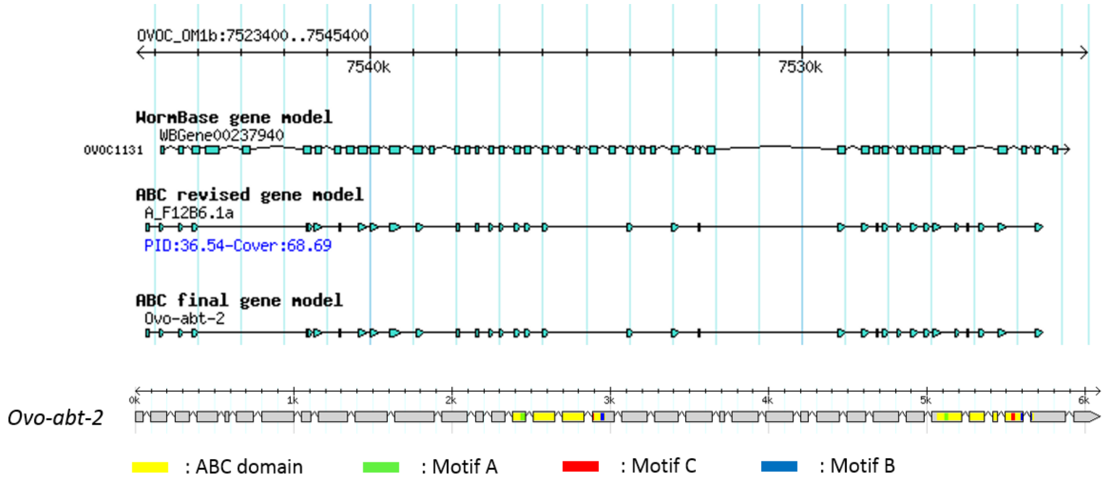


Figure 3.82: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. OVOC1131 was annotated in subfamily A and encoded two ABC domains, one of which was defective with a length of 98 aa. The revised gene model had the defective ABC domain improved to be a high-quality one (136 aa).

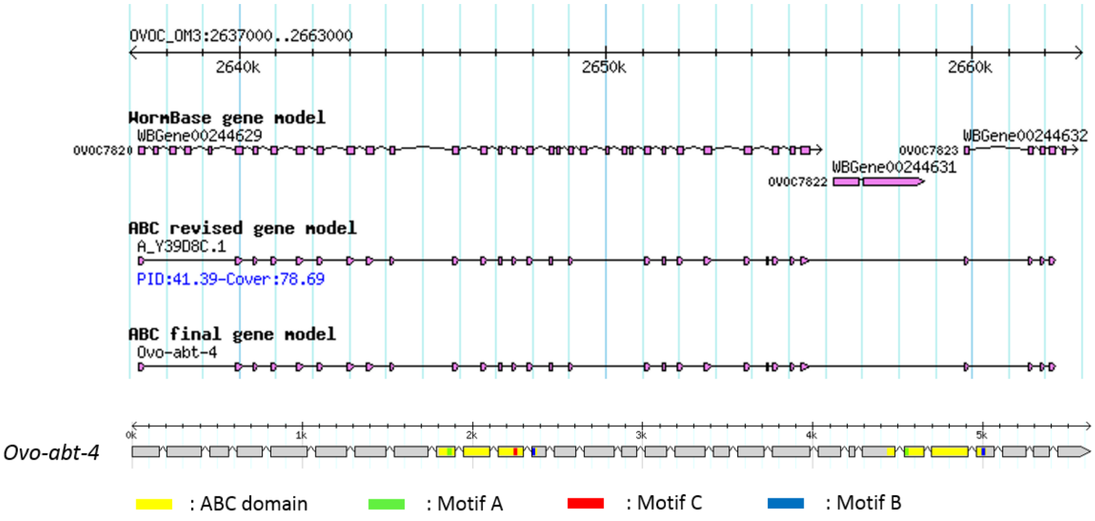


Figure 3.83: A representative case that two adjacent candidates were merged into one high-quality ABC transporter gene

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. OVOC7820 and OVOC7820 were merged into one high-quality ABC transporter in subfamily A

Through phylogenetic analysis, we found 18 out of 21 ABC transporter genes in *O. volvulus* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans* and we assigned the gene names for ABC transporter genes in *O. volvulus* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.84).

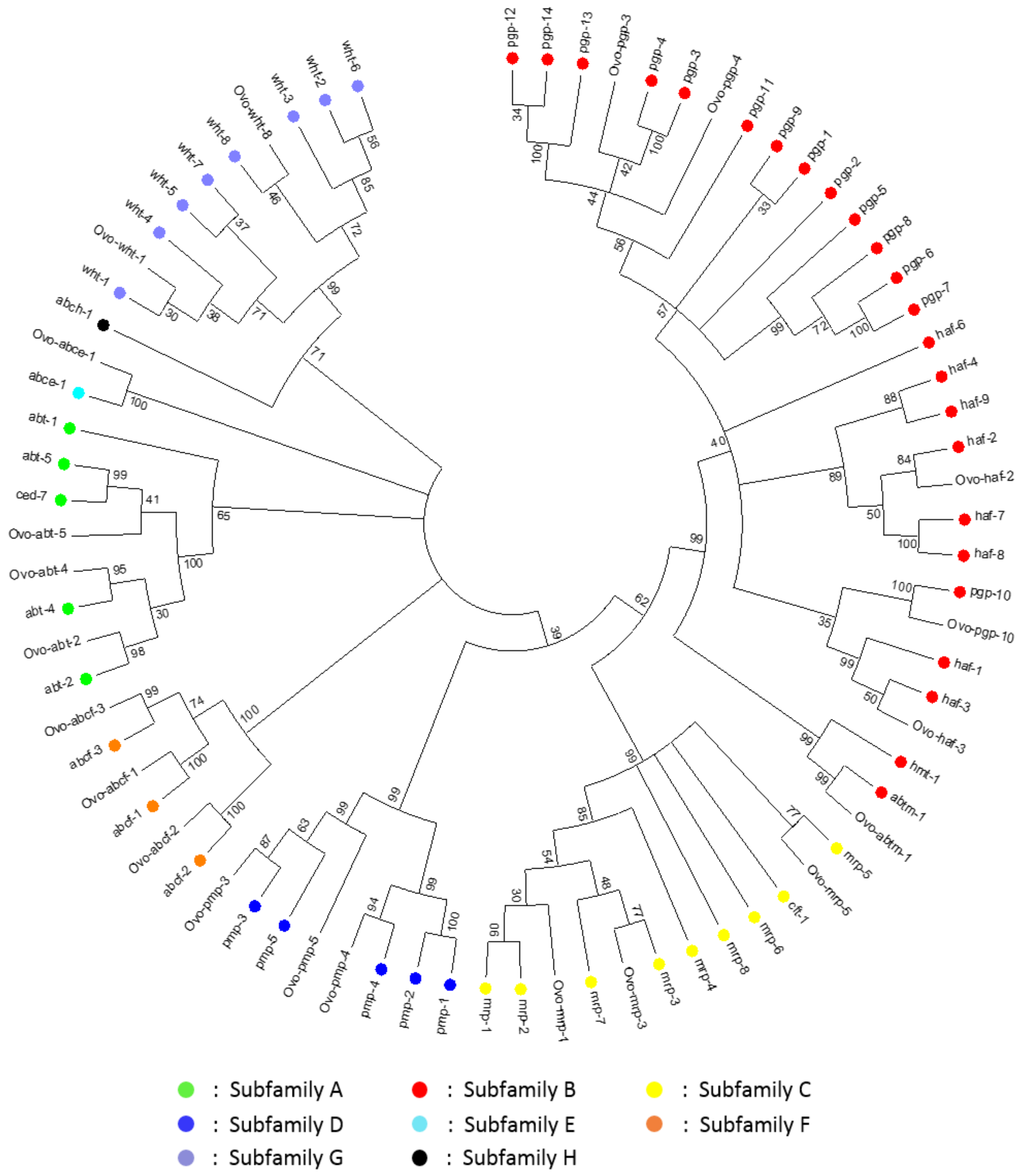


Figure 3.84: Phylogenetic analysis between *O. volvulus* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *O. volvulus* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *O. volvulus* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.23. Annotation of ABC transporter genes in *D. immitis*

Heartworm or also called dog heartworm (*D. immitis*) is a parasitic filarial nematode that is spread from host to host through the bites of mosquitoes (McCall et al. 2008). The definitive host is the dog, however, it also infects cats, foxes, coyotes, and, very rarely, humans (Lee et al. 2010), which is why this helminthiasis could be regarded as parasitic zoonosis. Although at one time heartworm was confined to the southern United States (Brown et al. 2012), it has now spread to nearly all locations where its vector is found (Genchi et al. 2009; Traversa et al. 2010). After applying the annotation pipeline to *D. immitis*, we obtained 32 ABC transporter gene candidates (26 candidates from InterProScan searches, six additional ones from BLAST searches). One candidate, nDi.2.2.2.t11270, was due to bacteria contamination. After examining the quality of the 31 candidates, 11 were high-quality ABC transporter genes. All these 11 genes also encoded appropriate TM domain (s). After trying to further improve the defective candidates, we generated five revised gene models of high-quality (Table 3.23). For example, two adjacent genes, nDi.2.2.2.t03276 and nDi.2.2.2.t03277, hit different parts of a full ABC transporter gene in subfamily B. Each of them encoded one high-quality ABC domain. Through running genBlastG, these two candidates were merged together into a single high-quality ABC transporter gene with two ABC domains (Figure 3.85). For the remaining 11 candidates, all of them could not be further improved to be high-quality ABC transporter genes. One of these 11 candidate genes nDi.2.2.2.t04931 had sequencing gaps within its genomic region. Therefore, this gene could be a complete ABC transporter when this genomic region is well sequenced and assembled (Figure 3.86). In summary, we annotated totally 18 high-quality ABC transporter genes in *D. immitis*, 17 of which had appropriate TM domain (s) (Table 3.23).

Table 3.23: High-quality ABC transporter genes in *D. immitis* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Dim-abt-5*	nDi.2.2.2.t06305	nDi.2.2.2.t06305	7TM-ABC-9TM-ABC	14709	30	1604	
B	*Dim-abtm-1*	nDi.2.2.2.t08117	nDi.2.2.2.t08117	6TM-ABC	4814	14	708	
	Dim-haf-1	nDi.2.2.2.t02044	nDi.2.2.2.t02044	7TM-ABC	6627	17	767	
	Dim-haf-2	nDi.2.2.2.t04722	nDi.2.2.2.t04722	10TM-ABC	4259	10	726	Exons were improved
	Dim-pgp-10	nDi.2.2.2.t03276	nDi.2.2.2.t03276	6TM-ABC-6TM-ABC	15383	31	1273	nDi.2.2.2.t03277 was merged with nDi.2.2.2.t03276; No start codon
	Dim-pgp-12	nDi.2.2.2.t00454	nDi.2.2.2.t00454	6TM-ABC-6TM-ABC	9493	25	1286	
	Dim-pgp-3	nDi.2.2.2.t03212	nDi.2.2.2.t03212	6TM-ABC-6TM-ABC	12242	29	1297	
C	*Dim-mrp-1*	nDi.2.2.2.t01111	nDi.2.2.2.t01111	11TM-ABC-10TM-ABC	16032	35	1690	
	Dim-mrp-3	nDi.2.2.2.t00532	nDi.2.2.2.t00532	9TM-ABC-5TM-ABC	11659	29	1314	nDi.2.2.2.t00533 was merged with nDi.2.2.2.t00532; No start codon
	Dim-mrp-5	nDi.2.2.2.t03446	nDi.2.2.2.t03446	9TM-ABC-7TM-ABC	11700	30	1474	
D	*Dim-pmp-3*	nDi.2.2.2.t08817	nDi.2.2.2.t08817	6TM-ABC	6645	16	672	
	Dim-pmp-4	nDi.2.2.2.t01123	nDi.2.2.2.t01123	5TM-ABC	6096	16	694	
E	*Dim-abce-1*	nDi.2.2.2.t07766	nDi.2.2.2.t07766	ABC-ABC	5250	13	610	
F	*Dim-abcf-1*	nDi.2.2.2.t04075	nDi.2.2.2.t04075	ABC-ABC	4539	16	615	Exons were improved
	Dim-abcf-2	nDi.2.2.2.t07483	nDi.2.2.2.t07483	ABC-ABC	4901	16	628	
	Dim-abcf-3	nDi.2.2.2.t05841	nDi.2.2.2.t05841	ABC-ABC	4700	18	695	Exons were improved
G	*Dim-wht-2*	nDi.2.2.2.t06583	nDi.2.2.2.t06583	ABC-5TM	5351	13	609	
	Dim-wht-4	nDi.2.2.2.t08524	nDi.2.2.2.t08524	ABC-6TM	12561	21	695	

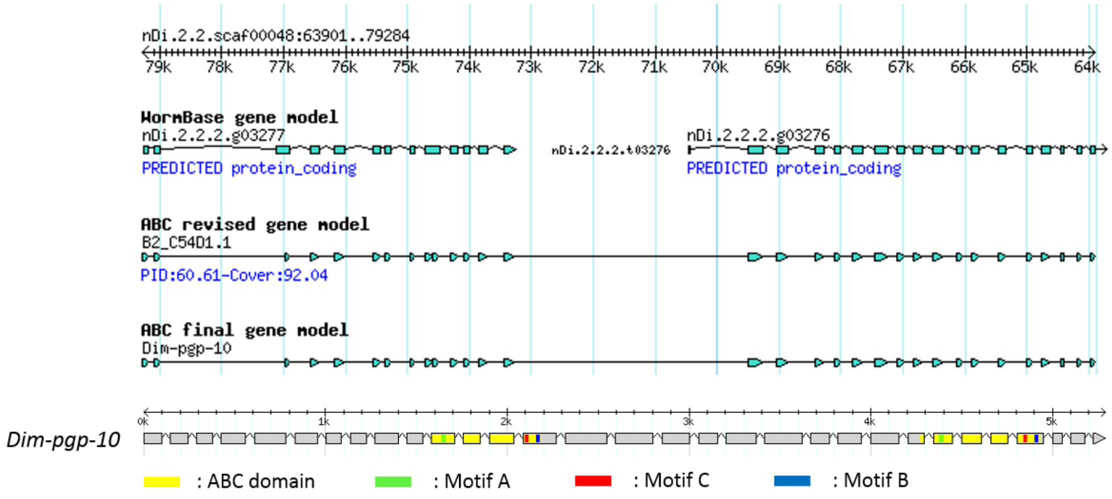


Figure 3.85: A representative case that two adjacent ABC candidates were merged into one high-quality ABC transporter gene

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Two adjacent genes, nDi.2.2.2.t03276 and nDi.2.2.2.t03277, were two fragments of a full ABC transporter gene in subfamily B, each of which encoded one high-quality ABC domain. The revised gene model was a result of merging these two candidates and was examined to be a high-quality full ABC transporter gene.

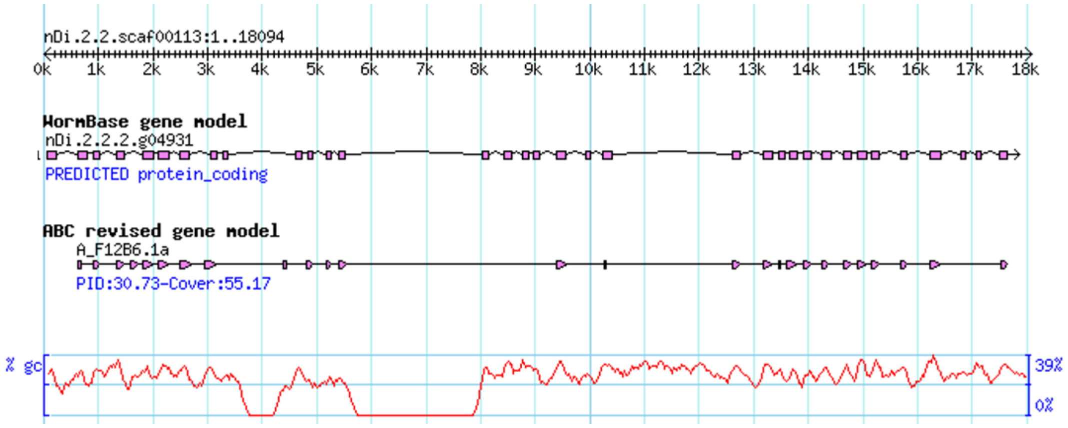


Figure 3.86: A representative case that sequencing error could result in incompleteness of an ABC transporter gene candidate

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. nDi.2.2.2.t04931 had sequencing gaps within its genomic region, leading to a defective ABC domain.

Through phylogenetic analysis, we found 13 out of 18 ABC transporter genes in *P. exspectatus* showed one-to-one orthologous relationship with ABC transporter gene in *C. elegans*. We assigned the gene names for ABC transporter genes in *D. immitis* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.87). *D. immitis* clustered with other three animal species, *L. loa*, *B. malayi* and *O. volvulus* from our evolitionary analysis. These four species contained similar total numbers of ABC transporter genes in total as well as similar numbers of ABC transporter genes in each subfamily, suggesting their common ancenster lost some of ABC transporter genes to adapt the living evironment.

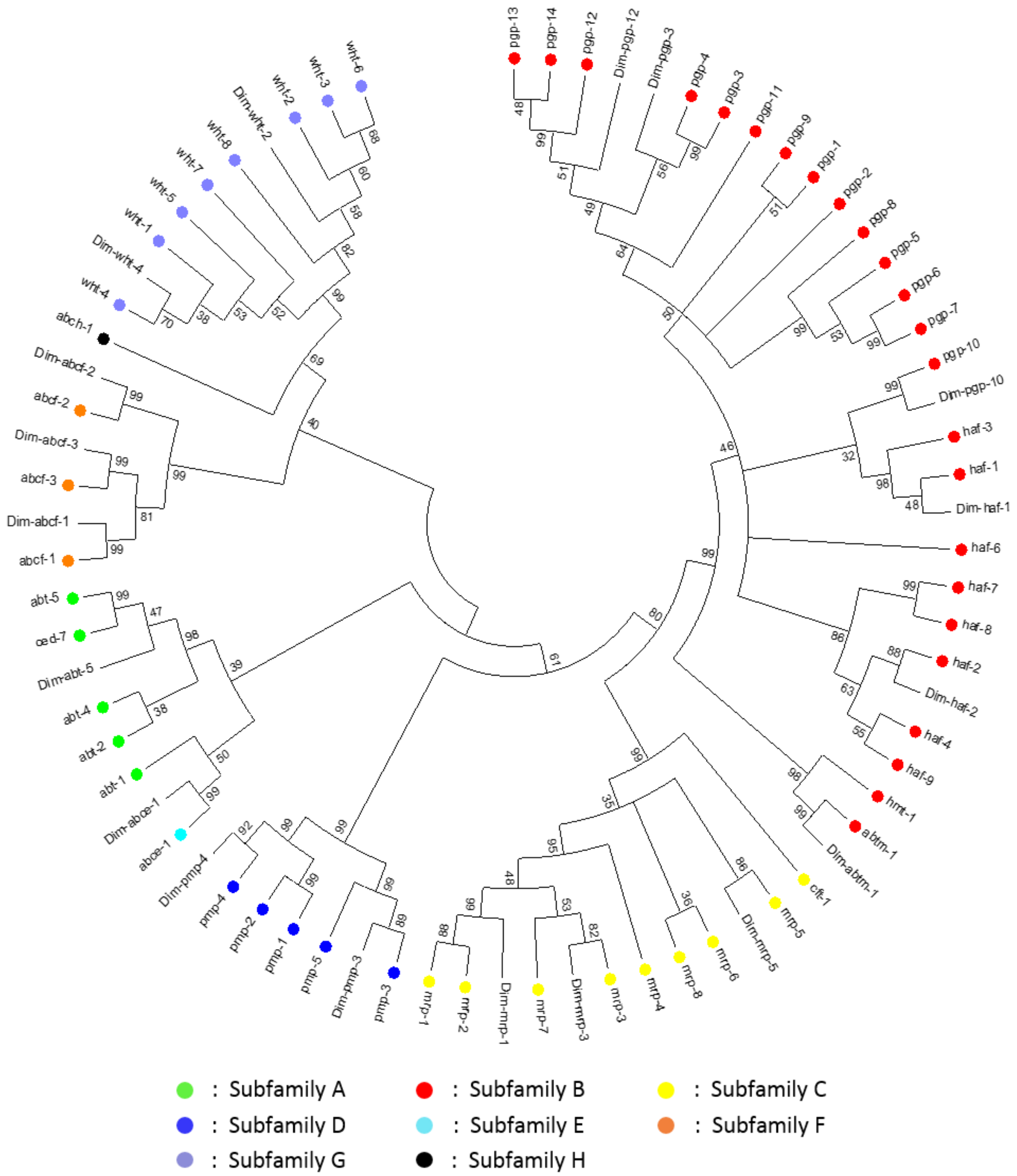


Figure 3.87: Phylogenetic analysis between *D. immitis* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *D. immitis* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *D. immitis* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.24. Annotation of ABC transporter genes in *T. spiralis*

The nematode *T. spiralis*, the most common cause of human trichinellosis, is a member of a clade that diverged early in the evolution of the Nematoda (Mitreva et al. 2011). After InterProScan and BLAST searches, we identified 20 ABC transporter gene candidates (17 candidates from InterProScan searches, three additional ones from BLAST searches), none of which were due to contamination. We checked the quality of all 20 candidates and only two of them were annotated to be good quality ABC transporter genes. And these two genes also encoded proper TM domain (s). For the 18 defective candidates, we successfully produced 12 improved gene models with high-quality, three of which with only TM domain improved (Table 3.24). For example, EFV59444 was annotated as a full transporter in subfamily F but had a length of 306 aa and had only one predicted ABC domain. By running genBlastG, we obtained a revised gene model in this region, encoding a longer protein (601 aa) with two high-quality ABC domains (Figure 3.88). Similarly, the length of EFV55152 increased from to 537 aa to 659 aa after improvement. The revised gene model encoded an ABC transporter with a complete TM domain with six helices (Figure 3.89). The remaining six candidates were most likely to be random hits that could not be improved. Taking together, we annotated 15 high-quality ABC transporter genes in *T. spiralis*, 14 of which had proper TM domain (s) (Table 3.24).

Table 3.24: High-quality ABC transporter genes in *T. spiralis* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Tsp-abt-1*	EFV56075b	EFV56075	7TM-ABC-4TM-ABC	10040	21	1236	EFV56075 was split; No start codon
	Tsp-abt-2	EFV56075a	EFV56075	7TM-ABC-6TM-ABC	9017	23	1256	EFV56075 was split; No start codon
	Tsp-abt-4	EFV54183	EFV54183	6TM-ABC-7TM-ABC	7002	19	1459	Exons were improved
B	*Tsp-abtm-1*	EFV54279	EFV54279	6TM-ABC	3693	16	665	TM helices were improved; No start codon
	Tsp-haf-1	EFV55152	EFV55152	TM-ABC	3419	15	660	TM helices were improved; No start codon
	Tsp-haf-9	EFV54367	EFV54367	10TM-ABC	3008	12	751	Exons were improved
C	*Tsp-mrp-1*	EFV53736	EFV53736	10TM-ABC-5TM-ABC	8062	25	1430	
	Tsp-mrp-2	EFV53848	EFV53848	9TM-ABC-6TM-ABC	7302	20	1397	TM helices were improved; No start codon
	Tsp-mrp-3	EFV55965	EFV55965	8TM-ABC-8TM-ABC	6881	25	1214	Exons were improved; No start codon
	Tsp-mrp-6	EFV59601	EFV59601	2TM-ABC-6TM-ABC	5319	18	1317	
E	*Tsp-abce-1*	EFV49731	EFV49731	ABC-ABC	2885	14	582	Exons were improved; No start codon
	Tsp-abce-2	EFV50509	EFV50509	ABC-ABC	2887	14	582	Exons were improved; No start codon
F	*Tsp-abcf-1*	EFV59444	EFV59444	ABC-ABC	3236	15	602	Exons were improved
	Tsp-abcf-2	EFV57419	EFV57419	ABC-ABC	3158	13	580	
	Tsp-abcf-3	EFV50642	EFV50642	ABC-ABC	3180	13	546	Exons were improved

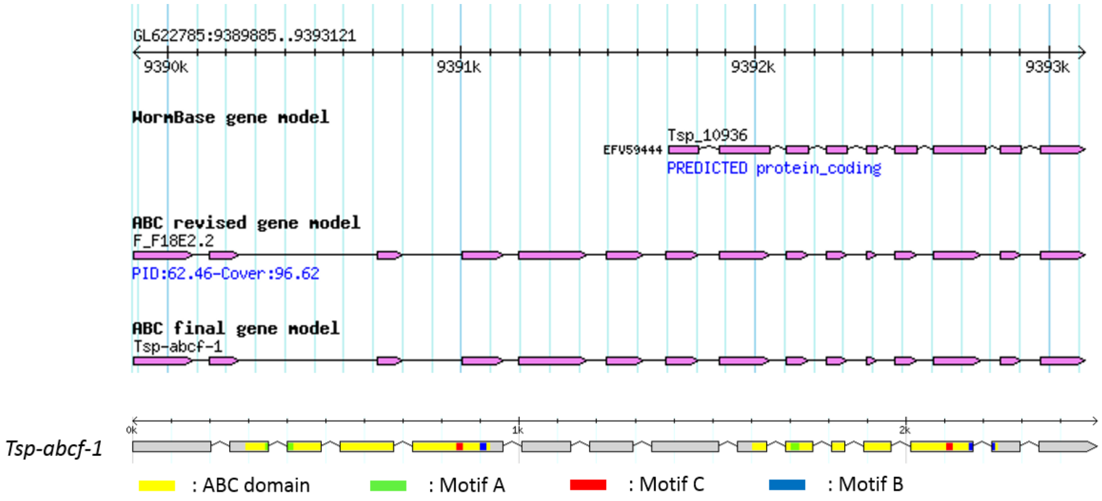
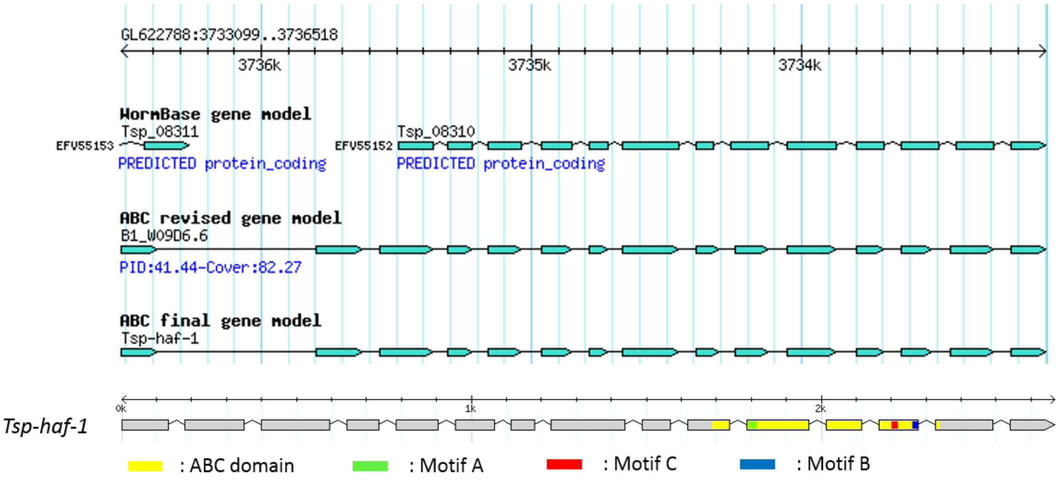


Figure 3.88: A representative case that the exons of one defective ABC transporter gene were improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. EFV59444 was annotated as a full transporter in subfamily F but it only had a length of 306 aa with only one predicted ABC domain. The revised gene model encoded a longer protein (601 aa) with two high-quality ABC domains.



TM domain prediction

```
EFV55152
537
YES
IN
3
1. 82-102, 2. 105-125, 3. 187-207
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooMMMMMMMMMMMMMMMM
MMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
```

```
Tsp-haf-1
659
YES
OUT
6
1. 16-36, 2. 68-88, 3. 110-130, 4. 206-226, 5. 229-249,
6. 311-331
oooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooooooooooooMMMMMMMMMMM
MMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooMMMMMMMMMMMM
MMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
```

Figure 3.89: A representative case that the TM domain of an ABC transporter gene was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B.The length of EFV55152 increased from to 537 aa to 659 aa after revision, leading a complete TM domain with six helices instead of three in the original gene model.

Through phylogenetic analysis, we found seven out of 15 ABC transporter genes in *T. spiralis* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans* and we assigned gene names for ABC transporter genes in *T. spiralis* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.90). Interestingly, in *T. spiralis*, there was no annotated ABC transporter gene in subfamily D, G and H. Besides, we observed different levels of gene contraction in other subfamilies, especially in subfamily B. These observation illustrated that *T. spiralis* had some differences compared to other animal parasites mentioned above, consistent with the evolutionary distance that *T. spiralis* had with others.

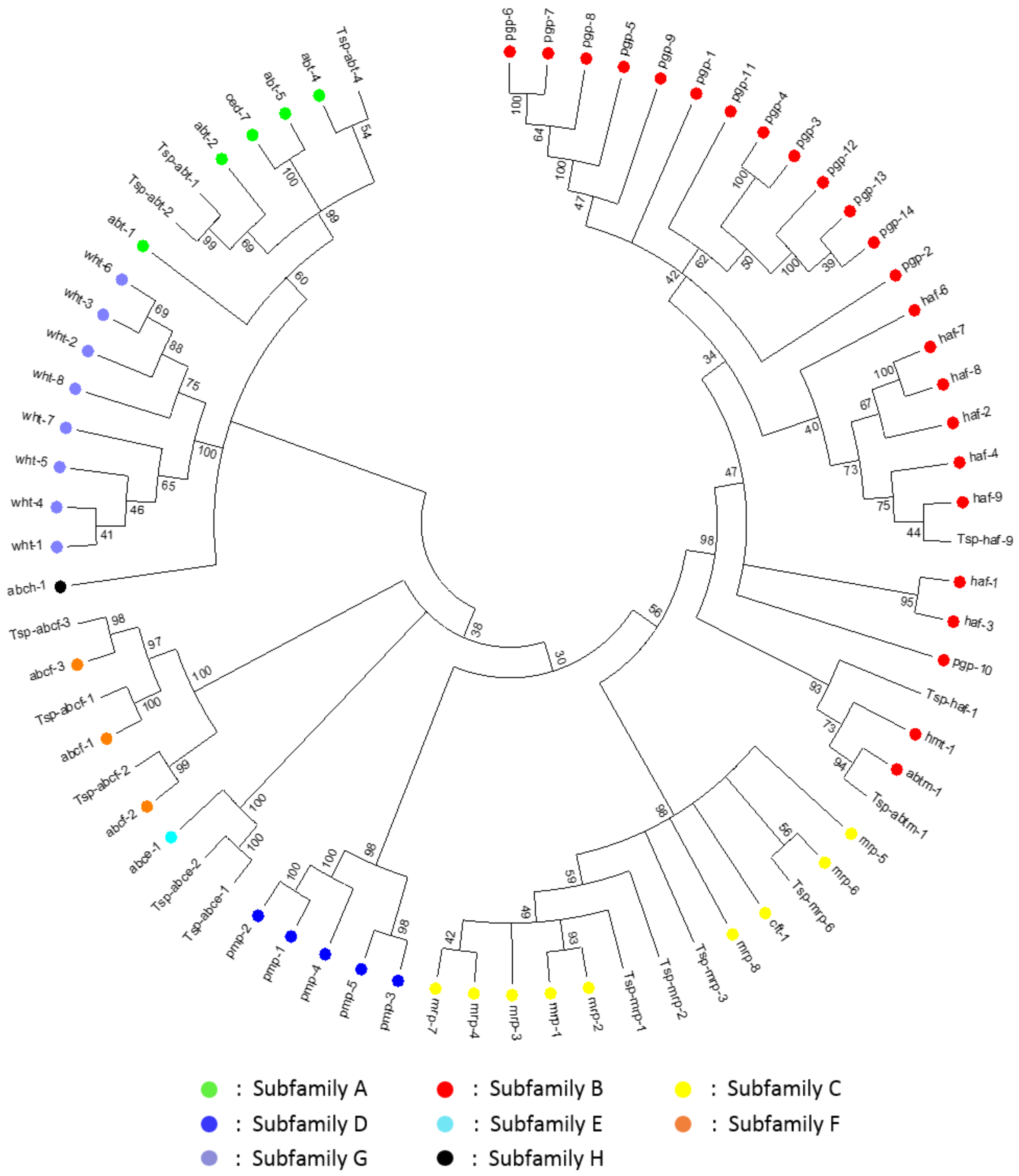


Figure 3.90: Phylogenetic analysis between *T. spiralis* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *T. spiralis* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *T. spiralis* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.25. Annotation of ABC transporter genes in *T. trichiura*

Trichuris (whipworm) infects one billion people worldwide and causes a disease (trichuriasis) that results in major socioeconomic losses in both humans and pigs (Stephenson et al. 2000). *T. trichiura* belongs to this genus and has evolved to occupy an unusual niche (Foth et al. 2014). By applying the annotation pipeline to *T. trichiura*, we obtained 54 ABC transporter gene candidates (53 candidates from InterProScan searches, one additional ones from BLAST searches). The relatively large number of candidates compared to that of *T. spiralis* reduced to 29 after identifying 25 candidates that were due to bacteria contamination. Among these 29 candidates, only three were high-quality ABC transporter genes. And all of these three genes also encoded appropriate TM domain (s). For 15 defective candidates, we tried to further improve each of them. We ended up with eight revised gene models of high-quality (Table 3.25), one of which with only TM domain improved. For example, TTRE_0000750401 is a representative case for exon improvement. The original model of TTRE_0000750401 was characterized to be a half ABC transporter gene in subfamily B. However, it encoded a short ABC domain (71aa 9.30E-10). After improvement, the revised gene model (Figure 3.91) containing a high-quality ABC domain (149 aa 3.8E-30). TTRE_0000120101 was annotated as a full ABC transporter in subfamily C but it encoded four predicted ABC domains. After improvement, two revised gene models was obtained as a result of splitting the original gene model. Each of them had two typical ABC domains (Figure 3.92), illustrating that both of them were high-quality ABC transporter genes. Six defective candidates that could not be improved. In total, we annotated 23 high-quality ABC transporter genes in *T. trichiura*, 19 of which had appropriate TM domain (s) (Table 3.25).

Table 3.25: High-quality ABC transporter genes in *T. trichiura* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Ttr-abt-1*	TTRE_0000247301	TTRE_0000247301	10TM-ABC-4TM-ABC	11386	39	1973	
	Ttr-abt-2	TTRE_0000422201	TTRE_0000422201	8TM-ABC-7TM-ABC	16146	23	1297	Exons were improved; No start codon
	Ttr-abt-3	TTRE_0000279301	TTRE_0000279301	12TM-ABC-9TM-ABC	20718	40	2300	
	Ttr-abt-4	TTRE_0000065401	TTRE_0000065401	7TM-ABC-9TM-ABC	8745	24	1503	TM helices were improved; No start and stop codon
	Ttr-abt-5	TTRE_0000281901	TTRE_0000281901	4TM-ABC-6TM-ABC	12254	16	884	Exons were improved; No start and stop codon
B	*Ttr-abtm-1*	TTRE_0000750401	TTRE_0000750401	6TM-ABC	11885	14	695	Exons were improved; No start codon
	Ttr-haf-4	TTRE_0000062901	TTRE_0000062901	10TM-ABC	5146	11	806	
	Ttr-hmt-1	TTRE_0000556301	TTRE_0000556301	4TM-ABC	4745	15	600	
	Ttr-pgp-1	TTRE_0000276701	TTRE_0000276701	7TM-ABC-4TM-ABC	7430	27	1235	
C	*Ttr-mrp-1*	TTRE_0000120101a	TTRE_0000120101	4TM-ABC-5TM-ABC	11243	15	1029	
	Ttr-mrp-2	TTRE_0000120101b	TTRE_0000120101	7TM-ABC-6TM-ABC	8605	21	1275	TTRE_0000120101 was split; No start and stop codon
	Ttr-mrp-3	TTRE_0000120301	TTRE_0000120301	9TM-ABC-6TM-ABC	11622	23	1514	TTRE_0000120101 was split; No start and stop codon
	Ttr-mrp-4	TTRE_0000767901	TTRE_0000767901	9TM-ABC-5TM-ABC	8658	26	1457	
	Ttr-mrp-5	TTRE_0000620601	TTRE_0000620601	2TM-ABC-5TM-ABC	10614	18	911	
	Ttr-mrp-6	TTRE_0000410801	TTRE_0000410801	6TM-ABC-5TM-ABC	6812	21	1264	
	Ttr-mrp-7	TTRE_0000255201	TTRE_0000255201	8TM-ABC-3TM-ABC	7764	25	1277	
D	*Ttr-pmp-3*	TTRE_0000419901	TTRE_0000419901	5TM-ABC	7793	14	590	Exons were improved
E	*Ttr-abce-1*	TTRE_0000416201	TTRE_0000416201	2TM-ABC-3TM-ABC	7043	13	1009	
F	*Ttr-abcf-1*	TTRE_0000097001	TTRE_0000097001	ABC-ABC	3203	15	726	
	Ttr-abcf-2	TTRE_0000522901	TTRE_0000522901	1TM-ABC-0TM-ABC	2889	10	638	
	Ttr-abcf-3	TTRE_0000735701	TTRE_0000735701	ABC-ABC	8146	14	541	
G	*Ttr-wht-1*	TTRE_0000623801	TTRE_0000623801	ABC-5TM	4562	13	508	
	Ttr-wht-2	TTRE_0000461701	TTRE_0000461701	ABC-5TM	2632	2	626	

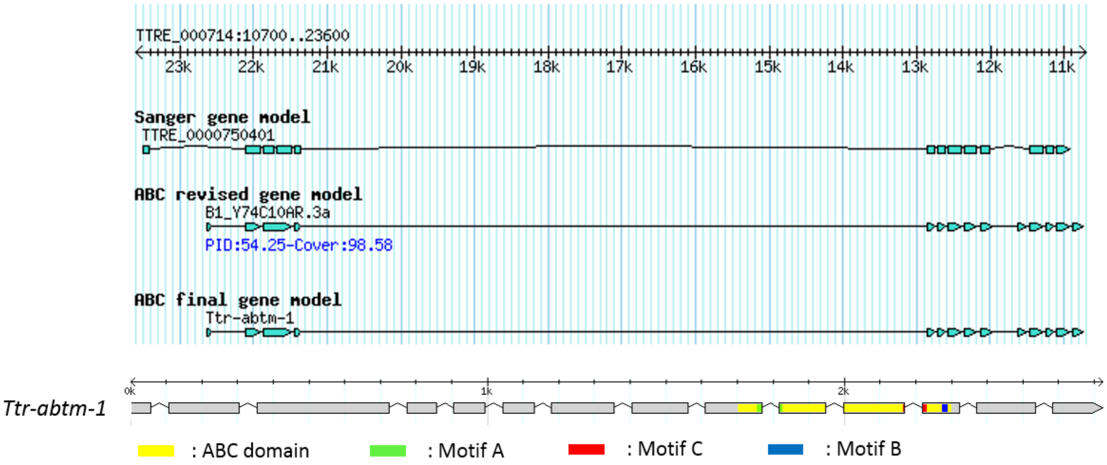


Figure 3.91: A representative case that the exons of one defective ABC transporter gene were improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The original model of TTRE_0000750401 was characterized to be a half ABC transporter gene in subfamily B. However, a short ABC domain (71aa, 9.30E-10) made this gene defective. The revised gene model contained a high-quality ABC domain (149 aa 3.8E-30), making it to be a high-quality ABC domain.

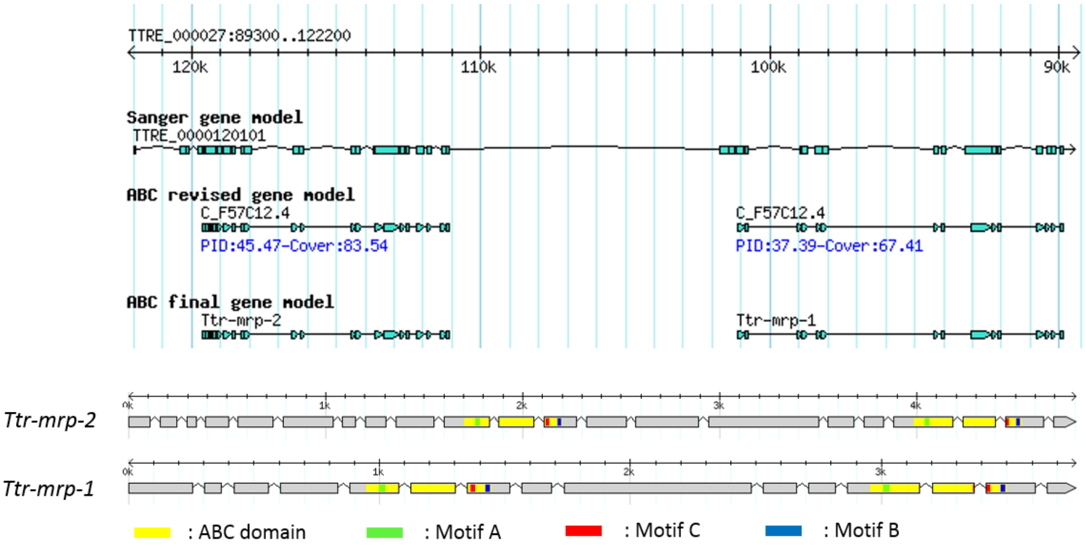


Figure 3.92: A representative case that the TM domain of an ABC transporter gene was improved

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. TTRE_0000120101 was annotated as a full ABC transporter in subfamily C but it contained four predicted ABC domains. After improvement, two revised gene models as a result of splitting the original gene model were obtained, each of which had two high-quality ABC domains.

Through phylogenetic analysis, we found only eight out of 23 ABC transporter genes in *T. trichiura* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned the gene names for ABC transporter genes in *T. trichiura* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.93). Compared to *C. elegans*, the large number of gene contraction happened in subfamily B (three), D (one) and G (two) in *T. trichiura*. In addition, *T. trichiura* had no annotated ABC transporter gene in subfamily H. Taking together, these gene losses in *T. trichiura* indicated that those genes might have unnecessary function or redundant function with other ABC transporters, leading to gene death during evolution.

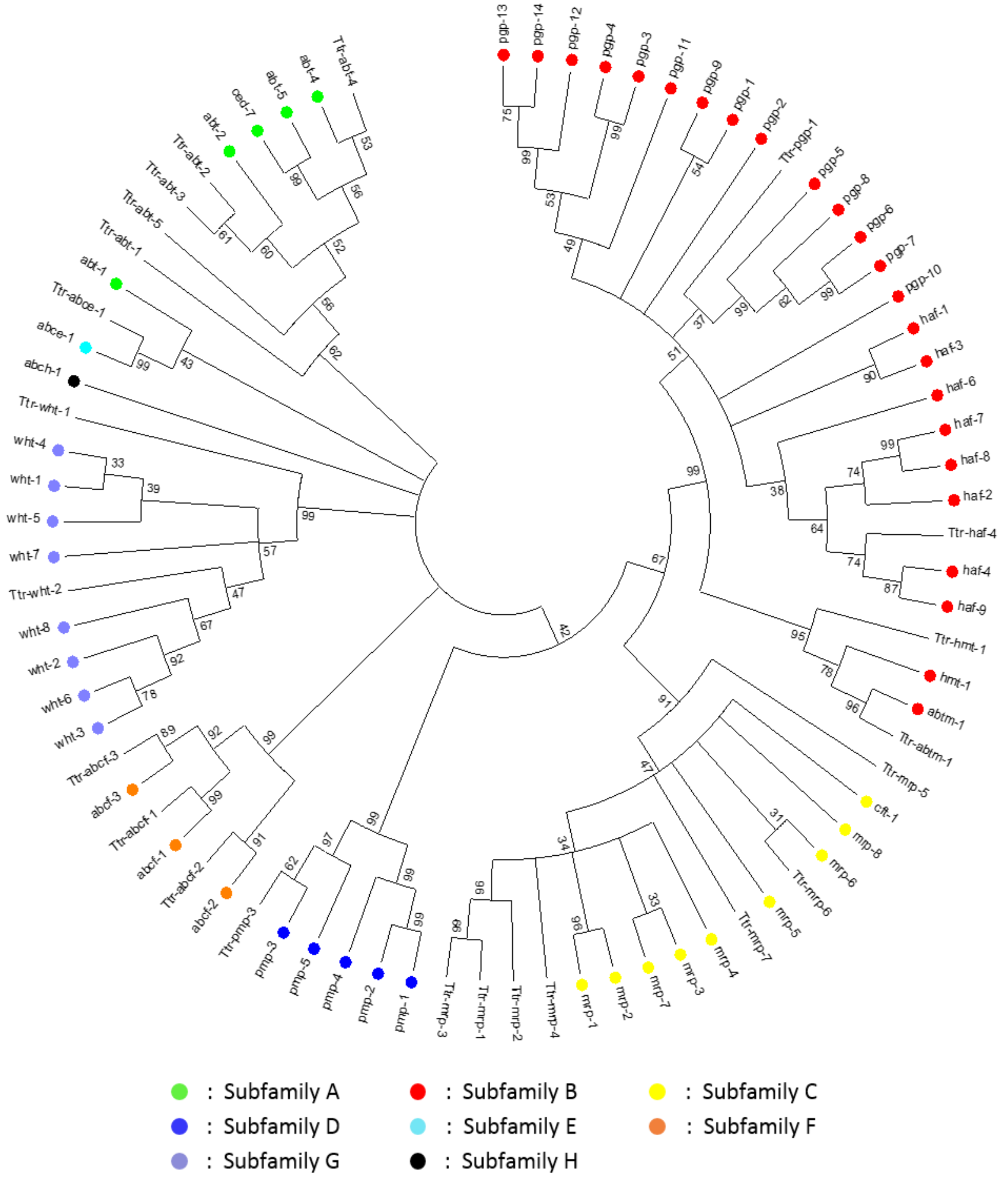


Figure 3.93: Phylogenetic analysis between *T. trichiura* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *T. trichiura* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *T. trichiura* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

3.26. Annotation of ABC transporter genes in *T. suis*

T. suis is closely related to *T. trichiura* and is a common mild intestinal pathogen of pigs, which is able to establish temporarily in the human caecum and colon (Beer 1973). After InterProScan and BLAST searches, we identified 24 ABC transporter gene candidates (22 candidates from InterProScan searches, two additional ones from BLAST searches). None of these candidates were due to contamination. Among the 24 putative candidates, nine were annotated to be good quality ABC transporter genes. And all these nine genes also encoded proper TM domain (s). For the 15 defective candidates, we tried to further improve their gene models. After examining the quality of revised gene models, we successfully produced 11 improved gene models with high-quality, four of which with only TM domain improved (Table 3.26). For example, M514_01565 was annotated as an ABC transporter gene in subfamily F and encoded two defective ABC domain (69 aa 8.70E-09; 124 aa 1.10E-20). By running genBlastG, a revised gene model encoding two typical ABC domains (132 aa 3.6E-21; 158 aa 1.9E-20) were obtained (Figure 3.94), making this new gene model a high-quality ABC transporter gene. The original gene model of M514_01755 was annotated as a half ABC transporter gene in subfamily G but it was extremely long (~40 kb), resulting in a protein with a length of 2744 aa. TM helices of M514_01755 clustered into three groups, which was unexpected. After improvement, we got a revised gene model encoding an ABC transporter with a length of 662 aa. This new gene model also encoded six TM helices in one TM domain as expected (Figure 3.95). There were three defective candidates were not able to improved. In summary, we annotated 21 high-quality ABC transporter genes in *T. suis*, 20 of which had proper TM domain (s) (Table 3.26).

Table 3.26: High-quality ABC transporter genes in *T. suis* after revision

Class	Gene name	ID	Original ID	Domain organization	Genomic span	Exon number	Protein length	Comments
A	*Tsu-abt-1*	M514_01480	M514_01480	4TM-ABC-6TM-ABC	11209	15	1018	Exons were improved
	Tsu-abt-2	M514_03649	M514_03649	6TM-ABC-5TM-ABC	15309	23	1166	Exons were improved
	Tsu-abt-4	M514_24960	M514_24960	9TM-ABC-6TM-ABC	16203	24	1333	TM helices were improved;
	Tsu-abt-5	M514_00301	M514_00301	8TM-ABC-8TM-ABC	8525	21	1481	Exons were improved
B	*Tsu-abtm-1*	M514_19593	M514_19593	5TM-ABC	16843	15	641	TM helices were improved; No start codon
	Tsu-haf-4	M514_04605	M514_04605	10TM-ABC	5836	11	792	No start and stop codon
	Tsu-hmt-1	M514_01758	M514_01758	5TM-ABC	5565	16	611	Exons were improved
	Tsu-pgp-10	M514_09538	M514_09538	6TM-ABC-6TM-ABC	7432	24	1325	No start codon
C	*Tsu-mrp-1*	M514_10741	M514_10741	11TM-ABC-5TM-ABC	26848	28	1541	No start and stop codon
	Tsu-mrp-2	M514_13250	M514_13250	11TM-ABC-6TM-ABC	14587	24	1506	No start codon
	Tsu-mrp-3	M514_09514	M514_09514	9TM-ABC-6TM-ABC	11846	21	1418	Exons were improved; No start codon
	Tsu-mrp-4	M514_02283	M514_02283	7TM-ABC-5TM-ABC	10579	21	1380	Exons were improved; No start codon
	Tsu-mrp-5	M514_14931	M514_14931	7TM-ABC-5TM-ABC	12004	25	1318	No start and stop codon
	Tsu-mrp-6	M514_22345	M514_22345	9TM-ABC-5TM-ABC	21882	25	1605	No start and stop codon
	Tsu-mrp-7	M514_00054	M514_00054	8TM-ABC-6TM-ABC	15682	22	1371	No start and stop codon
E	*Tsu-abce-1*	M514_18710	M514_18710	ABC-ABC	12815	14	674	No stop codon
F	*Tsu-abcf-1*	M514_17608	M514_17608	ABC-ABC	5925	12	667	No start and stop codon
	Tsu-abcf-2	M514_01565	M514_01565	1TM-ABC-0TM-ABC	3252	14	609	Exons were improved
	Tsu-abcf-3	M514_03079	M514_03079	ABC-ABC	14270	26	1060	No stop codon
G	*Tsu-wht-1*	M514_10577	M514_10577	ABC-5TM	3635	12	445	TM helices were improved; No start codon
	Tsu-wht-2	M514_01755	M514_01755	ABC-6TM	2019	2	663	TM helices were improved;

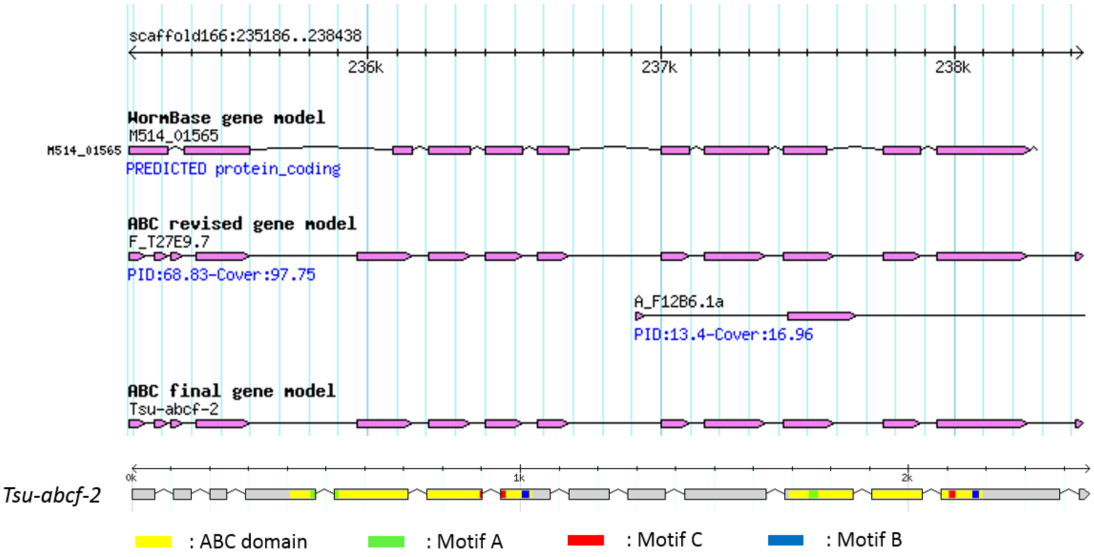
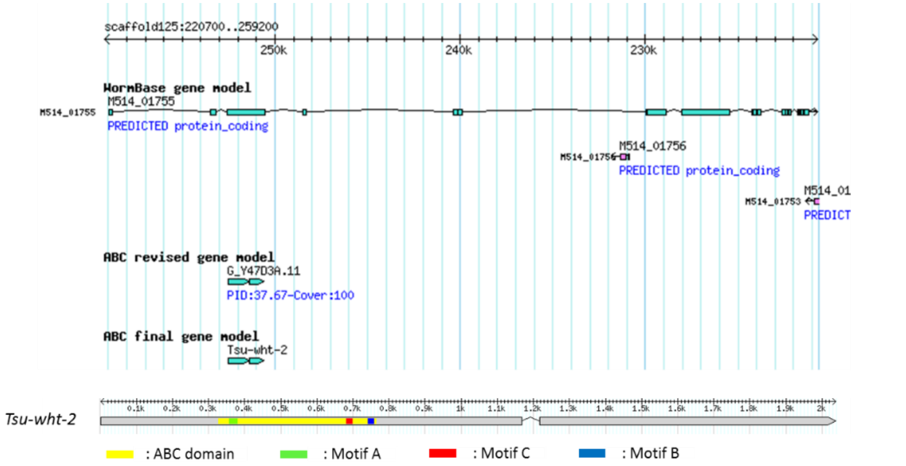


Figure 3.94: A representative case that the exons of one defective ABC transporter gene were improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. M514_01565, annotated as an ABC transporter gene in subfamily F, encoded two predicted domain but both of them were defective (ABC domains: 69 aa 8.70E-09; 124 aa 1.10E-20). The revised gene model with two typical ABC domains (132 aa 3.6E-21; 158 aa 1.9E-20), making this new gene model a high-quality ABC transporter gene.



TM domain prediction

```
M514_01755
2744
YES
OUT
10
1. 27-47, 2. 572-592, 3. 606-626, 4. 656-676, 5. 683-703,
6. 713-733, 7. 802-822, 8. 906-926, 9. 930-950, 10. 1525-1545
ooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooo
oooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMM
MMMMMMMMMMMMMMMMooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiMMMMMMMM
MMMMMMMMMMMMMooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiMMMMMMMMMMMMMMMMMMMMMoooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooo
```

```
Tsu-wht-2
662
YES
IN
6
1. 408-428, 2. 442-462, 3. 492-512, 4. 519-539, 5. 549-569,
6. 635-655
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMM
MMMMMMMMoooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiii
iiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooMMMMMMMMMMMMMMMMMMMMMi
iiiiiiiiMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiii
ii
```

Figure 3.95: A representative case that the TM domain of an ABC transporter gene was improved

“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; “ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. M514_01755 was annotated as a half ABC transporter gene in subfamily G but it was extremely long (~40 kb), resulting in a protein with 2744 aa. TM helices of M514_01755 clustered into three groups, which was unexpected. The revised gene encoded an ABC transporter with a length of 662 aa and six TM helices were present in this protein.

Through phylogenetic analysis, we found six out of 21 ABC transporter genes in *T. suis* showed one-to-one orthologous relationship with ABC transporter genes in *C. elegans*. We assigned gene names for ABC transporter genes in *T. suis* based on their relationship with ABC transporter genes in *C. elegans* (Figure 3.96). Similar to *T. trichiura*, when compared to *C. elegans*, the large number of gene contraction happened in subfamily B (4) and G (2) in *T. suis*. In subfamily D and H, there was no annotated ABC transporter gene *T. suis*. These gene losses were consistent in *T. suis and T. trichiura*, suggesting it happened in their common ancestor before speciation.

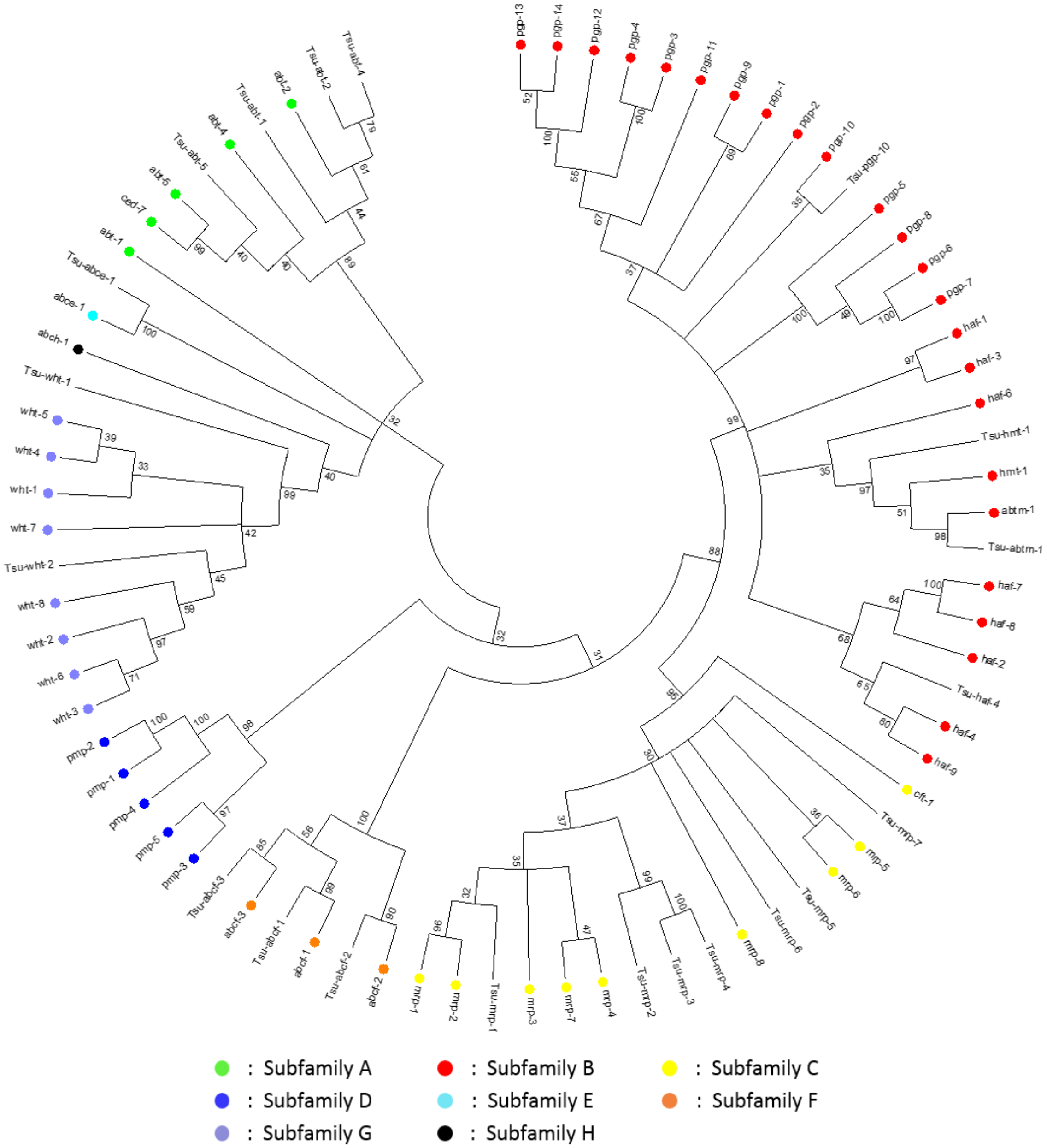


Figure 3.96: Phylogenetic analysis between *T. suis* and *C. elegans*
Phylogenetic tree was constructed using left ABC domain sequences of full transporters and the only ABC domain sequences of half transporters in *T. suis* and *C. elegans*. ABC transporter genes in *C. elegans* were highlighted by different color representing for different subfamilies. All ABC transporter gene names in *T. suis* were assigned based on ABC transporter genes in *C. elegans* by applying the name rule.

Chapter 4. Comparative analysis of ABC transporter genes in 29 nematode genomes

ABC systems show conservation in structure and function from bacteria to human (Higgins 1992) and the ABC genes are one of the few gene families that contain a large number of members in all eukaryotes, for which reason ABC genes are attractive for studying the evolution of gene families (Dean and Annilo 2005). In this study, we characterized high-quality ABC transporter genes in the genomes of 29 nematodes (Table 4.1). In general, the total number of ABC transporter genes in pathogenic nematodes generally are smaller than those found in non-pathogenic nematodes (Figure 4.1). The average number of ABC transporter genes in the non-pathogens is 58, while that in pathogens is 31. T-test results (2.39E-06) showed that the number of ABC transporters in pathogens vs non-pathogens are statistically significantly different, indicating that gene expansion in ABC transporter superfamily is not necessarily a mechanism for pathogenic nematodes to survive in their host environment, as proposed previously (Kikuchi et al. 2011). However, expansion of certain subfamilies could play a role it in pathogenicity.

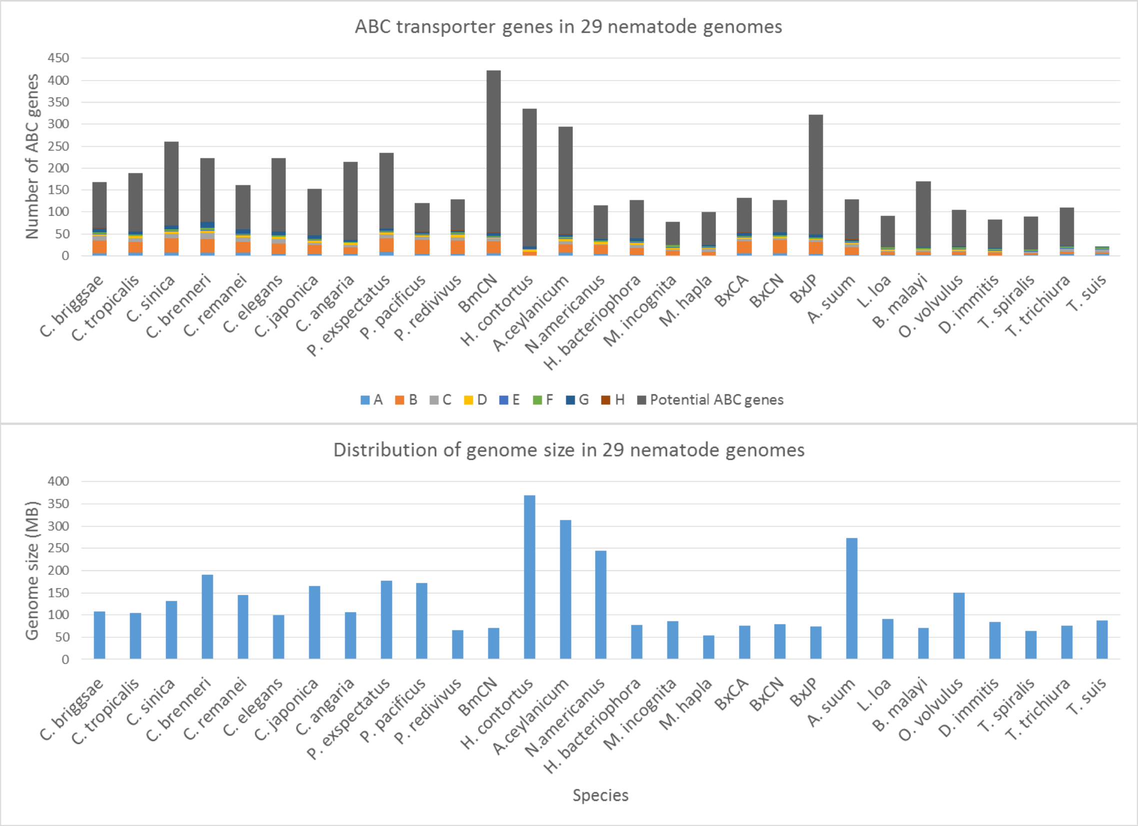


Figure 4.1: Distribution of ABC transporter genes in 29 nematode genomes. Different colors represent for different subfamilies. The total number of ABC transporter genes in pathogenic nematodes generally are smaller than those found in non-pathogenic nematodes.

Table 4.1: Subfamily information of high-quality ABC transporter genes in each nematode genome

Species	A	B	C	D	E	F	G	H	Total	Genome szie
C. briggsae	6	29	10	4	1	3	10	1	64	108MB
C. tropicalis	7	24	9	5	1	2	7	1	56	104MB
C. sinica	7	33	11	5	1	3	9	1	70	132MB
C. brenneri	8	30	14	6	1	5	13	1	78	190MB
C. remanei	8	24	10	5	1	3	9	1	61	145MB
C. elegans	5	24	9	5	1	3	8	1	56	100MB
C. japonica	4	20	6	5	1	3	8	0	47	166MB
C. angaria	5	15	5	5	0	1	6	0	37	106MB
P. exspectatus	9	32	8	5	1	2	5	0	62	177MB
P. pacificus	5	31	8	3	1	2	4	1	55	172MB
H. contortus	1	8	1	5	1	1	5	0	22	370MB
A.ceylanicum	7	19	7	6	1	3	6	1	50	313MB
N.americanus	5	19	3	4	1	3	4	0	39	244MB
H. bacteriophora	2	16	6	6	1	3	7	0	41	77MB
P. redivivus	5	30	7	7	1	3	5	1	59	65MB
M. incognita	1	11	3	3	2	4	0	0	24	86MB
M. hapla	1	9	6	2	1	3	2	0	24	54MB
BmCN	6	28	6	2	1	3	6	1	53	70MB
BxCA	6	28	5	3	1	3	6	1	53	76MB
BxCN	6	30	5	2	1	3	6	1	54	79MB
BxJP	5	26	5	2	1	3	6	1	49	74MB
A. suum	3	17	6	4	1	3	3	1	38	273MB
L. loa	3	6	3	3	1	3	1	0	20	91MB
B. malayi	2	7	4	1	1	3	2	0	19	71MB
O. volvulus	3	6	3	3	1	3	2	0	21	150MB
D. immitis	1	6	3	2	1	3	2	0	18	84MB
T. spiralis	3	3	4	0	2	3	0	0	15	64MB
T. trichiura	5	4	7	1	1	3	2	0	23	75MB
T. suis	4	4	7	0	1	3	2	0	21	87MB

The order of nematode species is arranged based on their orthologous relationship. The grey highlighted genomes represent for the genomes of pathogenic nematodes.

In order to compare our annotated ABC transporter genes in all 29 nematode genomes, we applied the program OrthoMCL (Li et al. 2003) to identify the ortholog groups among these ABC transporter genes

4.1. The conservation of subfamily E ABC transporter genes

ABCE gene is annotated as RNase L inhibitor in human (Bisbal et al. 1995). More recent data indicates that human ABCE protein has a central role in translation initiation (Chen et al. 2006). ABCE proteins are highly conserved among all eukaryotic species with over 90% identity in ABC domains of ABCE members across all eukaryotes (Kerr 2004) and over 65% identity in protein sequences of ABCE members between human and worm (Zhao et al. 2004a). Most of eukaryotic species have only one member, such as *C. elegans*, *D. rerio*, *D. melanogaster* and human, except for *Arabidopsis thaliana* (Zhao et al. 2004a)*,* suggesting the essential function of ABCE genes among different species.

To our expectation, we found that 26 out of 29 nematode genomes included in our analysis harbored only a single gene member in subfamily E. *Can-abce-1* in *C. anagria* was defective due to sequencing errors (mentioned in Charter 3). Interestingly, we found there were two expansions, one in *M. incognita*, another in *T. spiralis. Min-abce-1* and *Min-abce-2* were located in different contigs and their gene models showed some differences, especially in the intron region (Figure 4.2). To investigate whether these two genes in *M. incognita were* truly obtained from duplication, we extracted the upstream and downstream genomic region contained these two genes, as well as that contained *Mha-abce-2* in *M. hapla*. Surprisingly, not only the ABC transporter genes were conserved, proteins in the whole region shown in Figure 4.3, shared homologous relationship, indicating that there might be a duplication of a large region in *M. incognita*. Besides the similarity, the conserved regions also showed some variation, suggesting that after duplication, mutations were cumulated in these conserved region and differences emerged as well.

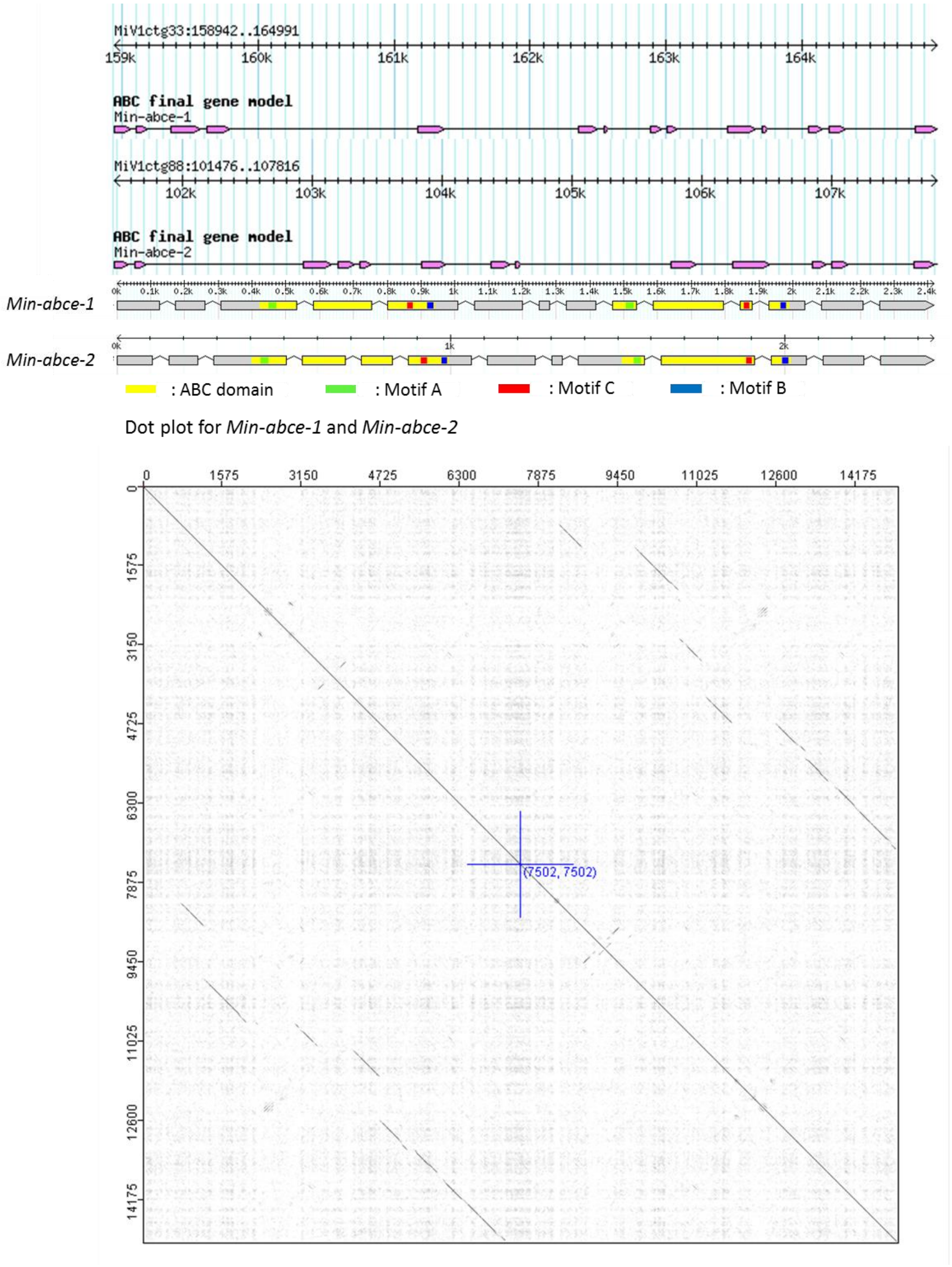


Figure 4.2: Expansion of ABCE subfamily in *M. incognita*
“ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. *Min-abce-1* and *Min-abce-2* did not share high similariy of their genomic region and gene structures were quite diverse.

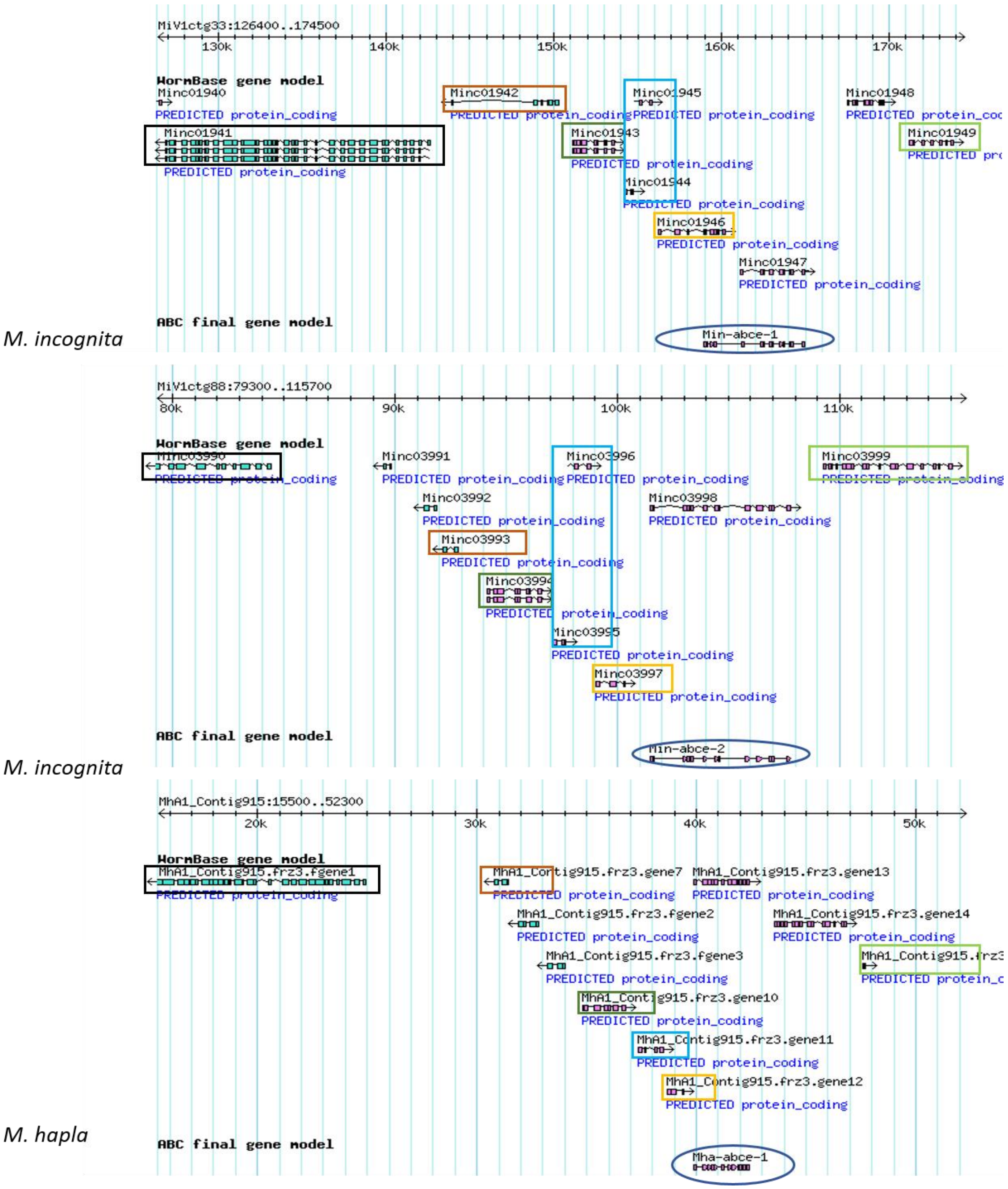


Figure 4.3: Duplication event in a region containing *Min-abce-1*
"WormBase gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. Genes highlighted with the same color shared homologous relationship. Two duplicated regions in *M. incognita* and their orthologous region in *M. hapla* showed that a number of genes duplicated in *M. incognita*.

For the ABCE expansion in *T. spiralis*, we checked the genomic region for each of them, and we found that *Tsp-abce-1* was located in a small contig only containing four genes. Gene structures of *Tsp-abce-1* and *Tsp-abce-2* were almost identical (Figure 4.4) and DNA alignment showed only few base differences between *Tsp-abce-1* and *Tsp-abce-2.* However, checking the upstream of *Tsp-abce-2*, identified a large sequencing gap. Together with the insufficient assembly of the contig containing *Tsp-abce-1*, it is not clearly demonstrated that the ABCE expansion in *T. spiralis* truly resulted from duplication, but the possibility still exists.

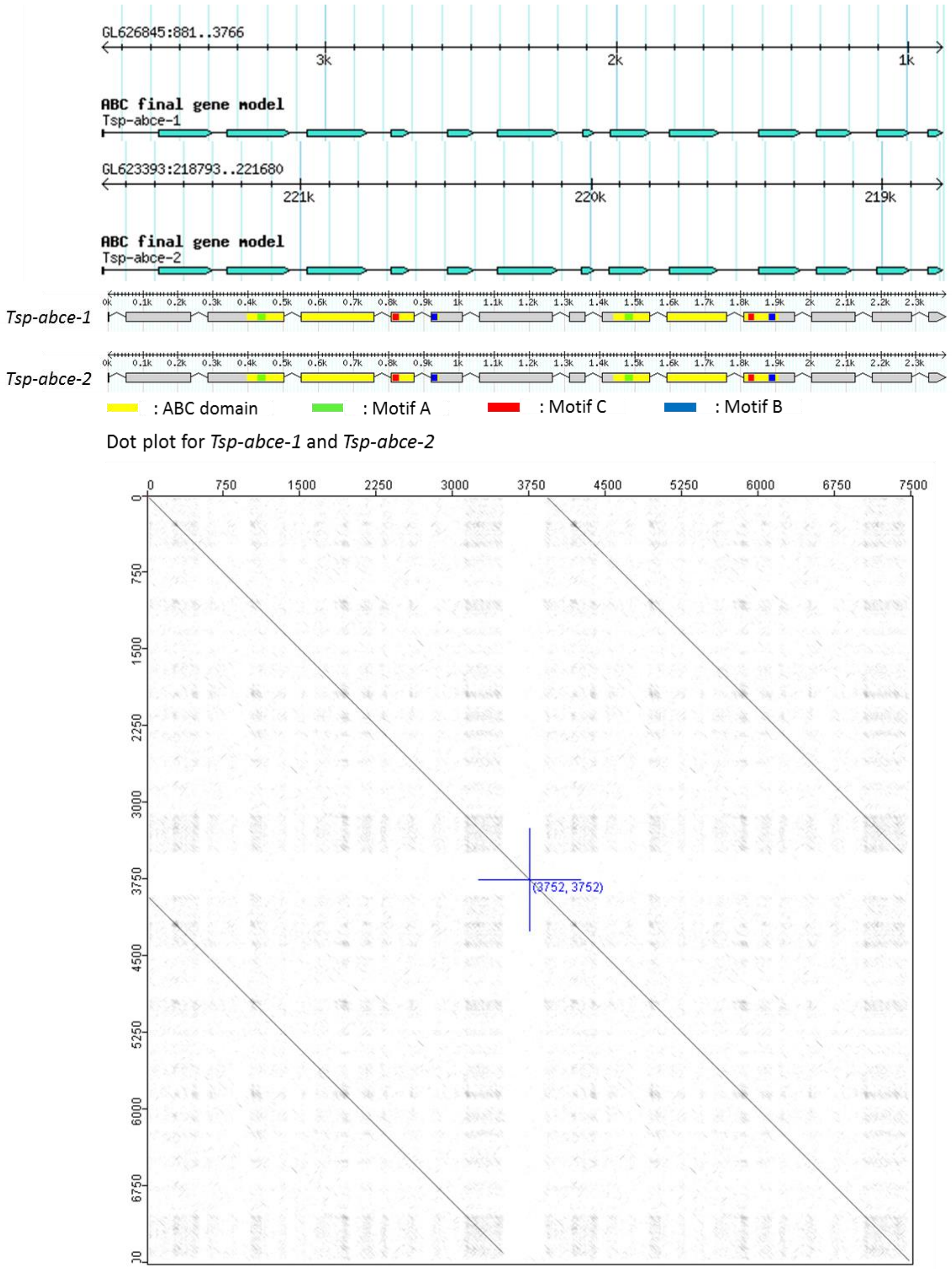


Figure 4.4: Expansion of ABCE subfamily in *T. spiralis*

“ABC final gene model” track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Gene structures of *Tsp-abce-1* and *Tsp-abce-2* were almost identical (Figure 4.3). And Dot plot showed high similarity between these two genes. *Tsp-abce-1* was located in a small contig while the upstream of *Tsp-abce-2* contained sequencing gap.

4.2. The conservation of subfamily F ABC transporter genes

ABCF proteins are important in ribosome assembly and protein translation (Marton et al. 1997; Tyzack et al. 2000). Subfamily F contains three conserved members in *C. elegans*, *A. gambiae*, *D. melanogaster*, all fish, and mammalian genomes examined (Dean and Annilo 2005), suggesting the non-redundant function that ABCF members harbored among different organisms during evolution.

As expected, most of the nematode genomes included in our analysis had three members in subfamily F: the orthologs of *abcf-1*, *abcf-2* and *abcf-3*, in *C. elegans*. All genomes contained a high-quality ortholog of *abcf-1* except that of *C. tropicalis* and that of *C. brenneri.* In *C. tropicalis*, there were two the defective candidates of *Ctr-abcf-1*, Csp11.Scaffold630.g21510.t1 had only one short ABC domain (91aa) and Csp11.Scaffold630.g21512.t2 had one of its ABC domain defective (36 aa). However, these defects were not due to technique issues (Figure 4.5). After comparing the genomic to those of *C. elegans*, *C. remanei* and *C. briggsae*, the strand of upstream and downstream conserved genes was reversed in *C. tropicalis*, making the whole region not conserved (Figure 4.5). Therefore, there could be an ABC transporter pseudogene in *C. tropicalis*. In contrast to *C. tropicalis*, three orthologs of *abcf-1* (*Cbn-abcf-1*, *Cbn-abcf-4* and *Cbn-abcf-5*) were found in *C. brenneri*. *Cbn-abcf-4* and *Cbn-abcf-5* were in the conserved region compared to those of *C. elegans*, *C. remanei* and *C. briggsae* (Figure 4.6). However, *Cbn-abcf-1* was in a small contig with only two genes. DNA alignment shows there three genes share a high similarity in most regions, suggesting that this ABCF expansion in *C. brenneri* was caused by tandem duplication (*Cbn-abcf-4* and *Cbn-abcf-5*) as well as heterozygosity (*Cbn-abcf-1*). Previous study found that 30% of *C. brenneri* genome are represented by two alleles in the assemblies (Barriere et al. 2009), which supports the hypothesis that the expansion of ABC transporter genes in *C. brenneri* resulted from technical problem, not real biological differences.

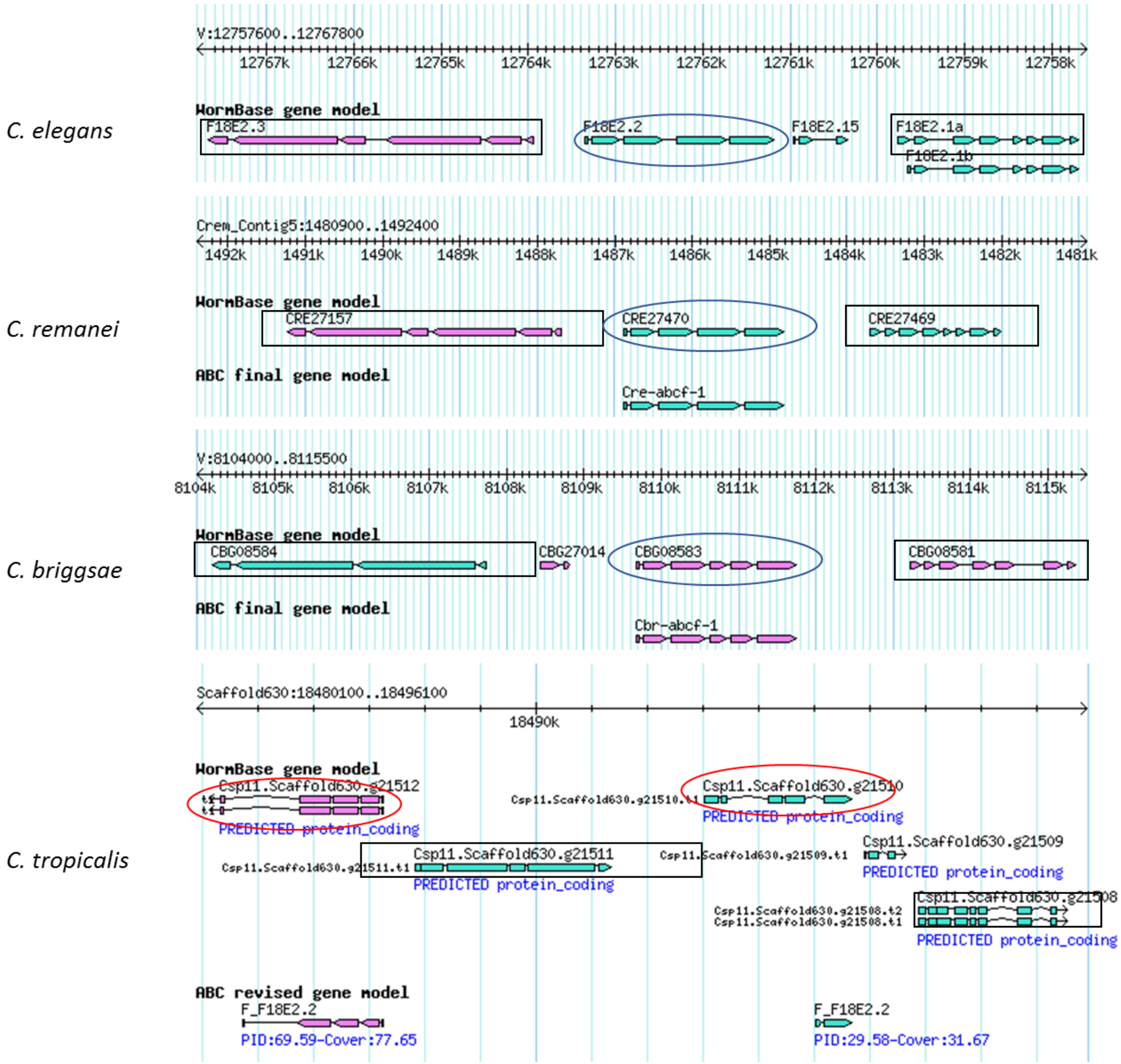


Figure 4.5: Ctr-abcf-1 was defective in *C. tropicalis*

“WormBase gene model” track includes original gene annotation; “ABC final gene model” track includes high-quality ABC transporter gene models. Genes in the black boxes shared orthologous relationship. In the conserved region, *C. elegans*, *C. remanei* and *C.briggsae*, all contained a high-quality ABC transporter genes. However, the orientation of Csp11.Scaffold630.g21511 was reverse, making this region not conserved. There were two defective candidates in this genomic region of *C. tropicalis*. Csp11.Scaffold630.g21510.t1 had only one short ABC domain (91aa) and Csp11.Scaffold630.g21512.t2 had one of its ABC domain defective (36 aa).

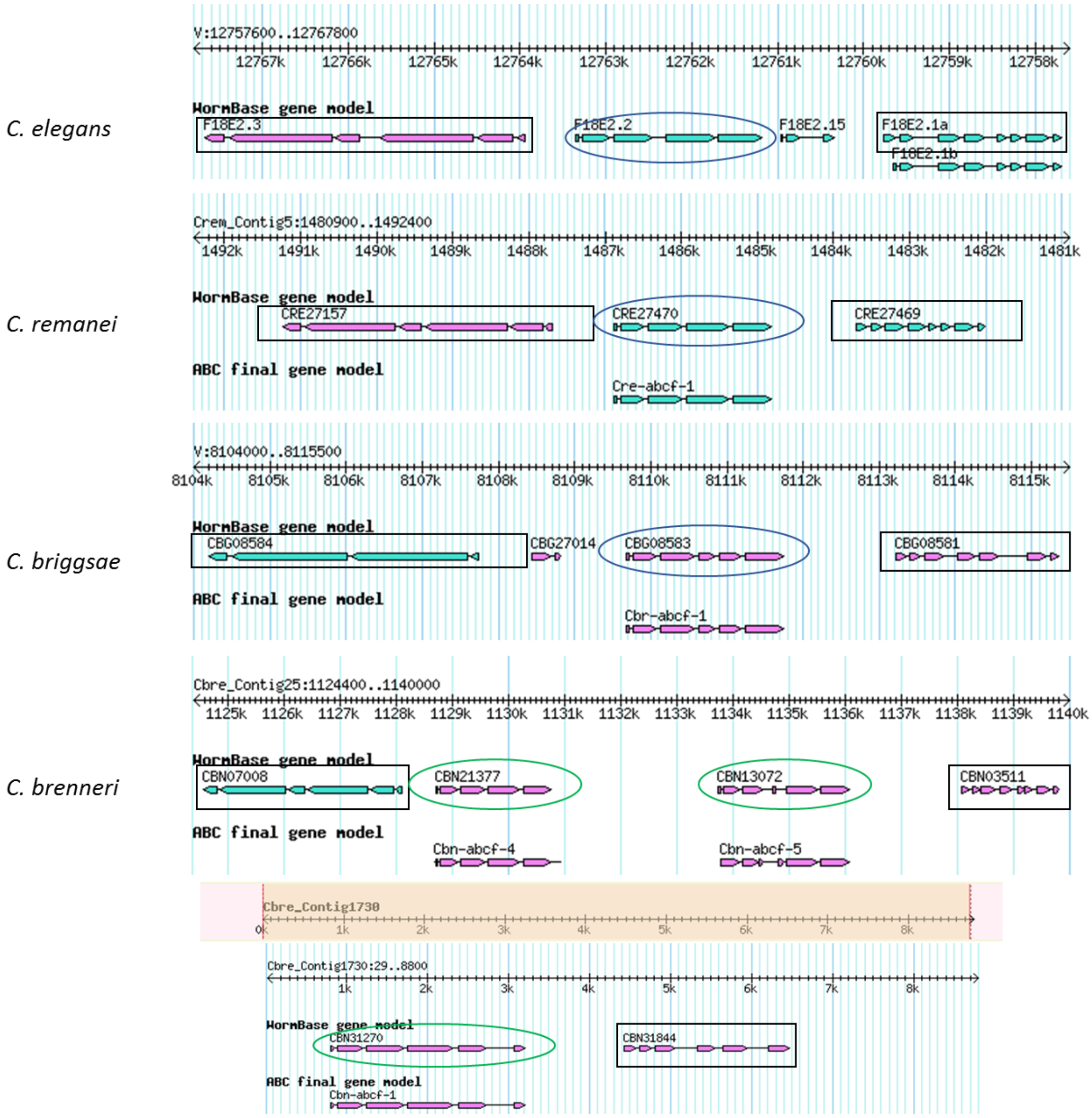


Figure 4.6: Expansion of ABCF subfamily caused by tandem duplication and heterozygosity in *C. brenneri*

“WormBase gene model” track includes original gene annotation; “ABC final gene model” track includes high-quality ABC transporter gene models. Genes in the black boxes shared orthologous relationship. Three orthologs of *abcf-1* were found in *C. brenneri*. *Cbn-abcf-4* and *Cbn-abcf-5* were in the conserved region compared to those of *C. elegans*, *C. remanei* and *C. briggsae*. *Cbn-abcf-1* was in a small contig with only two genes. DNA alignment shows there three genes share similarities, suggesting that this ABCF expansion in *C. brenneri* was caused by tandem duplication as well as heterozygosity.

We found single high-quality ortholog of *abcf-2* in all 29 nematode genomes*,* with the exception in *C. angaria*, *H. contortus* and *M. incognita.* The defective *Can-abcf-2* with one short predicted ABC domain (116 aa) was probably due to technique issue (Figure 4.7). Although without obviouse technical issues (Figure 4.8), *Hcon-abcf-2* candidate in *H. contortus* had two predicted ABC domains (44 aa, 6.4E-11; 99 aa, 3.1E-8), two of which were defective, suggesting that it could be a pseudogene in subfamily F. In *M. incognita*, expansion happened in subfamily F, leading to two orthologs of *abcf-2* (*Min-abcf-2 and Min-abcf-4*)*. Min-abcf-4* was a result of merging two adjacent genes and it showed high similarity to the genomic DNA of *Min-abcf-2* (Figure 4.9)*.* In addition, the gene structure and ABC domain sites of *Min-abcf-2* and *Min-abcf-4* were quite similar (Figure 4.9). Furthermore, similar to the expansion of ABCE subfamily in *M. incognita*, it seems that the duplication event was included a large region that also contained *Min-abcf-2* (Figure 4.10) when comparing *M. incognita* and *M. halpa.*

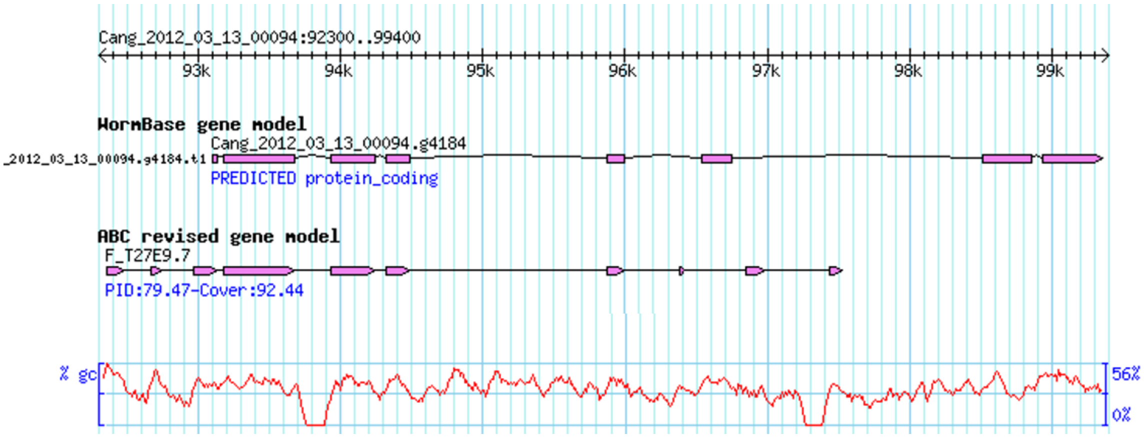


Figure 4.7: Incompleteness of *Can-abcf-2* due to sequencing errors
"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. The defective *Can-abcf-2* with one short predicted ABC domain (116 aa) was probably caused by sequencing gaps within the gene.

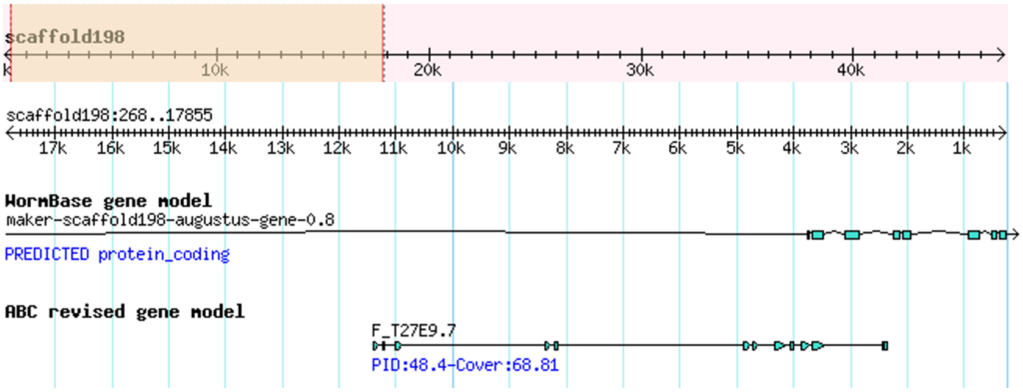


Figure 4.8: Pseudogene in *H. contortus* in subfamily F
"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. *Hcon-abcf-2* candidate in *H. contortus* had two predicted domains (44 aa, 6.4E-11; 99 aa, 3.1E-8), two of which were defective but there was no obvious technical issues, suggesting that it could be a pseudogene in subfamily F.

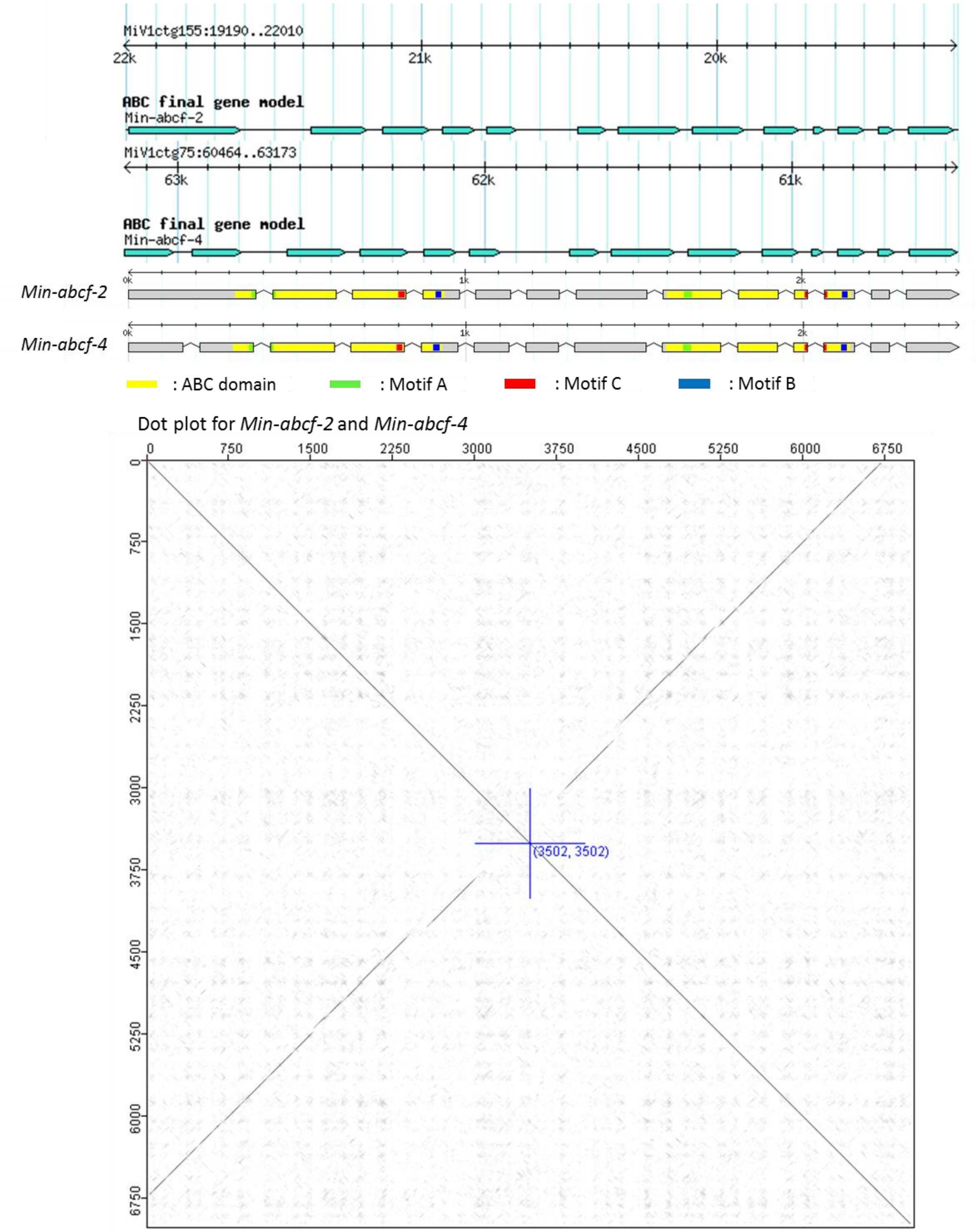


Figure 4.9: Expansion of ABCF subfamily in *M. incognita*

"ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. *Min-abcf-4* showed high similarity to the genomic sequences of *Min-abcf-2*. The gene structure and ABC domain sites of *Min-abcf-2* and *Min-abcf-4* were quite similar.

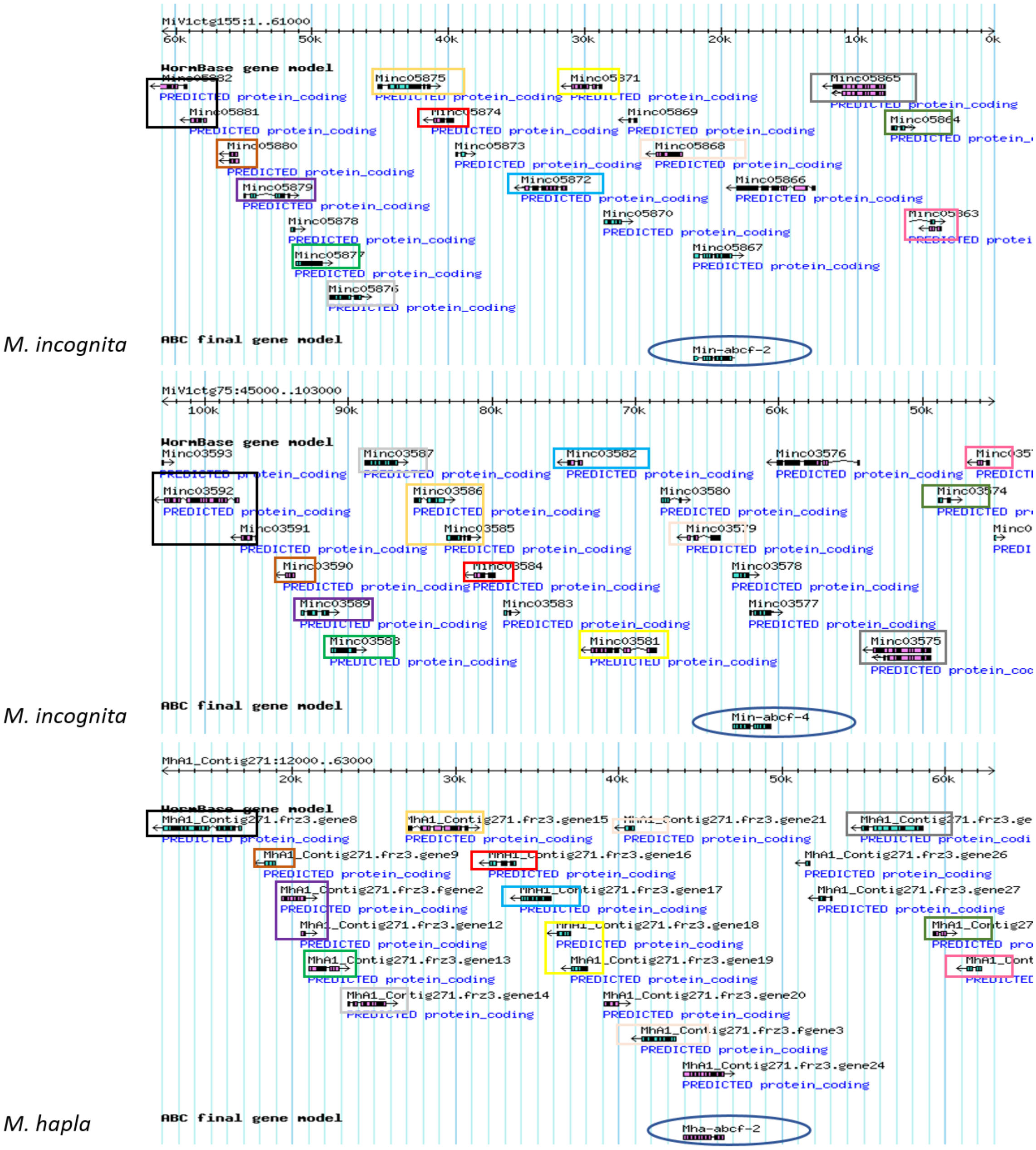


Figure 4.10: Duplication event in a region containing *Min-abcf-2*
"WormBase gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. Genes highlighted with the same color shared homologous relationship. Two duplicated regions in *M. incognita* and their orthologous region in *M. hapla* showed that a number of genes duplicated in *M. incognita*.

The last orthologous group in subfamily F clustered orthologs of *abcf-3.* We characterized high-quality ABC transporter genes in 25 nematode genomes. For the remaining four genomes, three of them (*C. angaria*, *P. pacificus* and *H. contortus*) had potential candidates which could be improved to be high-quality ABC transporter genes when the genome is well sequenced and assembled (Figure 4.11). Surprisingly, *P. expectatus* did not contain any annotated protein coding gene that could be improved to be *Pex-abcf-3.* To confirm that whether the gene loss is true in *P. expectatus,* genBlastG was applied using *abcf-3* protein sequence searching against entire genomic sequence of *P. expectatus*. Interestingly, we annotated a high-quality ABC transporter gene in a region that did not have any predicted gene model previously (Figure 4.12). This newly annotated gene was characterized as *Pex-abcf-3*. Therefore, 26 out of 29 genomes contained a single high-quality ortholog of *abcf-3* in each*.*

In conclusion, through our analysis, we found that every single ABC transporter gene in subfamily F were highly conserved during nematode evolution with only minor exceptions.

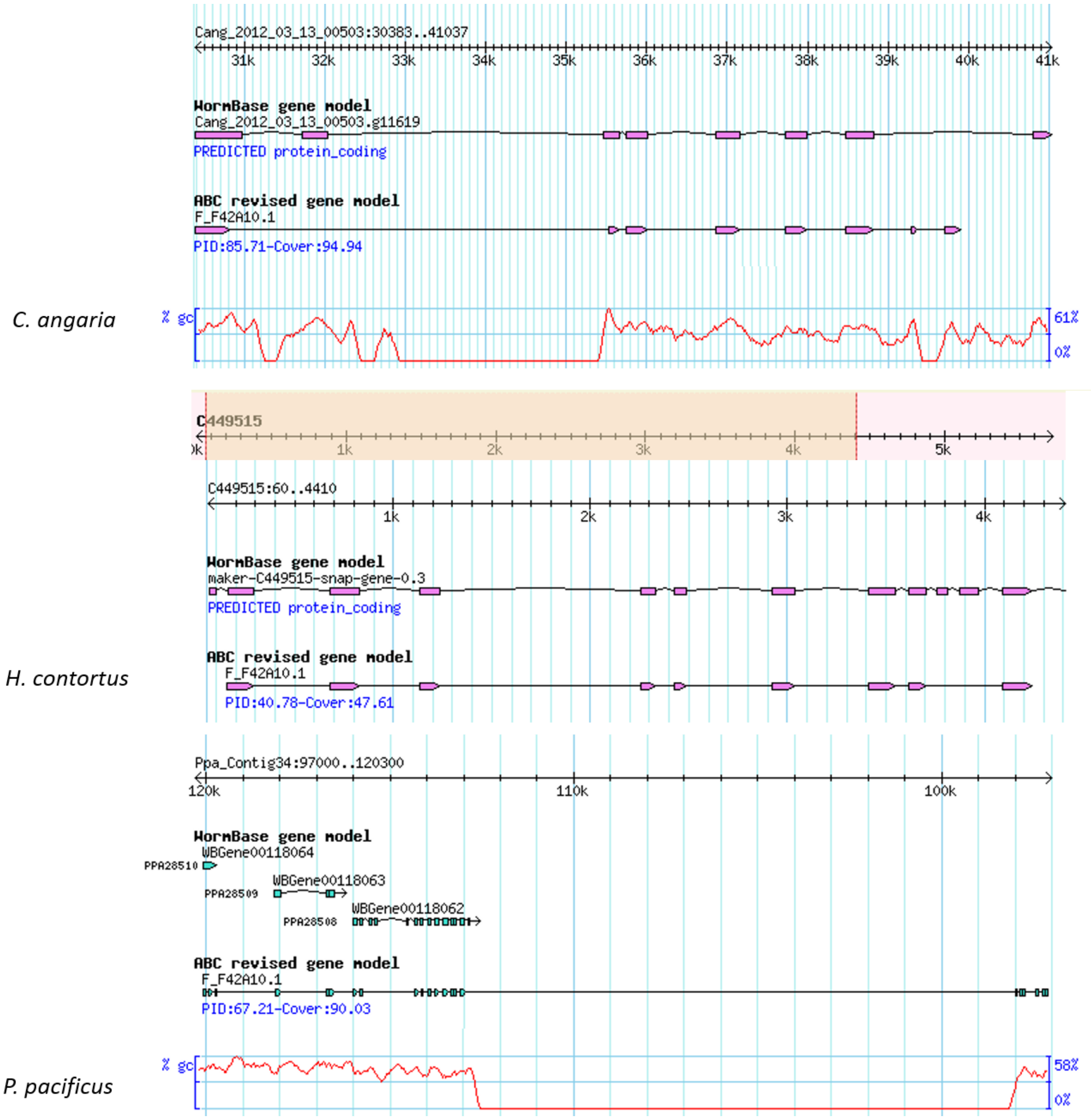


Figure 4.11: Incompleteness of *Can-abcf-3*, *Hco-abcf-3* and *Ppa-abcf-3* caused by technical issues

"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. *C. angaria*, *P. pacificus* and *H. contortus* had potential candidates which had defective ABC domains but might be improved to be high-quality ABC transporter genes when the genome is well sequenced and assembled.

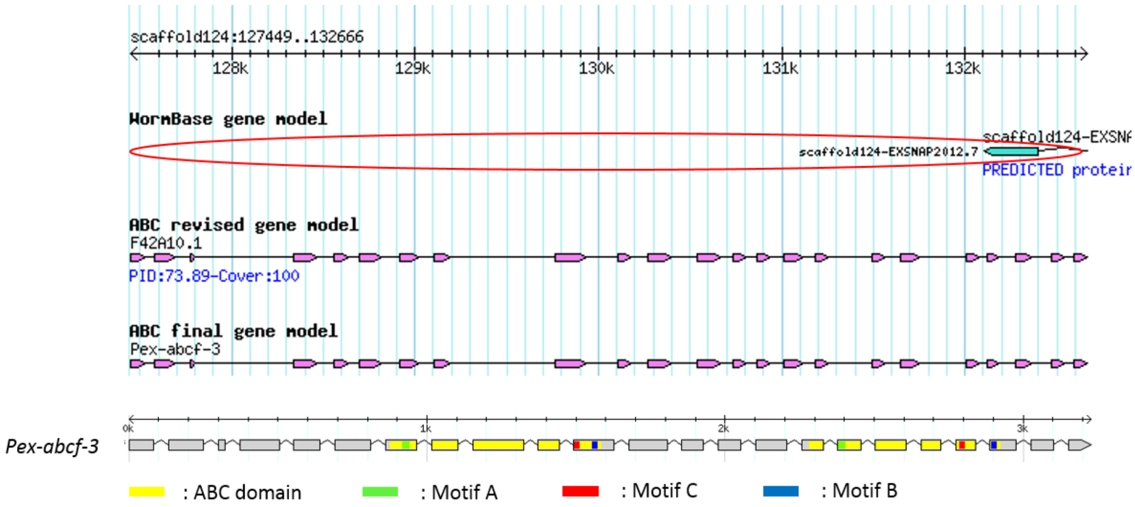


Figure 4.12: The gene model of *Pex-abcf-3* was annotated in our analysis
"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Original gene annotation failed to annotate the gene model of *Pex-abcf-3*. Through our genBlastG search, *Pex-abcf-3* was obtained and examined to be a high-quality ABC transporter gene.

4.3. *abtm-1* encodes a highly conserved ABC transporter of subfamily B

ATM is a half ABC transporters in subfamily B that has been identified in yeast, plants and mammals, which are required for cytosolic and nuclear Fe-S cluster assembly (Kispal et al. 1999; Pondarre et al. 2006; Bernard et al. 2009). In *C. elegans*, *abtm-1* was characterized as a widely expressed mitochondrial protein. It is an essential gene and knock down resulted in pleiotropic phenotypes (Gonzalez-Cabo et al. 2011). Worms with deleted *abtm-1* are arrested in embryonic stage and had morphogenetic defects and unusual premature, putative apoptotic events (Gonzalez-Cabo et al. 2011). The ortholog of *abtm-1* in human, ABCB7, is related to a disease called X-linked sideroblastic anemia with ataxia (XLSA/A) in humans (Maguire et al. 2001) Therefore, ATMs might be functionally essential in exporting compound from mitochondria during evolution.

Our analysis revealed that *abtm-1* was strongly conserved among most nematode species. Through our annotation pipeline, we found a single high-quality ortholog of *abtm-1* in 23 genomes, and two copies of *abtm-1* in three genomes (*C. brenneri*, *M. incognita*, and BmCN). In *C. brenneri*, *Cbn-abtm-1* and *Cbn-abtm-2* were found in different contigs both with sequencing gap in the downstream genomic region (Figure 4.13). However, the upstream genes were also quite similar to each other. Together with the other 12 ABC transporter gene expansions mentioned in Chapter 3 (Table 3.6), *Cbn-abtm-1* and *Cbn-abmt-2* were most likely caused by heterozygosity of *C. brenneri.*

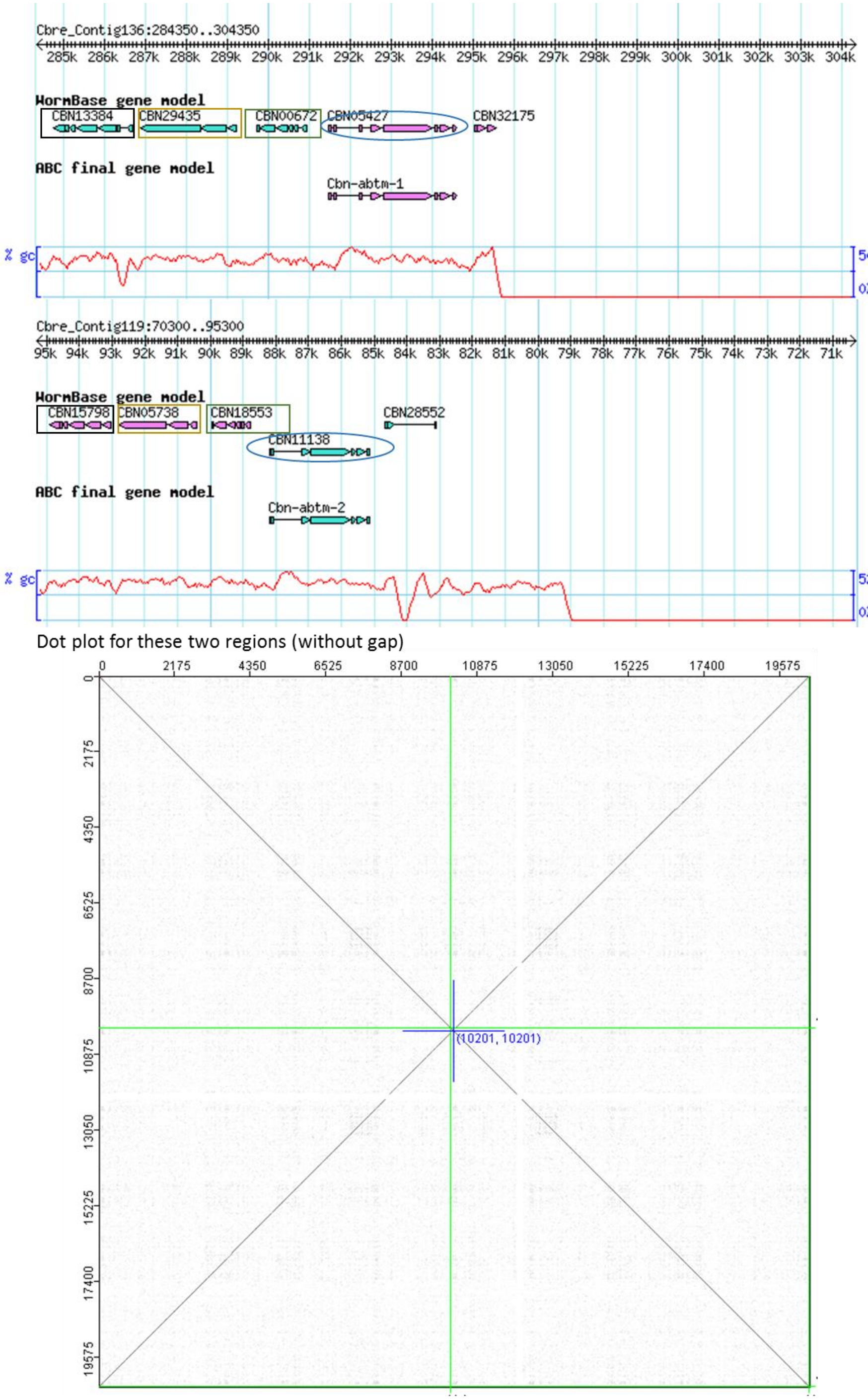


Figure 4.13: Two orthologs of *abtm-1* in *C. brenneri* most likely caused by heterozygosity

"WormBase gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. *Cbn-abtm-1* and *Cbn-abtm-2*, located in different contigs, had sequencing gap in the downstream genomic region. Dot plot showed that genomic sequences of the upstream genes were quite similar to each other, suggesting that this gene expansion was caused by heterozygosity.

M. incognita, also possessed two copies, *Min-abtm-1* and *Min-abtm-2* which shared both similar gene structure and similar genomic sequences (Figure 4.14). When compared with its closely related genome, *M. hapla*, we found not only ABC transporter gene was duplicated, but almost the entire contig containing *Min-abtm-2* (Figure 4.15). Together with the ABC transporter gene expansions in subfamily F and subfamily E, it is very surprising that southern root-knot nematode went through duplications in a large scale. Besides, both the genome size (86MB) and protein coding genes (19212) in *M. incognita* (Abad et al. 2008) is larger than those of *M. hapla* (54 MB and 16676, respectively) (Opperman et al. 2008), which could be explained by the large scale duplication events in *M. incognita* after speciation.

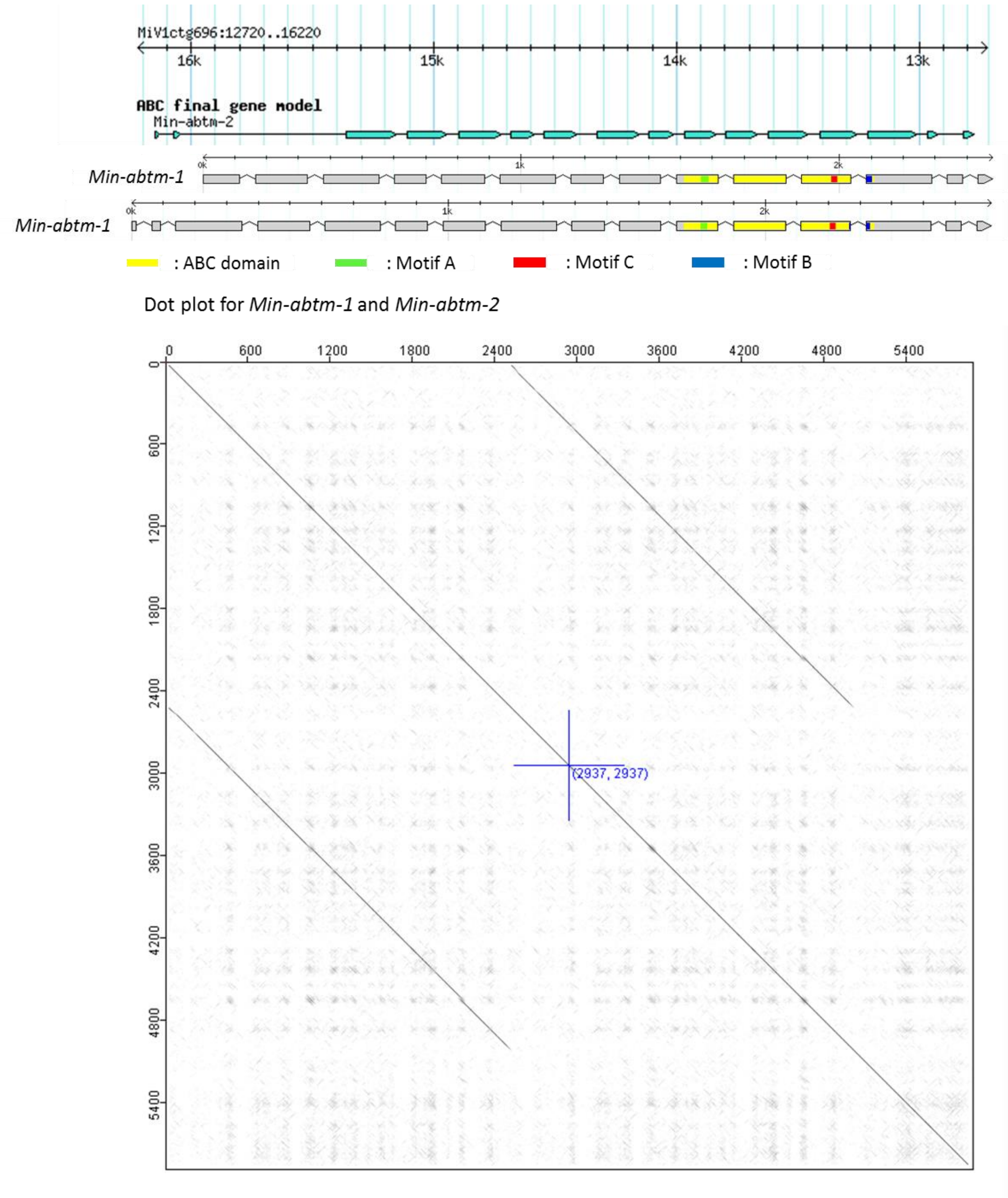


Figure 4.14: Two orthologs of *abtm-1* in *M. incognita* due to duplication
"ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. *Min-abtm-1* and *Min-abtm-2* share highly similar gene structures and genomic sequences.

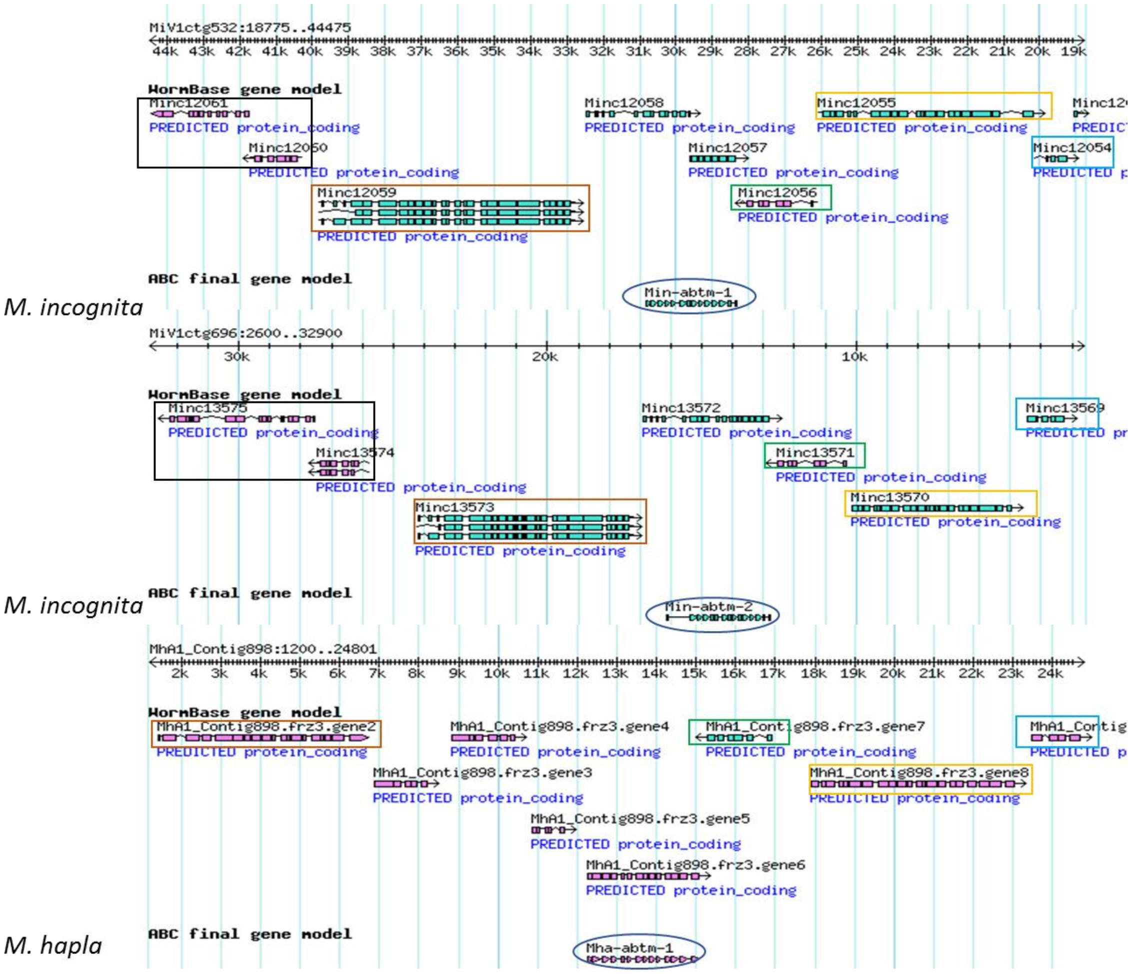
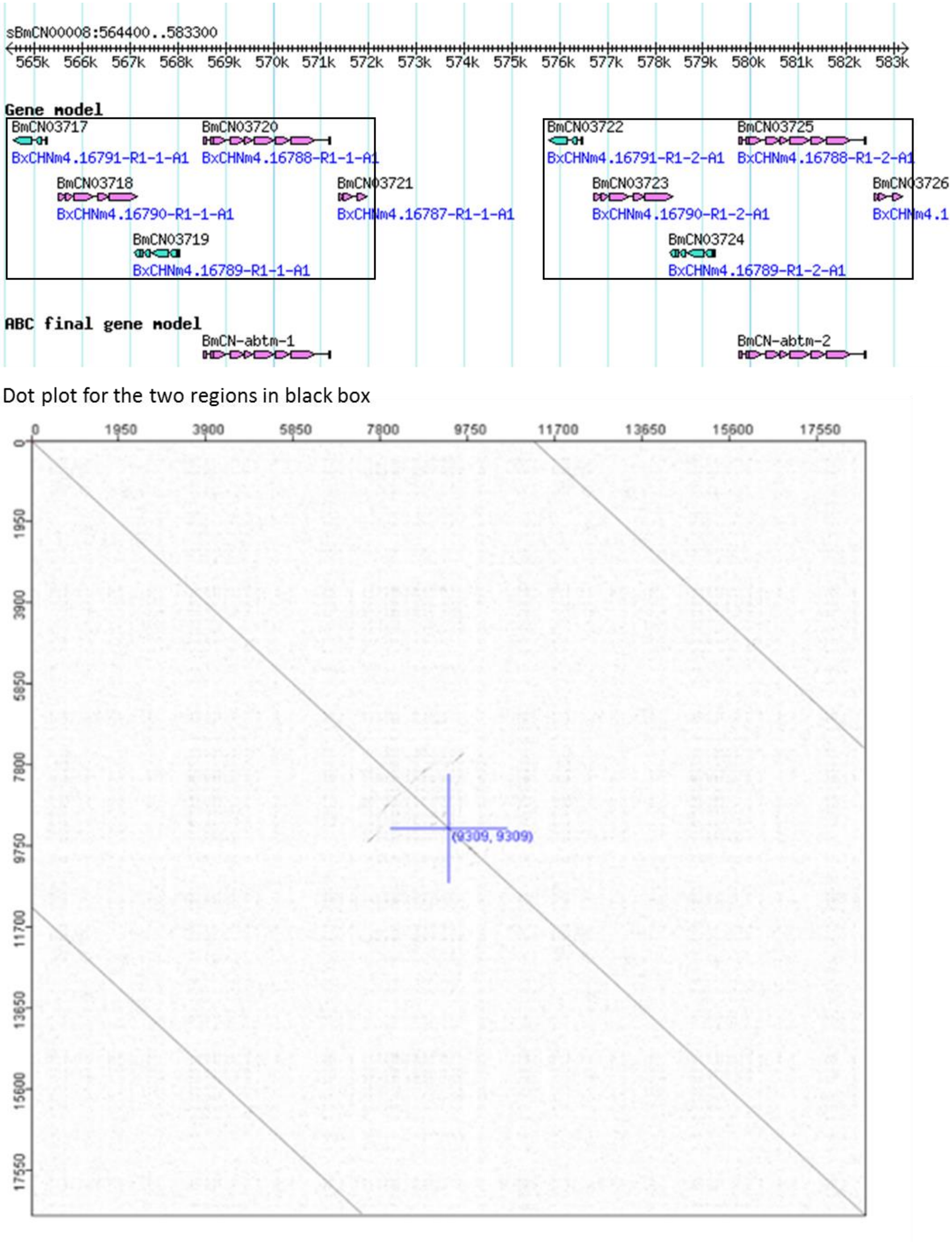


Figure 4.15: Duplication event in a region containing *Min-abtm-1*
"WormBase gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. Genes highlighted with the same color shared homologous relationship. Two duplicated regions in *M. incognita* and their orthologous region in *M. hapla* showed that a number of genes duplicated in *M. incognita*.

In *B. mucronatus*, we also identified two copies, *BmCN-abtm-1* and *BmCN-abtm-2*, which were in the same contig. When zooming out the genomic region, we found that two clusters, each of which contained five genes, were highly similar to each other (Figure 4.16). Dot plot suggested that there was a tandem duplication event that happened in five adjacent genes (Figure 4.16), leading to the expansion of ABC transporter gene in BmCN.



Dot plot for the two regions in black box

Figure 4.16: Tandem duplication in BmCN leading to two orthologs of *abtm-1*
“Gene model” track includes original gene annotation; “ABC final gene model” track includes high-quality ABC transporter gene models. There was one BmCN specific duplication event in which five adjacent genes in the black box duplicated tandem. *BmCN-abtm-1* was one of the five genes. Therefore, the duplication contributed one more gene (*BmCN-abtm-2*) of subfamily B in BmCN genome.nm

Three genomes (*C. remanei*, *P. pacificus* and *P. exspectatus*) did not have annotated high-quality ortholog of *abtm-1*. *C. remanei* had a defective candidate with sequencing gap its genomic region (Figure 4.17). It could be a high-quality ABC transporter genes after the genome is reconstructed. *P. pacificus* and *P. exspectatus* both did not contain any annotated protein coding gene that could be improved to be *Ppa-abtm-1* or *Pex-abtm-1.* Considering that this gene was conserved among 26 nematode genomes, we further searched for ortholog of *abtm-1* in the whole genomic sequence of *P. pacificus* and *P. exspectatus*. As we expected, the loss of *Ppa-abtm-1* and *Pex-abtm-1* was not true. It was due to the mis-annotation in the previous studies. After checking the quality of both gene models, each of them encodes one high-quality ABC domain, suggesting that they are high-quality ABC transporter genes (Figure 4.18).

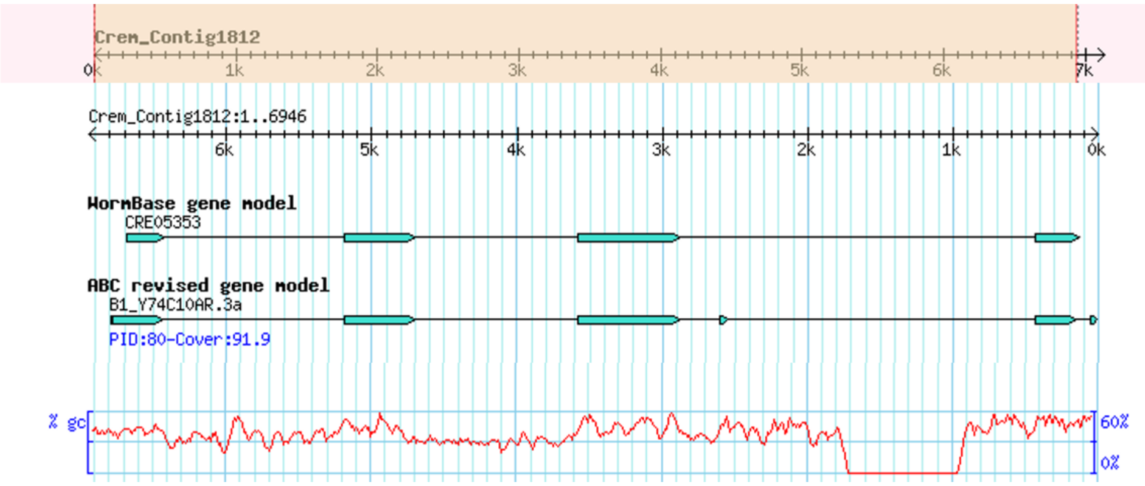


Figure 4.17: Incompleteness of *Cre-abtm-1* caused by technical issues
“WormBase gene model” track includes original gene annotation; “ABC revised gene model” track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins. CRE05353 was annotated to be *Cre-abtm-1* but it was located in a small contig with sequecing gap, which might cause the ABC domain to be defective (100 aa).

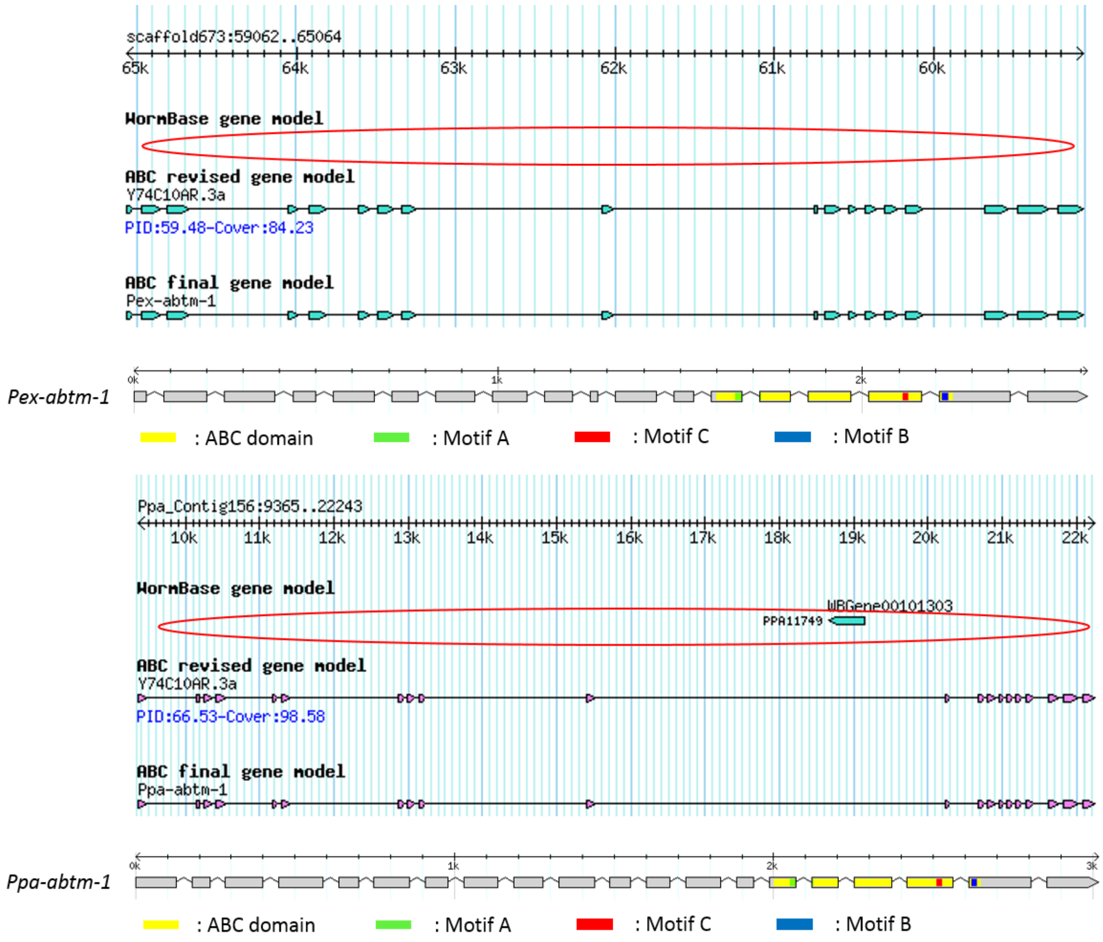


Figure 4.18: The gene model of *Pex-abtm-1* and *Ppa-abtm-1* was annotated in our analysis

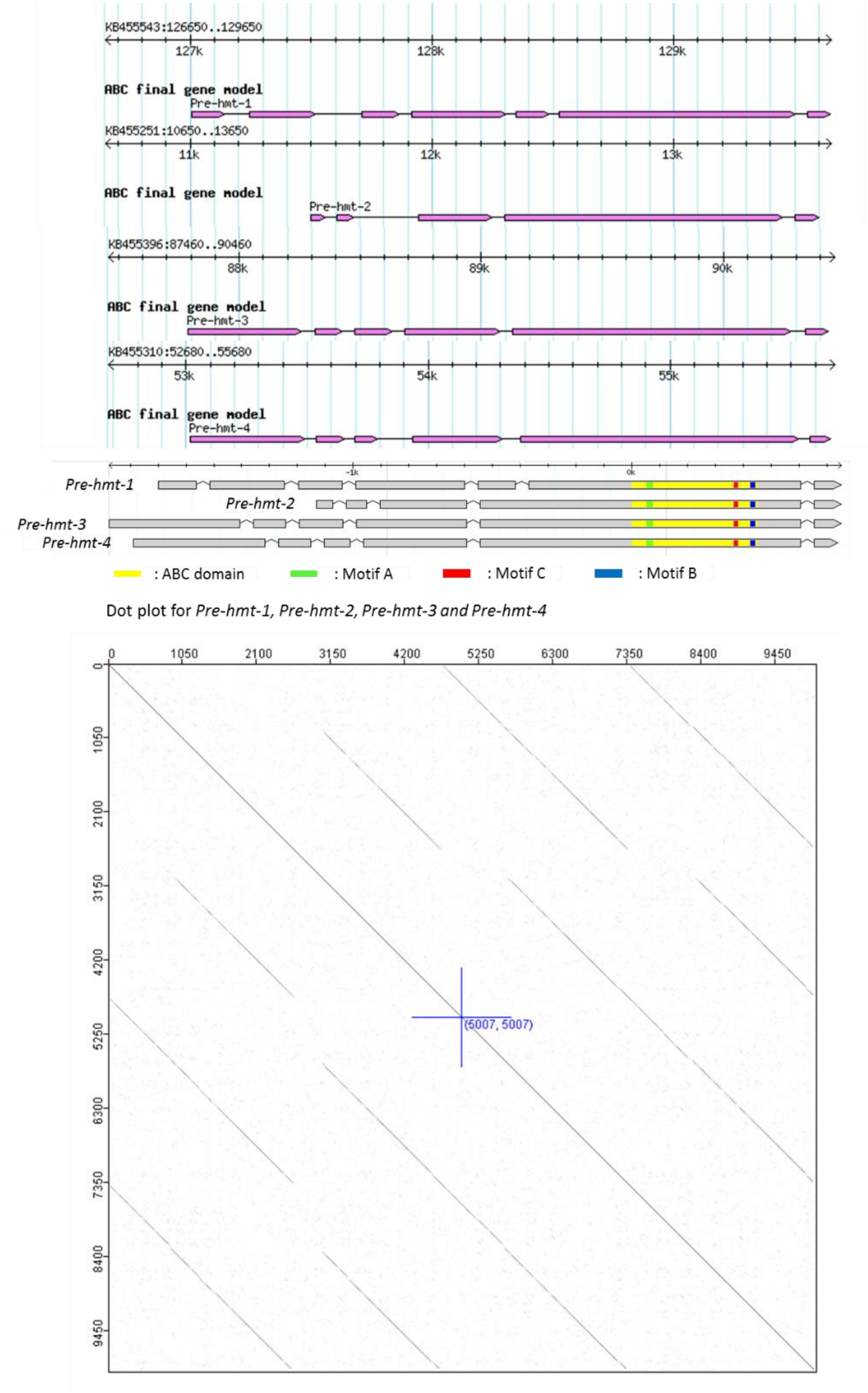
"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. Original gene annotation failed to annotate the gene model of *Pex-abtm-1* and *Ppa-abtm-1*. Through our genBlastG search, *Pex-abtm-1* and *Ppa-abtm-1* were obtained and examined to be a high-quality ABC transporter gene.

4.4. *hmt-1* encodes a conserved ABC transporter of subfamily B

In *C. elegans, hmt-1* is confirmed to be required for heavy metal detoxification and is expressed in coelomocytes, head neurons, and intestinal cells (Schwartz et al. 2010). HMT-1 counterpart of humans, ABCB6, is expressed in similar tissues and cell types which are affected by heavy metals (Mitsuhashi et al. 2000; Uriu-Adams and Keen 2005; Valko et al. 2005; Krishnamurthy et al. 2006; Bressler et al. 2007). Other than *C. elegans* and human, HMTs are identified in yeast, fly and mammals (Sooksa-Nguan et al. 2009), suggesting that HMTs were generally conserved during evolution.

Although no ortholog was found in *M. incognita*, *L. loa*, *B. malayi*, *O. volvulus* and *D. immitis* even after searching the entire genomic sequences, the remaining 24 nematode genomes contained at least one copy of this ABC transporter gene. Expansions occurred in *Bursaphelenchus* group with two copies in all four genomes, which will be explained together with the *Bursaphelenchus* specific ABC transporter genes later. Interestingly, there were four copies, *Pre-hmt-1*, *Pre-hmt-2*, *Pre-hmt-3* and *Pre-hmt-4* in *P. redivivus*, located in four different contigs. Their gene models were quite diverse but the ABC domain parts were conserved (Figure 4.19).

Our analysis found both HMT expansion and loss during nematode evolution. The expansion of HMT in *P. redivivus* may explain the high level of copper tolerance reported in *P. redivivus*, which has been shown to have higher tolerance to copper than *C. elegans* or *P. pacificus* (Boyd and Williams 2003). Gene loss in the five nematode mentioned above indicated that these nematodes might not need to detoxify heavy metals or have developed other mechanisms.



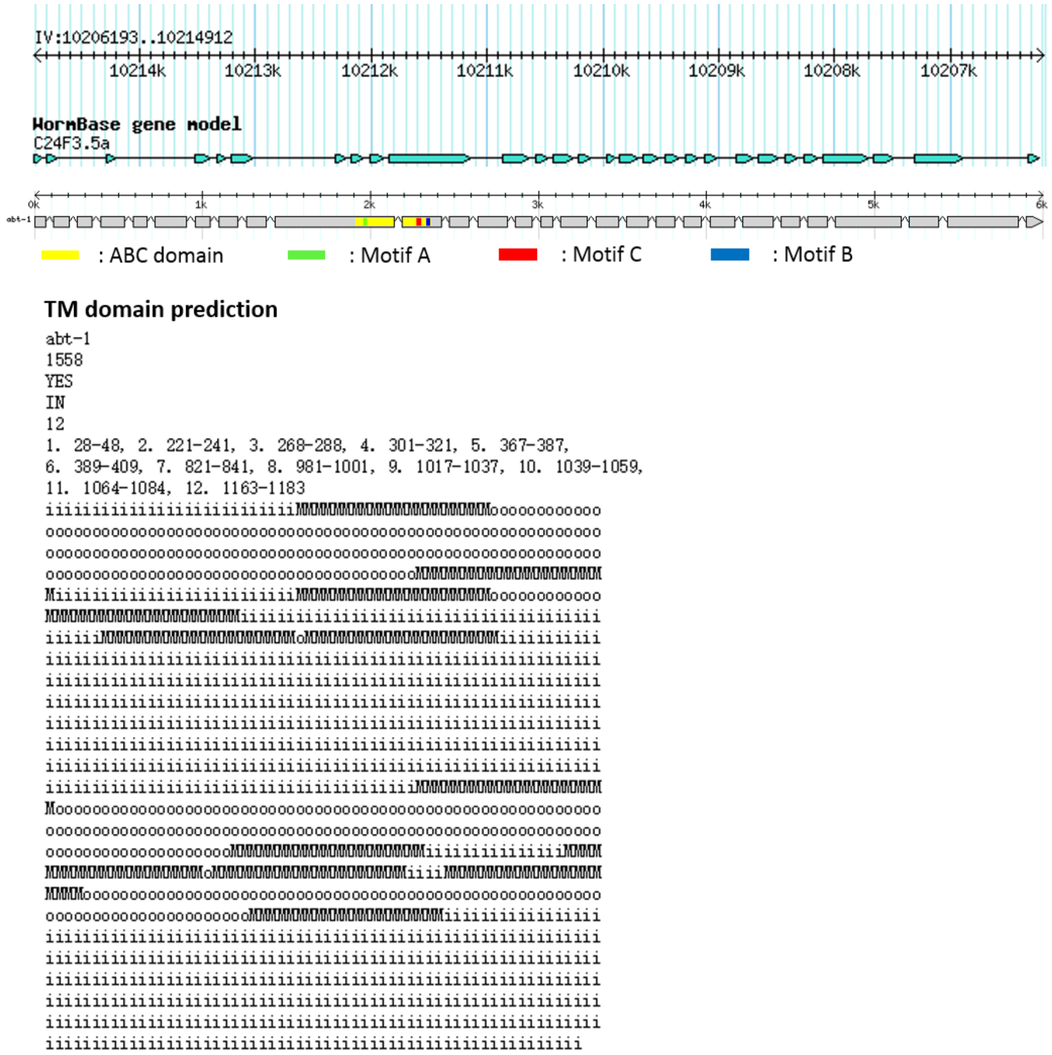
Dot plot for *Pre-hmt-1, Pre-hmt-2, Pre-hmt-3 and Pre-hmt-4*

Figure 4.19: Four orthologs of *abtm-1* in *P. redivivus*

"ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The genomic sequences of *Pre-abtm-1*, *Pre-abtm-2*, *Pre-abtm-3* and *Pre-abtm-4* showed similarities to each other. Their gene models were quite diverse but the ABC domain parts were conserved.

4.5. Genome-specific ABC transporters in *Caenorhabditis*

Our analysis identified a genomic specific ABC transporter genes in *Caenorhabditis*. We found that *abt-1* which was previously annotated in *C. elegans,* only contained one ABC domain but two TM domains (Figure 4.20). Interestingly, we found the orthologs of *abt-1* in other five *Caenorhabditis* species clustered in a single orthologous group in OrthoMCL result. After aligning their gene models, we noticed that *Cre-abt-7*, *Cbn-abt-1* and *Cja-abt-1* were not full length ABC transporter genes because of the technical issues (Figure 4.21). In *C. angaria*, there was one *Can-abt-1* candidate with the e-value of predicted ABC domain (3.0E-8) lower than our criteria. Therefore, we excluded it from our high-quality ABC transporter gene list in *C. angaria*. However, it could be a diverse ortholog of *abt-1*. In summary, we concluded that all the eight *Caenorhabditis* species should have this half transporter gene when the genome is well sequenced and assembled. The loss of second ABC domain might happen in the common ancestor of these eight *Caenorhabditis* species. As loss of *abt-1* activity via RNAi results in no obvious defective phenotype, the precise role of *abt-1* in *C. elegans* development and/or behavior is not yet known and it could be a pseudogene after losing the second ABC domain.



TM domain prediction

```
abt-1
1558
YES
IN
12
1. 28-48, 2. 221-241, 3. 268-288, 4. 301-321, 5. 367-387,
6. 389-409, 7. 821-841, 8. 981-1001, 9. 1017-1037, 10. 1039-1059,
11. 1064-1084, 12. 1163-1183
iiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMM
MiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMoooooooooooo
MMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiMMMMMMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMM
Mooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiMMMM
MMMMMMMMMMMMMMMMMoMMMMMMMMMMMMMMMMMMMMMiiiiMMMMMMMMMMMMMMMMM
MMMMoooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
```

Figure 4.20: Predicted ABC domain and TM domains in *abt-1* in *C. elegans*
“WormBase gene model” track includes original gene annotation. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. In *C. elegans*, *abt-1* were found to only contain one ABC domain but two TM domains.

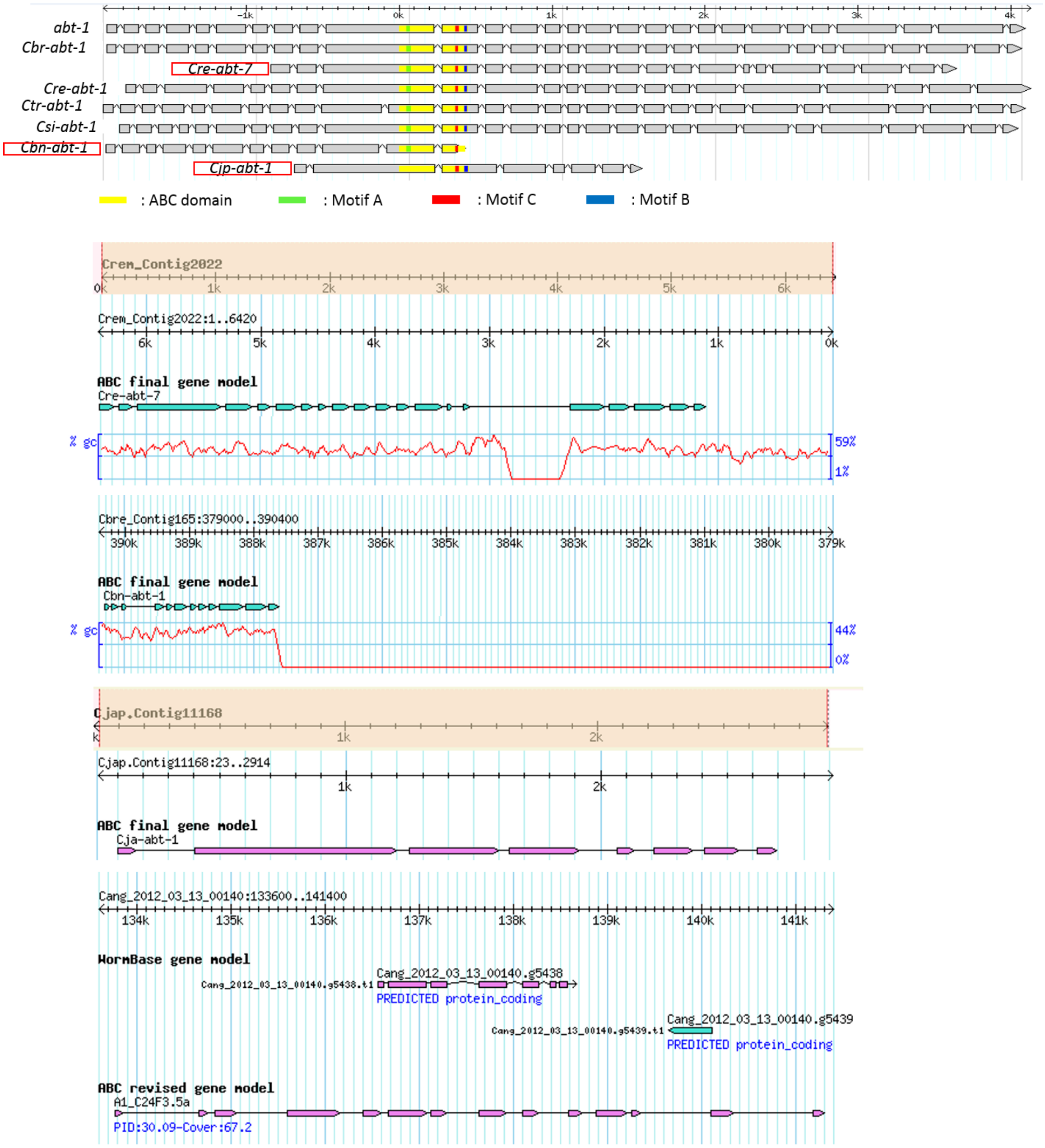


Figure 4.21: *Caenorhabditis* specific gene in subfamily A.
"WormBase gene model" track includes original gene annotation; "ABC revised gene model" track includes gene models annotated using genBlastG and *C. elegans* orthologs as query proteins; "ABC final gene model" track includes high-quality ABC transporter gene models. Yellow highlighted region represents for ABC domains; Green highlighted region represents for Walker A; Red highlighted region represents for Signature C; Blue highlighted region represents for Walker B. The gene structures of Caenorhabditis specific gene in subfamily A were quite similar. The incompleteness of *Cre-abt-7*, *Cbn-abt-1* and *Cja-abt-1* was caused by technical issues. In *C. angaria*, there was one *Can-abt-1* candidate with the e-value of predicted ABC domain lower (3.0E-8) than our criteria and the diverse gene structure.

4.6. Genome-specific ABC transporters in *Bursaphelenchus*

We identified genome specific genes in subgroup of subfamily B in four *Bursaphelenchus* genomes (Table 4.2). In the comparison to *C. elegans* (Chapter 2), there was a subgroup in subfamily B that did not contain any ABC transporter genes in *C. elegans*. When comparing to all ABC transporter genes in 29 nematode genomes using OrthoMCL, we found this subgroup in *Bursaphelenchus* clustered together without any additional genes from other species. After closely checking the genomic region, we identified that most of these genome specific genes were adjacent in each *Bursaphelenchus* genome (Figure 4.22). *BxCN-abcb-3*, *BxCN-abcb-3*, *BxCA-abcb-3* and *BmCN-abcb-3* were the orthologs of *hmt-1* mentioned in Chapter 4.4*. BxCN-abcb-8* and *BmCN-abcb-8* were two genes in this subgroup only specific to BmCN and BxCA (Table 4.2) since we could not found any potential candidates in BxCN and BxJP within the conserved genomic region (Figure 4.23). Similar to *abt-1*, we do not know what exact function of these *Bursaphelenchus* specific genes yet, but the merging of these genes demonstrated the rapid gene gain event during the evolution of ABC transporter genes in nematode genomes.

Table 4.2: *Bursaphelenchus* specific ABC transporter genes in subfamily B

Genome	Group1	Group2	Ortholog of *hmt-1*
BxCN	*BxCN-abcb-4* *BxCN-abcb-5* *BxCN-abcb-6* *BxCN-abcb-7*	*BxCN-abcb-1* *BxCN-abcb-2*	*BxCN-abcb-3*
BxJP	*BxJP-abcb-4* *BxJP-abcb-5* *BxJP-abcb-6*	*BxJP-abcb-1* *BxJP-abcb-2*	*BxJP-abcb-3*
BxCA	*BxCA-abcb-4* *BxCA-abcb-5* *BxCA-abcb-6*	*BxCA-abcb-1* *BxCA-abcb-2* *BxCA-abcb-7*	*BxJP-abcb-3*
BmCN	*BmCN-abcb-4* *BmCN-abcb-5* *BmCN-abcb-6*	*BmCN-abcb-1* *BmCN-abcb-2* *BmCN-abcb-7*	*BmCN-abcb-3*

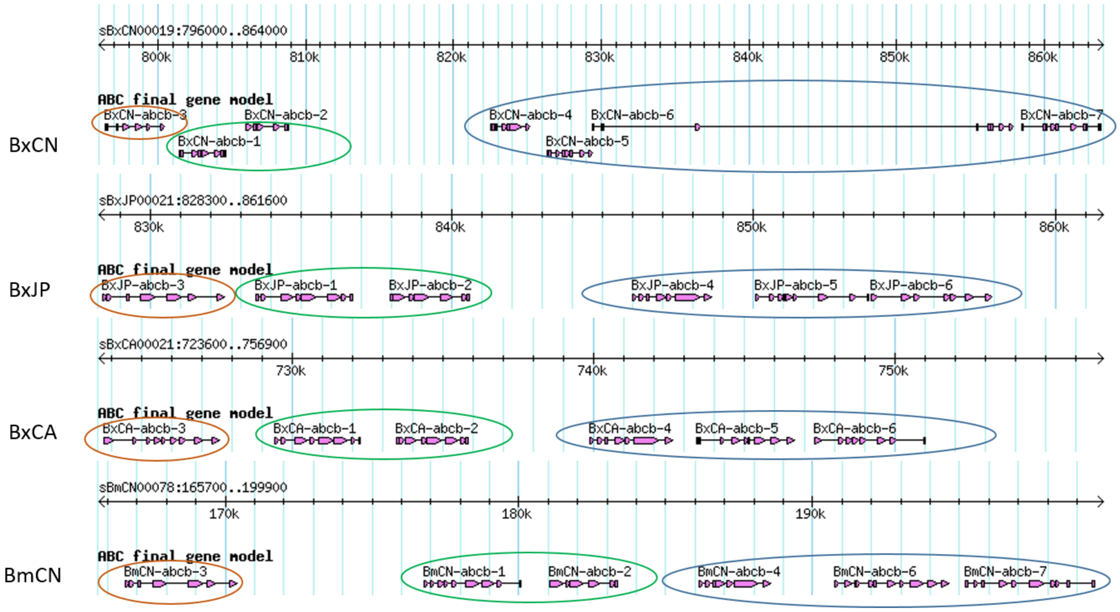


Figure 4.22: *Bursaphelenchus* specific ABC transporter genes in subfamily B
"ABC final gene model" track includes high-quality ABC transporter gene models. In each genome, *Bursaphelenchus* specific genes were clustered together. Genes circled by the same color were in the same orthologous group of OrthoMCL result.

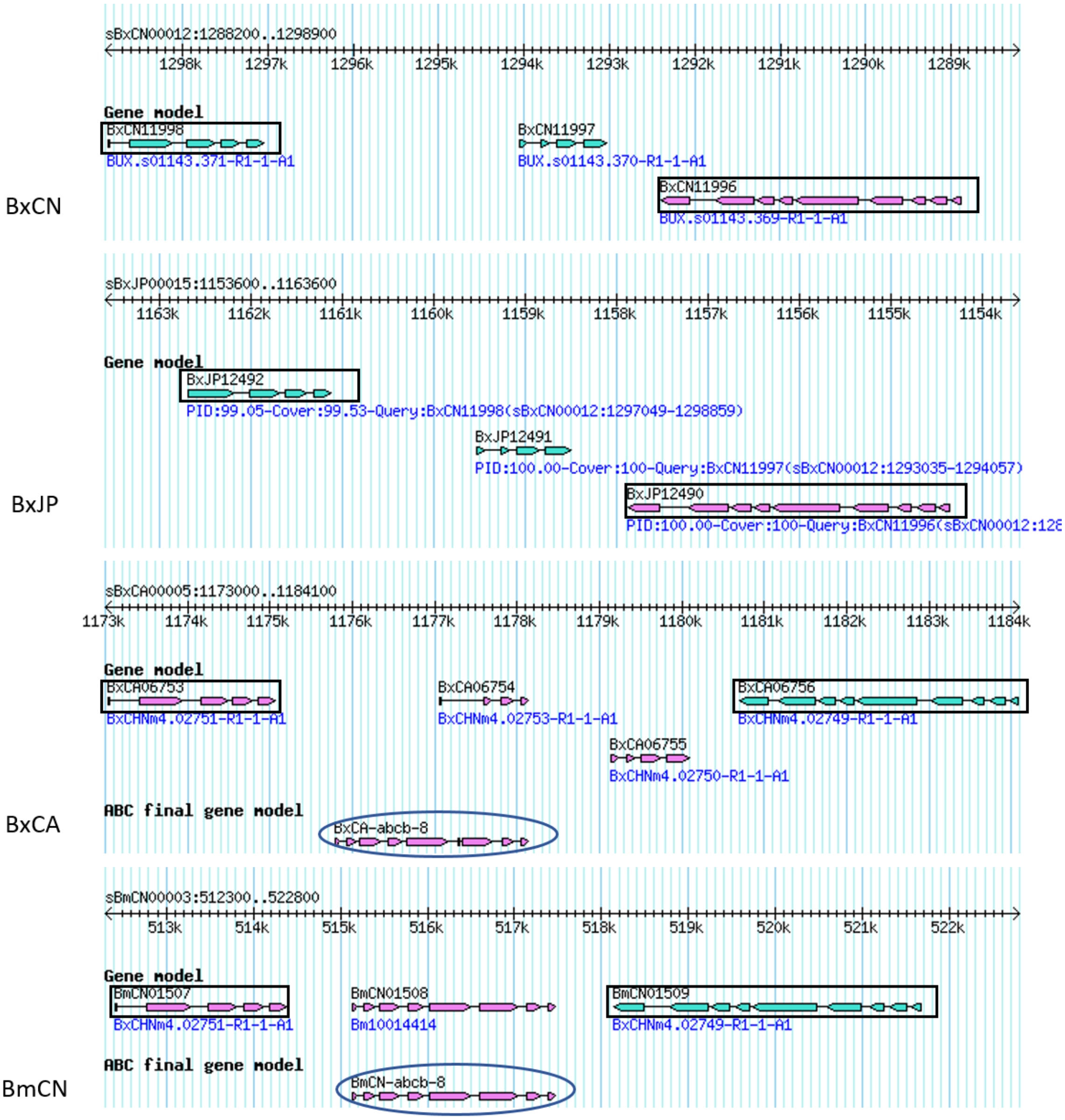


Figure 4.23: ABC transporter gene in subfamily B specific to BmCN and BxCA
"Gene model" track includes original gene annotation; "ABC final gene model" track includes high-quality ABC transporter gene models. The genes in the black boxes are conserved within four *Bursaphelechus* genomes. Within the conserved region, we cannot find any putative ABC transporter genes in BxCN and BxJP.

4.7. Conclusion

Based on the comparative analysis of ABC transporter genes between each nematode species and *C. elegans,* we know that although ABC transporter gene family is conserved, there are gene duplication and gene loss among these nematode genomes. Except for subfamily E and F, other subfamilies were quite diverse among different nematodes (Table 4.1). Some of these differences may reflect the genome assembly quality. We found that a subfamily B member *abtm-1* was strongly conserved among all 29 nematode species. *hmt-1* was relatively conserved with both expansion and contraction during nematode evolution. Species specific ABC transporter genes in *Caenorhabditis* and *Bursaphelenchus* showed the diversification of ABC transporter genes. In general, pathogenic nematodes contained less ABC transporter genes, probably because they have less variable and more protected environment and they lost some genes that are no longer needed. In contrast, free-living nematode would have many more of these transporters to cope an uncertain environment where they would likely be exposed to a large number of toxins and pathogens.

Chapter 5. Conclusion

In this study, we have developed, tested and successfully applied a robust bioinformatics pipeline that uses *C. elegans* ABC transporter genes as reference to annotate high-quality ABC transporter genes in nematode genomes. The bioinformatics pipeline uses InterProScan and BLAST to search for putative ABC transporters, followed by improving gene models using a homology-based gene finder, genBlastG. A high-quality ABC transporter gene contains appropriate number of ABC domains with appropriate length (130 aa – 165 aa) and necessary motifs (Walker A, Signature C, Walker B). We have demonstrated the effectiveness of this bioinformatics pipeline by using it to search for ABC transporter genes in *C. elegans*. Furthermore, we have applied this bioinformatics pipeline to search for ABC transporter genes in *C. briggsae* and *C. remanei*. In addition to finding almost all ABC transporter genes in these two genomes, we have found additional high-quality ABC transporter genes (Table 5.1).

Table 5.1: Comparison of ABC transporter genes in the genome of *C. briggsae* and *C. remanei* obtained from previous study and our analysis

		A	B	C	D	E	F	G	H	Total
Previous annotation	*C. briggsae*	6	23	9	5	1	3	9	2	58
	C. remanei	7	23	9	5	1	3	9	2	59
Annotation of this study	*C. briggsae*	6	29	10	4	1	3	10	1	64
	C. remanei	8	24	10	5	1	3	9	1	61

In addition to the genomes of *C. elegans*, *C. briggsae*, and *C. remanei*, we have applied this bioinformatics pipeline to search for ABC transporter genes in 26 additional nematode genomes. Among these 29 nematode organisms, 12 are non-pathogens, while 17 are pathogens. In general, the ABC transporter gene family sizes are larger in the non-pathogens than in the pathogens. The average number of ABC transporter genes in the non-pathogens is 58, while that in pathogens is 31. A previous study reported 106 ABC transporter genes in the genome of *B. xylophilus*, which is a pathogen of pine tree (Kikuchi et al. 2011). However, our study characterized only 49 high-quality ABC transporter genes in the same genome. Similar numbers of ABC transporter genes have been found in BxCN (54), which is a strain of *B. xylophilus* isolated in China and BxCA (53), which is a strain isolated in Canada. We found that many ABC transporter genes were annotated as partial genes, thus inflating the number of ABC transporter genes in the previous report. Therefore, the previous hypothesis that the highly expanded family of ABC transporter genes may facilitate the invasion and pathogenicity of PWN were proved to be incorrect. More importantly, the contradictory results illustrate that precise annotation of ABC transporter genes is required before we come up with any reasonable explanations.

Through phylogenetic analysis of ABC transporter genes, we found that some ABC transporter genes are very well conserved, while others show genome-specificity. Through comparative analysis using OrthoMCL, we found that ABC transporter genes in subfamily E and F are well conserved. Members of other subfamilies are quite diverse among different nematodes. Subfamily B showed substantial variations in the numbers of ABC transporter genes, ranging from three to 33 in different nematode genomes. Within subfamily B, the mitochondrial ABC transporter gene *abtm-1* is highly conserved among all 29 nematode genomes, with expansions only in *M. incognita* and BmCN. Thus, *abtm-1* may play an essential role in exporting compound from mitochondria during evolution of nematodes. Another subfamily B gene, *hmt-1*, also show strong conservation with at least one ortholog in 23 nematode genomes. *M. incognita*, and four animal parasites (*L. loa*, *B. malayi*, *O. volvulus* and *D. immitis*) do not harbor any putative ortholog of *hmt-1,* whereas small expansion with two copies occurrs in *Bursaphelenchus* group and large expansion with four copies in *P. redivivus* are observed. This case of gene loss and gain can reflect the adaptation of different life surroundings in nematodes. The lack of ABCH subfamily in some nematode genomes suggests that this subfamily is diverse and actively evolving.

In conclusion, this study provided a robust bioinformatics method to identified high quality ABC transporter genes in nematode genomes, which may contribute to understand the evolution of nematodes and how different inventory of ABC transporters could affect the interaction between nematodes and their surrounding environment. However, precise number of ABC transporter genes only can be obtained when the genomes are fully sequenced and assembled, as that of *C. elegans.* Therefore, more efforts needed to be done in getting complete genomes. Furthermore, because that ABC superfamily is a large and diverse family which plays a role in many cellular transport functions in nematodes as well as in anthelmintic resistance, future studies will involve the examination of tissue distribution of ABC transporter genes and then identification of function of ABC transporter genes in these nematodes. After that, we can even explore the gene networks of ABC transporters to provide a more comprehensive perspective to ABC transporter genes in different nematode species.

References

Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EG, Deleury E, Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC et al. 2008. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat Biotechnol* **26**: 909-915.

Abraham D, Leon O, Leon S, Lustigman S. 2001. Development of a recombinant antigen vaccine against infection with the filarial worm Onchocerca volvulus. *Infection and immunity* **69**: 262-270.

Akasaka T, Klinedinst S, Ocorr K, Bustamante EL, Kim SK, Bodmer R. 2006. The ATP-sensitive potassium (KATP) channel-encoded dSUR gene is required for Drosophila heart function and is regulated by tinman. *Proc Natl Acad Sci U S A* **103**: 11999-12004.

Allen JD, Brinkhuis RF, Wijnholds J, Schinkel AH. 1999. The mouse Bcrp1/Mxr/Abcp gene: amplification and overexpression in cell lines selected for resistance to topotecan, mitoxantrone, or doxorubicin. *Cancer Res* **59**: 4237-4241.

Allikmets R, Gerrard B, Hutchinson A, Dean M. 1996. Characterization of the human ABC superfamily: isolation and mapping of 21 new genes using the expressed sequence tags database. *Human molecular genetics* **5**: 1649-1655.

Ardelli BF. 2013. Transport proteins of the ABC systems superfamily and their role in drug action and resistance in nematodes. *Parasitology international* **62**: 639-646.

Ardelli BF, and Roger K. Prichard. 2008. Effects of Ivermectin and Moxidectin on the Transcription of Genes Coding for Multidrug Resistance Associated Proteins and Behaviour in Caenorhabditis elegans. *Journal of nematology* **40.4**: 290.

Ardelli BF, Prichard RK. 2013. Inhibition of P-glycoprotein enhances sensitivity of Caenorhabditis elegans to ivermectin. *Veterinary parasitology* **191**: 264-275.

Ardelli BF, Stitt LE, Tompkins JB. 2010. Inventory and analysis of ATP-binding cassette (ABC) systems in Brugia malayi. *Parasitology* **137**: 1195-1212.

Bai X, Adams BJ, Ciche TA, Clifton S, Gaugler R, Kim KS, Spieth J, Sternberg PW, Wilson RK, Grewal PS. 2013. A lover and a fighter: the genome sequence of an entomopathogenic nematode Heterorhabditis bacteriophora. *PloS one* **8**: e69618.

Ballatori N, Rebbeor JF, Connolly GC, Seward DJ, Lenth BE, Henson JH, Sundaram P, Boyer JL. 2000. Bile salt excretion in skate liver is mediated by a functional analog of Bsep/Spgp, the bile salt export pump. *Am J Physiol Gastrointest Liver Physiol* **278**: G57-63.

Bard SM. 2000. Multixenobiotic resistance as a cellular defense mechanism in aquatic organisms. *Aquat Toxicol* **48**: 357-389.

Barriere A, Yang SP, Pekarek E, Thomas CG, Haag ES, Ruvinsky I. 2009. Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. *Genome Res* **19**: 470-480.

Bartley DJ, McAllister H, Bartley Y, Dupuy J, Menez C, Alvinerie M, Jackson F, Lespine A. 2009. P-glycoprotein interfering agents potentiate ivermectin susceptibility in ivermectin sensitive and resistant isolates of Teladorsagia circumcincta and Haemonchus contortus. *Parasitology* **136**: 1081-1088.

Bartley DJ, Morrison AA, Dupuy J, Bartley Y, Sutra JF, Menez C, Alvinerie M, Jackson F, Devin L, Lespine A. 2012. Influence of Pluronic 85 and ketoconazole on disposition and efficacy of ivermectin in sheep infected with a multiple resistant Haemonchus contortus isolate. *Veterinary parasitology* **187**: 464-472.

Battu R, Verma A, Hariharan R, Krishna S, Kiran R, Jacob J, Ganapathy A, Ramprasad VL, Kumaramanickavel G, Jeyabalan N et al. 2015. Identification of Novel Mutations in ABCA4 Gene: Clinical and Genetic Analysis of Indian Patients with Stargardt Disease. *Biomed Res Int* **2015**: 940864.

Beckingham KM, Armstrong JD, Texada MJ, Munjaal R, Baker DA. 2005. Drosophila melanogaster--the model organism of choice for the complex biology of multi-cellular organisms. *Gravit Space Biol Bull* **18**: 17-29.

Beer RJ. 1973. Studies on the biology of the life-cycle of Trichuris suis Schrank, 1788. *Parasitology* **67**: 253-262.

Begun DJ, Whitley P. 2000. Genetics of alpha-amanitin resistance in a natural population of Drosophila melanogaster. *Heredity* **85 (Pt 2)**: 184-190.

Bernard DG, Cheng Y, Zhao Y, Balk J. 2009. An allelic mutant series of ATM3 reveals its key role in the biogenesis of cytosolic iron-sulfur proteins in Arabidopsis. *Plant Physiol* **151**: 590-602.

Biemans-Oldehinkel E, Doeven MK, Poolman B. 2006. ABC transporter architecture and regulatory roles of accessory domains. *FEBS Lett* **580**: 1023-1035.

Bisbal C, Martinand C, Silhol M, Lebleu B, Salehzada T. 1995. Cloning and characterization of a RNAse L inhibitor. A new component of the interferon-regulated 2-5A pathway. *J Biol Chem* **270**: 13308-13317.

Blackhall WJ, Prichard RK, Beech RN. 2008. P-glycoprotein selection in strains of Haemonchus contortus resistant to benzimidazoles. *Veterinary parasitology* **152**: 101-107.

Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF et al. 1997. The complete genome sequence of Escherichia coli K-12. *Science* **277**: 1453-1462.

Borst P, Evers R, Kool M, Wijnholds J. 2000. A family of drug transporters: the multidrug resistance-associated proteins. *J Natl Cancer Inst* **92**: 1295-1302.

Borst P, Schinkel AH, Smit JJ, Wagenaar E, Van Deemter L, Smith AJ, Eijdems EW, Baas F, Zaman GJ. 1993. Classical and novel forms of multidrug resistance and the physiological functions of P-glycoproteins in mammals. *Pharmacol Ther* **60**: 289-299.

Boyd WA, Williams PL. 2003. Comparison of the sensitivity of three nematode species to copper and their utility in aquatic and soil toxicity tests. *Environ Toxicol Chem* **22**: 2768-2774.

Brangi M, Litman T, Ciotti M, Nishiyama K, Kohlhagen G, Takimoto C, Robey R, Pommier Y, Fojo T, Bates SE. 1999. Camptothecin resistance: role of the ATP-binding cassette (ABC), mitoxantrone-resistance half-transporter (MXR), and potential for glucuronidation in MXR-expressing cells. *Cancer Res* **59**: 5938-5946.

Bressler JP, Olivi L, Cheong JH, Kim Y, Maerten A, Bannon D. 2007. Metal transporters in intestine and brain: their involvement in metal-associated neurotoxicities. *Human & experimental toxicology* **26**: 221-229.

Brodie R, Roper RL, Upton C. 2004. JDotter: a Java interface to multiple dotplots generated by dotter. *Bioinformatics* **20**: 279-281.

Broeks A, Janssen HW, Calafat J, Plasterk RH. 1995. A P-glycoprotein protects Caenorhabditis elegans against natural toxins. *The EMBO journal* **14**: 1858-1866.

Brown HE, Harrington LC, Kaufman PE, McKay T, Bowman DD, Nelson CT, Wang D, Lund R. 2012. Key factors influencing canine heartworm, Dirofilaria immitis, in the United States. *Parasites & vectors* **5**: 245.

Buss DS, McCaffery AR, Callaghan A. 2002. Evidence for p-glycoprotein modification of insecticide toxicity in mosquitoes of the Culex pipiens complex. *Med Vet Entomol* **16**: 218-222.

Caillaud MC, Dubreuil G, Quentin M, Perfus-Barbeoch L, Lecomte P, de Almeida Engler J, Abad P, Rosso MN, Favery B. 2008. Root-knot nematodes manipulate plant cell functions during a compatible interaction. *Journal of plant physiology* **165**: 104-113.

Campbell JL, Nash HA. 2001. Volatile general anesthetics reveal a neurobiological role for the white and brown genes of Drosophila melanogaster. *J Neurobiol* **49**: 339-349.

Casadevall A, Pirofski L. 2001. Host-pathogen interactions: the attributes of virulence. *J Infect Dis* **184**: 337-344.

Casadevall A, Pirofski LA. 1999. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. *Infection and immunity* **67**: 3703-3713.

Chakraburtty K. 2001. Translational regulation by ABC systems. *Research in microbiology* **152**: 391-399.

Chakraburtty K, Triana-Alonso FJ. 1998. Yeast elongation factor 3: structure and function. *Biological chemistry* **379**: 831-840.

Chang G. 2003. Multidrug resistance ABC transporters. *FEBS Lett* **555**: 102-105.

Chen YH, Wang CC, Xiao X, Wei L, Xu G. 2013. Multidrug resistance-associated protein 1 decreases the concentrations of antiepileptic drugs in cortical extracellular fluid in amygdale kindling rats. *Acta Pharmacol Sin* **34**: 473-479.

Chen ZQ, Dong J, Ishimura A, Daar I, Hinnebusch AG, Dean M. 2006. The essential vertebrate ABCE1 protein interacts with eukaryotic initiation factors. *J Biol Chem* **281**: 7452-7457.

Choi YH, Yu AM. 2014. ABC transporters in multidrug resistance and pharmacokinetics, and strategies for drug development. *Curr Pharm Des* **20**: 793-807.

Chu Y, Corey DR. 2012. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther* **22**: 271-274.

Ciche T. 2007. The biology and genome of Heterorhabditis bacteriophora. *WormBook : the online review of C elegans biology* doi:10.1895/wormbook.1.135.1: 1-9.

Cole SP, Bhardwaj G, Gerlach JH, Mackie JE, Grant CE, Almquist KC, Stewart AJ, Kurz EU, Duncan AM, Deeley RG. 1992. Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line. *Science* **258**: 1650-1654.

Consortium CeS. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**: 2012-2018.

Cordon-Cardo C, O'Brien JP, Casals D, Rittman-Grauer L, Biedler JL, Melamed MR, Bertino JR. 1989. Multidrug-resistance gene (P-glycoprotein) is expressed by endothelial cells at blood-brain barrier sites. *Proc Natl Acad Sci U S A* **86**: 695-698.

Currie E, King B, Lawrenson AL, Schroeder LK, Kershner AM, Hermann GJ. 2007. Role of the Caenorhabditis elegans multidrug resistance gene, mrp-4, in gut granule differentiation. *Genetics* **177**: 1569-1582.

Cutting GR. 2015. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet* **16**: 45-56.

Davidson AL, Dassa E, Orelle C, Chen J. 2008. Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiology and molecular biology reviews : MMBR* **72**: 317-364, table of contents.

Dean M. 2005. The genetics of ATP-binding cassette transporters. *Methods in enzymology* **400**: 409-429.

Dean M, Allikmets R. 2001. Complete characterization of the human ABC gene family. *Journal of bioenergetics and biomembranes* **33**: 475-479.

Dean M, Annilo T. 2005. Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. *Annual review of genomics and human genetics* **6**: 123-142.

Dean M, Hamon Y, Chimini G. 2001. The human ATP-binding cassette (ABC) transporter superfamily. *Journal of lipid research* **42**: 1007-1017.

Decottignies A, Goffeau A. 1997. Complete inventory of the yeast ABC proteins. *Nature genetics* **15**: 137-145.

Desjardins CA, Cerqueira GC, Goldberg JM, Dunning Hotopp JC, Haas BJ, Zucker J, Ribeiro JM, Saif S, Levin JZ, Fan L et al. 2013. Genomics of Loa loa, a Wolbachia-free filarial parasite of humans. *Nature genetics* **45**: 495-500.

Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, Fulton L, Fulton R, Godfrey J, Minx P et al. 2008. The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. *Nature genetics* **40**: 1193-1198.

Dong J, Lai R, Nielsen K, Fekete CA, Qiu H, Hinnebusch AG. 2004. The essential ATP-binding cassette protein RLI1 functions in translation by promoting preinitiation complex assembly. *J Biol Chem* **279**: 42157-42168.

Doyle LA, Yang W, Abruzzo LV, Krogmann T, Gao Y, Rishi AK, Ross DD. 1998. A multidrug resistance transporter from human MCF-7 breast cancer cells. *Proc Natl Acad Sci U S A* **95**: 15665-15670.

Dreesen TD, Johnson DH, Henikoff S. 1988. The brown protein of Drosophila melanogaster is similar to the white protein and to components of active transport complexes. *Mol Cell Biol* **8**: 5206-5215.

Erickson SM, Fischer K, Weil GJ, Christensen BM, Fischer PU. 2009. Distribution of Brugia malayi larvae and DNA in vector and non-vector mosquitoes: implications for molecular diagnostics. *Parasites & vectors* **2**: 56.

Ewart GD, Cannell D, Cox GB, Howells AJ. 1994. Mutational analysis of the traffic ATPase (ABC) transporters involved in uptake of eye pigment precursors in Drosophila melanogaster. Implications for structure-function relationships. *J Biol Chem* **269**: 10370-10377.

Felix MA, Braendle C, Cutter AD. 2014. A streamlined system for species diagnosis in Caenorhabditis (Nematoda: Rhabditidae) with name designations for 15 distinct biological species. *PloS one* **9**: e94723.

Ferenci T, Boos W, Schwartz M, Szmelcman S. 1977. Energy-coupling of the transport system of Escherichia coli dependent on maltose-binding protein. *Eur J Biochem* **75**: 187-193.

Fierst JL, Willis JH, Thomas CG, Wang W, Reynolds RM, Ahearne TE, Cutter AD, Phillips PC. 2015. Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes. *PLoS genetics* **11**: e1005323.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J et al. 2014. Pfam: the protein families database. *Nucleic Acids Res* **42**: D222-230.

Fischer S, Kluver N, Burkhardt-Medicke K, Pietsch M, Schmidt AM, Wellner P, Schirmer K, Luckenbach T. 2013. Abcb4 acts as multixenobiotic transporter and active barrier against chemical uptake in zebrafish (Danio rerio) embryos. *BMC biology* **11**: 69.

Flens MJ, Zaman GJ, van der Valk P, Izquierdo MA, Schroeijers AB, Scheffer GL, van der Groep P, de Haas M, Meijer CJ, Scheper RJ. 1996. Tissue distribution of the multidrug resistance protein. *Am J Pathol* **148**: 1237-1247.

Foth BJ, Tsai IJ, Reid AJ, Bancroft AJ, Nichol S, Tracey A, Holroyd N, Cotton JA, Stanley EJ, Zarowiecki M et al. 2014. Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. *Nature genetics* **46**: 693-700.

Fournier PE, Dubourg G, Raoult D. 2014. Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Med* **6**: 114.

Frech C, Chen N. 2010. Genome-wide comparative gene family classification. *PloS one* **5**: e13409.

Futai K. 2013. Pine wood nematode, Bursaphelenchus xylophilus. *Annual review of phytopathology* **51**: 61-83.

Gadsby DC, Vergani P, Csanady L. 2006. The ABC protein turned chloride channel whose failure causes cystic fibrosis. *Nature* **440**: 477-483.

Genchi C, Rinaldi L, Mortarino M, Genchi M, Cringoli G. 2009. Climate and Dirofilaria infection in Europe. *Veterinary parasitology* **163**: 286-292.

Gerlach JH, Endicott JA, Juranka PF, Henderson G, Sarangi F, Deuchars KL, Ling V. 1986. Homology between P-glycoprotein and a bacterial haemolysin transport protein suggests a model for multidrug resistance. *Nature* **324**: 485-489.

Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D et al. 2007. Draft genome of the filarial nematode parasite Brugia malayi. *Science* **317**: 1756-1760.

Gilleard JS. 2006. Understanding anthelmintic resistance: the need for genomics and genetics. *International journal for parasitology* **36**: 1227-1239.

Godel C, Kumar S, Koutsovoulos G, Ludin P, Nilsson D, Comandatore F, Wrobel N, Thompson M, Schmid CD, Goto S et al. 2012. The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **26**: 4650-4661.

Gonzalez-Cabo P, Bolinches-Amoros A, Cabello J, Ros S, Moreno S, Baylis HA, Palau F, Vazquez-Manrique RP. 2011. Disruption of the ATP-binding cassette B7 (ABTM-1/ABCB7) induces oxidative stress and premature cell death in Caenorhabditis elegans. *J Biol Chem* **286**: 21304-21314.

Goosen N, Moolenaar GF. 2001. Role of ATP hydrolysis by UvrA and UvrB during nucleotide excision repair. *Research in microbiology* **152**: 401-409.

Gottesman MM, Fojo T, Bates SE. 2002. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat Rev Cancer* **2**: 48-58.

Gregg RE, Connor WE, Lin DS, Brewer HB, Jr. 1986. Abnormal metabolism of shellfish sterols in a patient with sitosterolemia and xanthomatosis. *The Journal of clinical investigation* **77**: 1864-1872.

Gros P, Croop J, Housman D. 1986. Mammalian multidrug resistance gene: complete cDNA sequence indicates strong homology to bacterial transport proteins. *Cell* **47**: 371-380.

Haag ES, Chamberlin H, Coghlan A, Fitch DH, Peters AD, Schulenburg H. 2007. Caenorhabditis evolution: if they all look alike, you aren't looking hard enough. *Trends Genet* **23**: 101-104.

Haynes CM, Yang Y, Blais SP, Neubert TA, Ron D. 2010. The matrix peptide exporter HAF-1 signals a mitochondrial UPR by activating the transcription factor ZC376.7 in C. elegans. *Molecular cell* **37**: 529-540.

Higgins CF. 1992. ABC transporters: from microorganisms to man. *Annual review of cell biology* **8**: 67-113.

Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH. 2005. Genomics in C. elegans: so many genes, such a little worm. *Genome Res* **15**: 1651-1660.

Hipfner DR, Deeley RG, Cole SP. 1999. Structural, mechanistic and clinical aspects of MRP1. *Biochim Biophys Acta* **1461**: 359-376.

Holland IB. 2011. ABC transporters, mechanisms and biology: an overview. *Essays Biochem* **50**: 1-17.

Huang H, Lu-Bo Y, Haddad GG. 2014a. A Drosophila ABC transporter regulates lifespan. *PLoS genetics* **10**: e1004844.

Huang RE, Ren X, Qiu Y, Zhao Z. 2014b. Description of Caenorhabditis sinica sp. n. (Nematoda: Rhabditidae), a nematode species used in comparative biology for C. elegans. *PloS one* **9**: e110957.

Hyde SC, Emsley P, Hartshorn MJ, Mimmack MM, Gileadi U, Pearce SR, Gallagher MP, Gill DR, Hubbard RE, Higgins CF. 1990. Structural model of ATP-binding proteins associated with cystic fibrosis, multidrug resistance and bacterial transport. *Nature* **346**: 362-365.

Hyun HJ, Kim EM, Park SY, Jung JO, Chai JY, Hong ST. 2010. A case of severe anemia by Necator americanus infection in Korea. *Journal of Korean medical science* **25**: 1802-1804.

Janvilisri T, Venter H, Shahi S, Reuter G, Balakrishnan L, van Veen HW. 2003. Sterol transport by the human breast cancer resistance protein (ABCG2) expressed in Lactococcus lactis. *J Biol Chem* **278**: 20645-20651.

Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z et al. 2011. Ascaris suum draft genome. *Nature* **479**: 529-533.

Jex AR, Nejsum P, Schwarz EM, Hu L, Young ND, Hall RS, Korhonen PK, Liao S, Thamsborg S, Xia J et al. 2014. Genome and transcriptome of the porcine whipworm Trichuris suis. *Nature genetics* **46**: 701-706.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Jones PM, George AM. 2004. The ABC transporter structure and mechanism: perspectives on recent research. *Cellular and molecular life sciences : CMLS* **61**: 682-699.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27-30.

Kang J, Park J, Choi H, Burla B, Kretzschmar T, Lee Y, Martinoia E. 2011. Plant ABC Transporters. *Arabidopsis Book* **9**: e0153.

Kanzaki N, Ragsdale EJ, Herrmann M, Mayer WE, Sommer RJ. 2012. Description of three Pristionchus species (Nematoda: Diplogastridae) from Japan that form a cryptic species complex with the model organism P. pacificus. *Zoological science* **29**: 403-417.

Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* **35**: D55-60.

Kartner N, Evernden-Porelle D, Bradley G, Ling V. 1985. Detection of P-glycoprotein in multidrug-resistant cell lines by monoclonal antibodies. *Nature* **316**: 820-823.

Kawai H, Tanji T, Shiraishi H, Yamada M, Iijima R, Inoue T, Kezuka Y, Ohashi K, Yoshida Y, Tohyama K et al. 2009. Normal formation of a subset of intestinal granules in Caenorhabditis elegans requires ATP-binding cassette transporters HAF-4 and HAF-9, which are highly homologous to human lysosomal peptide transporter TAP-like. *Molecular biology of the cell* **20**: 2979-2990.

Kerboeuf D, Guegnard F, Le Vern Y. 2002. Analysis and partial reversal of multidrug resistance to anthelmintics due to P-glycoprotein in Haemonchus contortus eggs using Lens culinaris lectin. *Parasitology research* **88**: 816-821.

Kerr ID. 2004. Sequence analysis of twin ATP binding cassette proteins involved in translational control, antibiotic resistance, and ribonuclease L inhibition. *Biochem Biophys Res Commun* **315**: 166-173.

Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, Takanashi T, Tsai IJ, Assefa SA, Cock PJ et al. 2011. Genomic insights into the origin of parasitism in the emerging plant pathogen Bursaphelenchus xylophilus. *PLoS pathogens* **7**: e1002219.

Kispal G, Csere P, Prohl C, Lill R. 1999. The mitochondrial proteins Atm1p and Nfs1p are essential for biogenesis of cytosolic Fe/S proteins. *The EMBO journal* **18**: 3981-3989.

Klein I, Sarkadi B, Varadi A. 1999. An inventory of the human ABC proteins. *Biochim Biophys Acta* **1461**: 237-262.

Klein M, Mamnun YM, Eggmann T, Schuller C, Wolfger H, Martinoia E, Kuchler K. 2002. The ATP-binding cassette (ABC) transporter Bpt1p mediates vacuolar sequestration of glutathione conjugates in yeast. *FEBS Lett* **520**: 63-67.

Knopp S, Steinmann P, Hatz C, Keiser J, Utzinger J. 2012. Nematode infections: filariases. *Infectious disease clinics of North America* **26**: 359-381.

Kolaczkowski M, Kolaczowska A, Luczynski J, Witek S, Goffeau A. 1998. In vivo characterization of the drug resistance profile of the major ABC transporters and other components of the yeast pleiotropic drug resistance network. *Microbial drug resistance* **4**: 143-158.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**: 309-338.

Kovalchuk A, Driessen AJ. 2010. Phylogenetic analysis of fungal ABC transporters. *BMC genomics* **11**: 177.

Krishnamurthy PC, Du G, Fukuda Y, Sun D, Sampath J, Mercer KE, Wang J, Sosa-Pineda B, Murti KG, Schuetz JD. 2006. Identification of a mammalian mitochondrial porphyrin transporter. *Nature* **443**: 586-589.

Kroetz SM, Srinivasan J, Yaghoobian J, Sternberg PW, Hong RL. 2012. The cGMP signaling pathway affects feeding behavior in the necromenic nematode Pristionchus pacificus. *PloS one* **7**: e34464.

Kuchler K, Dohlman HG, Thorner J. 1993. The a-factor transporter (STE6 gene product) and cell polarity in the yeast Saccharomyces cerevisiae. *The Journal of cell biology* **120**: 1203-1215.

Kuchler K, Sterne RE, Thorner J. 1989. Saccharomyces cerevisiae STE6 gene product: a novel pathway for protein export in eukaryotic cells. *The EMBO journal* **8**: 3973-3984.

Kutil BL, Liu G, Vrebalov J, Wilkinson HH. 2004. Contig assembly and microsynteny analysis using a bacterial artificial chromosome library for Epichloe festucae, a mutualistic fungal endophyte of grasses. *Fungal Genet Biol* **41**: 23-32.

Laing R, Kikuchi T, Martinelli A, Tsai IJ, Beech RN, Redman E, Holroyd N, Bartley DJ, Beasley H, Britton C et al. 2013. The genome and transcriptome of Haemonchus contortus, a key model parasite for drug and vaccine discovery. *Genome biology* **14**: R88.

Lankat-Buttgereit B, Tampe R. 2002. The transporter associated with antigen processing: function and implications in human diseases. *Physiol Rev* **82**: 187-204.

Lee AC, Montgomery SP, Theis JH, Blagburn BL, Eberhard ML. 2010. Public health issues concerning the widespread distribution of canine heartworm disease. *Trends in parasitology* **26**: 168-173.

Leighton J, Schatz G. 1995. An ABC transporter in the mitochondrial inner membrane is required for normal growth of yeast. *The EMBO journal* **14**: 188-195.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178-2189.

Li Q, Sadowski S, Frank M, Chai C, Varadi A, Ho SY, Lou H, Dean M, Thisse C, Thisse B et al. 2010. The abcc6a gene expression is required for normal zebrafish development. *J Invest Dermatol* **130**: 2561-2568.

Lincke CR, Broeks A, The I, Plasterk RH, Borst P. 1993. The expression of two P-glycoprotein (pgp) genes in transgenic Caenorhabditis elegans is confined to intestinal cells. *The EMBO journal* **12**: 1615-1620.

Linton KJ, Higgins CF. 1998. The Escherichia coli ATP-binding cassette (ABC) proteins. *Molecular microbiology* **28**: 5-13.

Liu S, Zhou S, Tian L, Guo E, Luan Y, Zhang J, Li S. 2011. Genome-wide identification and characterization of ATP-binding cassette transporters in the silkworm, Bombyx mori. *BMC genomics* **12**: 491.

Long Y, Li Q, Cui Z. 2011. Molecular analysis and heavy metal detoxification of ABCC1/MRP1 in zebrafish. *Molecular biology reports* **38**: 1703-1711.

Luckenbach T, Fischer S, Sturm A. 2014. Current advances on ABC drug transporters in fish. *Comparative biochemistry and physiology Toxicology & pharmacology : CBP* **165**: 28-52.

Mack JT, Brown CB, Tew KD. 2008. ABCA2 as a therapeutic target in cancer and nervous system disorders. *Expert Opin Ther Targets* **12**: 491-504.

Mackenzie SM, Brooker MR, Gill TR, Cox GB, Howells AJ, Ewart GD. 1999. Mutations in the white gene of Drosophila melanogaster affecting ABC transporters that determine eye colouration. *Biochim Biophys Acta* **1419**: 173-185.

Maguire A, Hellier K, Hammans S, May A. 2001. X-linked cerebellar ataxia and sideroblastic anaemia associated with a missense mutation in the ABC7 gene predicting V411L. *Br J Haematol* **115**: 910-917.

Mahajan-Miklos S, Tan MW, Rahme LG, Ausubel FM. 1999. Molecular mechanisms of bacterial virulence elucidated using a Pseudomonas aeruginosa-Caenorhabditis elegans pathogenesis model. *Cell* **96**: 47-56.

Maliepaard M, van Gastelen MA, de Jong LA, Pluim D, van Waardenburg RC, Ruevekamp-Helmers MC, Floot BG, Schellens JH. 1999. Overexpression of the BCRP/MXR/ABCP gene in a topotecan-selected ovarian tumor cell line. *Cancer Res* **59**: 4559-4563.

Mamiya Y. 1988. History of pine wilt disease in Japan. *J Nematol* **20**: 219-226.

Mamiya Y, and Nobuo Enda. 1979. *Bursaphelenchus mucronatus n. sp.(Nematoda: Aphelenchoididae) from pine wood and its biology and pathogenicity to pine trees*.

Markov GV, Baskaran P, Sommer RJ. 2015. The same or not the same: lineage-specific gene expansions and homology relationships in multigene families in nematodes. *Journal of molecular evolution* **80**: 18-36.

Marton MJ, Vazquez de Aldana CR, Qiu H, Chakraburtty K, Hinnebusch AG. 1997. Evidence that GCN1 and GCN20, translational regulators of GCN4, function on elongating ribosomes in activation of eIF2alpha kinase GCN2. *Mol Cell Biol* **17**: 4474-4489.

Maugeri A, Klevering BJ, Rohrschneider K, Blankenagel A, Brunner HG, Deutman AF, Hoyng CB, Cremers FP. 2000. Mutations in the ABCA4 (ABCR) gene are the major cause of autosomal recessive cone-rod dystrophy. *Am J Hum Genet* **67**: 960-966.

Mayer F, Mayer N, Chinn L, Pinsonneault RL, Kroetz D, Bainton RJ. 2009. Evolutionary conservation of vertebrate blood-brain barrier chemoprotective mechanisms in Drosophila. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **29**: 3538-3550.

McCall JW, Genchi C, Kramer LH, Guerrero J, Venco L. 2008. Heartworm disease in animals and humans. *Advances in parasitology* **66**: 193-285.

Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, Taylor CM, Yin Y, Fulton L, Minx P et al. 2011. The draft genome of the parasitic nematode Trichinella spiralis. *Nature genetics* **43**: 228-235.

Mitsuhashi N, Miki T, Senbongi H, Yokoi N, Yano H, Miyazaki M, Nakajima N, Iwanaga T, Yokoyama Y, Shibata T et al. 2000. MTABC3, a novel mitochondrial ATP-binding cassette protein involved in iron homeostasis. *J Biol Chem* **275**: 17536-17540.

Miyake K, Mickley L, Litman T, Zhan Z, Robey R, Cristensen B, Brangi M, Greenberger L, Dean M, Fojo T et al. 1999. Molecular cloning of cDNAs which are highly overexpressed in mitoxantrone-resistant cells: demonstration of homology to ABC transport genes. *Cancer Res* **59**: 8-13.

Molento MB, Prichard RK. 1999. Effects of the multidrug-resistance-reversing agents verapamil and CL 347,099 on the efficacy of ivermectin or moxidectin against unselected and drug-selected strains of Haemonchus contortus in jirds (Meriones unguiculatus). *Parasitology research* **85**: 1007-1011.

Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**: 81-94.

Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, Sternberg PW. 2010. Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res* **20**: 1740-1747.

Mosser J, Douar AM, Sarde CO, Kioschis P, Feil R, Moser H, Poustka AM, Mandel JL, Aubourg P. 1993. Putative X-linked adrenoleukodystrophy gene shares unexpected homology with ABC transporters. *Nature* **361**: 726-730.

Nickle WR, Golden AM, Mamiya Y, Wergin WP. 1981. On the Taxonomy and Morphology of the Pine Wood Nematode, Bursaphelenchus xylophilus (Steiner &Buhrer 1934) Nickle 1970. *J Nematol* **13**: 385-392.

Ochman H, Davalos LM. 2006. The nature and dynamics of bacterial genomes. *Science* **311**: 1730-1733.

Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, Cromer J, Diener S, Gajan J, Graham S et al. 2008. Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. *Proc Natl Acad Sci U S A* **105**: 14802-14807.

Padgett JJ, Jacobsen KH. 2008. Loiasis: African eye worm. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **102**: 983-989.

Park JK, Sultana T, Lee SH, Kang S, Kim HK, Min GS, Eom KS, Nadler SA. 2011. Monophyly of clade III nematodes is not supported by phylogenetic analysis of complete mitochondrial genome sequences. *BMC genomics* **12**: 392.

Patel SB, Salen G, Hidaka H, Kwiterovich PO, Stalenhoef AF, Miettinen TA, Grundy SM, Lee MH, Rubenstein JS, Polymeropoulos MH et al. 1998. Mapping a gene involved in regulating dietary cholesterol absorption. The sitosterolemia locus is found at chromosome 2p21. *The Journal of clinical investigation* **102**: 1041-1044.

Paumi CM, Chuk M, Snider J, Stagljar I, Michaelis S. 2009. ABC transporters in Saccharomyces cerevisiae and their interactors: new technology advances the biology of the ABCC (MRP) subfamily. *Microbiology and molecular biology reviews : MMBR* **73**: 577-593.

Pearson WR. 1994. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol Biol* **24**: 307-331.

Percudani R. 2013. A Microbial Metagenome (Leucobacter sp.) in Caenorhabditis Whole Genome Sequences. *Bioinformatics and biology insights* **7**: 55-72.

Piper P, Mahe Y, Thompson S, Pandjaitan R, Holyoak C, Egner R, Muhlbauer M, Coote P, Kuchler K. 1998. The pdr12 ABC transporter is required for the development of weak organic acid resistance in yeast. *The EMBO journal* **17**: 4257-4265.

Pondarre C, Antiochos BB, Campagna DR, Clarke SL, Greer EL, Deck KM, McDonald A, Han AP, Medlock A, Kutok JL et al. 2006. The mitochondrial ATP-binding cassette transporter Abcb7 is essential in mice and participates in cytosolic iron-sulfur cluster biogenesis. *Human molecular genetics* **15**: 953-964.

Pondarre C, Campagna DR, Antiochos B, Sikorski L, Mulhern H, Fleming MD. 2007. Abcb7, the gene responsible for X-linked sideroblastic anemia with ataxia, is essential for hematopoiesis. *Blood* **109**: 3567-3569.

Popovic M, Zaja R, Loncar J, Smital T. 2010. A novel ABC transporter: the first insight into zebrafish (Danio rerio) ABCH1. *Marine environmental research* **69 Suppl**: S11-13.

Pujol A, Ferrer I, Camps C, Metzger E, Hindelang C, Callizot N, Ruiz M, Pampols T, Giros M, Mandel JL. 2004. Functional overlap between ABCD1 (ALD) and ABCD2 (ALDR) transporters: a therapeutic target for X-adrenoleukodystrophy. *Human molecular genetics* **13**: 2997-3006.

Remaley AT, Rust S, Rosier M, Knapper C, Naudin L, Broccardo C, Peterson KM, Koch C, Arnould I, Prades C et al. 1999. Human ATP-binding cassette transporter 1 (ABC1): genomic organization and identification of the genetic defect in the original Tangier disease kindred. *Proc Natl Acad Sci U S A* **96**: 12685-12690.

Robey RW, Medina-Perez WY, Nishiyama K, Lahusen T, Miyake K, Litman T, Senderowicz AM, Ross DD, Bates SE. 2001. Overexpression of the ATP-binding cassette half-transporter, ABCG2 (Mxr/BCrp/ABCP1), in flavopiridol-resistant human breast cancer cells. *Clin Cancer Res* **7**: 145-152.

Rodelsperger C, Neher RA, Weller AM, Eberhardt G, Witte H, Mayer WE, Dieterich C, Sommer RJ. 2014. Characterization of genetic diversity in the nematode Pristionchus pacificus from population-scale resequencing data. *Genetics* **196**: 1153-1165.

Ronchetti I, Boraldi F, Annovi G, Cianciulli P, Quaglino D. 2013. Fibroblast involvement in soft connective tissue calcification. *Front Genet* **4**: 22.

Rust S, Rosier M, Funke H, Real J, Amoura Z, Piette JC, Deleuze JF, Brewer HB, Duverger N, Denefle P et al. 1999. Tangier disease is caused by mutations in the gene encoding ATP-binding cassette transporter 1. *Nature genetics* **22**: 352-355.

Saint Andre A, Blackwell NM, Hall LR, Hoerauf A, Brattig NW, Volkmann L, Taylor MJ, Ford L, Hise AG, Lass JH et al. 2002. The role of endosymbiotic Wolbachia bacteria in the pathogenesis of river blindness. *Science* **295**: 1892-1895.

Sangster NC, Bannan SC, Weiss AS, Nulf SC, Klein RD, Geary TG. 1999. Haemonchus contortus: sequence heterogeneity of internucleotide binding domains from P-glycoproteins. *Experimental parasitology* **91**: 250-257.

Sarkadi B, Ozvegy-Laczka C, Nemet K, Varadi A. 2004. ABCG2 -- a transporter for all seasons. *FEBS Lett* **567**: 116-120.

Saurin W, Hofnung M, Dassa E. 1999. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *Journal of molecular evolution* **48**: 22-41.

Schinkel AH, Mayer U, Wagenaar E, Mol CA, van Deemter L, Smit JJ, van der Valk MA, Voordouw AC, Spits H, van Tellingen O et al. 1997. Normal viability and altered pharmacokinetics in mice lacking mdr1-type (drug-transporting) P-glycoproteins. *Proc Natl Acad Sci U S A* **94**: 4028-4033.

Schneider E, Hunke S. 1998. ATP-binding-cassette (ABC) transport systems: functional and structural aspects of the ATP-hydrolyzing subunits/domains. *FEMS microbiology reviews* **22**: 1-20.

Scholz S, Fischer S, Gundel U, Kuster E, Luckenbach T, Voelker D. 2008. The zebrafish embryo model in environmental risk assessment--applications beyond acute toxicity testing. *Environmental science and pollution research international* **15**: 394-404.

Schroeder LK, Kremer S, Kramer MJ, Currie E, Kwan E, Watts JL, Lawrenson AL, Hermann GJ. 2007. Function of the Caenorhabditis elegans ABC transporter PGP-2 in the biogenesis of a lysosome-related fat storage organelle. *Molecular biology of the cell* **18**: 995-1008.

Schuetz JD, Connelly MC, Sun D, Paibir SG, Flynn PM, Srinivas RV, Kumar A, Fridland A. 1999. MRP4: A previously unidentified factor in resistance to nucleoside-based antiviral drugs. *Nat Med* **5**: 1048-1051.

Schwartz MS, Benci JL, Selote DS, Sharma AK, Chen AG, Dang H, Fares H, Vatamaniuk OK. 2010. Detoxification of multiple heavy metals by a half-molecule ABC transporter, HMT-1, and coelomocytes of Caenorhabditis elegans. *PloS one* **5**: e9564.

Schwarz EM, Hu Y, Antoshechkin I, Miller MM, Sternberg PW, Aroian RV. 2015. The genome and transcriptome of the zoonotic hookworm Ancylostoma ceylanicum identify infection-specific gene families. *Nature genetics* **47**: 416-422.

Schwarz EM, Korhonen PK, Campbell BE, Young ND, Jex AR, Jabbar A, Hall RS, Mondal A, Howe AC, Pell J et al. 2013. The genome and developmental transcriptome of the strongylid nematode Haemonchus contortus. *Genome biology* **14**: R89.

Servos J, Haase E, Brendel M. 1993. Gene SNQ2 of Saccharomyces cerevisiae, which confers resistance to 4-nitroquinoline-N-oxide and other chemicals, encodes a 169 kDa protein homologous to ATP-dependent permeases. *Molecular & general genetics : MGG* **236**: 214-218.

Shani N, Valle D. 1996. A Saccharomyces cerevisiae homolog of the human adrenoleukodystrophy transporter is a heterodimer of two half ATP-binding cassette transporters. *Proc Natl Acad Sci U S A* **93**: 11901-11906.

She R, Chu JS, Uyar B, Wang J, Wang K, Chen N. 2011. genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**: 2141-2143.

Sheps JA, Ralph S, Zhao Z, Baillie DL, Ling V. 2004. The ABC transporter gene family of Caenorhabditis elegans has implications for the evolutionary dynamics of multidrug resistance in eukaryotes. *Genome biology* **5**: R15.

Shinya R, Morisaka H, Takeuchi Y, Futai K, Ueda M. 2013. Making headway in understanding pine wilt disease: what do we perceive in the postgenomic era? *Journal of bioscience and bioengineering* **116**: 1-8.

Sommer RJ, McGaughran A. 2013. The nematode Pristionchus pacificus as a model system for integrative studies in evolutionary biology. *Molecular ecology* **22**: 2380-2393.

Sonnhammer EL, Koonin EV. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* **18**: 619-620.

Sooksa-Nguan T, Yakubov B, Kozlovskyy VI, Barkume CM, Howe KJ, Thannhauser TW, Rutzke MA, Hart JJ, Kochian LV, Rea PA et al. 2009. Drosophila ABC transporter, DmHMT-1, confers tolerance to cadmium. DmHMT-1 and its yeast homolog, SpHMT-1, are not essential for vacuolar phytochelatin sequestration. *J Biol Chem* **284**: 354-362.

Srinivasan J, Dillman AR, Macchietto MG, Heikkinen L, Lakso M, Fracchia KM, Antoshechkin I, Mortazavi A, Wong G, Sternberg PW. 2013. The draft genome and transcriptome of Panagrellus redivivus are shaped by the harsh demands of a free-living lifestyle. *Genetics* **193**: 1279-1295.

St-Pierre MV, Serrano MA, Macias RI, Dubs U, Hoechli M, Lauper U, Meier PJ, Marin JJ. 2000. Expression of members of the multidrug resistance protein family in human term placenta. *Am J Physiol Regul Integr Comp Physiol* **279**: R1495-1503.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS biology* **1**: E45.

Stephenson LS, Holland CV, Cooper ES. 2000. The public health significance of Trichuris trichiura. *Parasitology* **121 Suppl**: S73-95.

Sternberg PW, Horvitz HR. 1981. Gonadal cell lineages of the nematode Panagrellus redivivus and implications for evolution by the modification of cell lineage. *Developmental biology* **88**: 147-166.

Stitt LE, Tompkins JB, Dooley LA, Ardelli BF. 2011. ABC transporters influence sensitivity of Brugia malayi to moxidectin and have potential roles in drug resistance. *Experimental parasitology* **129**: 137-144.

Sundaram P, Echalier B, Han W, Hull D, Timmons L. 2006. ATP-binding cassette transporters are required for efficient RNA interference in Caenorhabditis elegans. *Molecular biology of the cell* **17**: 3678-3688.

Sundaram P, Han W, Cohen N, Echalier B, Albin J, Timmons L. 2008. Caenorhabditis elegans ABCRNAi transporters interact genetically with rde-2 and mut-7. *Genetics* **178**: 801-814.

Szczypka MS, Wemmie JA, Moye-Rowley WS, Thiele DJ. 1994. A yeast metal resistance protein similar to human cystic fibrosis transmembrane conductance regulator (CFTR) and multidrug resistance-associated protein. *J Biol Chem* **269**: 22853-22857.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**: 2725-2729.

Tang YT, Gao X, Rosa BA, Abubucker S, Hallsworth-Pepin K, Martin J, Tyagi R, Heizer E, Zhang X, Bhonagiri-Palsikar V et al. 2014. Genome of the human hookworm Necator americanus. *Nature genetics* **46**: 261-269.

Tapadia MG, Lakhotia SC. 2005. Expression of mdr49 and mdr65 multidrug resistance genes in larval tissues of Drosophila melanogaster under normal and stress conditions. *Cell Stress Chaperones* **10**: 7-11.

Thiebaut F, Tsuruo T, Hamada H, Gottesman MM, Pastan I, Willingham MC. 1987. Cellular localization of the multidrug-resistance gene product P-glycoprotein in normal human tissues. *Proc Natl Acad Sci U S A* **84**: 7735-7738.

Toh S, Wada M, Uchiumi T, Inokuchi A, Makino Y, Horie Y, Adachi Y, Sakisaka S, Kuwano M. 1999. Genomic structure of the canalicular multispecific organic anion-transporter gene (MRP2/cMOAT) and mutations in the ATP-binding-cassette region in Dubin-Johnson syndrome. *Am J Hum Genet* **64**: 739-746.

Tompkins JB, Stitt LE, Morrissette AM, Ardelli BF. 2011. The role of Brugia malayi ATP-binding cassette (ABC) transporters in potentiating drug sensitivity. *Parasitology research* **109**: 1311-1322.

Traub RJ. 2013. Ancylostoma ceylanicum, a re-emerging but neglected parasitic zoonosis. *International journal for parasitology* **43**: 1009-1015.

Traversa D, Di Cesare A, Conboy G. 2010. Canine and feline cardiopulmonary parasitic nematodes in Europe: emerging and underestimated. *Parasites & vectors* **3**: 62.

Trip MD, Smulders YM, Wegman JJ, Hu X, Boer JM, ten Brink JB, Zwinderman AH, Kastelein JJ, Feskens EJ, Bergen AA. 2002. Frequent mutation in the ABCC6 gene (R1141X) is associated with a strong increase in the prevalence of coronary artery disease. *Circulation* **106**: 773-775.

Trudgill DL, Blok VC. 2001. Apomictic, polyphagous root-knot nematodes: exceptionally successful and damaging biotrophic root pathogens. *Annual review of phytopathology* **39**: 53-77.

Tsujii H, Konig J, Rost D, Stockel B, Leuschner U, Keppler D. 1999. Exon-intron organization of the human multidrug-resistance protein 2 (MRP2) gene mutated in Dubin-Johnson syndrome. *Gastroenterology* **117**: 653-660.

Turriziani O, Schuetz JD, Focher F, Scagnolari C, Sampath J, Adachi M, Bambacioni F, Riva E, Antonelli G. 2002. Impaired 2',3'-dideoxy-3'-thiacytidine accumulation in T-lymphoblastoid cells as a mechanism of acquired resistance independent of multidrug resistant protein 4 with a possible role for ATP-binding cassette C11. *Biochem J* **368**: 325-332.

Tyzack JK, Wang X, Belsham GJ, Proud CG. 2000. ABC50 interacts with eukaryotic initiation factor 2 and associates with the ribosome in an ATP-dependent manner. *J Biol Chem* **275**: 34131-34139.

Unnasch TR, Williams SA. 2000. The genomes of Onchocerca volvulus. *International journal for parasitology* **30**: 543-552.

Uriu-Adams JY, Keen CL. 2005. Copper, oxidative stress, and human health. *Mol Aspects Med* **26**: 268-298.

Valko M, Morris H, Cronin MT. 2005. Metals, toxicity and oxidative stress. *Curr Med Chem* **12**: 1161-1208.

van Herwaarden AE, Wagenaar E, Merino G, Jonker JW, Rosing H, Beijnen JH, Schinkel AH. 2007. Multidrug transporter ABCG2/breast cancer resistance protein secretes riboflavin (vitamin B2) into milk. *Mol Cell Biol* **27**: 1247-1253.

Vasiliou V, Vasiliou K, Nebert DW. 2009. Human ATP-binding cassette (ABC) transporter family. *Human genomics* **3**: 281-290.

Wang C, Lower S, Thomas VP, Williamson VM. 2010. Root-knot nematodes exhibit strain-specific clumping behavior that is inherited as a simple genetic trait. *PloS one* **5**: e15148.

Wang J, Near S, Young K, Connelly PW, Hegele RA. 2001. ABCC6 gene polymorphism associated with variation in plasma lipoproteins. *Journal of human genetics* **46**: 699-705.

Wilcox LJ, Balderes DA, Wharton B, Tinkelenberg AH, Rao G, Sturley SL. 2002. Transcriptional profiling identifies two members of the ATP-binding cassette transporter superfamily required for sterol uptake in yeast. *J Biol Chem* **277**: 32466-32472.

Wu CP, Calcagno AM, Ambudkar SV. 2008. Reversal of ABC drug transporter-mediated multidrug resistance in cancer cells: evaluation of current strategies. *Curr Mol Pharmacol* **1**: 93-105.

Wu CT, Budding M, Griffin MS, Croop JM. 1991. Isolation and characterization of Drosophila multidrug resistance gene homologs. *Mol Cell Biol* **11**: 3940-3948.

Wu YC, Horvitz HR. 1998. The C. elegans cell corpse engulfment gene ced-7 encodes a protein similar to ABC transporters. *Cell* **93**: 951-960.

Xie BB, Qin QL, Shi M, Chen LL, Shu YL, Luo Y, Wang XW, Rong JC, Gong ZT, Li D et al. 2014. Comparative genomics provide insights into evolution of trichoderma nutrition style. *Genome Biol Evol* **6**: 379-390.

Xie X, Cheng T, Wang G, Duan J, Niu W, Xia Q. 2012. Genome-wide analysis of the ATP-binding cassette (ABC) transporter gene family in the silkworm, Bombyx mori. *Molecular biology reports* **39**: 7281-7291.

Xu J, Liu Y, Yang Y, Bates S, Zhang JT. 2004. Characterization of oligomeric human half-ABC transporter ATP-binding cassette G2. *J Biol Chem* **279**: 19781-19789.

Yepiskoposyan H, Egli D, Fergestad T, Selvaraj A, Treiber C, Multhaup G, Georgiev O, Schaffner W. 2006. Transcriptome response to heavy metal stress in Drosophila reveals a new zinc transporter that confers resistance to zinc. *Nucleic Acids Res* **34**: 4866-4877.

Zhao Z, Fang LL, Johnsen R, Baillie DL. 2004a. ATP-binding cassette protein E is involved in gene transcription and translation in Caenorhabditis elegans. *Biochem Biophys Res Commun* **323**: 104-111.

Zhao Z, Sheps JA, Ling V, Fang LL, Baillie DL. 2004b. Expression analysis of ABC transporters reveals differential functions of tandemly duplicated genes in Caenorhabditis elegans. *Journal of molecular biology* **344**: 409-417.

Zhao Z, Thomas JH, Chen N, Sheps JA, Baillie DL. 2007. Comparative genomics and adaptive selection of the ATP-binding-cassette gene family in caenorhabditis species. *Genetics* **175**: 1407-1418.