GENE PREDICTION AND RFX TRANSCRIPTIONAL REGULATION ANALYSIS USING COMPARATIVE GENOMICS

by

Jeffrey Shih Chieh Chu
Bachelor of Science, University of British Columbia, 2004
Bachelor of Computer Science, University of British Columbia, 2006

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In the
Department of Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

Spring 2011

APPROVAL

Name:	**Jeffrey Shih Chieh Chu**
Degree:	**Doctor of Philosophy**
Title of Thesis:	**Gene prediction and RFX transcriptional regulation analysis using comparative genomics**

Examining Committee:

Chair: Dr. Nicholas Harden
Associate Professor,
Department of Molecular
Biology and Biochemistry

Dr. Nansheng Chen
Senior Supervisor
Associate Professor, Department of
Molecular Biology and Biochemistry

Dr. Ke Wang
Supervisor
Associate Professor, School of
Computing Science

Dr. David L. Baillie
Co-Senior Supervisor
Professor, Department of Molecular
Biology and Biochemistry

Dr. Steven Jones
Internal Examiner
Professor, Department of Molecular
Biology and Biochemistry

Dr. Fiona S. L. Brinkman
Supervisor
Professor, Department of Molecular
Biology and Biochemistry

Dr. Wyeth Wasserman
External Examiner
Professor, Department of Medical
Genetics, University of British Columbia

Date Defended/Approved: February 23, 2011

ABSTRACT

Regulatory Factor X (RFX) is a family of transcription factors (TF) that is conserved in all metazoans, in some fungi, and in only a few single-cellular organisms. Seven members are found in mammals, nine in fishes, three in fruit flies, and a single member in nematodes and fungi. RFX is involved in many different roles in humans, but a particular function that is conserved in many metazoans is its regulation of ciliogenesis. Probing over 150 genomes for the presence of RFX and ciliary genes led to the understanding of how RFX-cilia regulatory interaction occurred in evolution. Molecular phylogenetic analysis revealed that RFX is only found in metazoans, in some fungi, and in only one unicellular organism, *Monosiga brevicollis*. However, ciliary genes did not co-exist with RFX genes except in *Allomyces macrogynus* and *Monosiga brevicollis*. The data showed that RFX and cilia evolved independently until the time just before the establishment of metazoans. These results suggest that RFX TFs acquired the role of transcriptional regulation on ciliary genes before metazoans arose and such gain-in-function could be a driving force for metazoan evolution.

RFX regulate genes via a regulatory motif called the X-box motif. My laboratory, as well as others, has identified novel RFX target genes in *C. elegans*. However, accumulating evidence suggest more RFX genes could be uncovered and some of these genes could be regulated by divergent X-box motifs. Additional RFX target genes with divergent X-box motifs were identified in *C. elegans* by first revising the gene set in *C. briggsae*, *C. remanei*, and *C. brenneri* using a novel homology-based gene finder, genBlastG. Comparing the four genomes with the revised gene set revealed promoter regions with conserved X-box motif in all species except in *C. elegans*. Detailed examination revealed divergent X-box motifs in these regions. Mutagenesis experiments in the region upstream of F25B4.2 showed that divergent X-box motifs could drive gene expression and may repress gene expression as well. This study provides a deeper understanding regarding the evolution and mechanism of a conserved and important transcription factor.

"For God, who commanded the light to shine out of darkness, hath shined in our hearts, to give the light of the knowledge of the glory of God"

--II Cor. 4:6

ACKNOWLEDGEMENTS

I express my deepest appreciation for my supervisor Dr. Jack Chen who seemingly found me out of nowhere. His dedication, patience, and mentorship have made graduate school life exciting and productive. I also thank Dr. David Baillie, who has been extremely supportive from the first day, for his generosity of lab resources and his scientific insight. My supervisory committee, Dr. Fiona Brinkman and Dr. Jian Pei, has been supportive and guiding me in the right direction. Dr. Ke Wang and Dr. Rong She have been wonderful collaborators who programmed genBlastA and genBlastG, which became instrumental in all parts of my thesis.

I want to thank all the members, past and present, of the Chen lab. They made the Chen lab a wonderful place to work in and I am glad to be part of this great team. I also want to thank the members of the Baillie lab, especially Domena Tu who tirelessly performed all the microinjection procedures. I want to thank Shu Yi Chua and Rylan Fernandes for their friendship. Without them, I would not have experienced so many winnings in MBB social competitions.

Lastly, I thank my parents and my brothers and sisters in Christ, whose continual prayer led to this miracle.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

BBS	Bardet-Biedl Syndrome
BLAST	Basic Local Alignment Search Tool
BLS	Bare Lymphocyte Syndrome
Ciliary genes	Genes that function in IFT
DBD	DNA binding domain
FN	False negative. A correct feature that is not predicted.
FP	False positive. A predicted feature that is incorrect.
Genomic span similarity	A measure to determine how similar is the genomic coordinates of an HSP group to the annotated gene model.
HMM	Hidden Markov Model. Models each position of a feature with a probability of being a particular nucleotide or amino acid
HSP	High-scoring segment pair
HSP group	A set of HSPs returned by genBlastA to be part of one gene
IFT	Intraflagellar transport
MHCII	Major histocompatibility complex class II
PID	Percent identity. This is measured as the ratio of identical residues in a pairwise alignment over the length of the alignment
Promoter element	Specific DNA sequences that bind transcription factors
Promoter region	DNA sequence upstream of the translational start site

	defined in this thesis as 500 bp upstream.
Query Coverage	A measure to determine how much the query is covered by the selected set of HSPs. It is calculated as the proportion of the query sequence that is covered by the HSPs.
RFX	Regulatory Factor X
Sensitivity	Measures the proportion of features from the "Gold standard" is predicted out of all the features from the Gold standard. It is calculated as: TP / (TP+FN)
Specificity	Measures the proportion of features correctly predicted out of all predictions. It is calculated as: TP / (TP+FP)
Splice acceptor	The signal at the 3' end of the intron (usually AG)
Splice donor	The signal at the 5' end of the intron (usually GT)
TF	Transcription Factor
TP	True positive. A predicted feature that is also correct.
X-box motif	The DNA sequence that binds RFX TF

1: GENERAL INTRODUCTION

1.1 Transcription

The central dogma of molecular biology proposes that a gene on the DNA transfers protein-coding information by first transcribing to RNA and ultimately translating RNA to proteins (Crick 1958; Crick 1970). Yet how this process is controlled is an enduring question. The time and place that a gene turns on (expressed) or turns off (repressed) is crucial for development and homeostasis of an organism (Roeder 2003). To understand fully the regulation of gene expression is one of the major challenges in molecular biology today.

Transcription, the process of synthesizing RNA using DNA as template, is the first step of information transfer from DNA to proteins and it is, arguably, the primary step in gene regulation (Roeder 2003). In eukaryotic cells, RNA molecules are transcribed by three RNA polymerases: RNA polymerase I is primarily involved in transcribing 18S and 28S ribosomal RNAs (Roeder and Rutter 1970); RNA polymerase II transcribes protein-coding genes and microRNAs (Weinmann *et al.* 1974; Cai *et al.* 2004; Lee *et al.* 2004); and RNA polymerase III transcribes tRNA and 5S RNA (Weinmann *et al.* 1974; Weinmann and Roeder 1974). Prokaryotic transcription, on the other hand, involves only one multi-subunit RNA polymerase core enzyme (Borukhov and Nudler 2003).

The process of protein-coding gene transcription is nearly parallel between prokaryotes and eukaryotes. It involves three major steps: initiation, elongation, and termination.

1.1.1 Initiation

Transcription initiation in eukaryotes is a multi-step process that involves RNA polymerase II, general transcription factors (consisting of TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH), and numerous cofactors like the mediator complex and chromatin remodelling complexes (Figure 1-1). The first step is the assembly of pre-initiation complex (PIC) at the core promoter elements. Recognition begins with TFIID binding to the TATA-box, which often is located 25 to 30 bp upstream from the transcriptional start site. The binding of TFIID serves as a scaffold for the rest of the transcriptional machinery to bind. Binding of TFIID to the TATA-box is followed by TFIIA and TFIIB that further stabilizes this interaction. In addition to stabilizing TFIID-TATA-box interaction, TFIIB also binds to TFIIB-recognition elements (BRE) that are located upstream and downstream of the TATA-box (Thomas and Chiang 2006). These elements serve as additional anchor points for binding core promoter elements and orienting the polymerase to the proper direction. Next, TFIIF and RNA polymerase II binds to the existing TFIID-TFIIA-TFIIB complex. TFIIF is able to provide additional protein-DNA interaction sites by inducing changes to the DNA topology such that DNA is wrapped around RNA polymerase II. The change in conformation further stabilizes the complex and offers resistance to transcriptional repressors. Lastly, TFIIE and TFIIH are recruited to the complex by interacting with TFIIF. Together, TFIIE stimulates the ATPase, CTD kinase, and DNA helicase activities of TFIIH to allow promoter escape and transition from transcription initiation to transcription elongation.

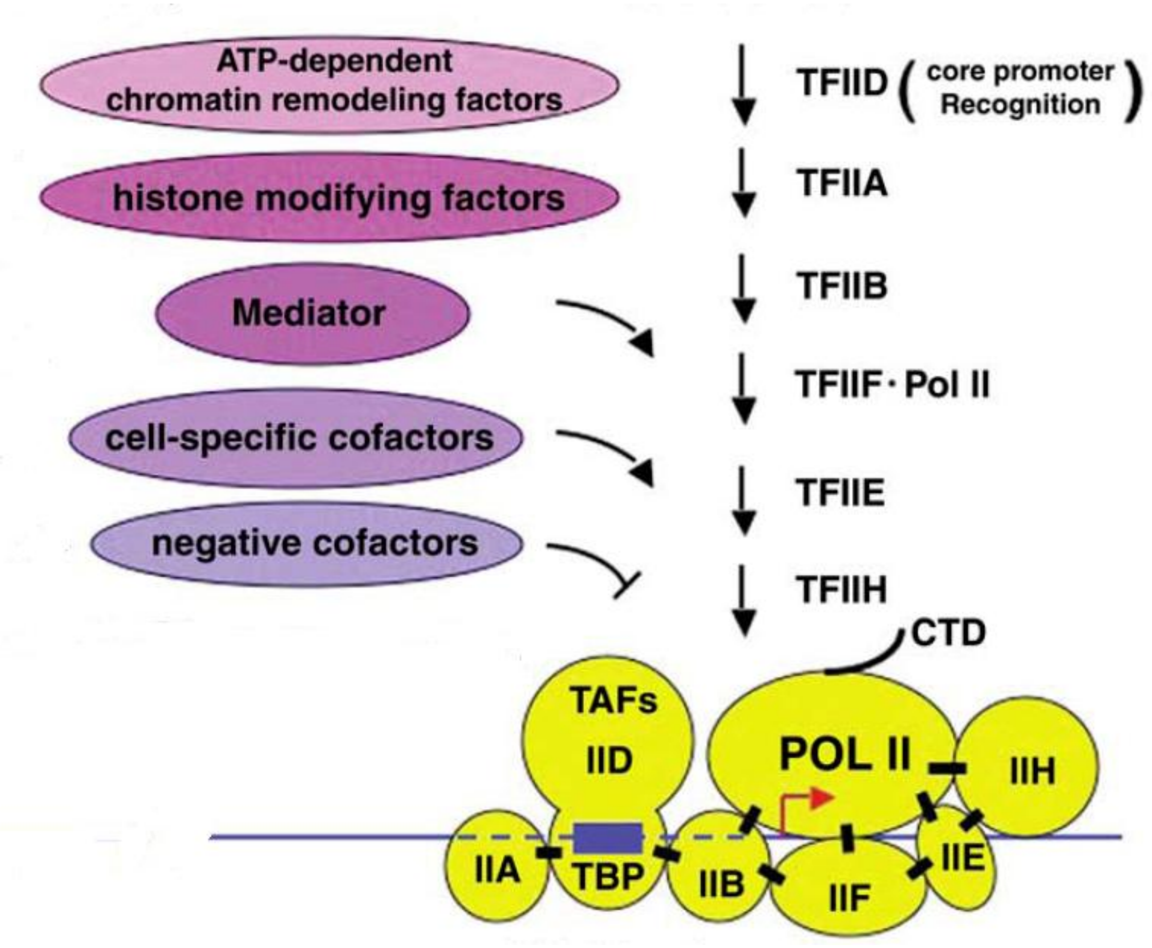


Figure 1-1 The model of PIC assembly and its regulation by other cofactors. This figure is adapted from (Roeder 2005)

The mediator complex can significantly influence transcription initiation. Over 31 subunits have been identified to play a wide variety of roles including physical interaction with RNA polymerase II, interacting with gene specific transcription factors to relay regulatory signals, and enhance phosphorylation of the RNA polymerase II C-terminal domain (CTD) (Thomas and Chiang 2006).

Due to its numerous subunits, mediator complex provides many opportunities for transcriptional regulation.

1.1.2 Elongation

Phosphorylation of CTD by TFIIH destabilizes interaction between RNA polymerase II and the rest of the initiation complex. However, this is not enough to release RNA polymerase II from the promoter into the gene. Initial elongation phase is a slow and inefficient process that is affected by DRB sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) (Nechaev and Adelman 2011). These negative regulators tend to pause and arrest transcription. However, this may provide the time to allow capping at the 5' end of the emerging RNA. Negative regulation by DSIF and NELF is reversed when positive transcription elongation factor b (P-TEFb) is recruited to the complex and phophorylates CTD, DSIF and NELF. The resulting complex is highly stable and can transcribe hundreds of thousands of bases (Nechaev and Adelman 2011).

1.1.3 Termination

Most protein-coding gene mRNA transcripts are terminated by polyadenylation. Pcf11 is the polyadenylation factor that preferentially binds to phosphorylated CTD and helps recruiting polyadenylation machinery. Polyadenylation machinery recognizes specific signals at the 3' end of the precursor mRNA and cleaves the RNA catalyzed by Cleavage and Polyadenylation Specific Factor (CPSF) (Beaudoing *et al.* 2000; Davila Lopez and Samuelsson 2008). Once cleaved, polyadenylate polymerase (PAP) extends the poly-A tail (Balbo and Bohm 2007).

1.2 Gene specific transcription factors

Gene specific transcription factors, in contrast to general transcription factors, are a class of transcription factors that directly interact with DNA at a very specific location in a sequence-dependent manner (Levine and Tjian 2003). These transcription factors are the major controllers of gene expression by interacting with a variety of factors and bringing them closer to the transcription initiation complex. These interactions include those with the mediator complex and chromatin remodelling complex. As the complexity of an organism increases, it would require a more diverse repertoire of regulation, both temporally and spatially. This is suggested by correlating the number of transcription factors with the complexity of different species where a less complex species, such as yeast, encodes about 300 transcription factors and a more complex species, such as humans, may have as many as 3000 transcription factors (Levine and Tjian 2003). For example, the regulatory factor X (RFX) transcription factor is a type of transcription factor that had undergone gene expansion in higher eukaryotes. Only one member can be found in yeast *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, but three copies can be found in *Drosophila melanogaster* and seven copies in humans (Emery *et al.* 1996; Aftab *et al.* 2008; Chu *et al.* 2010).

1.3 Regulatory Factor X (RFX) transcription factor

Regulatory Factor X (RFX) is a family of transcription factors consisting of seven members (RFX1-7) in mammals (Aftab *et al.* 2008; Chu *et al.* 2010). RFX genes are also extensively studied in *S. cerevisiae* (Huang *et al.* 1998), *D. melanogaster* (Durand *et al.* 2000; Otsuki *et al.* 2004), and *C. elegans* (Swoboda et al. 2000). The common feature among all RFX members is the highly conserved winged-helix DNA binding domain (DBD) (see Section 1.4.1) that shows more than 40% identity between *C. elegans*, yeast, and humans at the amino acid level (Emery *et al.* 1996; Chu *et al.* 2010). RFX transcription factors are responsible for diverse roles in humans and have been associated with an increasing number of disease conditions. I will summarize the function of each RFX transcription factor in humans, *D. melanogaster*, *C. elegans*, and *S. cerevisiae* and point out a conserved function in nearly all species.

1.3.1 RFX1

RFX1 is ubiquitously expressed in mammals with highest expression in the brain (Aftab *et al.* 2008; Feng *et al.* 2009). It is found to regulate a variety of genes but its main function is still unclear. It is known to activate interleukin-5 receptor α gene (IL5RA) (Iwama *et al.* 1999) and neuronal glutamate transporters type 3 (EAAT3) (Ma *et al.* 2006) and repress c-myc (Reinhold *et al.* 1995), proliferating cell nuclear antigen (PCNA) (Liu *et al.* 1999), collagen α2(I) (COL1A2) (Sengupta *et al.* 2002; Xu *et al.* 2006), and RFX1 itself (Katan-Khaykovich and Shaul 2001; Lubelsky *et al.* 2005). RFX1 is able to activate and repress genes, which suggests RFX1's activity is context dependent (Katan *et al.*

1997). For example, RFX1 can activate and repress immediate early gene Id2 (Wang *et al.* 2007). Id genes are important for cell fate determination, differentiation, and cell proliferation. These genes are inducible by serum very early on without *de novo* protein synthesis (Wang *et al.* 2007). RFX1 binds to the promoter region of Id2 as a homodimer before and after serum induction: represses Id2 before serum induction via its C-terminus and activates Id2 after serum induction via its N-terminus (Wang *et al.* 2007). This observation agrees with RFX1 activation/repression of EP elements (Katan *et al.* 1997) where RFX1 is thought to be in neutral state normally and changes its activity upon other signals. In the case of EP element, N-terminus is sufficient in driving reporter gene expression and the C-terminus is sufficient for repression (Katan *et al.* 1997).

1.3.2 RFX2

RFX2 is highly expressed in the testis and is suggested to play an important role in spermatogenesis by regulating testis-specific H1t gene (Horvath *et al.* 2004; Wolfe *et al.* 2004) and Alf, an important gene for spermatogenesis (Wolfe *et al.* 1995; Wolfe *et al.* 2006; VanWert *et al.* 2008; Horvath *et al.* 2009). Recent studies showed that RFX2 transcripts accumulate much more than RFX1, 3, and 4 in spermatocytes and spermatids and it is likely the main regulator of gene expression during pachytene (Kistler *et al.* 2009). A search in the RFX2 promoter revealed a number of GC boxes and three perfect MYB binding sites (Horvath *et al.* 2009). A-MYB is an essential transcription factor in pachytene and it is likely the source of high RFX2 accumulation in the testis (Toscani *et al.* 1997).

1.3.3 RFX3

RFX3 is expressed in the testis, pancreas, but most prominently in the brain. RFX3 is crucial for proper primary cilia development in embryo nodal cells (Bonnafe *et al.* 2004), brain ependymal cells (Baas *et al.* 2006) and pancreatic endocrine cells (Ait-Lounis *et al.* 2007). RFX3 knockout mice show severe defects in ciliary structure (Bonnafe *et al.* 2004) and general epithelial cell polarity defects (Baas *et al.* 2006). RFX3 actively regulate a dynein light chain gene, Dync2li1, which is a motor component for cilia assembly (Bonnafe *et al.* 2004). As a result, RFX3 knockout mice show high mortality rate and those that survive develop *situs inversus* and hydrocephalus. In addition to regulating primary cilia development, RFX3 also regulates forkhead box J1 (Foxj1), an important transcription factor in regulating genes in motile cilia by binding to the promoter of Foxj1 (El Zein *et al.* 2009). The regulatory interaction between RFX3 and Foxj1 suggests RFX3 is a prominent regulator for cilia development in general.

RFX3 also has repressive targets. Microtubule-associated protein 1A (MAP1A) is an abundant protein in the brain that stabilizes microtubules, mediates mRNA attachment to microtubules, and interacts with post synaptic proteins. RFX3 represses MAP1A in non-neuronal cells and is likely a major contributor to MAP1A tissue specificity (Nakayama *et al.* 2003).

1.3.4 RFX4

RFX4 was recently shown to regulate intraflagellar transport (IFT) via Ift172 in mouse for proper cilia formation in neuronal tissues such as the dorsal and ventral portions of the neural tube (Ashique *et al.* 2009). As a result, mutations in RFX4 cause ciliary defects in telencephalon and spinal cord (Ashique *et al.* 2009).

1.3.5 RFX5

RFX5 functions in the immune system by regulating the expression of major histocompatibility complex class II (MHCII) genes (Kara and Glimcher 1991). It was identified from a cDNA library screen for plasmids that can rescue MHCII gene expression in MHCII deficiency cell lines (Steimle *et al.* 1995). RFX5 is part of the RFX complex that also comprises of RFXB and RFXAP. All three members are required for proper expression of MHCII genes (Kara and Glimcher 1991). Within the complex, RFX5 homodimerizes to form RFX5$_{(2)}$-RFXB-RFXAP complex (Garvie *et al.* 2007). While only RFX5 can bind DNA, RFX5 requires the two other members to bind DNA with high affinity (Garvie and Boss 2008). RFX5 seems to autoinhibit its own DNA binding and RFXB and RFXAP relieves the autoinhibition (Garvie and Boss 2008). HMCII genes have a set of common *cis*-regulatory elements: W box, X1 box, X2 box, and the Y box. The X1, X2 and Y box binds the RFX complex, CREB, and NF-Y respectively (Boss and Jensen 2003). Together, these three elements form a stable complex at the promoter. The final activation of MHCII genes only happens upon binding of CIITA to the complex (Beresford and Boss 2001).

The transactivator CIITA, together with RFX5, can also repress a number of genes including collagens (Sengupta *et al.* 2002).It is thought that INF-γ response triggers expression of CIITA and represses COL1A2 expression via interacting with RFX5 (Piskurich *et al.* 1998; Piskurich *et al.* 1999; Sengupta *et al.* 2002). Histone modification is a likely mechanism of repression. RFX complex interacts with HDAC2 (Xu *et al.* 2006), BRG1(Mudhasani and Fontes 2005), and HDAC4 (Wang *et al.* 2005) that deacetylate histones in the first exon of COL1A2 gene (Xu *et al.* 2006).

Due to its function in regulating MHCII genes, RFX5 is associated with a human disease condition called the bare lymphocyte syndrome (BLS). This disorder is characterized by the absence of MHCII molecules on lymphocytes. Without MHCII molecules, lymphocytes are unable to present antigens to CD4 cells, thus impairing the immune response (Elhasid and Etzioni 1996; van Eggermond *et al.* 2008). BLS is caused by defects in any of the RFX complex. The genetic defects of these genes are inherited in an autosomal recessive fashion (van Eggermond *et al.* 2008).

1.3.6 RFX6

RFX6, which I identified in collaboration with my colleagues, is almost exclusively expressed in the pancreas (Aftab *et al.* 2008). Recent literatures have shown the importance of RFX6 in pancreatic islet formation by regulating many genes involved in β-cell maturation, such as islet hormone genes, zinc transporter Slc30a8, and G-protein coupled receptor Ffar1 (Smith *et al.* 2010;

Soyer *et al.* 2010). The involvement in islet formation suggests its function in insulin production.

1.3.7 RFX7

RFX7 is the the least studied member of the RFX family. While it shows ubiquitous expression (Aftab *et al.* 2008), its function is largely unexplored.

1.3.8 Yeast RFX1 (yRFX1)

Yeast RFX1, also called CRT1 to stand for constitutive RNR transcription, is the only RFX member in yeast. It functions in damaged induced DNA repair by regulating a set of genes including ribonucleotide reductase genes (RNR). Recent research shows that yRFX1 contains domains for repression and activation of its target genes which include RNR2, RNR3, RNR4, and yRFX1 itself (Huang *et al.* 1998; Zhang and Reese 2005). yRFX1 does not have repression activity on its own but requires co-repressors such as TUP1 and SSN6 to bind at the N-terminal domain (Huang *et al.* 1998; Zhang and Reese 2005). Surprisingly, the activating part of the protein, where it binds TFIID and SWI/SNF nucleosome remodelling complex, is also in the N-terminus overlapping with the repression domain (Zhang and Reese 2005). Because yRFX1 binds its own promoter, it undergoes a negative self feedback loop. During DNA damage or under damaging agent, such as hydroxyurea (HU) or methyl methane sulphate (MMS), yRFX1 transcript is increased. During replication block, Dun1 phosphorylates yRFX1 to reduce binding of its own promoter and derepresses the target genes (Lubelsky *et al.* 2005). As DNA damage is repaired and yRFX1 protein level decrease, repression is restored (Huang *et al.* 1998). Another study have suggested phosphorylation changes from repression to activation by changing N-terminus partner to TFIID (Zhang and Reese 2005). After TFIID binds DNA, yRFX1 is free to exit from the promoter (Zhang and Reese 2005).

1.3.9 *Drosophila* RFX (dRFX)

This is the first of the three RFX protein found in *Drosophila* by using hybridization of human RFX1 DBD sequence to a *D. melanogaster* genomic library and confirmed by 5'-RACE (Durand *et al.* 2000). dRFX shows homology to mammalian RFX1-3. Many functional domains (See section 1.4) are strongly conserved (Durand *et al.* 2000). dRFX is expressed throughout all developmental stages (Durand *et al.* 2000). It is first observed in 10-11 cell stage where two sensory organ precursor cells of each thoracic and abdominal segment are seen. At stage 12, 2nd order precursor cells are expressed. From stage 14 onwards, the expression is restricted more in the brain and chordotonal organs (lateral and ventral). At the end of embryogenesis, dRFX is only found in type I neurons of thoracic and abdominal segments and all sensory neurons in the head (Vandaele *et al.* 2001). Mutation in dRFX causes 85% lethality during first instar larvae. The surviving larvae grow to pupae that show defects in sensing odor, chemotaxis, mechanotransduction, and auditory function (Dubruille *et al.* 2002). The cilia of type I neurons show developmental defects (Dubruille *et al.* 2002). Taken together, dRFX is important for cilia development, similar to RFX3 in mammals.

1.3.10 *Drosophila* RFX1 (dRFX1)

dRFX1 was identified by homology based search using the highly conserved DNA binding domain (DBD) from mammalian RFX genes as queries. dRFX1 showed closest similarity to mammalian RFX5 (Durand *et al.* 2000; Chu *et al.* 2010), however its function has not been explored.

1.3.11 *Drosophila* RFX2 (dRFX2)

dRFX2 was found through yeast-1-hybrid using an upstream regulatory element from the PCNA gene against a *Drosophila* cDNA library (Otsuki *et al.* 2004). The clone is confirmed by 5'-RACE and hybridizing the cDNA sequence to a genomic library (Otsuki *et al.* 2004). The 842 intronless amino acid sequence was found to have 57% identity with dRFX (Otsuki *et al.* 2004). According to northern blot analysis, dRFX2 is expressed in embryos and larvae but not found in pupae and adults (Otsuki *et al.* 2004). Expression can be seen in the salivary gland in larvae.

1.3.12 *C. elegans* DAF-19

DAF-19 is the first RFX transcription factor identified to regulate ciliary functions. The first critical evidence linking RFX and ciliary genes was reported by Swoboda and colleagues (Swoboda et al. 2000) where they cloned *daf-19* in *C. elegans* and found that it is the first and only RFX gene in *C. elegans*. They showed that in the absence of a functional DAF-19, ciliated neurons in *C. elegans* lost their cilia and displayed chemosensory defects (Che), dye filling defect (Dyf), and constitutive dauer formation (Daf-c) (Swoboda et al. 2000).

Furthermore, they demonstrated that DAF-19 regulates the expression of ciliary genes, including *che-2*, *osm-1*, *osm-6* and many Bardet-Biedl Syndrome (BBS) genes through binding to a DNA element called the X-box motif (Emery et al. 1996; Swoboda et al. 2000).

The expression pattern of DAF-19 is rather complex. Currently, four alternative transcripts have been identified: *daf-19a*, *daf-19b*, *daf-19c*, and *daf-19d* (Figure 1-2). Based on antibody staining, *daf-19a* and *daf-19b*, were shown to express in non-ciliated neurons (Senti and Swoboda 2008). *daf-19d* (labelled as *daf-19c* in Senti *et al*. but annotated as *daf-19d* in WormBase), on the other hand, is solely expressed in the 60 ciliated neurons based on antibody staining (Senti and Swoboda 2008). The shortest transcript, *daf-19c* (labelled as *daf-19m* in Wang *et al*. but annotated as *daf-19c* in WormBase), is specifically expressed in many male-specific neurons, such as CEM, HOB, and RnB neurons (Wang *et al.* 2010).

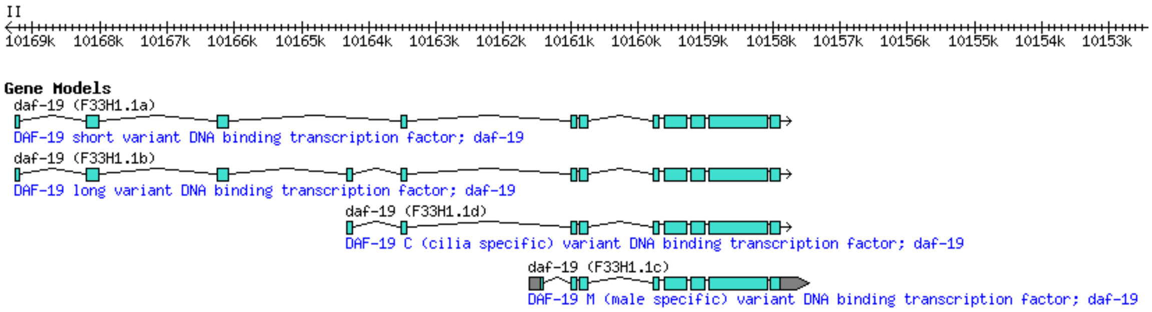

II
10169k
10168k
10167k
10166k
10165k
10164k
10163k
10162k
10161k
10160k
10159k
10158k
10157k
10156k
10155k
10154k
10153k
Gene Models
daf-19 (F33H1.1a)
DAF-19 short variant DNA binding transcription factor; daf-19
daf-19 (F33H1.1b)
DAF-19 long variant DNA binding transcription factor; daf-19
daf-19 (F33H1.1d)
DAF-19 C (cilia specific) variant DNA binding transcription factor; daf-19
daf-19 (F33H1.1c)
DAF-19 M (male specific) variant DNA binding transcription factor; daf-19

Figure 1-2 Four alternative splice forms annotated in WormBase WS221

The differential expression pattern of *daf-19* isoforms also correlates with their function. DAF-19a/b were suggested to function in synapse of non-ciliated neurons and affect synaptic protein expression (Senti and Swoboda 2008). DAF-19d functions in cilia formation and it is sufficient to rescue Dyf and Daf-c phenotype (Senti and Swoboda 2008). DAF-19c, when mutated, disrupts sensory signalling genes in IL2 and male specific ciliated neurons but not the genes required for ciliogenesis (Wang *et al.* 2010). *daf-19c* specific mutant worms display location of vulva (Lov) and response (Rsp) phenotype suggesting *daf-19c* plays an important role in mating (Wang *et al.* 2010).

1.4 Functional domains of RFX

Different members of RFX can function in drastically different pathways and processes. However, RFX transcription factors share many functional domains, of which all RFX members share the DNA binding domain. The other domains that exist in some RFX members are the activation domain and the dimerization domains (Figure 1-3).

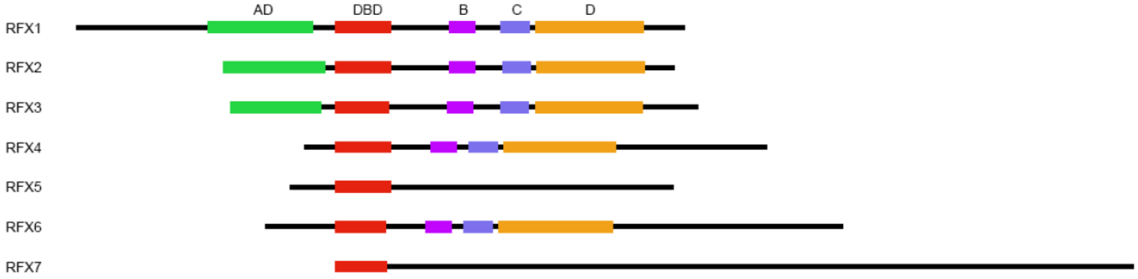


Figure 1-3 The domains of RFX transcription factors. AD (activation domain) is found only in RFX1-3; DBD (DNA binding domain) is found in all members; B, C, and D domains are the dimerization and extended-dimerization domains.

1.4.1 DNA binding domain (DBD)

RFX DBD spans 76 amino acids and shows more than 40% identity between yeast, fly, worm, and humans at the protein level (Emery *et al.* 1996; Gajiwala *et al.* 2000; Aftab *et al.* 2008; Chu *et al.* 2010).

RFX DBD has a winged helix type structure. Winged helix structure was first characterized in hepatocyte nuclear factor 3 (HNF-3), which shared high conservation with fork head protein in *Drosophila* (Costa *et al.* 1989; Weigel *et al.* 1989). Winged helix DBD from HNF-3 is characterized by three α-helices (H1-3), three β-strands (S1-3), and two wings (W1 and W2) (Gajiwala and Burley 2000). The two wing loops flanks H3, resembling a butterfly and hence the name, winged helix (Lai *et al.* 1993). RFX DBD was crystallized using RFX1 DBD with a palindromic sequence (CGTTACCATGGTAACG) called the X-box motif (See section 1.6) (Gajiwala *et al.* 2000). Crystallized structure revealed that RFX1 DBD lacks W2, similar to another winged helix protein family E2F (Zheng *et al.* 1999; Gajiwala *et al.* 2000) (Figure 1-4). All winged helix proteins, including the HNF family and the E2F family, use H3 to bind to the major groove and makes most of the contact with DNA. However, RFX1 DBD only makes a single DNA interaction in the minor groove with H3, while most of the DNA contact comes from W1 (Gajiwala *et al.* 2000). Arg 58, Gly 60, and Arg 62 in the W1 region make direct hydrogen bonding interactions with the last three nucleotides of the X-box motif in the reverse strand. In addition, Ser 65 and Tyr 67 in the W1 region make water-mediated interactions with the 9th nucleotide (Gajiwala *et al.* 2000) (Figure 1-5).

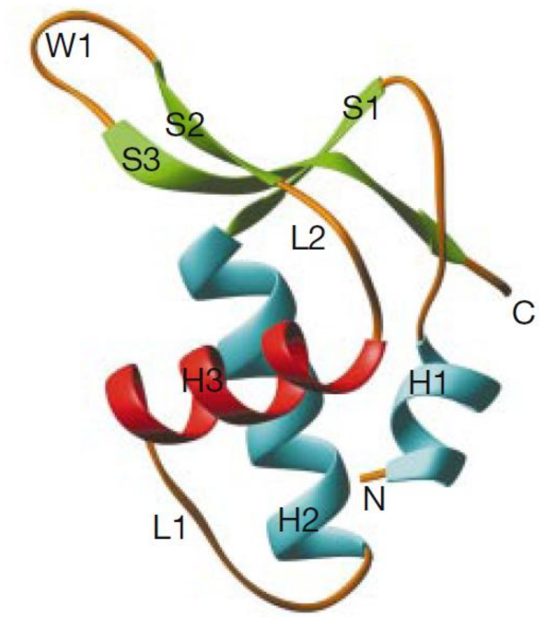


Figure 1-4 The winged helix structure of human RFX1 DBD. This figure is adapted from (Gajiwala *et al.* 2000)

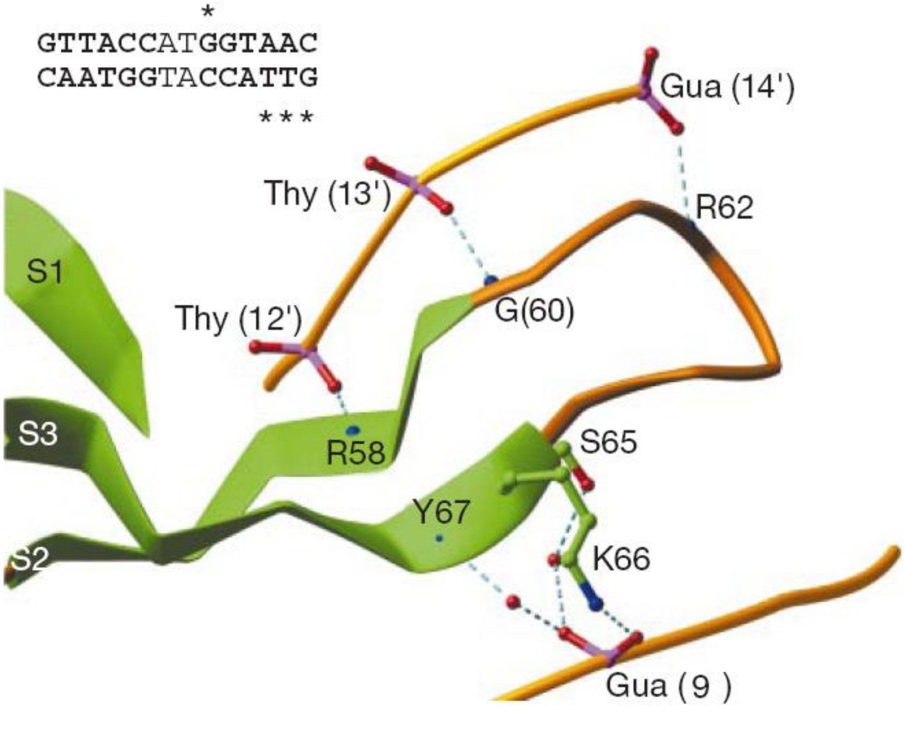


Figure 1-5 The W1 domain of RFX1. Residue R58, G60, and R62 have direct interaction with the last three nucleotides of the X-box motif in the minus strand. S65 and Y67 also make water-mediated interactions with guanine at the ninth position. This figure is adapted from (Gajiwala *et al.* 2000).

1.4.2 Activation domain

The N terminus of RFX1, RFX2, and RFX3 are rich in proline and glutamine, which are the hallmarks of activation domains (Emery *et al.* 1996). Gene cut down studies showed that RFX1 with glutamine-rich region (residue 233-351) removed have dramatically reduced activity (Katan *et al.* 1997). Furthermore, these residues, when fused with GAL4 DBD, can activate reporter gene independently from other parts of RFX1 (Katan *et al.* 1997). This region is conserved between RFX1, 2, and 3 but it is not found in all other RFX members, including those in *S. cerevisiae*, *D. melanogaster* and *C. elegans* (Emery *et al.* 1996; Aftab *et al.* 2008; Chu *et al.* 2010). The members that lack activation domains may serve as co-activators in a protein complex. RFX5 is a good example of its role as a co-activator forming a stable complex with RFXB and RFXAP. The complex then, interacts with CIITA to activate MHCII genes (Beresford and Boss 2001; Garvie and Boss 2008). Similarly, RFX4 requires GPS2 to promote the transcription of *Cx3cl1* in COS-1 cell lines (Zhang *et al.* 2008).

1.4.3 Dimerization domain

Typical transcription factors work as a dimer and RFX is no exception. Several studies have shown, using gel electrophoresis mobility shift assays (EMSA) and co-immunoprecipitation, that RFX1, RFX2, RFX3, RFX4, and RFX6 can homodimerize and heterodimerize with itself or with each other (Reith *et al.* 1994; Iwama *et al.* 1999; Katan-Khaykovich *et al.* 1999; Morotomi-Yano *et al.* 2002; Smith *et al.* 2010). Dimerization of RFX proteins occurs at the highly conserved dimerization domain near the C-terminus. Dimerization domain found in RFX1-4 and 6 is also found in dRFX and DAF-19 suggesting that these protein can also homodimerize with itself or heterodimerize with other factors (Efimenko *et al.* 2005). RFX5 and RFX7 do not possess the dimerization domain found in the other members, yet, RFX5 is known to dimerize and interact with many other factors, including RFXB, RFXAP, and CIITA. In contrast, the N-terminus of RFX5 has a stretch of leucine rich region that is critical for dimerization (Jabrane-Ferrat *et al.* 2002). In fact, mutating one leucine in this region will abolish dimerization (Jabrane-Ferrat *et al.* 2002). Without dimerization, RFX complex does not assemble at MHCII gene promoters highlighting the importance for functioning in dimers.

Our understanding of the difference between homodimers and heterodimers of RFX is still incomplete. For instance, whether RFX3 homodimers and RFX3-RFX4 heterodimers would regulate different set of genes is still unknown. Only a recent study on RFX3-RFX6 interaction suggests the two may cooperate in regulating a subset of genes for pancreatic islet formation but not the genes for cilia formation (Smith *et al.* 2010). It is possible that RFX interacts with different partners to achieve spatial and temporal specificity in gene regulation.

1.5 Conservation of RFX and its role in ciliogenesis

Identification and elucidation of the RFX DBD revealed that it is the defining feature of RFX proteins. Such characteristic domain opened the way to identify additional RFX genes in other species. In 2008, my colleagues and I generated a hidden Markov model (HMM) of the DBD from human RFX1-5 and used it to probe the human genome (Aftab *et al.* 2008). Not only did we recover all five RFX genes, we also uncovered RFX6 and RFX7. Taking advantage of many sequenced mammalian genomes, we further probed five mammalian species (chimpanzee, monkey, dog, mouse, and rat) and found all seven RFX is conserved in all species. Using the DBD sequences and phylogenetic analysis, we were able to categorize RFX genes into three subgroups: (1) RFX1-3, (2) RFX4 and 6, (3) RFX5 and 7. Not only DBD sequences show three groups, the members in each subgroup share the same functional domains in addition to the DBD: RFX1-3 possess both activation domain and dimerization domain; RFX4 and RFX6 possess only dimerization domain; RFX5 and RFX7 possess none of these two domains. The three subgroups suggest there might be multiple independent incidences of gene duplication prior to the radiation of mammals.

Although each RFX gene has a different function, it appears that its role in cilia development is conserved in all species that are studied (all except yeast). RFX3 and RFX4 are important for cilia development in humans and mice; DAF-19 in *C. elegans*; and dRFX in *D. malenogaster*. It seems that the regulation of ciliary genes by RFX is established very early on in metazoan evolution. As I will present in Chapter 2 of my thesis, RFX and cilia have evolved independently but may have coincided and interacted with ciliary genes some point in time just before the establishment of metazoans. In contrast, the yeast RFX may have coincided with ciliary genes yet did not establish any regulatory interaction.

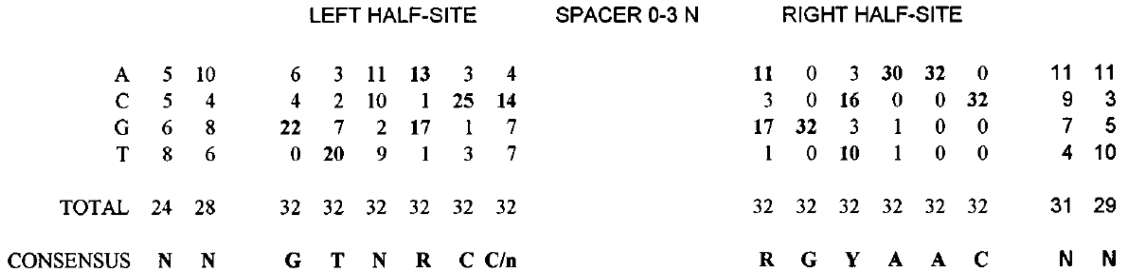
Therefore, through evolution, some fungi lost cilia, some fungi lost RFX, and some fungi lost both.

1.6 RFX binding motif

A conserved DNA binding domain across all RFX genes suggests a conserved DNA binding site. The first RFX binding motif was identified in the promoters of MHCII genes. When studying MHCII gene structure in the 80s, two conserved elements were found in the 5’ UTR. These elements were found in different Human MHCII genes and mouse genes Eα and Eβ (Mathis *et al.* 1983; Saito *et al.* 1983; Kelly and Trowsdale 1985; O'Sullivan *et al.* 1986). These two elements are named X-box and Y-box (Dorn *et al.* 1987). The X-box was found to be essential for driving reporter gene expression in B-cells (Sherman *et al.* 1987; Sloan and Boss 1988) and human fibroblasts (Boss and Strominger 1986). Furthermore, X-box is important for driving expression in tissue specific manner. Constructs with X-box only drives expression in B-cells while constructs without X-box drives strong expression in T-cells but not B-cells (Sloan and Boss 1988).

Several studies at the time have shown that X-box binds nuclear factors (Dorn *et al.* 1987; Miwa *et al.* 1987; Sherman *et al.* 1987; Reith *et al.* 1988). By comparing EMSA data from normal B-cells and B-cells established from severe combined immunodeficiency (SCID) patients, Reith *et al*. found the protein that specifically binds to the X-box and named it “RF-X” (Reith *et al.* 1988). They were subsequently able to clone the cDNA that encodes the first RFX protein (RFX1) by screening a λ gt11 library (Reith *et al.* 1989).

Since then, many instances of X-box motifs have been found in the promoter of MHCII genes (Steimle *et al.* 1995), hepatitis B virus surface antigen (Siegrist *et al.* 1993), and c-myc (Reinhold *et al.* 1995). However, it is not until 1996 when Emery *et al*. used site selection procedure with random oligonucleotides and determined the preferred binding consensus motif for RFX1 (Emery *et al.* 1996). The X-box consensus motif is characterized by two 6-bp half sites separated by zero to three nucleotides of spacing: GTNRCC/n-(N_{0-3})-RGYAAC (Figure 1-6). That being said, there are still many sequences that diverge quite a bit from the consensus sequence and bind RFX1, albeit at a lower affinity (Emery *et al.* 1996). However, in all cases, at least one half site closely resembles RGYAAC motif (Emery *et al.* 1996). Their observations suggest X-box motif can be quite degenerate and yet remain functional as long as one of the two half sites matches the consensus.



			LEFT HALF-SITE						SPACER 0-3 N	RIGHT HALF-SITE							
A	5	10	6	3	11	13	3	4		11	0	3	30	32	0	11	11
C	5	4	4	2	10	1	25	14		3	0	16	0	0	32	9	3
G	6	8	22	7	2	17	1	7		17	32	3	1	0	0	7	5
T	8	6	0	20	9	1	3	7		1	0	10	1	0	0	4	10
TOTAL	24	28	32	32	32	32	32	32		32	32	32	32	32	32	31	29
CONSENSUS	N	N	G	T	N	R	C	C/n		R	G	Y	A	A	C	N	N

Figure 1-6 The preferred binding consensus of human RFX1. This figure is adapted from (Emery *et al.* 1996)

1.7 Application of X-box motif: finding ciliary genes

Identification of X-box motifs in RFX regulated genes and elucidation of its consensus sequence invites the question whether additional target genes, especially those involved in cilia development, can be found by simply looking for the presence of X-box motifs. Four studies applied this idea using a combination of bioinformatic and molecular biology approaches: three performed in *C. elegans* and one in *D. melanogaster*.

In 2005, Efimenko *et al.* (Efimenko *et al.* 2005) searched the *C. elegans* promoter regions (defined in this project as the 1Kb genomic sequence upstream of the translational start site of each gene) for candidate X-box motifs that resemble a “relaxed” X-box consensus (RYYNYY WW RRNRAC), a “refined” X-box consensus (GTHNYY AT RRNAAC), and an “average” X-box consensus (RTHNYY WT RRNRAC). Their search returned 1927, 128, and 758 candidate target genes, respectively. They examined the expression pattern of a subset of 758 candidates and were able to subdivide those target genes into two groups. Group 1 genes are those strongly regulated by DAF-19 and critical for cilia development. Genes belonging to this group include members of the dynein motor *xbx-1* and *dylt-2*, members of the IFT complex B, as well as members of the BBS complex (Efimenko *et al.* 2005; Ou *et al.* 2007). These genes are expressed in most, if not all, of the ciliated neurons in *C. elegans*. Group 2 genes are those required in certain ciliated neurons for specialized functions. Genes belonging to this group include IFT complex A gene *che-11* and many *xbx* genes (e.g. *xbx-3* – *xbx-7*). Their expression is usually localized to a subset of ciliated neurons. In some cases, expression in other tissues was also observed (e.g. *xbx-6*).

In the same year, Blacque *et al.* (Blacque *et al.* 2005) looked for putative DAF-19 target genes in *C. elegans* by employing a HMM based approach. The HMM, trained with 22 known X-box motifs, was used to examine 1500 bp upstream promoter region of each gene for putative X-box motifs. To focus on finding ciliary genes, they cross matched the candidates to genes that show enriched expression in ciliated neurons based on SAGE tags. 46 candidate genes were found that (1) have enriched expression in ciliated neurons in comparison to pan-neuronal, muscle, and gut and (2) have a putative X-box motif within 250 bp of the translational start site. Expression analysis of 27 of these 46 candidate genes showed that they are expressed in ciliated neurons. One of the genes identified was *dyf-13* (C27H5.7) which has been suggested to dock *osm-3* kinesin motor to IFT complex B (Ou *et al.* 2007).

The third project in *C. elegans*, which was done in my laboratory, took advantage of additional *Caenorhabditis* genomes available at the time and carried out an X-box motif search using comparative genomics (Chen *et al.* 2006). We looked for putative X-box motifs using HMM based approach in three *Caenorhabditis* species (*C. elegans*, *C. briggsae*, and *C. remanei*) and screened for candidate genes that contain X-box motifs in all orthologous promoters regions. We used a longer promoter region sequences (2000 bp) than the previous two studies because there are cases in which X-box motifs are found outside of the 1 kb genomic sequence upstream of the translational start site

(Fan *et al.* 2004). Our search returned 94 genes where their promoter regions contain X-box motifs in all three species. One of the genes in our list is M04C9.5 (*dyf-5*) and it is *daf-19* dependent (Chen *et al.* 2006) as well as X-box dependent (Wang and Chen, unpublished results).

Finally, the most recent genome-wide search for X-box motifs and RFX target genes was performed in *D. malenogaster* using a similar comparative genomics approach (Laurencon *et al.* 2007). The authors looked for putative X-box motifs in *D. melanogaster* and *D. pseudoobscura* using a Perl based sequence matching algorithm with varying degree of sequence degeneracy. They identified 83 candidate genes using their most stringent criteria (matching GYTRYY N1-3 RRHRAC within 1000 bp of upstream promoter region). Examining 25 candidates revealed 16 are down-regulated. Some of these genes include CG15161 (homolog of *dyf-6*), CG4536 (homolog of *osm-9*), CG3259 (homolog of *dyf-11*), and CG9227 (homolog of *tza-1*).

1.8 From ciliary genes in model organisms to ciliopathy genes in humans

Identifying ciliary genes in model organism, like *C. elegans* and *D. melanogaster,* can facilitate identification of ciliary genes in humans, which can be a challenging and time-consuming task. Many ciliary genes in humans are associated with ciliopathies, which are a group of disorders associated with ciliary defects (Badano *et al.* 2006). Specific disorders classified as ciliopathy include polycystic kidney disease (PKD), immotile cilia syndrome, Bardet-Biedl syndrome (BBS), Meckel-Gruber syndrome (MKS), Oral-Facial-Digital syndrome, Nephronophthisis, Retinitis pigmentosa, and *situs inversus* (Leitch *et al.* 2008; Marshall 2008). The identification of many ciliopathy genes, especially BBS genes, were greatly facilitated by studies in *C. elegans* (BBS3, BBS7 and BBS8), zebra fish (BBS2, BBS11 and BBS12), and *D. malenogaster* (BBS5) (Beales 2005; Blacque and Leroux 2006; Stoetzel *et al.* 2007).

1.9 Additional search needed

Even with our current advances, the known 14 human BBS genes only constitute 25% to 50% of the ciliopathy cases (Yang *et al.* 2008). Therefore, the search for additional target genes is needed. A few reasons indicate RFX target genes can still be found: First, an important caveat in previous genome-wide projects is that they were designed to find X-box motifs similar to the known 14 bp consensus. Consequently, more diverged X-box motifs would be missed and therefore candidate ciliary genes missed. For example, the X-box motif found in the promoter of *nph-1* is 15 bp, which was missed in all three *C. elegans* projects (Winkelbauer et al. 2005). Second, some X-box motifs found in humans can be even more variable in length in the middle spacer region than initially reported (Emery et al. 1996; Lubelsky et al. 2005). For example, RFX1 binds to its own upstream promoter region with two halves of an X-box motif separated by 60 bps (Lubelsky et al. 2005). These “divergent motifs” cannot be found in consensus-based searches. Finally, HMM profile build for searching X-box motifs in the *C. elegans* genome, which retrieves many ciliary genes in *C. elegans*, fail to identify X-box motifs in most known ciliary genes in the human genome despite the high conservation of RFX DBDs. One possibility is that these ciliary genes contain X-box motifs that are different from the consensus sequence.

In Chapter 5 of my thesis, I will present candidate genes in *C. elegans* identified by having divergent X-box motifs in its promoter region using comparative genomics. These divergent X-box motifs may function differently than what has been found before in *C. elegans*.

1.10 Comparative genomics

Comparative genomics is becoming an indispensible tool in the age where whole-genome sequencing is readily accessible and genomic data is abundant. The principle behind comparative genomics is that functional elements are conserved in different genomes and these conserved elements will share similarities (Hardison 2003). It is a tool to separate the similarities from the differences between genomes. The higher the number of genomes used in comparison, the greater the comparative power to tackle many biological questions. For example, the availability of hundres of genomes across different phylum allows me to study the evolutionary history of RFX genes and ciliary genes at a greater resolution (Chapter 2). In another question, I can study the conservation of X-box motifs in four different *Caenorhabditis* species and thereby predicting its importance. Furthermore, comparing the orthologous promoter region in each *Caenorhabditis* species can tell me whether a gene contains a conserved X-box motif or may harbor a divergent X-box motif (Chapter 5).

One of the early steps, if not the first step, in comparative genomics studies is the proper correspondence of orthologous sequences (Kellis *et al.* 2004). However, accurate calling of orthology relies heavily on the quality of genome annotation. While the *C. elegans* genome is well annotated by WormBase curators and the *C. elegans* community during the last dozen years since it was published (*C. elegans* Sequencing Consortium 1998), the genomes of *C. briggsae* has not been revised since its publication in 2003 (Stein *et al.* 2003), and the genome annotation of *C. remanei* and *C. brenneri* have not been published. In order to apply comparative genomics effectively to *Caenorhabditis* species for studying regulatory motif conservation/divergence, I will need to revise gene annotations in *C. briggsae*, *C. remanei*, and *C. brenneri* so they are of similar quality. To this end, I have co-developed a suite of programs, genBlastA and genBlastG, to predict the orthologous genomic region and the gene model that shows the highest similarity to the *C. elegans* ortholog. In Chapter 3 and 4, I will briefly describe the algorithm of genBlastA and genBlastG, but mainly focus on the performance and their application to annotating *Caenorhabditis* species.

1.11 Thesis organization

My thesis is organized as follows. First in Chapter 2, I will propose a model for the evolution of RFX transcription factors and ciliary genes based on computational homology searches in over a hundred species. These results suggest a possible acquisition of regulatory interaction between RFX genes and ciliary genes in evolution. Then in Chapter 3 and 4, I will describe the homology-based gene finder, genBlastA and genBlastG, and their performance in predicting orthologous gene structures in comparison to other popular homology based gene finders. I will use *C. briggsae* genome to show case how genBlastG is applied to gene annotation and show examples of gene models revised by genBlastG. Chapter 5 will describe the computational discovery of divergent X-box motifs using revised gene sets generated by genBlastG and comparative genomics over four *Caenorhabditis* species. I will also experimentally characterize the function for some of the divergent X-box motifs in RFX mediated transcription regulation. In the final chapter, I will provide a general discussion highlighting key findings and their implications.

2: CONVERGENT EVOLUTION OF RFX TRANSCRIPTION FACTORS AND CILIARY GENES PREDATED THE ORIGIN OF METAZOANS

Note regarding contributions

This chapter has been published in BMC Evolutionary Biology. The full citation is shown below:

Chu JSC, Baillie DL, and Chen N (2010). Convergent evolution of RFX transcription factors and ciliary genes predated the origin of metazoans. BMC Evolutionary Biology 10:130.

As the first author, I acquired all genomic and protein sequence data from public databases and performed all the bioinformatics analyses including running and parsing BLAST results, domain identification, multiple sequence alignment and phylogenetic analysis. D.L. Baillie and N. Chen conceived the study. N. Chen and I wrote the manuscript.

2.1 Abstract

Background: Intraflagellar transport (IFT) genes, which are critical for the development and function of cilia and flagella in metazoans, are tightly regulated by the Regulatory factor X (RFX) transcription factors (TFs). However, how and when their evolutionary relationship was established remained unknown.

Results: We have evidence suggesting that RFX TFs and IFT genes evolved independently but converged before the first appearance of metazoans. Both ciliary genes and RFX TFs exist in all metazoans and some unicellular eukaryotes. However, while RFX TFs and IFT genes are found in all sequenced metazoan genomes, RFX TFs do not co-exist with IFT genes in most pre-metazoans. For example, neither the budding yeast nor the fission yeast possess cilia although both have well-defined RFX TFs. Conversely, most unicellular eukaryotes, including the green alga *Chlamydomonas reinhardtii*, have typical cilia and well conserved IFT genes but lack RFX TFs. Outside of metazoans, RFX TFs and IFT genes co-exist only in choanoflagellates including *M. brevicollis*, and only one fungus *Allomyces macrogynus* of the 51 sequenced fungus genomes. *M. brevicollis* has two putative RFX genes and a full complement of ciliary genes.

Conclusions: The evolution of RFX TFs and IFT genes were independent in pre-metazoans. We propose that their convergence in evolution, or the acquired transcriptional regulation of IFT genes by RFX TFs, played a pivotal role in the establishment of metazoan.

2.2 Introduction

All metazoans and many unicellular eukaryotes have functional cilia (also known as flagella) (Satir *et al.* 2008). Both motile and immotile cilia (also known as sensory or primary cilia) hold many receptors for sensing environmental signals. Cilia may offer competitive advantages to ciliated organisms by allowing them to avoid predation and also to track nutritionally rich resources (Mitchell 2007). It is thus not surprising that cilia and most ciliary genes are deeply conserved, both in structure and in function. Such high levels of conservation suggest a common evolutionary origin (Satir *et al.* 2008). Ciliary defects have been associated with defective development in the nematode *Caenorhabditis elegans* (Swoboda *et al.* 2000) as well as a growing list of devastating human genetic disease conditions collectively called ciliopathies, including polycystic kidney disease (PKD), Bardet-Biedl syndrome (BBS), Alstrome syndrome, Joubert syndrome, Meckel-Gruber syndrome, and primary ciliary dyskinesia (Badano *et al.* 2006; Chen *et al.* 2006). In mammals, cilia are found on essentially all cell types, highlighting the critical role cilia play (Wheatley *et al.* 1996). One essential cellular process in cilia is the intraflagellar transport (IFT) that is responsible for the assembly and maintenance of eukaryotic cilia. The IFT machinery consists of four basic molecular modules: (a) motors, (b) Complex A, (c) Complex B, and (d) BBS complex (Ou *et al.* 2007; Pedersen and Rosenbaum 2008).

How IFT genes are regulated at the transcriptional level remained largely unknown until Swoboda and colleagues discovered in *C. elegans* that many IFT genes are regulated by DAF-19, a RFX type transcription factor (Swoboda *et al.* 2000). Mutations in *daf-19* resulted in defects in cilia development and constitutive dauer formation (Swoboda *et al.* 2000). DAF-19 binds to X-box motif, which is a highly conserved *cis*-regulatory element first discovered in mammals (Dorn *et al.* 1987; Swoboda *et al.* 2000). Ciliary genes in *C. elegans* often contain one or more putative X-box motifs 100 bp – 250 bp upstream of the coding sequences (Swoboda *et al.* 2000; Blacque *et al.* 2005; Efimenko *et al.* 2005; Chen *et al.* 2006). In addition, ciliary genes and cilia development in the fruit fly *Drosophila melanogaster* were also suggested to be regulated by RFX TFs (Laurencon *et al.* 2007). Two RFX genes dRFX (Durand *et al.* 2000) and dRFX2 (Otsuki *et al.* 2004) have been identified in *D. melanogaster*. dRFX was identified through a homology search for the RFX DNA binding domain (DBD) and dRFX2 was identified through yeast-one-hybrid (Y1H) screening for transcription factors that bind to a putative promoter sequence (Durand *et al.* 2000; Otsuki *et al.* 2004). Notably, dRFX2 has not been found in the *D. melanogaster* genome sequences, suggesting that it is likely located within the heterochromatin regions (William Gelbart, *personal communication*).

RFX TFs were first identified in mammals as binding proteins of the X-box motif (Reith *et al.* 1988). Through bioinformatics searches and molecular characterization, seven RFX genes—RFX1-7 have been found in mammals (Emery *et al.* 1996; Aftab *et al.* 2008). Different mammalian RFX genes show differential but overlapping expression patterns (Aftab *et al.* 2008), suggesting that they have complementary and cooperative roles in regulating genes in many different biological pathways. Indeed, mammalian RFX TFs have been shown to interact with each other and with many additional co-factors (Aftab *et al.* 2008). Accumulating evidence confirms that RFX genes regulate development and function of cilia in mammals as well. For instance, RFX3 knockout in mice led to abnormal cilia development in both brain (Bonnafe *et al.* 2004) and pancreas (Ait-Lounis *et al.* 2007).

Outside of metazoans, however, there is no evidence suggesting that IFT genes are regulated by RFX TFs. No RFX TFs have been reported in the green alga *Chlamydomonas reinhardtii*, a popular model organism for studying cilia biology. Conversely, RFX TFs exist in organisms including the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe* that do not have cilia (Emery *et al.* 1996), suggesting that RFX TFs do not regulate ciliary genes in these organisms. Based on these observations, we hypothesize that IFT genes and RFX TFs evolved independently and that their evolution converged at some point. To test this hypothesis, we have identified and examined IFT genes and RFX TFs in hundreds of fully sequence genomes that have become available recently.

2.3 Methods and Materials

2.3.1 Data sources

All sequence data (both genomic DNA sequences and gene annotation data including cDNA and protein sequences) were downloaded from public databases. The list of genomes and the data source are described in Appendix B. Briefly, the number species examined include 32 mammals, 6 fishes, 21 arthropods, 4 other vertebrates, 12 nematodes and other invertebrates, 51 fungi, 20 protists, and 8 plants and algae. The initial set of DNA binding domains that were used as queries for BLAST searches were taken from Human RFX1-7 (Aftab *et al.* 2008), *C. elegans* DAF-19 (Swoboda *et al.* 2000), *D. melanogaster* dRFX (Durand *et al.* 2000), and yeast RFX1 (Huang *et al.* 1998).

2.3.2 Identification of RFX TFs

We carried out similarity search using WU-BLAST (version 2.2.6; http://blast.wustl.edu) with e-value 0.01 and sequence filter (option –F) turned off. An initial set of DBDs was used as query to search against all the mammalian proteomes (entire collection of protein peptides). The resulting DBDs were added to the query list and then used to search against arthropods. The iteration of adding DBD and blasting continues until all species have been searched. A hit is accepted as a candidate DBD if the corrected percent identity over the entire domain length is >= 40%. The corrected percent identity was calculated as the number of identical positions divided by total length of the query. We also searched for candidate RFX TFs in genomic sequences (DNA sequences) in case that RFX TFs have been missed in the gene annotations.

2.3.3 Identification of ciliary genes

We carried out similarity searches using WU-BLAST (version 2.2.6; http://blast.wustl.edu) with e-value 0.01 and without sequence filter (without –F). Human protein sequences were taken from NCBI and used as queries (Table 2-1). PID was calculated as the number of identical amino acids reported by WU-BLAST over the entire length of the query.

Table 2-1 Human IFT genes and their associated Accession ID

Gene Name	NCBI Accession
IFT88	NP_783195.2
BBS5	NP_689597.1
IFT80	NP_065851.1
IFT52	NP_057088.2
IFT172	NP_056477.1
CLUAP1	NP_055856.1
TTC8	NP_653197.2
DYNC2H1	NP_001073932.1
IFT122	NP_443711.1
IFT57	NP_060480.1
KIF3A	NP_008985.3
WDR35	NP_001006658.1
WDR19	NP_079408.3
BBS2	NP_114091.3
IFT20	NP_777547.1
KIF3B	NP_004789.1
IFT81	NP_054774.2
ARL6	NP_115522.1
IFT74	NP_079379.2
IFT140	NP_055529.2
BBS1	NP_078925.3
BBS7	NP_789794.1
KIFAP3	NP_055785.2
KIF17	NP_065867.2

2.3.4 Phylogenetic analysis

Phylogenetic analysis was done using MEGA4 (Kumar *et al.* 2008). Multiple sequence alignment was done using CLUSTALW (included in MEGA4) with default settings. Phylogenetic trees were inferred using the Neighbor-Joining method.

2.3.5 Functional domain identification and analysis

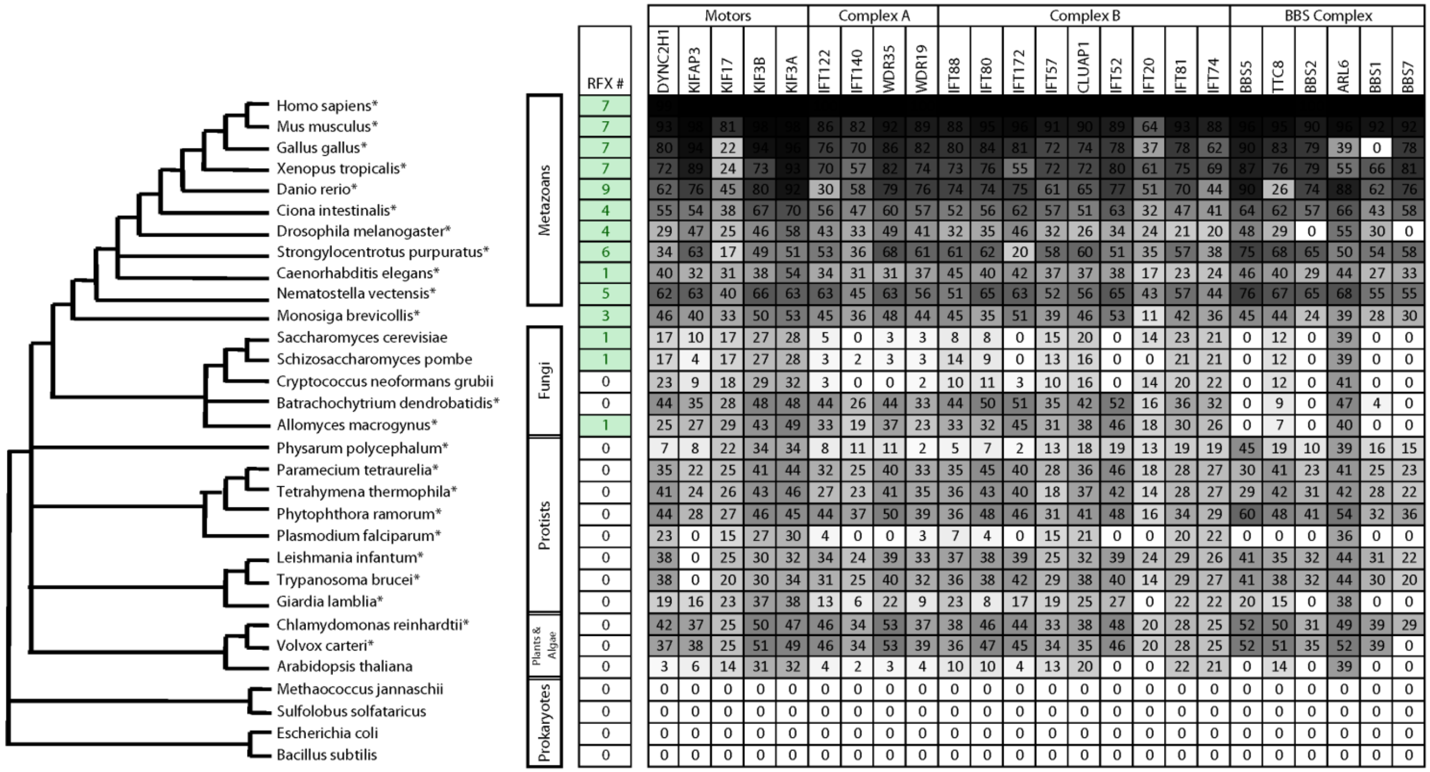
Sequences for activation, B, C, and D domains were taken from previous publications (Emery *et al.* 1996; Aftab *et al.* 2008). The multiple sequence alignment was performed for each domain and used as input for hmmbuild to generate a HMM profile for each domain. hmmsearch was used to scan the proteome of selected species to find regions of similar profile. Both hmmbuild and hmmsearch are programs part of the HMMER suite (Durbin *et al.* 1998) (http://hmmer.janelia.org).

2.4 Results

2.4.1 Molecular evolution of ciliary genes

Cilia have been observed to exist in many organisms including mammals, fruit flies, and *C. elegans*. Here, we examine the conservation of cilia by examining the ciliary components identified through searches for human orthologs. In total, we have examined the sequenced genomes of 153 species ranging from metazoans to fungi and plants. The ciliary components examined here include: (1) Five genes from the Motor module (DYNC2H1, K1FAP3, KIF17, KIF3B, and KIF3A); (2) Four from the Complex A module (IFT122, IFT140, WDR35, and WDR19); (3) Nine from the Complex B module (IFT88, IFT80, IFT172, IFT57, CLUAP1, IFT52, IFT20, IFT81, and IFT74); (4) Six from the BBS complex (BBS5, TTC8, BBS2, ARL6, BBS1, and BBS7) (Ou *et al.* 2007; Pedersen and Rosenbaum 2008) (Figure 2-1). We present results from 31 representative species in Figure 2-1. Most of the ciliary genes examined are strongly conserved in all metazoans ranging from the sea anemone (*Nematostella vectensis*) to human (*Homo sapiens*) (Figure 2-1). The unicellular choanoflagellate *Monosiga brevicollis*, which have been regarded as the closest extant relative of the last unicellular ancestor of metazoans (King *et al.* 2008), also have well conserved ciliary genes. Many ciliated protists, including *Paramecium tetraurelia*, *Tetrahymena thermophila*, and *Phytophthora ramorum* have most of the ciliary genes, consistent with previous reports (Wickstead and Gull 2007). Also in agreement with previous reports (Pan 2008; Pedersen *et al.*

2008), we have identified conserved ciliary genes in the unicellular algae *Chlamydomonas reinhardtii* and its closely related multicellular organism *Volvox carteri*. Protists *Giardia lamblia* and *Physarum polycephalum* have ciliary features that are similar to cilia development in mammals (Aldrich 1968; Wakasugi and Ohta 1973; Wright *et al.* 1979; Dawson *et al.* 2007). However, we observe reduced similarity for all ciliary genes in these two species, suggesting that these ciliary genes in protozoa are fast evolving (Ginger *et al.* 2008). The apicomplexan parasite *Plasmodium falciparum* lacks many ciliary genes, consistent to the idea that the apicomplexan parasites may have an entirely different ciliary assembly mechanism (Briggs *et al.* 2004). Among the 51 sequenced fungi identified to date, we found only two species, *Allomyces macrogynus* and *Batrachochytrium dendrobatidis*, have conserved IFT genes (Figure 2-1). Interestingly, both species lack most of the BBS complex components. These observations are consistent with previous proposal that cilia were lost independently in many fungal species in evolution (Cracraft and Donoghue 2004; James *et al.* 2006). Taken together, our comparative identification and analysis of IFT genes suggest that IFT genes are deeply conserved and can be found in all metazoans, most unicellular eukaryotes, and some fungi, but they do not exist in plants such as *Arabidopsis thaliana* and prokaryotes (Wickstead and Gull 2007; Satir *et al.* 2008) (Figure 2-1).

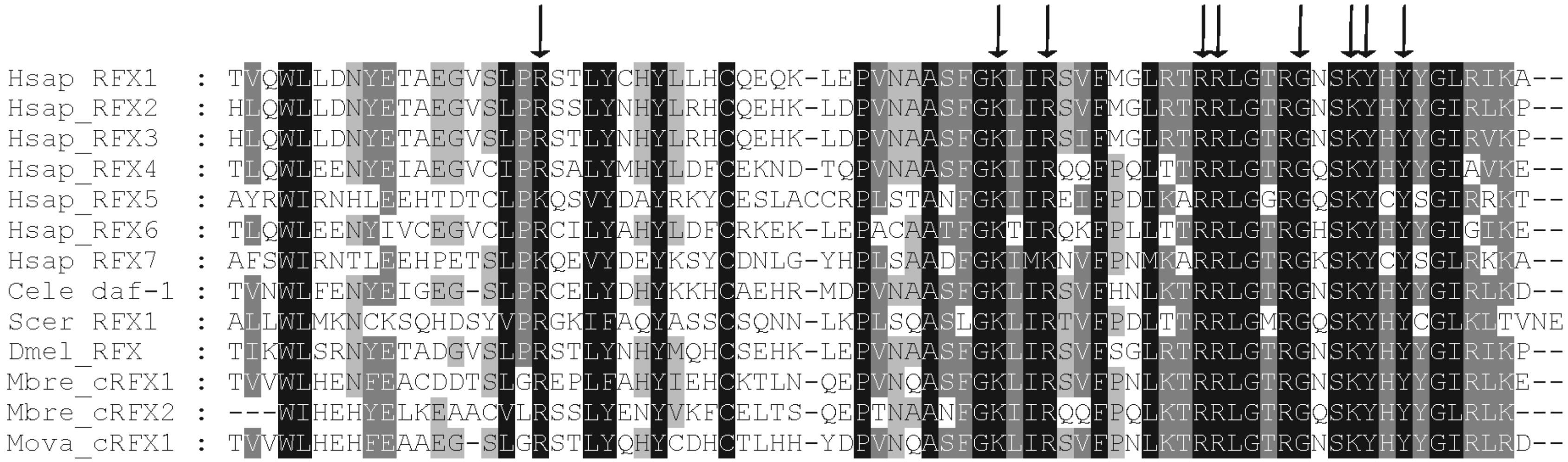


			Motors					Complex A				Complex B									BBS Complex					
		RFX #	DYNC2H1	KIFAP3	KIF17	KIF3B	KIF3A	IFT122	IFT140	WDR35	WDR19	IFT88	IFT80	IFT172	IFT57	CLUAP1	IFT52	IFT20	IFT81	IFT74	BBS5	TTC8	BBS2	ARL6	BBS1	BBS7
Homo sapiens*	Metazoans	7	99	[illegible]	[illegible]	[illegible]	[illegible]	100	[illegible]	[illegible]	100	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	100	[illegible]	[illegible]	[illegible]
Mus musculus*		7	93	98	81	98	98	86	82	92	89	88	95	96	91	90	89	64	93	88	96	95	90	96	92	92
Gallus gallus*		7	80	94	22	94	96	76	70	86	82	80	84	81	72	74	78	37	78	62	90	83	79	39	0	78
Xenopus tropicalis*		7	72	89	24	73	93	70	57	82	74	73	76	55	72	72	80	61	75	69	87	76	79	55	66	81
Danio rerio*		9	62	76	45	80	92	30	58	79	76	74	74	75	61	65	77	51	70	44	90	26	74	88	62	76
Ciona intestinalis*		4	55	54	38	67	70	56	47	60	57	52	56	62	57	51	63	32	47	41	64	62	57	66	43	58
Drosophila melanogaster*		4	29	47	25	46	58	43	33	49	41	32	35	46	32	26	34	24	21	20	48	29	0	55	30	0
Strongylocentrotus purpuratus*		6	34	63	17	49	51	53	36	68	61	61	62	20	58	60	51	35	57	38	75	68	65	50	54	58
Caenorhabditis elegans*		1	40	32	31	38	54	34	31	31	37	45	40	42	37	37	38	17	23	24	46	40	29	44	27	33
Nematostella vectensis*		5	62	63	40	66	63	63	45	63	56	51	65	63	52	56	65	43	57	44	76	67	65	68	55	55
Monosiga brevicollis*		3	46	40	33	50	53	45	36	48	44	45	35	51	39	46	53	11	42	36	45	44	24	39	28	30
Saccharomyces cerevisiae	Fungi	1	17	10	17	27	28	5	0	3	3	8	8	0	15	20	0	14	23	21	0	12	0	39	0	0
Schizosaccharomyces pombe		1	17	4	17	27	28	3	2	3	3	14	9	0	13	16	0	0	21	21	0	12	0	39	0	0
Cryptococcus neoformans grubii		0	23	9	18	29	32	3	0	0	2	10	11	3	10	16	0	14	20	22	0	12	0	41	0	0
Batrachochytrium dendrobatidis*		0	44	35	28	48	48	44	26	44	33	44	50	51	35	42	52	16	36	32	0	9	0	47	4	0
Allomyces macrogynus*		1	25	27	29	43	49	33	19	37	23	33	32	45	31	38	46	18	30	26	0	7	0	40	0	0
Physarum polycephalum*	Protists	0	7	8	22	34	34	8	11	11	2	5	7	2	13	18	19	13	19	19	45	19	10	39	16	15
Paramecium tetraurelia*		0	35	22	25	41	44	32	25	40	33	35	45	40	28	36	46	18	28	27	30	41	23	41	25	23
Tetrahymena thermophila*		0	41	24	26	43	46	27	23	41	35	36	43	40	18	37	42	14	28	27	29	42	31	42	28	22
Phytophthora ramorum*		0	44	28	27	46	45	44	37	50	39	36	48	46	31	41	48	16	34	29	60	48	41	54	32	36
Plasmodium falciparum*		0	23	0	15	27	30	4	0	0	3	7	4	0	15	21	0	0	20	22	0	0	0	36	0	0
Leishmania infantum*		0	38	0	25	30	32	34	24	39	33	37	38	39	25	32	39	24	29	26	41	35	32	44	31	22
Trypanosoma brucei*		0	38	0	20	30	34	31	25	40	32	36	38	42	29	38	40	14	29	27	41	38	32	44	30	20
Giardia lamblia*		0	19	16	23	37	38	13	6	22	9	23	8	17	19	25	27	0	22	22	20	15	0	38	0	0
Chlamydomonas reinhardtii*	Plants & Algae	0	42	37	25	50	47	46	34	53	37	38	46	44	33	38	48	20	28	25	52	50	31	49	39	29
Volvox carteri*		0	37	38	25	51	49	46	34	53	39	36	47	45	34	35	46	20	28	25	52	51	35	52	39	0
Arabidopsis thaliana		0	3	6	14	31	32	4	2	3	4	10	10	4	13	20	0	0	22	21	0	14	0	39	0	0
Methaococcus jannaschii	Prokaryotes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sulfolobus solfataricus		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Escherichia coli		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bacillus subtilis		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2-1 The conservation of RFX TFs and ciliary IFT components in selected species. These species were selected to provide a wide sampling of the "tree of life". The phylogenetic relationship between each species was derived from the "Tree of Life Web Project" (Maddison and Schulz 2007). Species indicated with "*" have ciliated cells based on published evidence. The 'RFX #' column shows the number of putative RFX TFs identified in this project or reported previously. The grey scale table shows the sequence conservation of individual ciliary components in each species. Darker shade represents higher sequence similarity and conservation. The numbers in each box indicate the percent identity revealed by the alignments between IFT genes and their corresponding human orthologs. The data used to generate this figure was collected by J. Chu.

2.4.2 Molecular evolution of RFX TFs

Using well defined RFX DBD peptide sequences (76 amino acids long) (Figure 2-2) from human (Aftab *et al.* 2008), *C. elegans* (Swoboda *et al.* 2000), *D. Melanogaster* (Durand *et al.* 2000), and *S. Cerevisae* (Huang *et al.* 1998) as queries, we searched the genomes of the same 153 species for RFX TFs. Because the known RFX DBDs in yeast as well as humans show very high similarity, we used very stringent criteria to look for new RFX TFs. We only consider proteins whose putative RFX DBD show at least 40% percentage identity (PID) to the queries (see Methods). RFX DBD has been shown to contain nine residues that have direct contact with DNA sequences (X-box motifs) (Gajiwala *et al.* 2000). All nine residues are highly conserved in all known RFX DBDs (Figure 2-2).



```
Hsap RFX1   : TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQK-LEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA--
Hsap_RFX2   : HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHK-LDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP--
Hsap_RFX3   : HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHK-LDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP--
Hsap_RFX4   : TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKND-TQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE--
Hsap_RFX5   : AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT--
Hsap_RFX6   : TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEK-LEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE--
Hsap RFX7   : AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLG-YHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA--
Cele daf-1  : TVNWLFENYEIGEG-SLPRCELYDHYKKHCAEHR-MDPVNAASFGKLIRSVFHNLKTRRLGTRGNSKYHYYGIRLKD--
Scer RFX1   : ALLWLMKNCKSQHDSYVPRGKIFAQYASSCSQNN-LKPLSQASLGKLIRTVFPDLTTRRLGMRGQSKYHYCGLKLTVNE
Dmel_RFX    : TIKWLSRNYETADGVSLPRSTLYNHYMQHCSEHK-LEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP--
Mbre_cRFX1  : TVVWLHENFEACDDTSLGREPLFAHYIEHCKTLN-QEPVNQASFGKLIRSVFPNLKTRRLGTRGNSKYHYYGIRLKE--
Mbre_cRFX2  : ---WIHEHYELKEAACVLRSSLYENYVKFCELTS-QEPTNAANFGKIIRQQFPQLKTRRLGTRGQSKYHYYGLRLK---
Mova_cRFX1  : TVVWLHEHFEAAEG-SLGRSTLYQHYCDHCTLHH-YDPVNQASFGKLIRSVFPNLKTRRLGTRGNSKYHYYGIRLRD--
```

Figure 2-2 DBDs of RFX TFs are highly conserved. Representative DBDs from humans (hRFX1-7), *C. elegans* (DAF-19), *D. melanogaster* (dRFX), *S. serevisiae* (sRFX1), and *M. brevicollis* (Mbre_cRFX1 and Mbre_cRFX2). DBDs from different species show high similarity at the peptide level. Nine residues of DBD that directly contact DNA (indicated by arrows) are essentially identical for all RFX TFs. The data used to generate this figure was collected by J. Chu. DNA contact sites is taken from (Gajiwala *et al.* 2000).

We found candidate RFX TFs in all sequenced metazoan genomes (Figure 2-1). In addition to the RFX TFs that have been reported previously, including seven RFX TFs found in mammals (Aftab *et al.* 2008), DAF-19 in *C. elegans* (Swoboda *et al.* 2000), and dRFX (Durand *et al.* 2000), we found many RFX genes that have not been described previously. We have identified seven RFX genes (RFX1-7) in all vertebrate genomes except fish genomes, which have nine putative RFX genes (RFX1-9). We have also identified four RFX genes in *Ciona intesttinalis*, six in the purple sea urchin (*Strongylocentrotus purpuratus*), and five in the sea anemone (*Nematastella vectensis*). In *D. melanogaster*, in addition to the two RFX genes reported previously—dRFX and dRFX2, we have identified a novel RFX TF, which we named *dRFX1*. Interestingly, among all metazoans examined, nematodes including *C. elegans* are the only organisms that possess just one RFX gene.

RFX TFs are also found in some non-metazoans. Of the 51 fungal species examined, we identified single RFX TFs in 44 species, including the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe*, as previously reported (Emery *et al.* 1996), as well as a ciliated fungus *Allomyces macrogynus*, whose genome was recently sequenced by the Fungal Genome Initiative of the Broad Institute (http://www.broadinstitute.org/). All unicellular organisms we have examined possess either one RFX gene (fungi) or none, except for the choanoflagellates. For example, *M. brevicollis*, which was recently sequenced (King *et al.* 2008), contain two genes (Mbre_cRFX1 and Mbre_cRFX2) with well-defined RFX DBDs.

RFX DBD sequences are the defining features of all known RFXs and show high similarity (>40% PID) to each other. However, there are a small number of additional proteins that contain domains that show weaker similarity (<30% PID) to known RFX DBDs. In particular, a gene (ARID2) in the human genome contains a RFX-like domain that shows 29% PID to the human RFX1 DBD. Among the nine residues that have direct contact with DNA sequences, five can be found in the RFX-like domain found in ARID2. ARID2, whose function as a transcription factor has not been well studied, has orthologs in all mammals as well as other vertebrates (data not shown). Additionally, a gene in *M. brevicollis* also shows weak similarity (27%) to known RFX DBDs (five of the nine residues that have direct contact with DNA are conserved). We name this novel gene Mbre_cRFX3. Because of their low similarity to known RFX DBDs, these RFX like genes—ARID2 and Mbre_cRFX3—are not regarded as RFX TFs in this project and thus are not examined further. No RFX genes have been found in any bacteria, ancient bacteria, or plants (Figure 2-1).

DBDs in the two putative RFX TFs in *M. brevicollis* are essentially indistinguishable from the DBDs in previously characterized RFX TFs with ~70% PID at the peptide level. All nine residues that make direct contacts with X-box motifs are conserved (Gajiwala *et al.* 2000) (Figure 2-2, residues indicated with arrows). In addition to the DBDs, Mbre_cRFX1 also shares other functional domains within known RFX TFs including the dimerization domains (DD), and the extended dimerization domains (B and C domains), which exist in all mammalian RFX TFs except RFX5 and RFX7 (Emery *et al.* 1996; Aftab *et al.* 2008) (Figure

2-3). Aligning Mbre_cRFX1 to human RFX1-3 shows clear alignment for conserved DBD, DD, and extended dimerization domains (B and C domains) (Figure 2-4). None of the *M. brevicollis* RFX TFs have readily identifiable activation domains (AD). The lack of typical AD in RFX TFs in *M. brevicollis*, *C. elegans*, *D. melanogaster*, and sea anemone (Figure 2-3) suggests that AD might have been acquired later in metazoan evolution. Alternative hypothesis is that their ADs have yet to be identified and characterized. Mbre_cRFX2 has a readily identifiable DBD but lacks other conserved domains, which is similar to the human RFX5 and RFX7 that lack other domains (Figure 2-3). The presence of DBD (in both Mbre_cRFX1 and Mbre_cRFX2) and other conserved protein domains (in Mbre_cRFX1) suggest that they may function in transcriptional regulation of gene expression in *M. brevicollis*. However, their target genes remain to be identified.

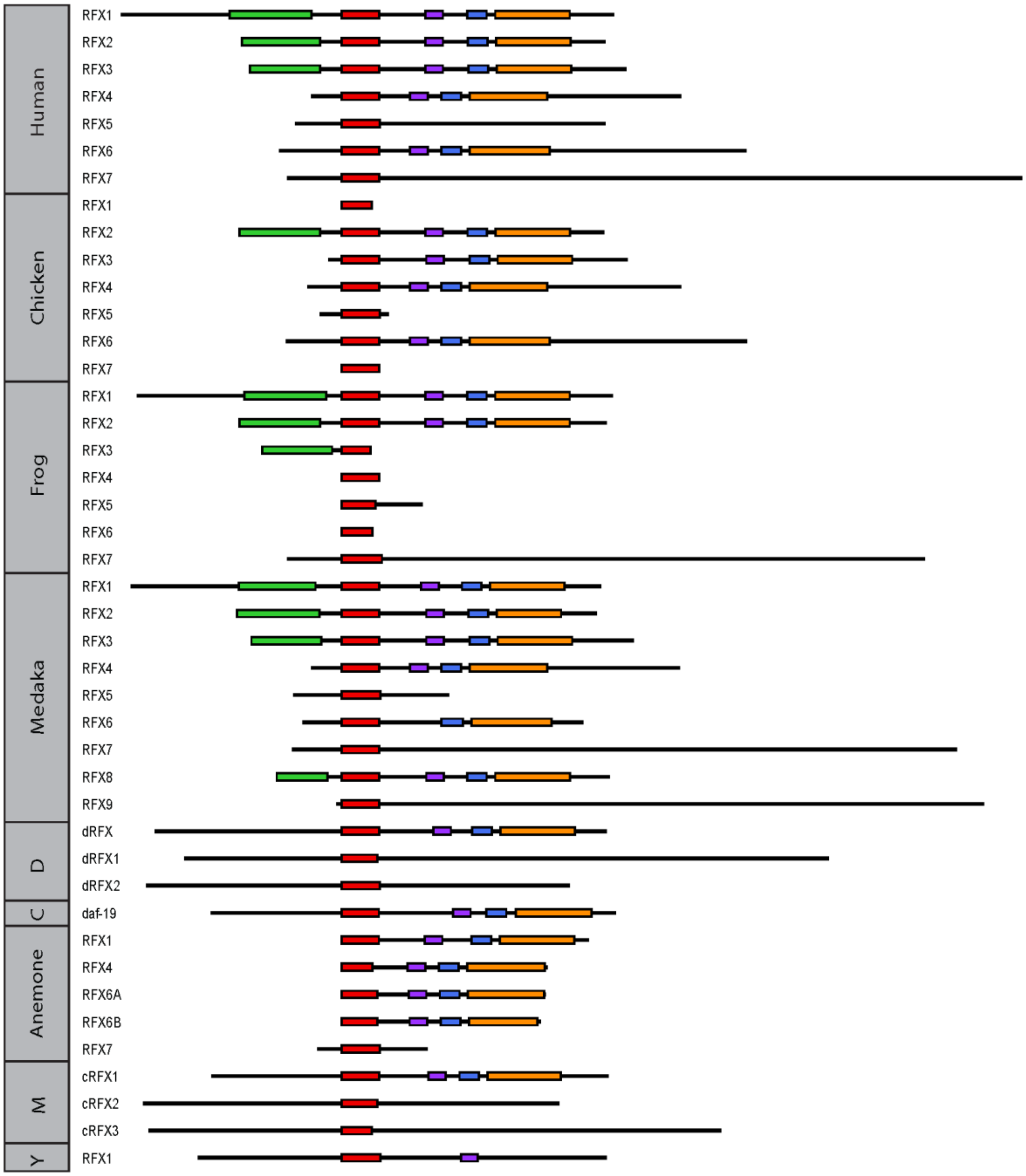


Figure 2-3 Predicted protein domains of RFX TFs in representative species. The defining domain of all RFX TFs—DBD—is shown in red. Other domains including the activation domain (green), the B domain (purple), C domain (blue), and D domain (orange) are not present in all RFX TFs. In the left column, Y stands for the budding yeast *S. cerevisiae*, M for *Monosiga brevicollis*, C for *C. elegans*, and D for *D. melanogaster*. The data used to generate this figure was collected by J. Chu.

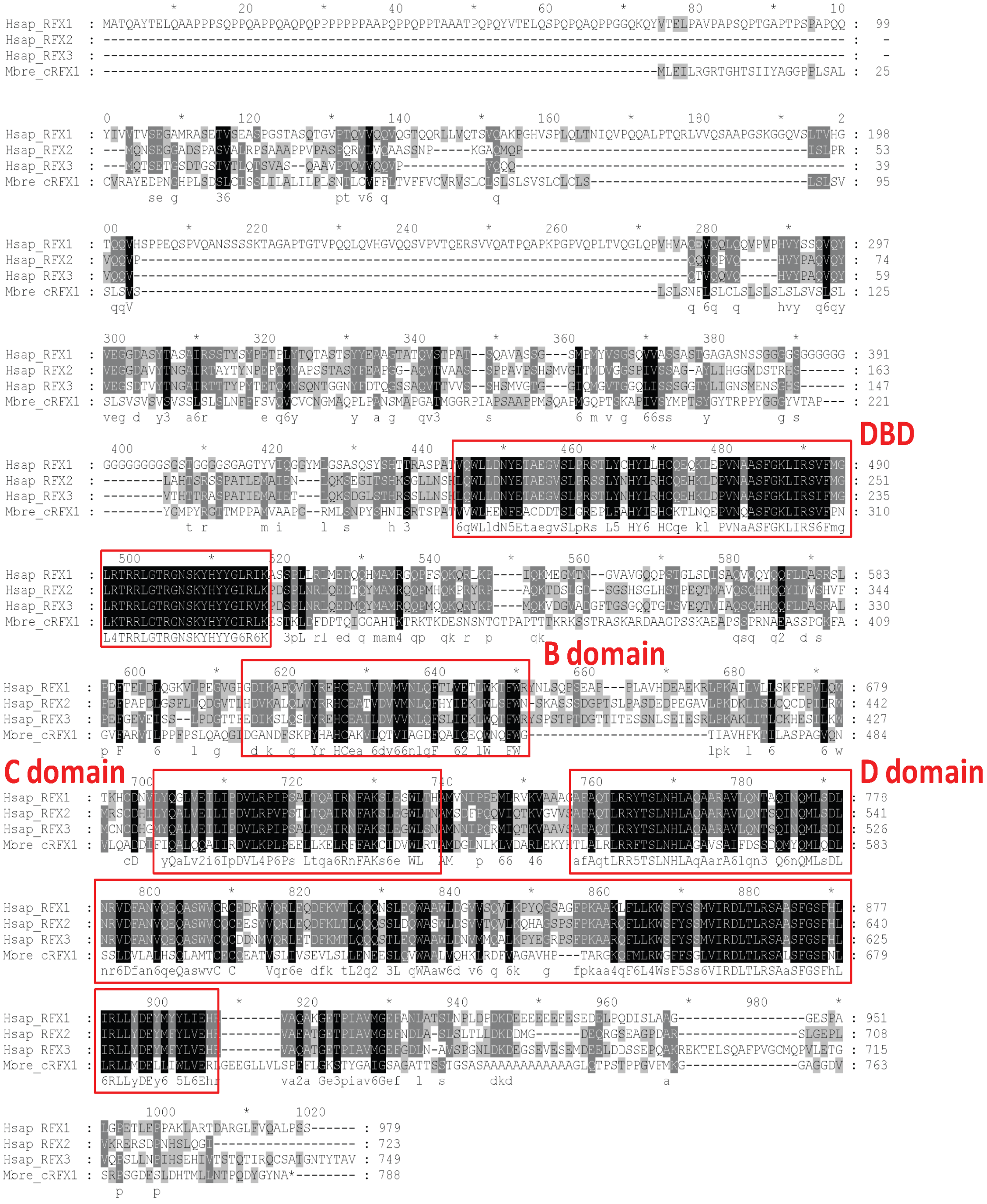


Figure 2-4 Sequence alignment between *M. brevicollis* Mbre_cRFX1 and Human RFX1-3 with all functional domains highlighted. Amino acid residues are color coded with darker color representing higher conservation. Putative functional domains are boxed and labeled. The data used to generate this figure was collected by J. Chu.

To further examine the relationship between the *M. brevicollis* RFX TFs and those identified in mammals and other species, we constructed a phylogenetic tree that contains all known and putative RFX TFs based on the similarity between the DBD domains (Figure 2-5). Sequences outside of the DBDs are excluded from analysis since they are often very diverse and are not readily alignable. Previous analysis of mammalian RFX TFs revealed three groups: RFX1-3 (bootstrap value = 49), RFX4-RFX6 (bootstrap value = 38), and RFX5-RFX7 groups (bootstrap value = 98) (Aftab *et al.* 2008), which is generally consistent with this phylogenetic tree with newly identified members (Figure 2-5). The phylogenetic tree contains an additional clade (shown in black), which contains RFX TFs identified in fungal genomes and, interestingly, *dRFX2* in *D. melanogaster* (Otsuki *et al.* 2004) (bootstrap value = 51). Fungal RFX TFs (members in the Fungi clade) and RFX5-RFX7 TFs show similar domain compositions with all members lacking B, C, and D domains, which are found in the RFX1-3 and RFX4-RFX6 TFs (Figure 2-3). The inferred phylogenetic tree clearly shows that the Mbre_cRFX1 fits into the RFX1-3 group (bootstrap value = 49), while Mbre_cRFX2 fits into the RFX4-6 group (bootstrap value = 14). Mbre_cRFX3, which show weaker similarity to known DBDs, clusters closer to DBDs of the RFX5-7 groups. However, as mentioned before, we did not include Mbre_cRFX3 in the phylogenetic tree. The phylogenetic relationship between *M. brevicollis* and previously identified RFX TFs suggest that these three RFX TFs families were established before the split between choanoflagellates and metazoans.

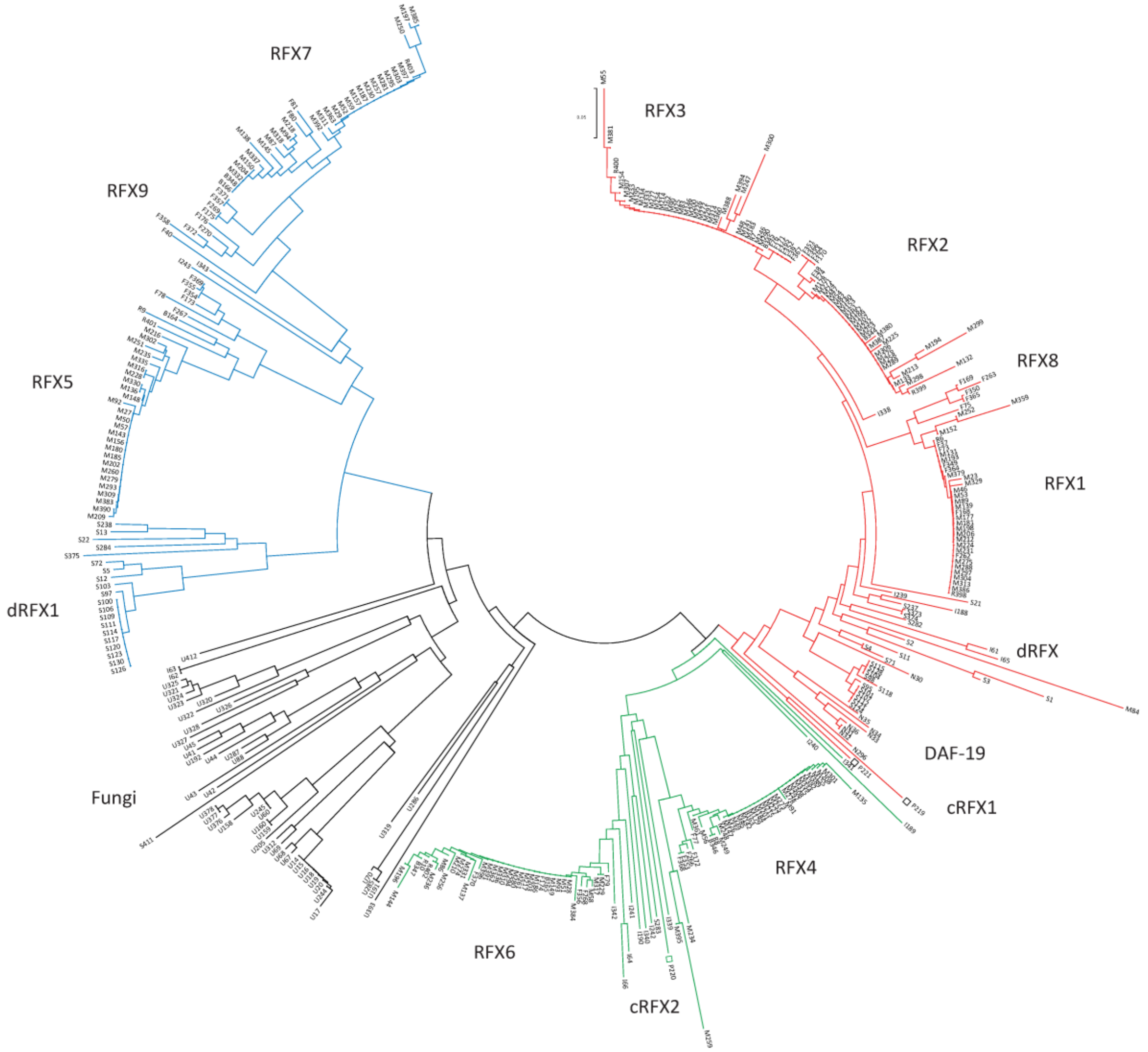


Figure 2-5 The phylogenetic tree of all RFX DBDs found in this study. Each distinct group of RFX is labeled. The label for each putative RFX corresponds to records in Additional file 1. The colored branches indicate three major groups of RFX TFs: RFX1-3 in red, RFX4-6 in green, and RFX5-7 in blue. The fish RFX8 TFs cluster with the RFX1-3 group, while the fish RFX9 TFs cluster with the RFX5-7 group. All nematodes are grouped together (labeled DAF-19) with the RFX1-3 group. Some insects RFX TFs (labeled as dRFX) group with RFX1-3, while others (labeled dRFX1) with RFX5-7. *M. brevicollis* (cRFX1 and cRFX2) are shown in open squares (□). *Drosophila* dRFX2 (Otsuki *et al.* 2004) is shown in the tree but it is not found in the sequenced *D. melanogaster* genome nor any sequenced *Drosophila* genome. It is likely located in the heterochromatic region (William Gelbart, *personal communication*). The phylogenetic tree was inferred using the Neighbor-Joining method (Saitou and Nei 1987). Phylogenetic analysis was performed using MEGA4 (Tamura *et al.* 2007). The data used to generate this figure was collected by J. Chu.

In the inferred phylogenetic tree, the nematodes are the only metazoans that have only one RFX TF, which groups together with the mammalian RFX1-3 group (Figure 2-5). It was proposed previously that prior to the complete sequencing of the *C. elegans* genome, more RFX TFs should exist in *C. elegans* (Emery *et al.* 1996). However, exhaustive searches of the completed *C. elegans* genome revealed no traces of additional RFX genes, suggesting other RFX groups (RFX4-6 and RFX5-7) were lost in the last common ancestor of the nematode species. In fact, none of the seven sequenced nematode genomes has more than one RFX TF.

2.4.3 Evolutionary relationship between ciliary genes and RFX TFs

The above comprehensive identification of IFT genes and RFX TFs shows clearly that all metazoans have both ciliary genes and RFX genes. Since IFT genes have been demonstrated to be regulated by RFX TFs in *C. elegans*, *D. melanogaster*, and humans, IFT genes in all metazoans are likely regulated by RFX TFs. Our analysis strongly suggests that IFT genes and RFX TFs evolved independently. We have identified 44 out of 51 fungal species that have single RFX genes but no IFT genes. We believe RFX genes in these species do not regulate ciliary genes expression. Indeed, Crt1/RFX in the budding yeast plays a role in DNA damage response (Lubelsky *et al.* 2005). Outside of metazoans, only two sequenced genomes have IFT and RFX genes, the choanoflagellate *M. brevicollis* and the fungus *A. macrogynus*. Outside of metazoans, choanoflagellates, and fungi, none of the sequenced genomes possesses a single RFX gene, regardless of the possession of IFT genes.

2.5 Discussion

This is the first project to comprehensively identify and compare RFX TFs in the entire “tree of life” since Emery and colleagues described RFXs in humans (RFX1-5), mice (RFX1-3 and RFX5), *C. elegans*, and the budding and fission yeasts domains more than a decade ago (Emery *et al.* 1996). In this paper, we identified for the first time (1) nine RFX genes in all sequenced fish genomes; (2) two RFX genes in the choanoflagellate *M. brevicollis* genome; (3) single RFX genes in many fungal genomes. Additionally, we have identified RFX genes in many vertebrates. Furthermore, we have identified a third RFX (*dRFX1*) in the fruit fly *D. melanogaster*. Based on our phylogenetic analysis of all RFX TFs identified in the “tree of life”, we have confirmed the hypothesis proposed by Emery and colleagues that *C. elegans* has lost RFX genes as it evolved (Emery *et al.* 1996). On the other hand, we could not identify dRFX2 in *D. melanogaster* genome nor could we identify orthologs in any other species. It is possible dRFX2 reside in the heterochromatin region. Alternatively, our phylogenetic anlaysis where putative dRFX2 DBD branch together with fungi may suggest dRFX2 is a yeast-1 hybrid experimental artifact. Since the putative dRFX2 DBD is not found in the 51 fungal species, we can only hypostulate it is from a fungal species not yet sequenced.

Comparative analysis of the molecular evolution of IFT genes and RFX genes revealed a compelling converging relationship between these two gene groups, which is summarized in a model illustrated in Figure 2-6. We propose that the common ancestor of metazoans, choanoflagellates, and fungi was ciliated and had one RFX gene. Thorugh evolutionary time, some fungal species, including *Batrachochytrium dendrobatidis*, inherited cilia but lost RFX; some species, like budding yeast and fission yeast, lost their cilia but retained RFX; some species, like *Cryptococcus neoformans grubii*, lost both cilia and RFX TFs; only a few fungal species identified to date, including *Allomyces macrogynus*, retained both RFX TFs and cilia (Figure 2-1 and Figure 2-6). In ciliated fungi, which do not have RFX genes, ciliary genes are likely regulated by factors other than RFX TFs. In contrast, the common ancestor of metazoans and choanoflagellates was ciliated and had multiple RFX genes.

Some transcription factors exihibit similar evolutionary pattern as RFX. LSF/Grainyhead transcription factor family are conserved amongst metazoans, choanoflagellate, and fungi but not in algae and amoebozoan (Traylor-Knowles *et al.* 2010). However, a large number of transcription factors, including ETS family, POU family, PAX family, and LIM-HD family, only arose after the emergence of metazoans (Degnan *et al.* 2009). This suggest the convergence of RFX genes and ciliary genes before the estabilishment of choanoflagellates is likely meaningful. Nevertheless, we do not rule out the fact that RFX genes may not have any transcriptional regulation on ciliary genes in choanoflagellates, which would be interesting to investigate further.

The plurality of RFX genes was probably due to gene duplication event (Figure 2-6). The expansion of the RFX gene family, in the common ancestor of metazoans and choanoflagellates, might have provided a platform for the development of interactions between RFX TFs and IFT genes and the establishment of transcriptional regulatory relationships between RFX TFs and IFT genes in metazoans. The convergent molecular evolution of IFT genes and RFX TFs might have provided a pivotal driving force in the emergence and evolution of metazoans.

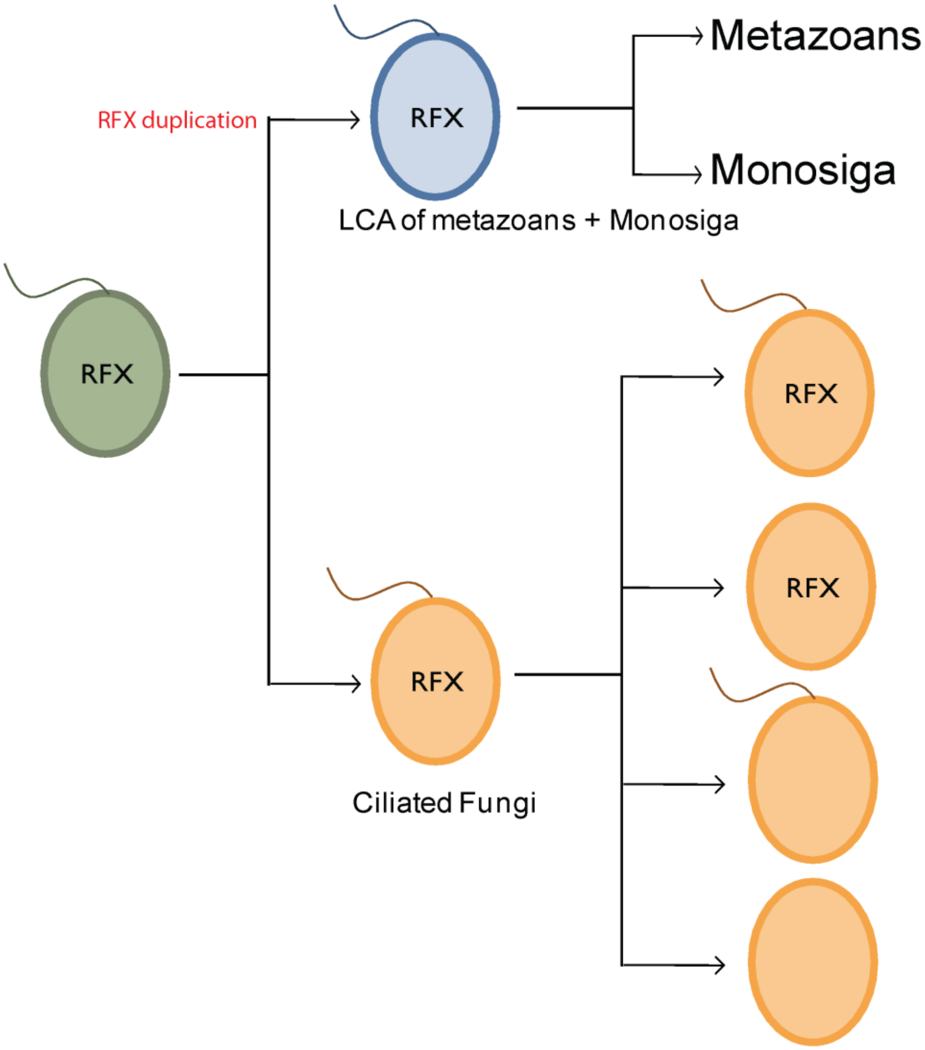


Figure 2-6 RFX TF-mediated transcription and the origin of metazoans. The common ancestor of metazoans, choanoflagellates, and fungi was likely a ciliated unicellular eukaryote with a single RFX TF. Over the course of evolution, some fungi lost RFX TFs while preserving cilia, some lost cilia but kept RFX, and some lost both. Only a few fungi identified to date kept both cilia and RFX. The last common ancestor (LCA) of *Monosiga* and metazoans preserved both cilia and RFX. This figure is illustrated by J. Chu

The evolution of multicellular metazoans from a unicellular protozoan ancestor represents a major and the most spectacular transition in the "history of life". This transition is demonstrated by the abrupt appearance of a huge variety of metazoans in the fossil record approximately 560 million years ago during the Cambrian explosion (Conway-Morris 2003). Many environmental, ecological, and other evolutionary factors have been proposed to contribute to this transition (King 2004; Ruiz-Trillo *et al.* 2007). Great efforts have been made to understand this transition by studying protein-coding regions of numerous genes and gene families that are ubiquitous in and limited to metazoans. Findings obtained in these studies showed that many genes and gene families previously found to be expressed only in metazoans are also found in choanoflagellates giving evidence that metazoans arose from choanoflagellates. For example, work by King and colleagues clearly demonstrated that choanoflagellates have a receptor tyrosine kinase that is found in metazoans but not in other eukaryotes (Conway-Morris 2003). Manning and colleagues searched the sequenced choanoflagellates *M. brevicollis* genome (King *et al.* 2008), and identified a highly elaborate tyrosine kinase signaling network (Manning *et al.* 2008). Many additional genes are shared by *M. brevicollis* and metazoans, including cadherin, which are essential for metazoan development (Abedin and King 2008), and transcription factors such as P53 and Myc (King *et al.* 2008). These findings encouraged additional large scale searches, including the UNICORN (unicellular opisthokont research initiative) project (Ruiz-Trillo *et al.* 2007), for genes and gene families critical for the transition from unicellularity to multicellularity. However, accumulating evidence is showing that these genes predated the origin of metazoans and played different roles from their counterparts in metazoans. Thus these genes, even though some have been co-opted to perform novel functions in metazoans, are probably not be the main driving force underlying the transition from unicellular protozoans to multicelluar metazoans.

What then was the main factor driving this transition? In contrast to coding sequences of genes, which are usually under strong purifying selection, regulatory sequences show much more rapid evolution. Compelling evidence suggests that changes in *cis*-regulatory sequences and transcriptional regulation in general play a pivotal role in evolution (King 2004; Wray 2007). Kingsley and colleagues recently identified changes in *cis*-regulatory modules that dictate dramatic changes in pigmentation in sticklebacks and humans (Miller *et al.* 2007). Thus, the transition from unicellular flagellates to multicellular metazoans may have been driven by innovations at the transcriptional level.

The evolution of ciliary genes and RFX transcription factors were independent prior to metazoans and converged in choanoflagellates. The convergent evolution of RFX TFs and ciliary genes (IFT genes in particular) in the common ancestor of metazoans and choanoflagellates prompt us to propose that the acquired tight control of ciliary genes at the transcription level by RFX TFs served as one of the critical driving forces in the establishment of multicellularity and the rise of metazoans.

3: GENBLASTA: ENABLING BLAST TO IDENTIFY HOMOLOGOUS GENE SEQUENCES

Note regarding contributions:

This chapter has been published in Genome Research. The full citation is shown below:

She R*, Chu JSC*, Wang K, Pei J, Chen N (2009) genBlastA: Enabling BLAST to identify homologous gene sequences. Genome Research 19:143-149.

(*) Equal contributions

As the co-first author, I contributed to both the developmental phase and the experimental phase. I was responsible for testing the program with different cases to ensure the correctness of the algorithm. N. Chen and I prepared all the test gene sets needed. I set up all the databases for visualization and performance testing. I collected and compared all the experimental results from genBlastA, WU-BLAST, and ML. N. Chen, J. Pei, and K. Wang conceived the study. N. Chen, K. Wang, and R. She designed the algorithm. R. She implemented the algorithm. R. She, K. Wang, N. Chen and I wrote the manuscript.

3.1 Abstract

BLAST is an extensively used local similarity search tool for identifying homologous sequences. When a gene sequence (either protein sequence or nucleotide sequence) is used as a query to search for homologous sequences in a genome, the search results, represented as a list of high-scoring segment pairs (HSPs), are fragments of candidate genes rather than full-length candidate genes. Relevant HSPs (“signals”), which represent candidate genes in the target genome sequences, are buried within a report that contains also hundreds to thousands of irrelevant HSPs (“noises”). Consequently, BLAST results are often overwhelming and confusing even to experienced users. For effective use of BLAST for gene finding, a program is needed for extracting relevant HSPs that represent candidate homologous genes from the entire HSP report. To achieve this goal, we have designed a graph-based algorithm, genBlastA, which automatically filters HSPs into well-defined groups, each representing a candidate gene in the target genome. The novelty of genBlastA is an edge length metric that reflects a set of biologically motivated requirements so that each shortest path corresponds to an HSP group representing a homologous gene. We demonstrate that this novel algorithm is both efficient and accurate for identifying homologous sequences, and that it outperforms existing approaches with similar functionalities.

3.2 Introduction

Genome sequencing projects, such as the human genome project (Lander *et al.* 2001; Venter *et al.* 2001), have produced an enormous amount of nucleotide sequence. With recent advances in sequencing technologies (Margulies *et al.* 2005; Bentley 2006), the volume of the nucleotide sequences is expanding at an exponential pace, further enriching genomic sequence resources. To exploit these resources effectively for biological and medical research, many homology based similarity search and alignment tools have been developed and optimized in the past 20 years. Representative similarity search and alignment tools include BLAST(Altschul *et al.* 1990), FASTA (Pearson and Lipman 1988), sim4 (Florea *et al.* 1998), WU-BLAST (Lopez *et al.* 2003), and BLAT (Kent 2002). Some homology based search and alignment tools, such as GeneWise (Birney *et al.* 2004), Exonerate (Slater and Birney 2005), and a recent program developed by Cui and colleagues (Cui *et al.* 2007), have also been developed. These tools have been extremely useful especially for comparative genomics, in which genomes of both closely and distantly related species are compared so that knowledge gained in the genome of one species can be used to understand the genome of other species (Hardison 2003).

In general, these search tools work by identifying a list of sequence segments that show similarity to a query sequence. For example, BLAST detects regions of similarity between a query sequence and target sequences in a database. As illustrated in Figure 3-1, each match between a query sequence fragment and a target sequence fragment is reported as a high-scoring segment pair (HSP), which consists of a pair of sequences: [Q,T], where Q is a segment from the query sequence (i.e. query segment) and T is a matching segment from a target sequence in the target database (i.e. target segment). When a BLAST search returns many HSPs for a query in a genome, it suggests the existence of one or more homologous genes in the database, with each HSP potentially corresponding to a coding exon of the gene. BLAST assigns each HSP a bit score, an expectation value (e-value), as well as a percentage of identity (PID) and similarity values. Among these HSPs, some may represent candidate bona fide genes and can provide biologists with a meaningful starting point for further research, while others may be irrelevant hits. Although BLAST and other similarity searching tools produce lists of HSPs, they do not reveal which HSPs represent candidate genes, let alone reveal how many homologous genes exist in the target genome.

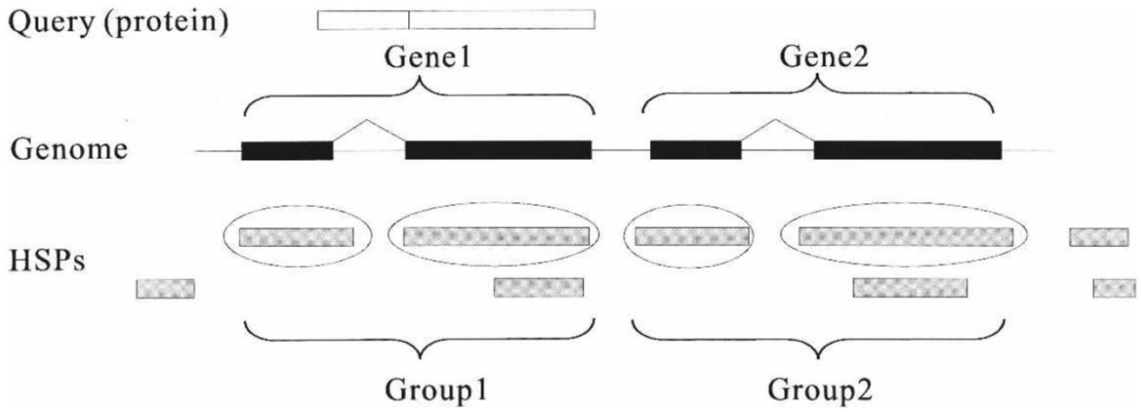


Figure 3-1 Grouping of HSPs into groups representing paralogs in tandem. For simplicity, this figure shows only a small portion of the HSPs returned by BLAST. Each HSP may correspond to a coding segment (likely an exon) of a gene, thus a group of HSPs may collectively represent a full-length gene. Each shaded box at the bottom of the figures represents an HSP at its corresponding genomic position. Candidate genes are shown on the genome, with exons (black boxes) connected by introns (lines). The HSP groups that best represent the genes are shown under the corresponding genes, with relevant HSPs in the groups circled. Two paralogous genes in tandem (gene1 and gene2) are shown. The boundary of the two genes must be correctly resolved. This figure was illustrated by R. She.

Over the past years, *ad hoc* solutions have been developed to filter and group HSPs that are produced using BLAST or other similarity-based searching tools to represent genes. The problem is that these *ad hoc* solutions can resolve some genes but fail in many cases. Earlier programs developed with the functionality of grouping HSPs include HSPcrunch (Sonnhammer and Durbin 1994), WU-BLAST (Lopez *et al.* 2003), LIS (Zhang 2003), and BLAST2GENE (Suyama *et al.* 2004). HSPcrunch reorganizes BLAST output by sorting HSPs in sequential order. HSPcrunch groups adjascent HSPs together based on linearity and the distance threshold between two HSPs. Although HSPcrunch can identify putative orthologous gene regions, it provides no ranking to indicate which group is most probable. Additonally, the grouping can fail if HSPs overlap. WU-BLAST is a BLAST program derivative. It can categorize HSPs into groups when users enable the “topcomboE” option. Within each group produced by WU-BLAST, HSPs are usually adjacent and collinear. Although WU-BLAST can successfully group many HSPs into gene-like structures, for genes within tandem clusters in the target genome, WU-BLAST inevitably fails. For these cases, WU-BLAST tends to group HSPs corresponding to different genes into the same group, as discussed later. A program based on the longest increasing subsequence algorithm (LIS) was developed to filter and group BLAST HSPs (Zhang 2003). Similar to the WU-BLAST program, it does not reliably interpret HSPs representing multiple paralogous genes. Another program, BLAST2GENE (Suyama *et al.* 2004), was developed to specifically solve the multiple paralogous gene problem; however, because it relies on many arbitrary thresholds and matrix usage, its application may be limited.

More recently, Cui *et al*. developed a new filtering and grouping algorithm that processes BLAST results, which was in turn used for identifying homologous genes (Cui *et al.* 2007). The investigators applied a three-step procedure to filter and group HSPs that represent candidate genes: (1) Filter all HSPs by discarding HSPs with scores lower than a heuristic value; (2) group HSPs based on their physical distance along the chromosomes; and (3) further filter HSPs by estimating the genomic span of target regions. All HSPs that fall outside of the target regions are excluded from further analysis. Comparing to WU-BLAST, which fails in filtering and grouping HSPs representing all tandem homologous genes, this program correctly filters and groups HSPs representing some tandem homologous genes. However, this program has an important weakness, which is its dependence on the physical distances (step 2) between gene structures (groups of HSPs) to separate groups. It assumes that the distance between different genes are significantly larger than the distance between HSPs within a group, which is not true, especially for paralogous genes in tandem clusters. Due to the usage of *ad hoc* distance thresholds to separate adjacent genes, the program by Cui *et al.* fails to resolve individual paralogous genes within tandem clusters. On one hand, if the threshold value is too large, HSPs corresponding to multiple genes will be lumped together into a large group. On the other hand, if the threshold value is too small, HSPs corresponding to a same gene could be divided into different HSP groups. In addition to this important weakness, the program by Cui *et al*. cannot be applied to filter HSPs that represent genes because this program does not remove random HSPs that fall into the genomic region that contain the candidate gene.

The filtering and grouping task is particularly challenging when the query gene has a large number of paralogous genes in tandem in the target genome, as illustrated in Figure 3-1. Figure 3-1 shows that a query gene could have two (or more) homologous genes (Gene1 and Gene2) that are located in adjacent genomic regions. A large number of genes in almost all sequenced genomes to date are parts of tandem homologous gene clusters. For example, in the nematode *C. elegans* genome, more than 1400 chemosensory genes form many tandem gene clusters, each of which contains two or more homologous genes (Robertson and Thomas 2006). Therefore, a program that is capable of filtering and assembling HSPs representing genes in tandem clusters is very important.

In this project, we developed a new graph-based algorithm, genBlastA, to directly address the above described challenge in filtering and assembling HSPs into genomic gene regions. A distinctive feature of genBlastA is that it does not rely on using *ad hoc* thresholds for filtering noise HSPs and on physical distance between target genes. Instead, genBlastA models the relationships and constraints among HSPs as a directed graph—designated the HSP graph—and models the HSP filtering and assembling problem as a search for the shortest paths in this graph. The novelty of this graph-based algorithm is an innovative edge length metric that reflects a set of biologically motivated requirements so that each shortest path corresponds to an HSP group representing a homologous gene. Unlike existing *ad hoc* grouping methods, this method filters and assembles HSPs on the basis of optimizing the path length to best capture the quality of a group of HSPs as a candidate gene. Consequently, our method is more robust, and it finds an optimal solution (with respect to a given length metric) without imposing a prior constraint on gene structures.

We have tested the performance of genBlastA extensively in filtering and assembling HSPs found in the genomes of two closely related nematode species: *Caenorhabditis elegans* (C. elegans Sequencing Consortium 1998) and *Caenorhabditis briggsae* (Stein *et al.* 2003). These genomes were selected for testing because both have been fully sequenced and extensively annotated. Our study shows that the performance of genBlastA is significantly better than that of WU-BLAST and the program by Cui *et al.*

3.3 Methods and Materials

3.3.1 Problem definition

In this work, we study the following problem: given a query (protein or DNA) sequence and a database of target genomic sequences, we want to identify all homologous genomic regions containing target genes (genes in the target sequences that are homologous to the query gene). First, as a preprocessing step, we apply BLAST to find local alignments between the query sequence and the target sequences. This step produces a list of HSPs, with each HSP containing the following information: (1) the target segment T and its location in the target sequence, and the corresponding query segment Q and its location in the query sequence, (2) an E-value, and (3) a PID value. In the second step, we filter and group the HSPs such that each group of HSPs forms a candidate region containing the target gene, called candidate gene region. genBlastA focuses on the second step.

An example of a list of HSPs is shown in Figure 3-2A, where the correspondence between the target segment (T) and query segment (Q) in a HSP is illustrated by dotted lines. For example, [Q_1,T_1] and [Q_1,T_2] are two different HSPs. HSPs may overlap in terms of their genomic positions and/or their query correspondences. Note the HSPs shown in this figure are only for illustration purposes, through which we will show that our algorithm is able to handle HSPs with all kinds of relationships.

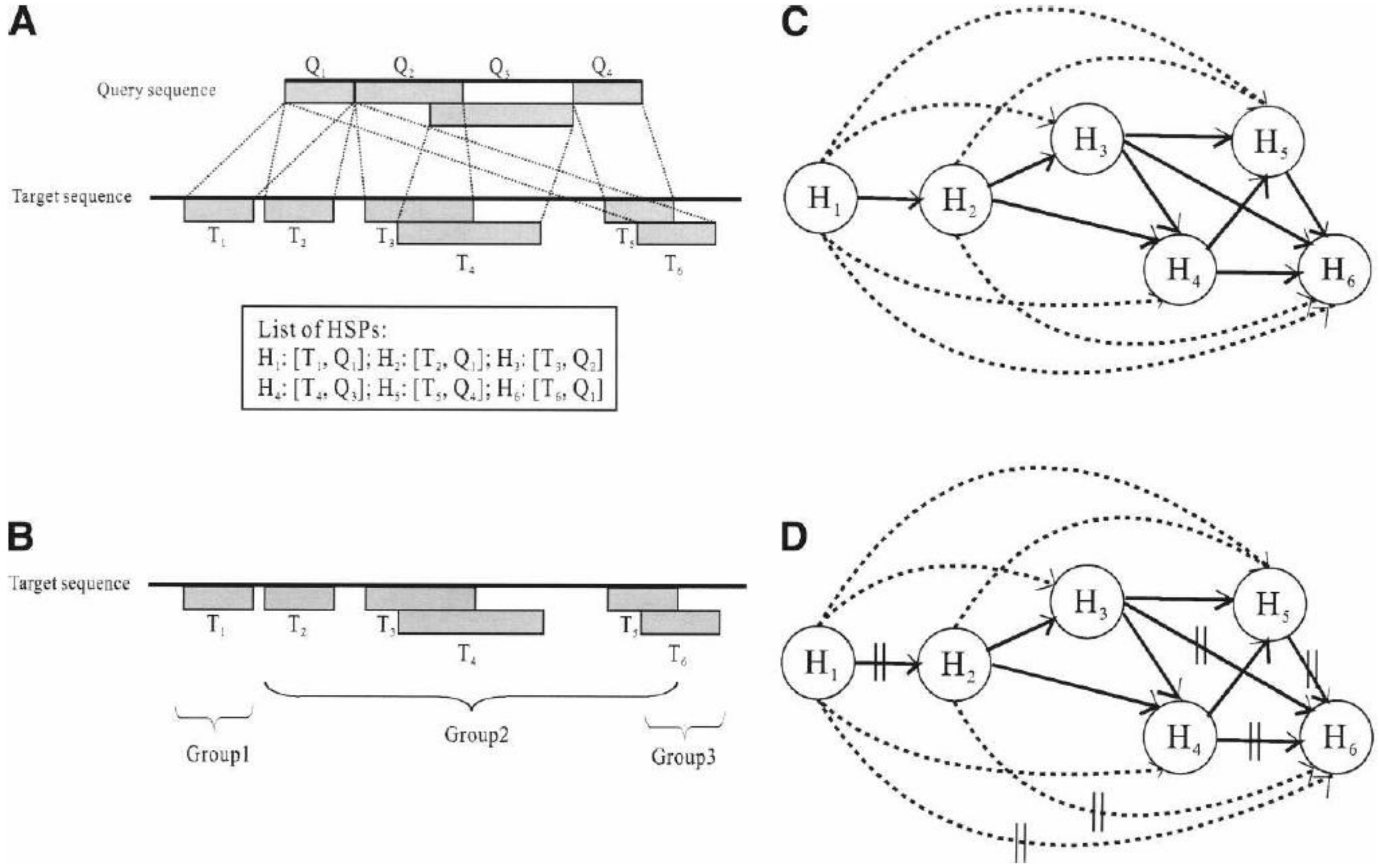


Figure 3-2 (A) HSPs returned by BLAST. Q1, Q2, Q3, and Q4 represent query segments, while T1, T2, T3, T4, T5, and T6 represent target segments. (B) Example of groups of HSPs. (C) The HSP graph, with solid lines representing edges and dotted edges indicating skip edges. (D) The HSP graph, with vertical bars indicating separating edges. This figure was illustrated by R. She.

Each genomic sequence has two strands: positive and negative. Each strand is considered a separate target sequence by genBlastA. Their only difference is the direction of alignment between the target gene and the query gene. Because each target sequence is independent and has its own list of HSPs, we process each target sequence separately in order to obtain the candidate gene regions for that sequence. Finally, all candidates for all target sequences are ranked into a single ranked list by their score as computed by our algorithm (discussed later). From now on, for brevity, all discussions will be based on query sequence and a single positive-strand target sequence.

3.3.2 HSP groups

With each HSP target segment that matches a query segment, a sequential group of HSP target segments can collectively match a larger piece of the query sequence. We are interested in those groups of HSPs, which correspond to genes that are homologous to the query gene. Such groups are termed HSP groups. In general, there are different numbers of HSP groups in the target sequence for each query gene. If the query gene is not conserved in the target genome then no HSP group can be found. If the query gene belongs to a multi-gene family (or the query gene has many paralogous genes), there will be multiple HSP groups in the target sequence, each representing a candidate region encoding a paralogous gene.

Consider the example in Figure 3-2A: T_3 and T_4 are in the same order as their query segments and therefore [Q_3,T_4] can be in the same group as [Q_2,T_3]. In fact, by merging T_3 and T_4 into one continuous target region, and merging their query segments into one continuous query region, we have a larger and better alignment. Figure 3-2B shows a possible grouping of HSPs that satisfies the sequential ordering and co-linearity requirements. Note that Group 1 and Group 3 have incomplete query gene coverage because a large portion of the query sequence is not covered by their query segments. In contrast, Group 2 covers the entire query sequence. A good HSP group should have large query coverage.

3.3.3 Graph modeling

An HSP graph is a graph representation that captures the above requirements on HSP groups. Each HSP is represented by a node, with edges that model the sequential ordering of the HSP target segments and edges that skip HSPs. An HSP grouping is modelled by grouping the nodes on a path such that each group covers as many query segments as possible while preserving co-linearity. By using a length metric, we will show that an optimal HSP group is a shortest path in the HSP graph.

Figure 3-2C shows the HSP graph for the HSPs in Figure 3-2A. The dotted edges are skip edges. Each path in the graph represents a way of selecting HSPs along the path. With skip edges, the HSP graph provides a complete search space for all possible groupings of HSPs. The number of skip edges can be very large. However, after introducing a length metric on edges (Appendix B), we will show that many skip edges can be removed without affecting the result. Our program genBlastA will not construct such skip edges thus dramatically increasing the efficiency of genBlastA. To distinguish these two types of edges, we add a vertical bar to each separating edge (Figure 3-2D). For example, $H_1 \rightarrow H_2$ is a separating edge, which means that its source node and destination node should belong to different groups. The skip edge $H_1 \rightarrow H_3$ is an extension edge, and the skip edge $H_1 \rightarrow H_6$ is a separating edge.

Having extension edges and separating edges in place, each path in the HSP graph represents a way of filtering and grouping HSPs. As we traverse a path, following an extension edge extends the current HSP group to include the destination node, and following a separating edge ends the current HSP group at its source node and starts a new HSP group at its destination node. If an extension edge is a skip edge, following the edge will skip over the nodes on the paths that are shortcut by the edge. In this sense, the HSP graph provides a complete search space for filtering and grouping HSPs.

The single-source shortest path algorithm for a directed acyclic graph can be done efficiently in O(E) time, where E is the number of edges (Manber 1989). Running this algorithm once for each possible starting node H_1, the total running time is O(E•V), where V is the number of end nodes of separating edges and is bounded by the number of HSPs.

3.4 RESULTS

In this project, we developed the program genBlastA (described in Methods and Materials) that uses a novel graph based algorithm with excellent capability for identifying HSP groups that represent orthologs (genes in different species but with same origin in evolution), paralogs (genes duplicated within a species), as well as novel genes (genes that have not yet been identified).

3.4.1 Test gene set preparation and test strategy

The datasets used for evaluation were obtained from WormBase release WS170 (http://www.wormbase.org/), an integrated database for the biology and genomics of *C. elegans* and other nematode species including *C. briggsae* (Chen *et al.* 2005). For testing the performance of genBlastA, we have selected a test gene set of 464 *C. elegans* genes that are representative of the *C. elegans* genome. To achieve this representation, the majority (300 genes) of these genes were taken from three representational contiguous regions of *C. elegans* Chromosome I. These three regions are the left arm (containing 100 genes), the middle region (containing 100 genes), and the right arm (containing 100 genes) of chromosomal regions. To ensure that the test gene set contains representative genes of different complexities, we further included 164 additional genes, including genes with internal repetitive regions (Pfam domains) and genes that belong to large paralogous tandem clusters. The test gene set is available at http://genome.sfu.ca/projects/genBlastA/.

To evaluate the capability of genBlastA to identify and group HSPs into gene-like structures and the capability of identifying novel genes, we selected *C. elegans* genome as the target database for *C. elegans* query genes (called “EvsE test”). To evaluate the performance of genBlastA in identifying homologous sequences in genomes of different but related species, we used *C. briggsae* genome as the target database for the same set of *C. elegans* query genes (called “EvsB test”). These two species split approximately 80-120 million years ago (Coghlan and Wolfe 2002; Stein *et al.* 2003), around the same time as the human/mouse split (Waterston *et al.* 2002).

In our experiments, genBlastA was able to process all 464 test genes (with over 43,000 HSPs reported by BLAST in EvsE test) within only one minute on a medium-speed PC (with a Pentium-IV 2.6GHz CPU). Since these 464 genes are representative of the entire *C. elegans* genome and comprise 2% of the genome, we calculate that it would take less than 1 hour to process the entire genome (which contains approximately 20,000 genes).

We compared the performance of genBlastA with two existing programs with similar functionalities—WU-BLAST (Lopez *et al.* 2003) and the program by Cui *et al*. (Cui *et al.* 2007). WU-BLAST is available by an academic license. Since the HSP grouping functionality of the program by Cui *et al.* is not readily available, we implemented this program, called “ML” in the following text, based on their publication (Cui *et al.* 2007). ML requires a distance threshold to resolve different HSP groups. This threshold is not described in detail in their publication and therefore, we derived an optimal distance value based on experimenting with different threshold values. In our experiments, we found that ML performs best for our test cases described below when the distance threshold is set to 1,000 bp. Therefore, this distance was used for ML throughout our analysis.

For each query gene in the test gene set, we first ran TBLASTN against the *C. elegans* genome (for EvsE test) and the *C. briggsae* genome (for EvsB test) with two different BLAST settings: “ungapped” and “gapped”. While the gapped HSPs are generally longer with more gaps and mismatches, ungapped HSPs are generally shorter with much higher PIDs. We then carried out three sets of experiments, each with a different purpose.

(1) Resolving paralogous genes in tandem clusters

This first experiment was designed to test the capability of these programs in addressing the major challenge that we have identified—resolving HSP groups that correspond to target gene families in the target genome. For this purpose, we selected 30 genes from the test gene set that belong to large gene families and these family members form tandem gene clusters.

(2) Searching for Orthologous Groups

In this test, each gene in the test gene set was used as a query to identify the top-ranked HSP group, i.e., the candidate ortholog of the query gene. Since the top-ranked group is expected to be the most similar to the query gene, in the EvsE test, it is expected to map to the query gene itself; in the EvsB test, it should map to its *C. briggsae* ortholog.

(3) Identifying Novel Genes

In the third experiment, we explored the utility of genBlastA for identifying novel (paralogous) genes, i.e., the genomic regions that show high similarity to known genes but have no gene annotations.

3.4.2 Resolving paralogous genes in tandem clusters

To test the three programs' abilities to resolve tandem duplicate genes, we examined the HSP groups produced for 30 query genes that all are members of large gene families. For our comparison, after we identified HSP groups using genBlastA, WU-BLAST, and ML, we retained all candidate regions with query coverage ≥ 50%. The HSP groups were then examined and divided into two categories: "specific" and "non-specific" groups. An HSP group is called "specific" if the corresponding genomic region contains only one annotated gene, "non-specific" if the region has multiple annotated genes. HSP groups with high similarity to the query and contain only single genes are likely to be true paralogs. The programs' performance in resolving single genes are evaluated by comparing the ratio of "specific" groups (the number of "specific" HSP groups over the total number of HSP groups examined). Figure 3-3 illustrates an example, in which five paralogous genes are in a tandem gene cluster. As expected, WU-BLAST correctly identified only one target gene and failed to produce HSP groups corresponding to the four other genes. ML produced three groups, two of which erroneously contain HSPs corresponding to other adjacent genes. ML missed groups for two target genes (T27B7.3 and T27B7.6a), and mistakenly grouped HSPs corresponding to T27B7.6a to the HSP group corresponding to T27B7.5 (Figure 3-3). In contrast, genBlastA successfully resolved all five genes, producing five groups of HSPs.

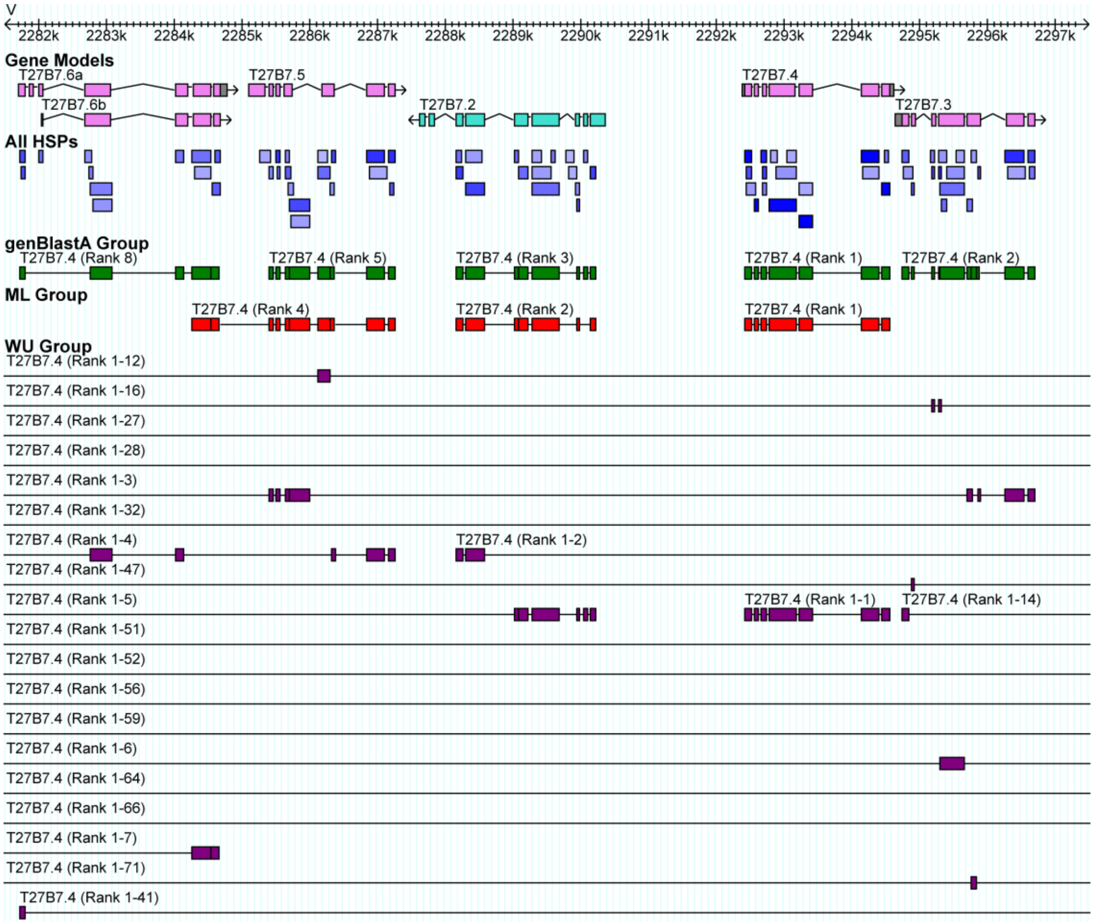


Figure 3-3 Grouping HSPs into groups representing individual genes. genBlastA was able to resolve all 5 members while ML resolved only 2 and WU only 1. Gene models are shown in the "Gene Models" track. HSPs are shown as blue boxes in the "All HSPs" track. The color indicates different PID for the HSPs. Darker color indicates higher PID. The "genBlastA Group", "ML Group", and "Wu Group" tracks show HSPs groupings that are returned by genBlastA, ML and WU-BLAST, respectively. The data used to generate this figure was collected by J. Chu.

In summary, when BLAST was executed with the "ungapped" setting in the EvsE sets, the average ratio of "specific" HSP groups by genBlastA is around 80%, which is significantly higher than that produced by WU-BLAST (~20%) or ML (~40%) (Figure 3-4). Similar results were observed when WU-BLAST was performed with the “gapped” setting. Thus, in all cases, whether BLAST was executed with “ungapped” or “gapped” setting, genBlastA was able to resolve more “specific” HSP groups in tandem duplicates compared to either WU-BLAST or ML. WU-BLAST usually generated numerous HSP groups but also spanned regions with multiple genes (therefore nonspecific). Consequently, WU-BLAST groups together tandem paralogous genes, leading to poor performance in resolving tandem paralogous genes. ML had poor performance due to its use of a distance threshold. In particular, as the distance threshold increases, the ability of ML to resolve closely spaced paralogous groups decreases.

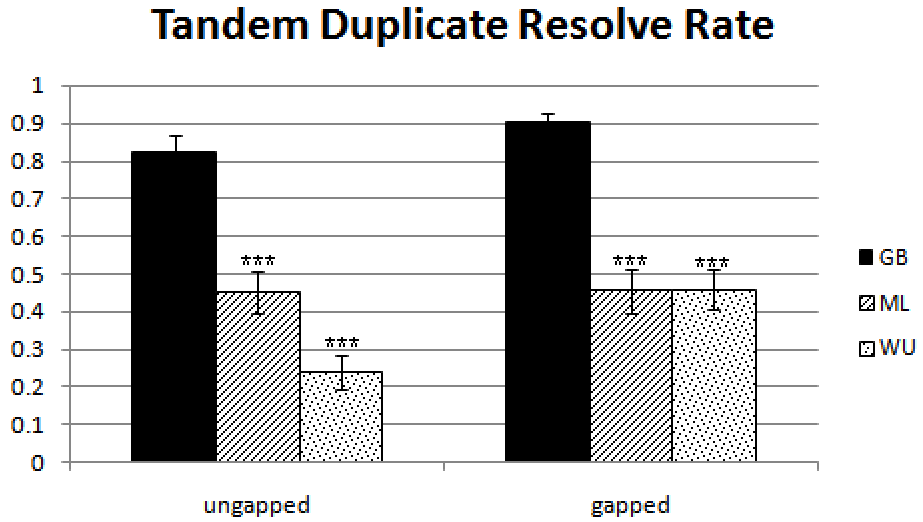


Figure 3-4 Grouping of HSPs to represent individual homologous genes in tandem clusters. This figure shows average resolve rate for a total of 30 tandem duplicated gene clusters in the EvsE dataset for genBlastA (GB), Cui *et al* (2007) (ML), and Wu-Blast (WU). Ratio of specific groups was calculated as the number of genes resolved over the total number of genes in each tandem gene cluster. A gene is considered resolved if the HSP group overlaps with only one single gene in WormBase and the span similarity is ≥ 50%. Gapped and Ungapped represent two independent BLAST results using either gapped setting or ungapped setting. GB alpha value is 0.5. ML distance threshold is 1000. Error bars = standard error. (***) shows statistical significance (p < 0.001) by paired Student’s T Test. The data used in this figure was collected by J. Chu.

3.4.3 Searching for Orthologous Groups

In this test, the top-ranked HSP group corresponding to each query gene is evaluated by comparing to the expected gene as annotated in WormBase (WS170). First, we compared the accuracy rates of three programs when *C. elegans* genes were used as query genes to search for top-ranked genes in *C. elegans* genome. The accuracy rate is defined as the percentage of correctly assembled HSP groups. The accuracy rate for genBlastA is 97.2%, much higher than those of WU-BLAST and ML, which are 67.0% and 82.8%, respectively. For more accurate comparisons, the similarity or overlap between the HSP group and the expected gene were quantified. We used the following two criteria to evaluate the top-ranked HSP groups: (1) query coverage and (2) genomic span. Query coverage measures the similarity between the HSP group and the query gene. It is defined as the proportion of the query sequence covered by the HSPs in the HSP group identified by each of the three programs. A program should identity the HSP group that best covers the query gene. Genomic span measures the extent of overlap between the genomic region given by the HSP group and the expected gene region in the target genome. We evaluated this using the Jaccard similarity: For the annotated target gene region R_A and the reported gene region R_R, their similarity is ($|R_A \cap R_R|/|R_A \cup R_R|$). This result is zero when two regions do not overlap.

Query Coverage Test

Figure 3-5 shows the average query coverage for 464 query genes in the test gene set. When WU-BLAST was executed using the “ungapped” setting in the EvsE test (Figure 3-5a) and the EvsB test (Figure 3-5c), genBlastA identifies HSP groups with close to 100% query coverage and significantly outperformed both WU-BLAST and ML. Similarly, when WU-BLAST was executed using the “gapped” setting, genBlastA significantly outperformed both WU-BLAST and ML in the EvsE test (Figure 3-5a) and the EvsB test (Figure 3-5c).

Genomic Span Test

As shown in Figure 3-5b, when WU-BLAST was executed using the “ungapped” setting, for both EvsE and EvsB tests, genBlastA significantly outperformed both WU-BLAST and ML by a large margin, suggesting that genomic regions predicted by WU-BLAST and ML are dramatically different from the real genomic regions. Similarly, when BLAST was executed using the “gapped” setting, for both EvsE and EvsB tests, genBlastA outperformed both WU-BLAST and ML significantly, while WU-BLAST outperformed ML.

Taken together, genBlastA outperformed both WU-BLAST and ML in identifying orthologous HSP groups.

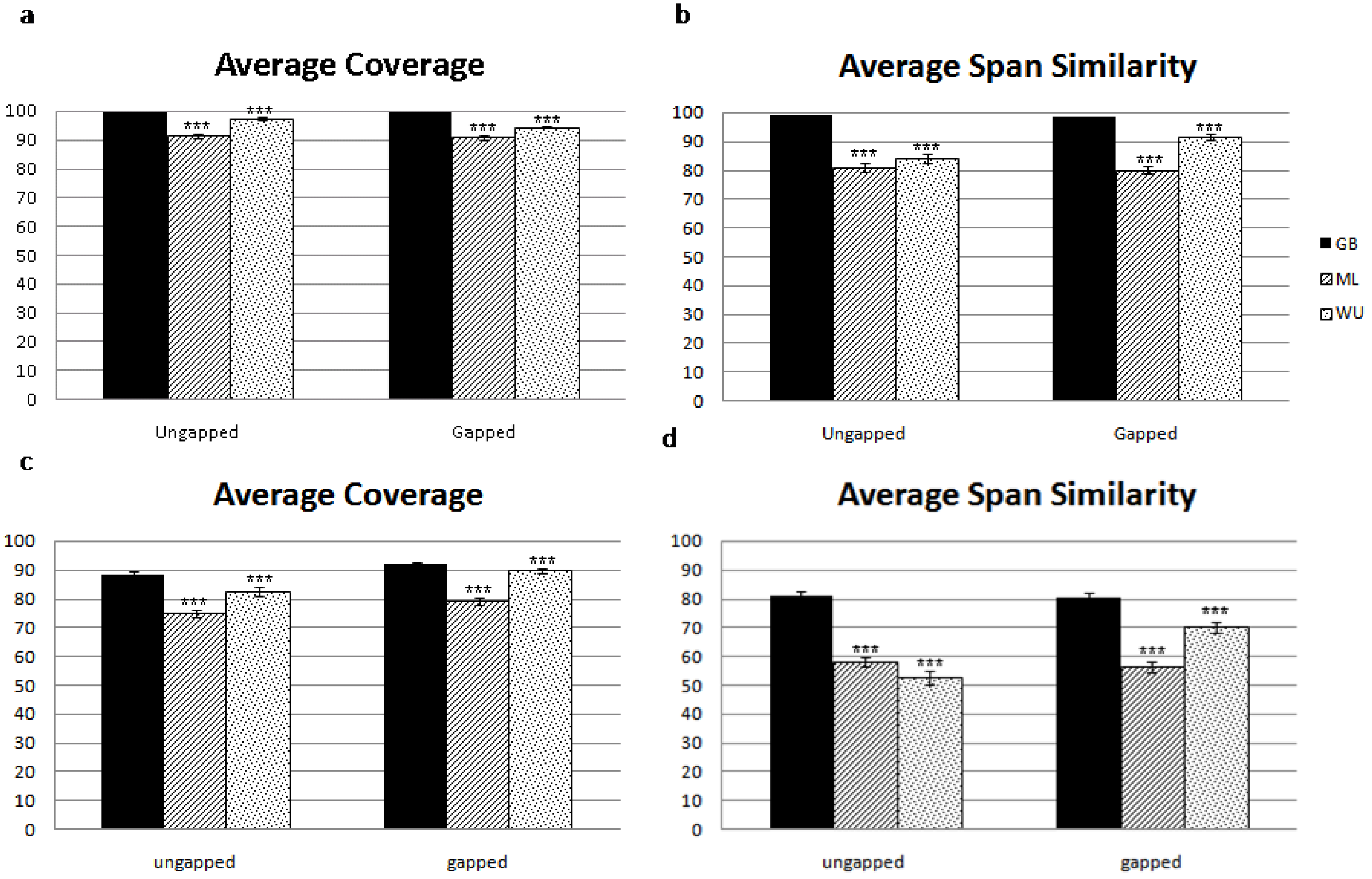


Figure 3-5 (a) Average coverage for EvsE dataset. (b) Average span similarity for EvsE dataset. (c) Average coverage for EvsB dataset. (d) Average span similarity for EvsB dataset. In all cases, figures represent averaged results over 464 test genes for three different programs genBlastA (GB), Cui *et al* (2007) (ML), and Wu-Blast (WU). Gapped and Ungapped represent two independent BLAST results using either gapped setting or ungapped setting. Span similarity is calculated by Jaccard similarity. GB alpha value is 0.5. ML distance threshold is 1000. The error bars represent standard error and (***) show statistical significance (p < 0.001) by paired Student's T Test. The data used in this figure was collected by J. Chu.

3.4.4 Identifying Novel Genes

Since genBlastA can be applied to effectively identify homologous genomic regions in a target genome, we reasoned that it can be used for identifying novel paralogous genes that have been missed by other approaches. To demonstrate this, we examined whether genBlastA can be used to identify HSP groups in *C. elegans* genome that are homologous to the test genes and that do not overlap with any existing gene annotation, therefore, identifying putative novel genes or novel pseudogenes.

We evaluated all candidate homologous gene regions for the 464 query genes for ones that show both significant query gene coverage (> 80%) and do not correspond to known genes. We found eight candidates in our search. In particular, four of them contain putative novel genes that are relatively long (>300 amino acids) (Table 3-1). Based on recent RNAseq data from modENCODE project, only the region V:4432685..4432152 have read coverage during mid-L4 stage. The lack of RNA read data in other regions suggest the presence of pseudogenes.

Table 3-1 genBlastG models that map to empty genomic regions in *C. elegans*

Model	Coverage (%)	Coordinate	Length (bp)
C17F3.3.gb-5.2	61.6601	II:13721734..13722244	510
C17F3.3.gb-60.2	70.3557	V:4432685..4432152	533
C17F3.3.gb-71.2	53.3597	I:11675930..11675384	546
F47B7.4.gb-11.1	66.3677	X:3755586..3754542	1044
M199.5.gb-89.1	66.0606	II:3466361..3467668	1307
VH15N14R.1.gb-10.2	61.7117	II:1448130..1449831	1701
VH15N14R.1.gb-9.2	62.1622	II:7445359..7446354	995
Y45G12A.1.gb-36.1	59.9398	V:4971049..4970151	898

3.5 DISCUSSION

BLAST and related search programs have been widely used for identifying homologous sequences since they are sensitive and effective in finding homologous fragments for query genes. However, BLAST results often contain a large number of HSPs and can be challenging, if not overwhelming, for the end users. Our program genBlastA provides an effective way to interpret the large list of HSPs reported by BLAST in order to allow users to focus targets they find interesting. genBlastA enables users to effectively identify homologous genomic regions because each genomic region represents a full-length candidate gene rather than fragments of a gene (HSPs). Thus, genBlastA empowers users by allowing them to effectively identify candidate genes in target genomes. This will make BLAST and related programs even more useful.

Our analysis has clearly shown that genBlastA outperforms existing programs developed previously with similar objectives. In particular, genBlastA is very effective in grouping HSPs corresponding individual genes within tandem clusters of homologous genes. Both WU-BLAST and the program developed by Cui *et al*. (2007) failed in this task. Although ML performs better than WU-BLAST in resolving multiple paralogous genes in tandem clusters, the current ML program is not ready for this job because the current ML program is not capable of removing random HSPs in the genomic regions.

The ability of effectively resolving HSP groups by genBlastA will enable users to take advantage of HSP groups, which are useful in several ways. First, genBlastA can be used by researchers to quickly locate candidate gene structures in the identified homologous genomic regions in the target genomes. Compared with the large collection of HSPs reported by BLAST and similar programs, ranked HSP groups provide much more useful information relevant to full-length target gene structures, instead of fragments of target genes. In fact, any program (not just BLAST) that generates a list of local alignments can be used as input for genBlastA, as we have demonstrated using WU-BLAST. With appropriate parsers, programs such as Crossmatch (P. Green, unpublished), FASTA (Lipman and Pearson 1985), and PatternHunter (Ma *et al.* 2002) can be used. Since end users such as experimental biologists are usually more interested in genes, genBlastA makes search results from local alignment programs more accessible and meaningful to them.

Second, genBlastA can be used to preprocess genomic DNA sequences for gene finding programs including genewise (Birney *et al.* 2004) and exonerate (Slater and Birney 2005). Both GeneWise and Exonerate are widely used for homology based gene prediction programs. However, both programs, especially GeneWise, are computationally expensive when used to search for candidate genes in entire genomes. Their performance can be dramatically enhanced if their genomics search spaces are reduced. genBlastA, which is capable of identifying candidate genomic regions, can be used effectively to preprocess the genomic sequences in order to reduce search spaces. It can also be integrated in the program by Cui *et al*. to identify homologous genes.

Third, these HSPs can be used to resolve gene structures, either manually or computationally. Candidate gene models can be accurately defined by HSPs in each HSP group, intron-exon splicing information at the edges of HSPs, as well as the similarity between query and candidate genes.

4: REVISING *C. BRIGGSAE* GENE ANNOTATION USING GENBLASTG, A BLAST-BASED HIGH PERFORMANCE GENE FINDER

Note regarding contributions:

A portion of this study has been submitted to Bioinformatics:

She R., Chu J.S.-C., Uyar B., Wang J., Wang K., Chen N. genBlastG: extending BLAST to be a high performance gene finder. Bioinformatics (submitted).

In this study, I prepared all the necessary dataset and databases. I also evaluated runtime, accuracy, and PID of predicted gene models from genBlastG, GeneWise and Exonerate for *Caenorhabditis* species and plant species. B. Uyar evaluated genBlastG performance using human genomic sequences. I predicted the gene models in *C. briggsae* and identified the improved gene models. I designed the primers used for experimental validation. M. Tang, C. Wang, and J. Wang performed RNA extraction and RT-PCR from *C. briggsae*. R. She, K. Wang, and N. Chen designed the genBlastG algorithm. R. She implemented the algorithm. R. She, K. Wang, N. Chen, and I wrote the manuscript.

4.1 Abstract

We present in this report a new homology-based gene prediction algorithm genBlastG. Taking advantage of the homologous genomic regions defined by the program genBlastA that we have recently developed, genBlastG defines gene models by specifying gene start and stop signals, splicing donor and acceptor sequences, as well as the alignment identity between candidate genes and their corresponding queries. Comparing to GeneWise and Exonerate, two popular homology-based gene prediction programs, genBlastG predicts gene models with higher accuracy in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*, and in the human genome. Additionally, genBlastG runs up to 1,000-fold faster for gene prediction in the worm genomes and in the human genome. Using genBlastA and genBlastG, we have revised 1,805 *C. briggsae* gene models and identified 85 gene models that have been missed in previous annotation efforts. In conclusion, genBlastG represents a powerful and easy-to-use homology-based gene prediction program with both high accuracy and speed.

4.2 Introduction

Basic Local Alignment Search Tool (BLAST) (Altschul *et al.* 1990) is one of the most popular bioinformatics tools ever developed. Frequently, BLAST users expect to identify homologous genes for comparative analysis. For example, following the discovery of a previously unknown gene in a human genome, a biologist will typically perform a BLAST search of the publicly accessible databases such as the mouse genome database to see if another species carries a similar gene, hoping to gain insights into the function and regulatory signals of the newly found gene. BLAST, however, returns the user a (usually large) collection of local alignments called high-scoring segment pairs (HSPs). Such HSPs provide no indication how they are structured into the gene model because they are usually only isolated regions of similarity with some being simply noise.

In the last decade, many homology-based gene predictors have been developed (Brent 2005). Homology-based programs make use of extrinsic evidence such as protein sequences, mRNAs, ESTs or other genomic sequences in finding the genes on the DNA sequence. The availability of genome sequences of related species has created a growing demand for better and faster homology-based gene prediction programs, which gives rise to many developments in this area, such as GeneWise (Birney *et al.* 2004), Projector (Meyer and Durbin 2004), TwinScan (Korf *et al.* 2001), Exonerate (Slater and Birney 2005), SGP2 (Parra *et al.* 2003), SLAM (Pachter *et al.* 2001). In general, it has been shown that homology-based gene prediction methods outperform the *ab initio* methods in terms of accuracy when there are extrinsic evidences available (Zhang 2002; Coghlan *et al.* 2008), with GeneWise being one of the most widely used homology-based gene prediction programs. These algorithms are independent of BLAST or use BLAST only as a pre-processing tool to narrow down gene search space (Cui *et al.* 2007). Our work is motivated by the following question: can we extract the gene structure directly from the HSPs found by BLAST, which are expected to represent high quality local alignments? The rationale is to delegate the expensive local alignment search to the well developed BLAST and focus on extracting and defining the best gene structure that such HSPs represent. Here we present a novel algorithm, genBlastG, which takes HSPs identified by BLAST as the input and defines the gene models that they represent. Unlike previous gene finding algorithms, genBlastG is able to leverage the vast improvement in speed and search quality of BLAST made in last 20 years since its first publication (Altschul *et al.* 1990) and benefit from the wide acceptance and availability of BLAST. Although BLAST is used in this study for detecting HSPs, many other local alignment search programs, including BLAT (Kent 2002), Crossmatch (P. Green, unpublished), FASTA (Lipman and Pearson 1985), PatternHunter (Ma *et al.* 2002), can also be used as an input for genBlastG given appropriate parsers is provided.

4.3 Methods and Materials

4.3.1 Input and output of genBlastG

Each input to genBlastG is an HSP group defined by genBlastA, including all HSPs alignments, as well as the query protein and the target genome sequences. The output generated by genBlastG is the predicted gene models containing exact positions of coding protein-coding exons and introns, in addition to the predicted mRNA and protein sequences.

4.3.2 Challenges and genBlastG algorithm overview

Each protein-coding gene is composed of one or more exons separated by introns, which are flanked by splicing signals (Breathnach and Chambon 1981). The task of protein-coding gene prediction can thus be defined as determining the start codon (“ATG” for all genes), splice sites, and a stop codon (one of the three alternative codons “TAG”, “TGA”, or “TAA”) of a gene. A canonical intron starts with the base pairs “GT” and ends with the base pairs “AG” (Burset *et al.* 2000), which are referred to as the splice “donor” and splice “acceptor” sites, respectively. However, the presence of these signals is not sufficient to identify the splice sites because there could be numerous random pairs of “GT/AG” signals present in the neighbourhood of a gene. This problem may be alleviated with the additional information conveyed in HSPs.

Within each HSP group, the majority of HSPs correspond to individual coding exons in a one-to-one relationship, and the genomic regions between adjacent HSPs represent introns. However, there is not a simple one-to-one correspondence between HSP gaps and introns in a significant number of cases. Due to the threshold-based alignments generated by most local alignment tools (including BLAST), HSPs are extended alignments that allow gaps and mismatches. Therefore, it is possible that *bona fide* introns reside inside a single HSP, especially when an intron is small. As illustrated in Figure 4-1, there are three major challenges in identifying candidate introns:

Challenge 1: One HSP corresponding to multiple exons.

Challenge 2: Multiple HSPs corresponding to one exon.

Challenge 3: Exon regions (especially small exons) are not represented by HSPs at all.

We tackle all of these challenges by examining the similarity between the query gene and the target genes. Specifically, our program consists of four main steps. First, we determine the approximate regions for each intron. Second, for each approximate intron region, we find the candidate splice sites (donors and acceptors) in that region. Third, we determine the optimal combination of donor and acceptor sites from all splice site candidates in each intron region. Lastly, we implement a post-processing procedure to finalize the gene model by maximizing the query coverage and percentage identity to the query gene.

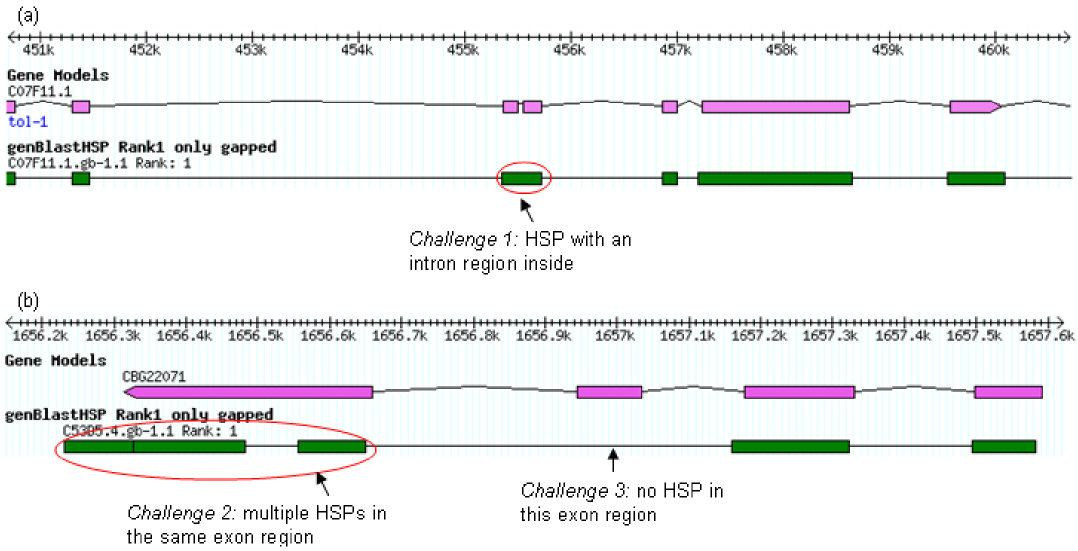


Figure 4-1 Three challenges of mapping HSPs to exons. The purple boxes are exons and the connecting lines between exons are introns. The green boxes are HSPs aligned to their corresponding genomic location. Most HSPs correspond to exactly one exonic region, but there are also challenges. (a) Challenge 1: six HSPs are shown in C07F11.1 region with the third HSP covering two exons and the intron in between. (b) Challenge 2: the first exon of CBG22071 is covered by more than one HSPs. Challenge 3: the second exon of CBG22071 does not have any corresponding HSP. The data used to generate this figure was collected by J. Chu and R. She.

4.3.3 The algorithm

Step 1: Determine the approximate intron regions

In this step, we locate the genomic regions indicated by the HSP group (generated using genBlastA) that represent introns. There are three types:

Type 1: Introns between adjacent HSPs. This is the simplest and the most common case. Here, we define a minimum intron length that may differ for different species (MIN_INTRON_REGION_LEN) (Deutsch and Long 1999). We consider the genomic region between two adjacent HSPs to be a candidate intron region if it is more than the MIN_INTRON_REGION_LEN (Figure 4-2a).

Type 2: Introns within a HSP. This is the case described by challenge 1. To identify candidate intron regions inside one HSP, we examine the alignment of each HSP. Since we expect the intron region to have no query correspondence, the intron region inside a HSP should be aligned with gaps on the query sequence. Therefore, if there is a region in the HSP alignment where the query segment consists of continuous gaps that are longer than MIN_INTRON_REGION_LEN, that region is considered to be a candidate intron region. For example, in Figure 4-2b, the gap on the query in the HSP <T4, Q4> leads to the intron region 3.

Type 3: Introns between adjacent HSPs with overlapping query segments. The borders of the intron region between two HSPs may need adjusting if the two HSPs contain overlapping query segments. Because the entire group of HSPs is expected to represent one complete gene that is homologous to the query, the spliced sequence (obtained by concatenating HSP target segments) should align well with the query sequence. Consider the two adjacent HSPs with overlapping query segments shown in Figure 4-2c, the overlapped part of the query segment should only be aligned with one of the HSP target segments. We chose the HSP with higher identity to keep the overlapping region and truncate the other HSP.

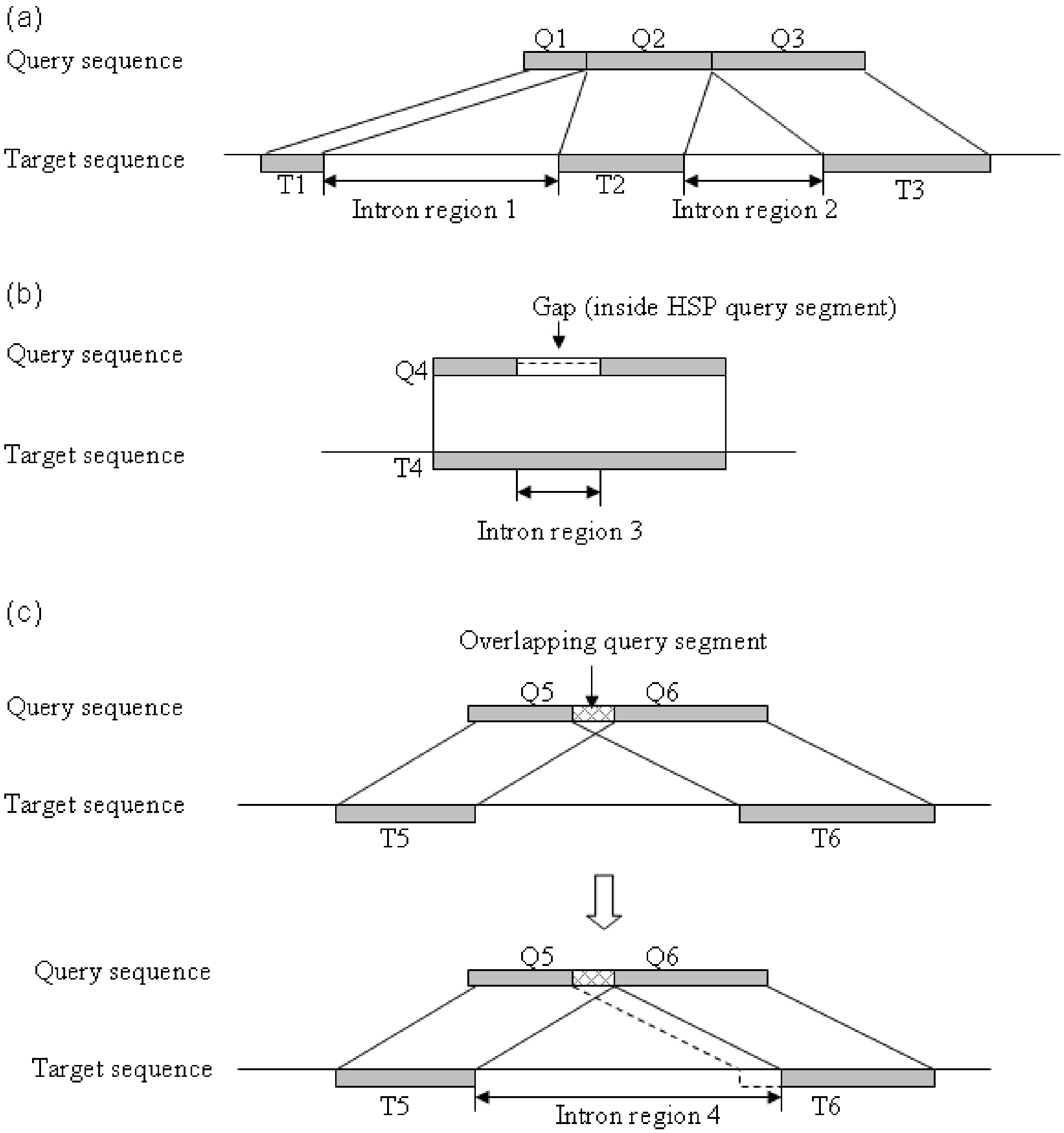


Figure 4-2 Three scenarios for identifying candidate intron regions. (a) Two intron regions(region 1 and 2) are between adjacent HSPs. (b) Intron region 3 is inside an HSP region. (c) Intron region 4 is between two adjacent HSP with an overlapping query segment. This figure is illustrated by R. She.

Step 2: Select candidate splice sites

Once putative intron regions are defined, splice sites are selected so that neighboring exons form appropriate reading frames. A candidate intron region represents approximate boundaries of introns, therefore we must examine multiple choices of splice sites around the borders of candidate intron region.

Splice site detection problem is simplified to selecting splice sites that are close to the borders of intron regions. For each intron region, we search for donor and acceptor signals (GT/AG) independently, and a number of splice signals that are closest to the borders are selected as the candidate splice sites. We use a user-defined threshold (MAX_NUM_SPLICE_SITES) to control the number of candidates selected around each border, i.e. the number of donors or acceptors is at most MAX_NUM_SPLICE_SITES. The selection of donors depends on the existence of “GT” signals within the given region and their relative distances to the 5’ border of the intron region. For example, if MAX_NUM_SPLICE_SITES is 20, then the candidate donor sites are identified by at most 20 “GT” signals that are closest to the 5’ border (could be at either side of the border). Similarly, the selection of acceptors depends on “AG” signals and the 3’ border of the intron region. Therefore, for each intron region, we get at most MAX_NUM_SPLICE_SITES donors and MAX_NUM_SPLICE_SITES acceptors.

Step 3: Find the best pair of splice sites

An intron is flanked by a pair of donor and acceptor. The best pair of donor and acceptor should maintain the reading frame in the spliced sequence (by joining the 5’ exon and the 3’ exon) and maximize percent identity (PID) of the translated peptide alignment called “spliced alignment”. The procedure here enforces that adjacent exons are in-frame and there is no in-frame stop codon in the spliced sequence. Consider an intron region I and its associated set of donors (d_1, …, d_n) and acceptors (a1, …, a_n) as given by Step 2. A pair of donor and acceptor (d_i, a_j) is considered a valid pairing if the donor and acceptor are in -frame with each other and there is no in-frame stop codon in the corresponding spliced sequence *S*, which is formed by connecting 5' HSP target segment with the 3' HSP target segment. It is possible that there is no valid pair of donor and acceptor exists in an intron region, in which case no intron will be predicted for this region, and consequently the genomic region will be treated as an exon.

We considered the possibility of artifactual intron regions between HSPs (as described by Challenge 2). Thus, we also consider the case in which there is no intron in the region I. In this case, no splice site is selected and the spliced sequence is simply the genomic sequence from the beginning of 5' HSP target segment to the end of 3' HSP target segment.

PID of all valid spliced sequences with the corresponding query segments is computed and the donor-acceptor pair that gives the spliced sequence with highest PID is returned. Note that it is possible for more than one spliced alignments (from different donor-acceptor pairs) to have the identical highest PID, in which case the alignments will be further compared in terms of their alignment scores computed based on a substitution matrix that scores the alignment significance between amino acids, such as BLOSUM62 (Henikoff and Henikoff 1992).

Figure 4-3 shows an example of the spliced sequence induced from two HSPs for donor D1 and acceptor A1. The corresponding query segment Q is the part of the query from the beginning of the 5' HSP query segment to the end of the 3' HSP query segment. The quality of the spliced alignment, i.e. the alignment between S and Q, determines the selection of the best pair of donor and acceptor for the current intron region, i.e. the valid pair that results in the highest PID will be selected.

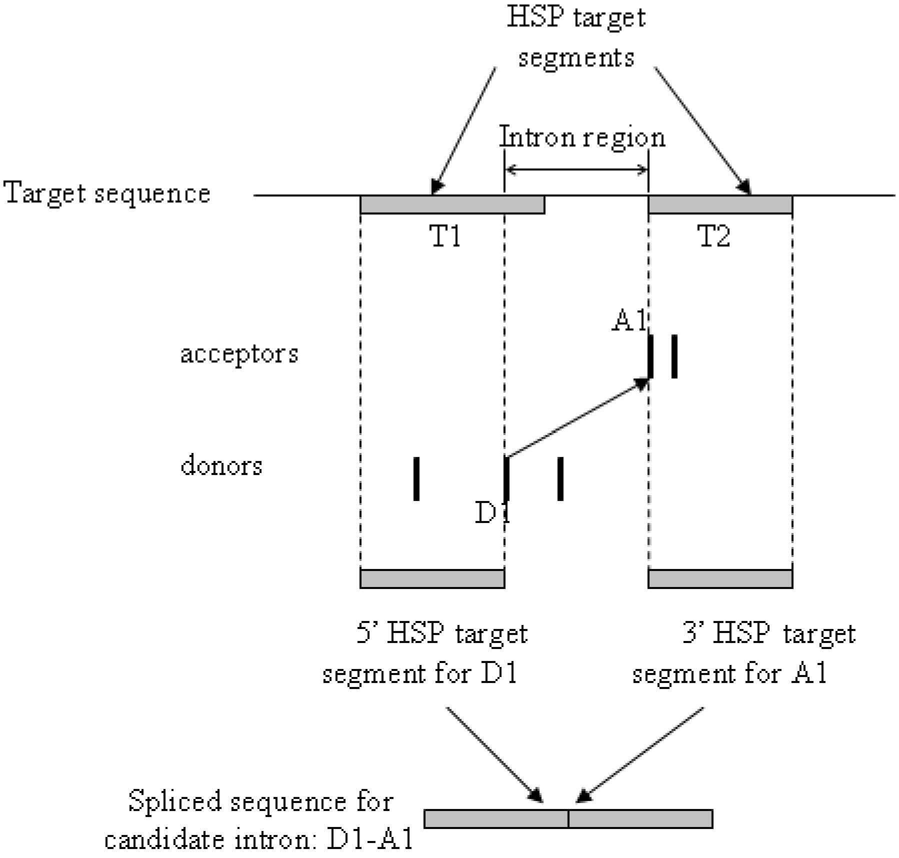

HSP target segments
Intron region
Target sequence
T1
T2
A1
acceptors
donors
D1
5' HSP target segment for D1
3' HSP target segment for A1
Spliced sequence for candidate intron: D1-A1

Figure 4-3 Finding the best pair of donor and acceptor for each intron region. This figure shows one intron region with three candidate donors and two candidate acceptors. For donor site D1, the 5' HSP target segment begins at the starting position of T1. For acceptor site A1, the 3' HSP target segment ends at the last position of T2. The pairing of D1 and A1 results in the spliced sequence that is formed by connecting the 5' HSP target segment and the 3' HSP target segment. The pair that leads to the best alignment between such spliced sequence and the corresponding query region is chosen. This figure is illustrated by R. She.

Step 4: Post-processing of candidate gene model

Selecting the best pair of donor and acceptor for each intron region produces an initial gene structure. However, gene models could still be missing one or more exons because local alignment programs often fail to pick up weak and short similarities as described by challenge 3. The purpose of the post-processing step is attempting to repair the initial prediction by searching for additional local alignments within this DNA region. We check for possible missing query segments between adjacent exons, before the first exon, and after the last exon. Any query segment that is not covered by initially-predicted exons are adjusted.

To find possible missing alignments, we performed local alignment in genomic regions between exons (starts from first position of the 5’ exon to the last position of the 3’ exon), in region before the first exon (starts from a user defined length to the last position of the first exon), and in region after the last exon (starts from the first position of the last exon to a user defined length). The new local alignments are then used to locate the possible new set of splice sites within this region, for appropriate incorporation of the newly recovered exon as follows. The local alignments (including the original HSPs and the newly-found alignments) that fall within this region will be used to find a possibly new set of splice sites in this region by following the same three steps as described above (Steps 1 to 3). The resulting new set of exons in this region is then compared with the initially-predicted exons in the same region. The set of exons that leads to higher PID in the spliced alignment is chosen as the final exons. Figure 4-4 shows an example where a missing query segment is identified between Exon2 and Exon3, which is not covered by any HSP in the HSP group (Figure 4-4a). A local alignment with the missing query segment is performed in this genomic region to find any additional alignment (Figure 4-4b). The additional alignment is used to identify new donor-acceptor sites and ultimately a new set of exons (Figure 4-4c).

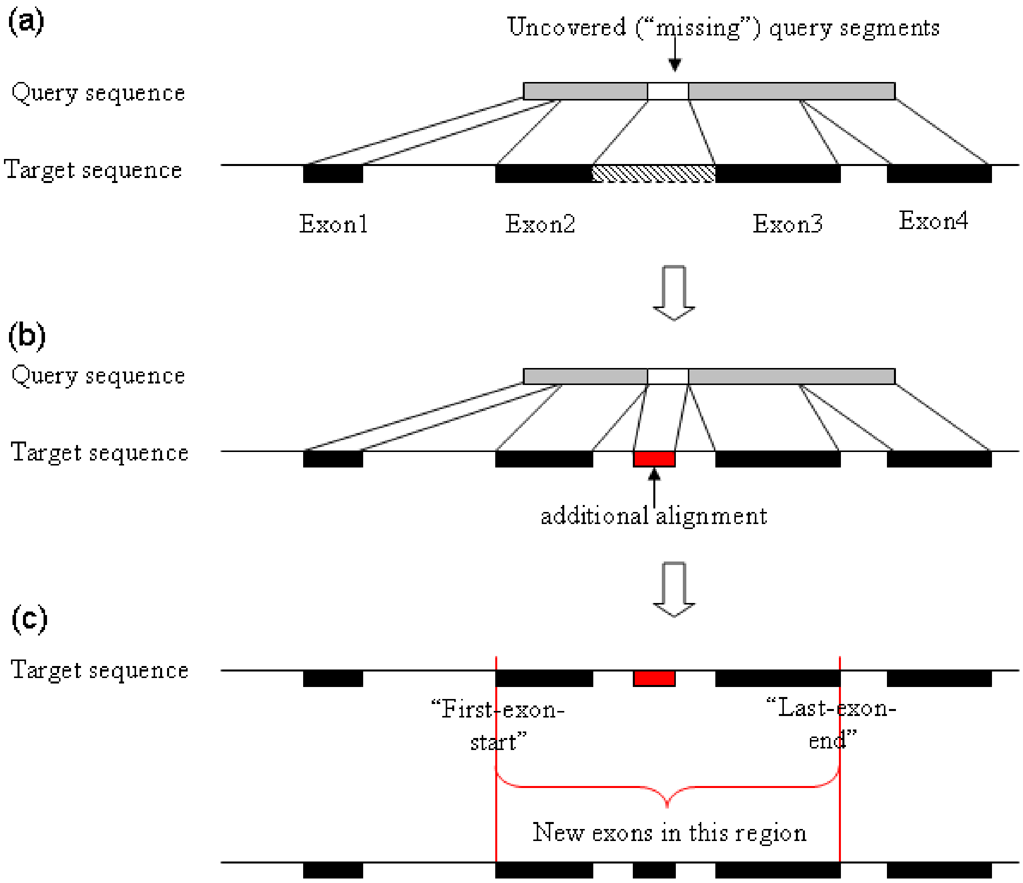


Figure 4-4 Repairing the initial gene structure for missing alignments (Challenge 3). Black boxes represents predicted exons. (a) A missing query coverage is found on the initial gene structure. The shaded region on the target DNA is searched for new local alignment that aligns with the missing query piece. (b) A new alignment (red box) is found. (c) The new alignment is used to produce a new set of exons. This figure is illustrated by R. She.

4.3.4 Genomic and protein sequence data

The genomic sequences of *C. elegans* and *C. briggsae* were downloaded from WormBase (WS200, http://www.wormbase.org). The genomic sequence of *A. thaliana* was obtained from The Arabidopsis Information Resource (TAIR9, http://www.arabidopsis.org). Query proteins for *C. elegans* and *O. sativa ssp japonica* were obtained from WormBase (WS200) and the Rice Annotation Project Database (RAP-DB build 5, http://rapdb.dna.affrc.go.jp), respectively. Proteins sequences used as queries were filtered to contain only those confirmed by cDNAs. The total number of protein used as queries for *C. elegans* and *O. sativa ssp japonica* were 6844 and 23762, respectively.

4.3.5 Program parameters for genBlastG, GeneWise, and Exonerate

genBlastG was executed to examine the top 10 genBlastA HSP groups with score > 0 (-r 10, -s 0). Analyses in Chapter 3 indicate that a positive score gives likely homologous regions. We examined top 10 groups for each query to ensure any gene family expansion in *C. briggsae* would be captured. The genBlastG parameter MIN_INTRON_REGION_LEN (see Step 1 in 4.3.3) was set to 40, which satisfied 95% of all confirmed introns in *C. elegans*.

GeneWise was executed to identify gene modes globally with only GT-AG splice signals (-init global, -splice flat). Since GeneWise also used genBlastA region as input, only one gene model was expected in that region and therefore a global search is appropriate.

Exonerate was executed using either heuristic or exhaustive search. In both situations, Exonerate only uses GT-AG splice signals to identify gene models (--forcegtag true).

4.3.6 Specificity, sensitivity, and runtime evaluation

Specificity and sensitivity is calculated at the transcript, exon, and nucleotide level. Specificity (Sp) is defined for each type of feature as how many predicted feature(s) match exactly to the confirmed annotation. The number of predicted features matching the confirmed annotation is the true positive (TP) while the number of features not in confirmed annotation is false positive (FP). Specificity is calculated as: Sp = TP / (TP+FP). For example, specificity at the nucleotide level is calculated as the number of nucleotides in the predicted model that are also part of the confirmed annotation divided by the length of the predicted model. Sensitivity is defined for each type of feature as how many annotated feature is identified in the predicted feature. The number of features in confirmed annotation that are not identified in the predicted model is false negative (FN). Sensitivity is calculated as: Sn = TP / (TP+FN). For example, sensitivity at the nucleotide level is calculated as the number of nucleotide in the confirmed annotation that are also part of the predicted model divided by the length of confirmed annotation.

Runtime evaluation is based on the number of user seconds a process takes to complete. The start and end time is obtained using the Perl “time” command.

4.3.7 Percent identity (PID) calculation

PID was calculated for each predicted model. PID is the percentage of identical amino acids over the length of the global alignment between the predicted peptide and the query. For each predicted gene model, global alignment was done using a dynamic programming script running the Needleman-Wunsch algorithm (Needleman and Wunsch 1970). Our alignment script uses a scoring matrix that maximizes PID such that a match = 1, mismatch = 0, and gap = 0.

4.3.8 Identifying and removing redundant gene models

Two models are considered overlapped if (1) the genomic span of the two gene models overlap by more than 15% of the shorter gene length and (2) more than 25% of the overlapped region are overlapping exons. All pair-wise combinations were considered when there are three or more models. Overlapping models were filtered based on the value = PID * (MAXIMUM_RANK – RANK + 1)/MAXIMUM_RANK. MAXIMUM_RANK is the "r" parameter from genBlastA/G (in this case 10) and RANK is the actual rank of the model. Whichever model has the higher value is kept and all others filtered away. All models were loaded to MySQL database and visualized with GBrowse (Stein *et al.* 2002).

4.3.9 *C. briggsae* cDNA preparation

C. briggsae worms were grown on five 6cm NGM plates with OP50 lawn until saturated (Stiernagle 2006). The worms were washed with M9 buffer and collected by centrifuge in 3000rpm for 2 minutes in 4^{o}C. Worm were then resuspended in 150µl of M9 and seeded 15 10cm plates with 15µl of worms per plate. Worms were collected by washing and centrifugation as above when saturated (usually after 3 days in 25^{o}C). We added 800µl of Trizol and 200µl of glycogen to every 1ml of packed worms and vortex for 2 minutes. To crack the worm, we froze the tube in liquid nitrogen and let it thaw completely in 42^{o}C water bath and repeated the freeze-thaw cycle again. We then added 200µl of chloroform for every 1ml of Trizol used and mixed vigorously. After centrifuge at 12,000g for 15 minutes in 4^{o}C and without disturbing the interface, we retrieved the upper aqueous phase to a fresh tube and incubated on ice. Then we added 1ml of isopropanol to every 1ml of Trizol used and mix vigorously. The solution was centrifuged at 7,500g for 5 minutes in 4^{o}C and the supernatant was carefully removed. The pellet was washed with 75% DEPC ethanol and centrifuged and decanted as before. We dissolved the pellet in equal worm volume of DEPC water. We cleaned up the RNA using RNeasy kit (Qiagen) and selected polyadenylated RNA using Poly(A)Purist (Ambion). Polyadenylated RNA was dephosphorylated with C.I.P and decapped with T.A.P. A 5' RNA adaptor was ligated to the decapped RNA and the final RNA was reverse transcribed using SuperScript III (Invitrogen).

4.4 Results

We have extensively tested the performance of genBlastG in predicting genes in the model organism *Caenorhabditis elegans* genome (*C. elegans* Sequencing Consortium 1998) and in the genome of its sister species *Caenorhabditis briggsae* (Stein *et al.* 2003). The *C. elegans* genome, which is the only animal genome that is complete without remaining gaps, is arguably the best annotated animal genome with most of its genes curated by WormBase curators as well as the *C. elegans* research community (Chen *et al.* 2005; Hillier *et al.* 2005). In contrast, the *C. briggsae* genome was more recently sequenced (Stein et al. 2003) with limited annotation. We have also tested genBlastG on the human and the Arabidopsis genome to test its performance in other species. All experiments are done on a computer with Intel Xeon E5405 2.00 GHz CPU (6M Cache) and 16G memory.

To evaluate the performance of genBlastG, we have compared it against GeneWise (Birney *et al.* 2004) and Exonerate (Slater and Birney 2005), two popular homology-based gene prediction programs, in two aspects: runtime and accuracy. We also compared Exonerate running in heuristic mode or exhaustive mode. While heuristic mode runs faster, it may not find the best fitting model that can be found in exhaustive mode (Slater and Birney 2005). To run genBlastG, we first run genBlastA to identify the genomic regions that contain candidate genes. Similarly, to run GeneWise or Exonerate, it is required to narrow the genome sequence down to the genomic regions that contain the candidate genes. Therefore, in this test, for every query gene, we run genBlastA to identify the desired genomic region(s) before running either genBlastG, GeneWise, or Exonerate. Hence, for runtime, we only measure the time used for running genBlastG, GeneWise, and Exonerate, excluding the time used for running genBlastA. For simplicity, the following experiments focus on the top-ranked region reported by genBlastA (She *et al.* 2009).

4.4.1 Performance evaluation in the same genome (*C. elegans* queries against *C. elegans* target)

First, we tested the ability of genBlastG, GeneWise and Exonerate to reconstruct the gene models using confirmed *C. elegans* queries in its own genome. We have chosen to test all confirmed genes in *C. elegans*, which contains 6,844 genes in the WS200 release. For query genes with multiple isoforms, only the longest isoform is chosen for testing. Each program was tested for sensitivity, specificity, and runtime.

Sensitivity and specificity

We evaluated each predicted gene models for sensitivity and specificity at the transcript, exon, and nucleotide levels (Burset and Guigo 1996). genBlastG outperformed GeneWise as well as Exonerate (Slater and Birney 2005) in generating gene models with high accuracy at transcript and exon level and performs similarly at the nucleotide levels (Table 4-1). The high accuracy largely comes from the high quality of HSPs returned by BLAST and genBlastG's effort of maximizing similarity to the query gene in defining exons.

Table 4-1 Sensitivity (Sn) and specificity (Sp) comparisons of genBlastG, GeneWise, and Exonerate in predicting genes in *C. elegans* using *C. elegans* proteins as queries (*n* = 6,844 genes)

	Transcript		**Exon**		**Nucleotide**	
	Sp. (%)	Sn. (%)	Sp. (%)	Sn. (%)	Sp. (%)	Sn. (%)
genBlastG	94.10	94.10	98.31	97.85	99.79	99.74
GeneWise	91.07	91.07	97.50	96.92	99.87	99.68
Exonerate (exhaustive)	93.73	93.73	98.15	97.77	99.85	99.83
Exonerate (heuristic)	91.03	91.03	97.41	96.26	99.89	99.40

Runtime comparison

To examine the effect of gene length on the runtime of different algorithms, we divided the 2,731 *C. elegans* genes on Chromosome I into five categories depending on their peptide sequence lengths. Figure 4-5 shows the average runtime of genBlastG, GeneWise, and Exonerate (heuristic and exhaustive) on genes with different lengths. genBlastG runs considerably faster than GeneWise and Exonerate, especially for longer queries. This result is expected considering the algorithms underlying these programs. GeneWise uses an exhaustive dynamic-programming approach to align the query protein to the target DNA sequence, thus it drastically slows down when the query or target sequence length increases (Birney and Durbin 2000). Similarly, Exonerate (exhaustive mode) searches all sub-optimal alignments in order to find the best fitting model (Slater and Birney 2005). In contrast, genBlastG makes extensive use of existing local alignments from BLAST HSPs and does not need to perform exhaustive sequence alignments. The main effort of genBlastG was spent on evaluating candidate splice sites by examining “spliced alignments”, which is based on existing HSPs and do not incur actual sequence alignment.

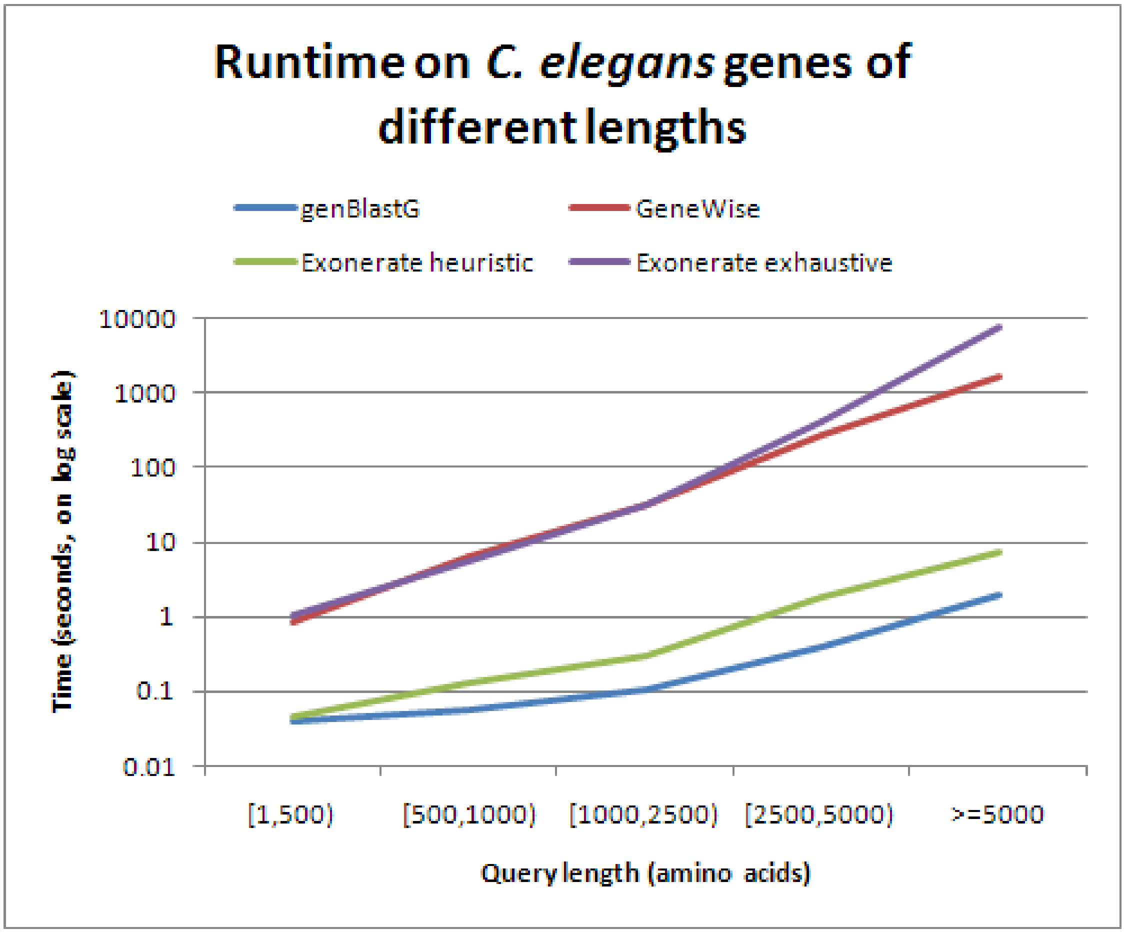


Figure 4-5 Runtime comparison between genBlastG, GeneWise, and Exonerate heuristic mode and exhaustive mode. The data used to generate this figure was collected by J. Chu and R. She.

4.4.2 Performance evaluation in the closely related genome (*C. elegans* queries against *C. briggsae* target)

To test how genBlastG performs in predicting genes in closely related genomes, we have tested its ability in predicting genes in *C. briggsae* genome (Stein *et al.* 2003) using *C. elegans* proteins as queries. These two species have been estimated to diverge from the common ancestor about 30 million years ago (MYA) (Cutter 2008), although other studies suggested that these two species could have diverged much earlier (Stein *et al.* 2003). Again, we used all fully confirmed protein sequences as queries. However, essentially all *C. briggsae* genes are predicted without experimental validation, we cannot assess specificity and sensitivity. Therefore, percentage identity (PID) between a query protein and the corresponding predicted protein was used as a measurement of the quality of predicted genes in *C. briggsae*. The average PID by genBlastG (78.51%, *n* = 6,344 genes) is similar to that by GeneWise (78.26%), but clearly higher than that by Exonerate (76.13%). genBlastG performs faster than GeneWise and Exonerate. genBlastG resolves gene structures with an average of 0.16s per gene, where as GeneWise, Exonerate (exhaustive), and Exonerate (heuristic) resolve gene structures with an average of 3.5s, 3.4s, 0.25s per gene, respectively. The scatter plot shown in Figure 4-6 indicates many gene models show higher PID (dots under the diagonal) than the gene models generated by either GeneWise or Exonerate. There are 283 cases (4.46%) where genBlastG predicts a gene model with higher PID (by more than 10% points) than GeneWise and there are 119 cases (1.87%) where GeneWise predicts a gene model with higher PID (by more than 10% points) than genBlastG. Examining some of these 119 cases revealed that many of these GeneWise gene models do not end of a stop codon. genBlastG aims to predict full length models and will extend the last exon to the nearest stop codon. By extending beyond the homologous sequences, genBlastG models resulted in lower PID. We choose to implement this functionality because recent work shows that genes can incorporate non-coding sequences to form novel genes after the stop codons were mutated (Knowles and McLysaght 2009). Experimental validation suggests that at least two genBlastG-predicted gene models that include extended sequences beyond the mutated stop codons are *bona fide* (see Section 4.5.4).

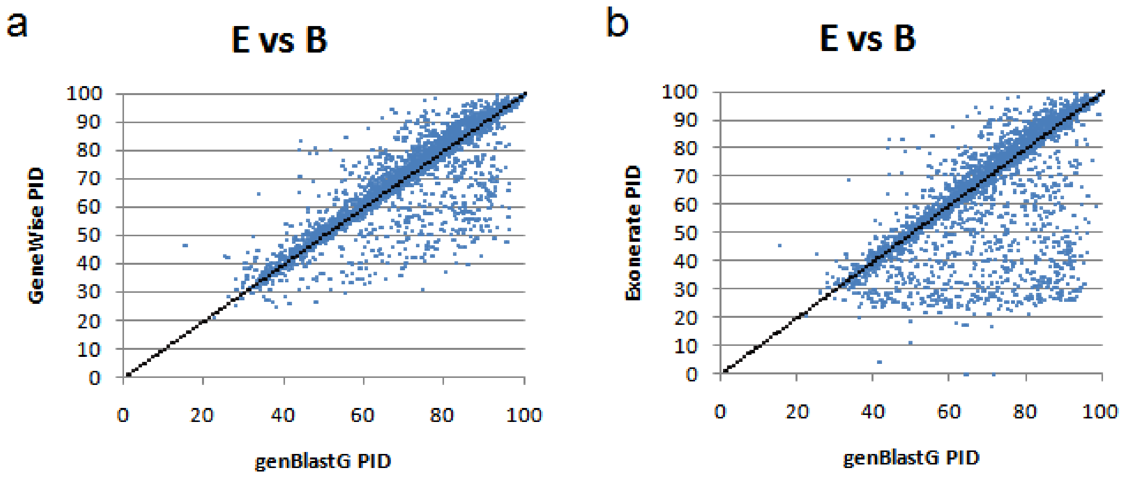


Figure 4-6 Scatter plot shows PID comparison between (a) genBlastG and GeneWise and (b) genBlastG and Exonerate. Gene models were generated using *C. elegans* protein as queries and *C. briggsae* genomic DNA as target (E vs B). Dots below the diagonal line indicate genBlsatG models have higher PID. Dots above the diagonal indicate genBlastG models have lower PID. Dots near the diagonal indicate genBlsatG models have similar PID than GeneWise or Exonerate models. The data used to generate this figure was collected by J. Chu and R. She.

4.4.3 Performance evaluation in the distantly related genome (*Oryza sativa ssp japonica* queries against *Arabidopsis thaliana* target)

genBlastG also favorably performs in predicting genes in distantly related genome. Specifically, we search genes in *Arabidopsis thaliana* genome using rice (O*ryza sativa ssp japonica*) proteins that are fully supported by cDNAs as queries. These two species diverged from their common ancestor more than 100 MYA (Itoh *et al.* 2007). We have evaluated sensitivity and specificity of 9,175 gene model predictions using 23,762 *O. sativa* proteins. As shown in Table 4-2, genBlastG clearly outperforms GeneWise and exonerate at the transcript and exon levels, while their performances at the nucleotide level are similar.

genBlastG also outperforms GeneWise and Exonerate (exhaustive) in running time. genBlastG resolves gene structures with an average of 0.39s per gene, where GeneWise and Exonerate (exhaustive) resolve gene structures with an average of 1.83s and 2.5s per gene, respectively. Exonerate using heuristics runs the fastest with an average of 0.07s per gene, but also performs the worst in terms of sensitivity and specificity.

Table 4-2 Sensitivity (Sn) and Specificity (Sp) comparisons of genBlastG, GeneWise, and Exonerate in predicting genes in *A. thaliana* using rice proteins as queries (n = 9,175 genes)

	Transcript		**Exon**		**Nucleotide**	
	Sp. (%)	Sn. (%)	Sp. (%)	Sn. (%)	Sp. (%)	Sn. (%)
genBlastG	21.10	21.10	55.56	55.90	96.27	86.72
GeneWise	12.63	12.63	48.85	47.98	97.39	89.14
Exonerate (exhaustive)	6.29	6.29	50.01	48.06	95.43	89.13
Exonerate (heuristic)	3.11	3.11	43.12	34.39	99.08	66.17

4.5 Revising the entire *C. briggsae* gene set using genBlastG

The power for comparative genomics relies largely on the accurate annotation of the gene sets. The *C. elegans* research community and the WormBase curators have actively curated the C. elegans genome since its publication about a decade ago (C. elegans Sequencing Consortium 1998; Chen *et al.* 2005; Hillier *et al.* 2005). In the WS200 release, 85.2% of all *C. elegans* genes either are confirmed or partially confirmed (http://www.wormbase.org/wiki/index.php/WS200). In contrast, very few *C. briggsae* genes have been scrutinized since its publication (Stein *et al.* 2003) and, not surprisingly, many gene models have been found to be defective. Recently, the nGASP effort has attempted to improve the *C. briggsae* gene set (Coghlan *et al.* 2008). Despite the nGASP effort, many genes are still obviously defective when they are compared with their orthologous genes in *C. elegans*. We have attempted to revise *C. briggsae* gene models based on their homology to *C. elegans* genes by using genBlastG.

We used PID to evaluate the quality of the predicted gene models, assuming that the gene model that shows the highest PID to its query gene is more likely to be the better gene model. To compare genBlastG gene models with the current WormBase gene models and the nGASP gene models, we plotted the PID of genBlastG gene models against that of the corresponding WormBase gene models and the nGASP gene models (Figure 4-7), respectively. A point on the diagonal line (shown as the dotted line) indicates that the genBlastG gene model is similar (in terms of PID) to a WormBase or a nGASP gene model, while a point below (or above) the diagonal line indicates that a genBlastG gene model has higher (or lower) PID than that of a corresponding WormBase/nGASP gene model. As indicated in Figure 4-7, PIDs of most gene models generated using genBlastG are similar (9,545, or 83%) to those of the WormBase gene models. However, a considerably large number of genBlastG gene models (1,805, or 16%) show higher PID by at least 10% compared to the current WormBase *C. briggsae* gene models. In contrast, only < 1% (44) of the genBlastG models show lower PID by at least 10% to the current WormBase *C. briggsae* models (Figure 4-7a). Comparing to the results from nGASP *C.*

briggsae gene models, 83% (9,319) of the genBlastG models are similar, 17% (1,887) are better by at least 10% and with only < 1% (36) of the gene models show lower PID (Figure 4-7b). To be conservative, we propose to use genBlastG gene models that show at least 10% improvement in PID than their corresponding WormBase gene models to replace the current *C. briggsae* gene models in WormBase. The comparison of *C. briggsae* gene models generated using genBlastG with the current WormBase gene models suggests that there are four categories of changes: (1) gene model split, (2) gene model merge, (3) gene model trimming/extension, and (4) internal exon alteration (Figure 4-8).

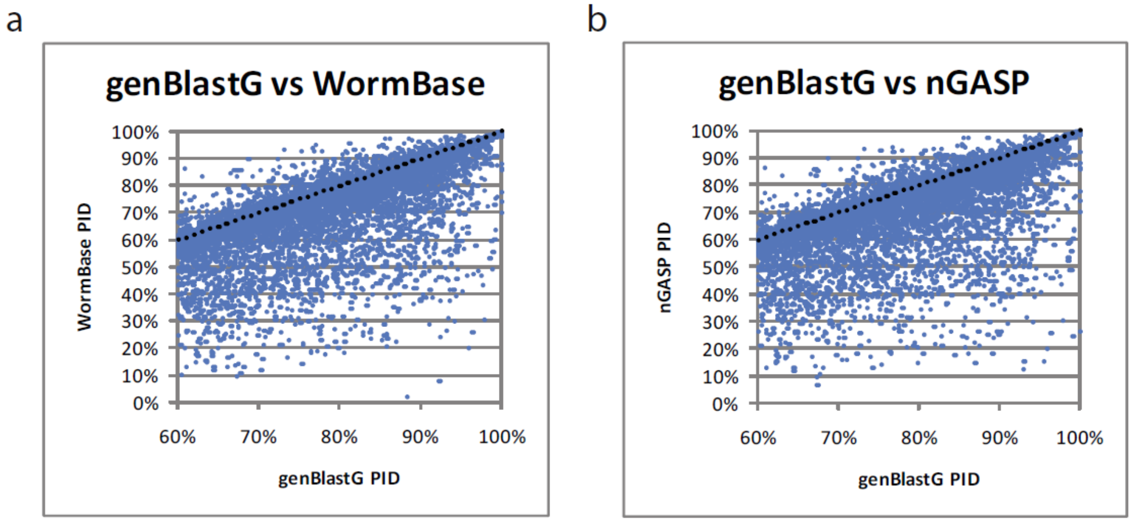


Figure 4-7 PID scatter plots for different types of gene models. (a) genBlastG models vs. current WormBase models. (b) genBlastG models vs. nGASP models. Each dot represents the PID of genBlastG model and WormBase/nGASP model of one query gene. The dots that are at or close to the diagonal represent genBlastG models that have comparable similarity to WormBase models or nGASP models. Dots that are below the diagonal represent genBlastG models that have higher similarity compared to WormBase or nGASP models. Dots that are above the diagonal represent genBlastG models that have lower similarity compared to WormBase or nGASP models. The data used to generate this figure was collected by J. Chu and R. She.

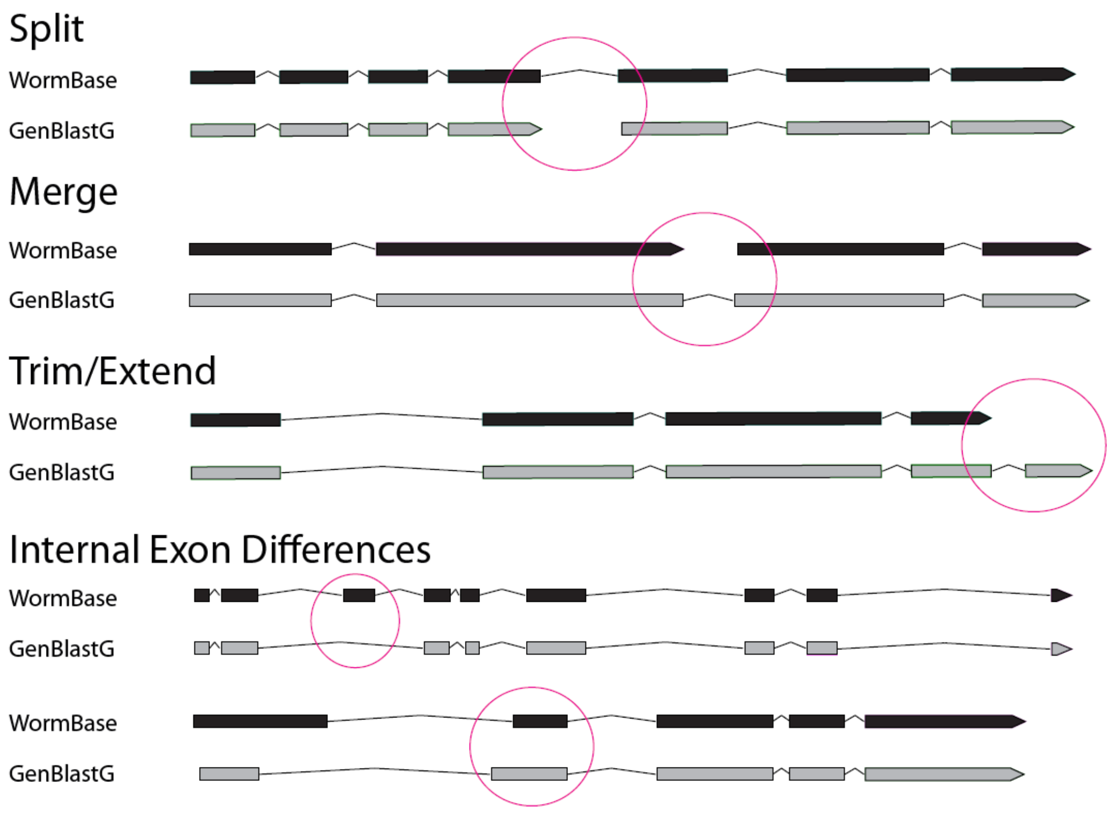


Figure 4-8 Four categories of gene structure differences between the two genomes. Split: A current WormBase gene model that should be split into two different gene models; Merge: Two current WormBase gene models that should be merged into one single gene model; Trim/Extend: A current WormBase gene models that should be extended or trimmed at the ends; Internal Exon differences: internal exons should be added/removed. This figure was illustrated by J. Chu.

4.5.1 Gene model split suggested by genBlastG.

In the WS200 WormBase annotation of *C. briggsae*, many gene models are in fact the false merge of separate adjacent gene models. According to the homology to their corresponding *C. elegans* gene models, these gene models should be split to 2 or more gene models. Altogether, we have found 398 such cases. Some (158) of these cases have been fixed by the nGASP project (Figure 4-9a), while many (240) have not (Figure 4-9b). After the split, new gene models show significantly improved PIDs to their corresponding query genes. This is not unexpected since unrelated sequences are removed. Figure 4-9c and Figure 4-9d indicates that nearly all split cases show improved PID with 340 cases showing at least 10% improvement. The average PID of genBlastG models from these split cases is 77% whereas WormBase is 52% and nGASP is 60% ($p < 0.0001$). Experimental validation by PCR amplification for the two examples shows that while individual gene models can be PCR-amplified, the whole length suggested by WormBase cannot be PCR-amplified despite numerous attempts (Figure 4-9e and f).

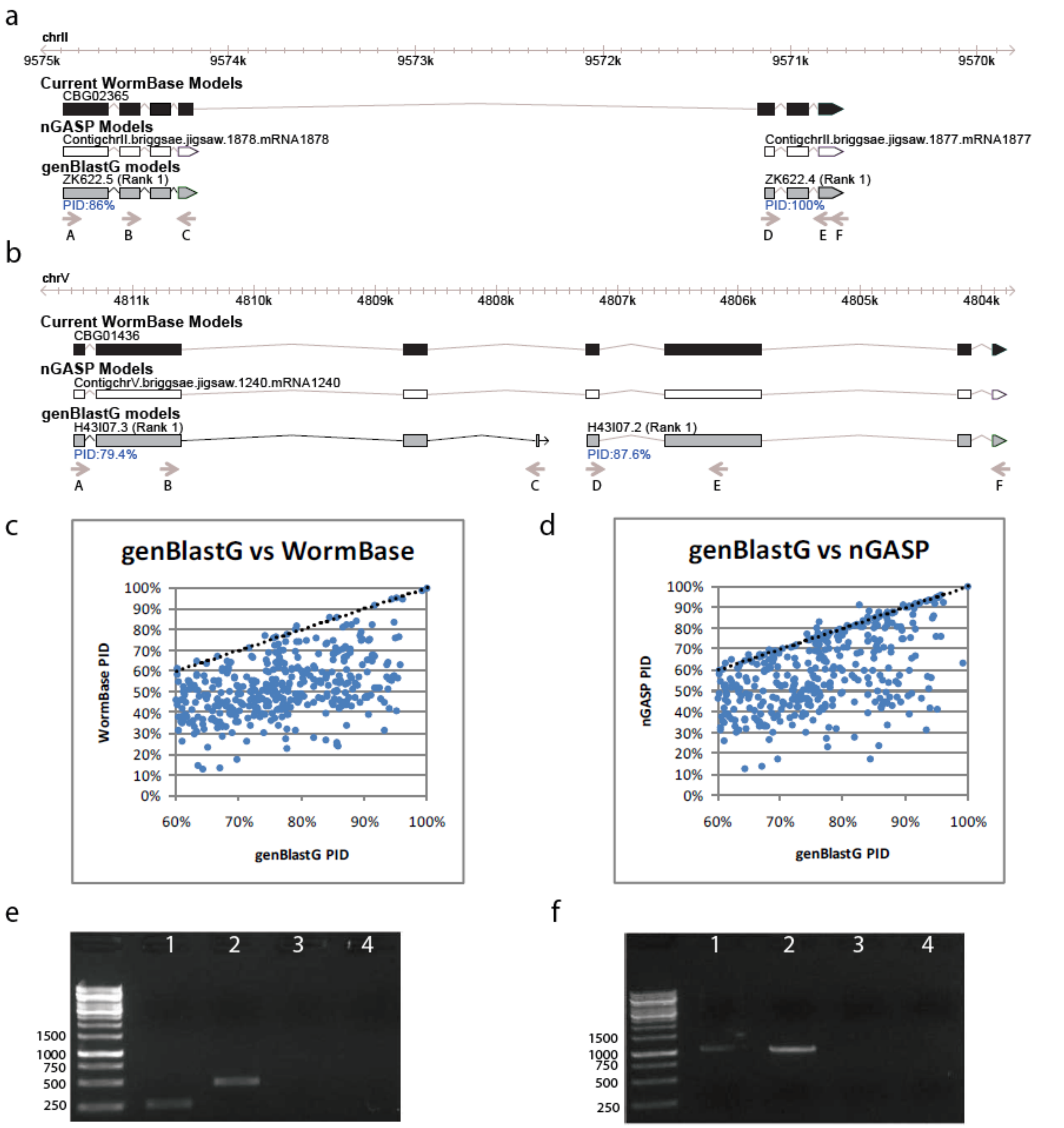


Figure 4-9 Representative gene model split cases. (a) WormBase model CBG02365 is split to two models (ZK622.5 and ZK622.4) with more than 85% PID each. nGASP predictions support this split. (b) WormBase model CBG01436 is split into two models (H43I07.3 and H43I07.2) with more than 75% PID for each. This case was not fixed by nGASP. The labeled arrows in panel (a) and (b) indicate the primer positions for PCR verification. (c) Scatter plot showing PID comparison between genBlastG and WormBase. (d) Scatter plot showing PID comparison between genBlastG and nGASP. (e) The split CBG02365 into two separate genes shown in (a) is experimentally validated. Numbers in prentices indicate expected band size if there is any. Lane 1: D→F (263bp); Lane 2: A→C (486bp); Lane 3: A→F (762bp); Lane 4: B→E (516bp). Expected bands are shown in lanes 1 and 2, while no bands are found in lanes 3 and 4, supporting the gene model split. (f) The split of CBG01436 into two separate genes shown in (b) is experimentally validated as well. Lane 1: D→F (1004bp); Lane 2: A→C (1044bp); Lane 3: A→F (2081bp); Lane 4: B→E (860bp). Expected bands are observed in lanes 1 and 2, while no bands are found in lanes 3 and 4, supporting the gene model split. The numbers beside panel (e) and (f) indicate ladder sizes in bp. The data used to generate this figure was collected by J. Chu, R. She, and J. Wang.

4.5.2 Gene model merge suggested by genBlastG.

genBlastG predicts a number of gene models in the current WormBase annotation of *C. briggsae* may have been erroneously split into two or more gene models. Based on the homology to their corresponding genes in *C. elegans*, these gene models should be merged to form a single gene model. We have found 239 merge cases in the current WormBase models. Some of these cases (40) have been fixed by the nGASP project (Figure 4-10a), but many (199) have not (Figure 4-10b). Nearly all merge cases show improved PID with 179 cases showing at least 10% improvement (Figure 4-10c and d). The PID of genBlastG gene models on average is 79% whereas the average PIDs for WormBase and nGASP models are 58% and 60%, respectively ($p < 0.0001$). Experimental validation indicates that many merges are real since the full length or the junction can be amplified from a *C. briggsae* cDNA library (Figure 4-10e and f).

The mRNA transcripts of many eukaryotic genes are alternatively spliced to generate different protein coding transcript from a single gene. Our current analysis only focused on transcript that gives the longest protein sequence. However, genBlastG can be applied to predict the structure of different isoforms. We observed in our analysis that predicting different isoforms can improve gene annotation quality such that different isoforms merges different WormBase gene models. Figure 4-11 shows an example of 3 small gene models (CBG26024, CBG20187, and CBG20185) should be part of a larger model (CBG20190) in separate isoforms. This example shows that having isoform information will improve gene model prediction dramatically especially to the single exons genes.

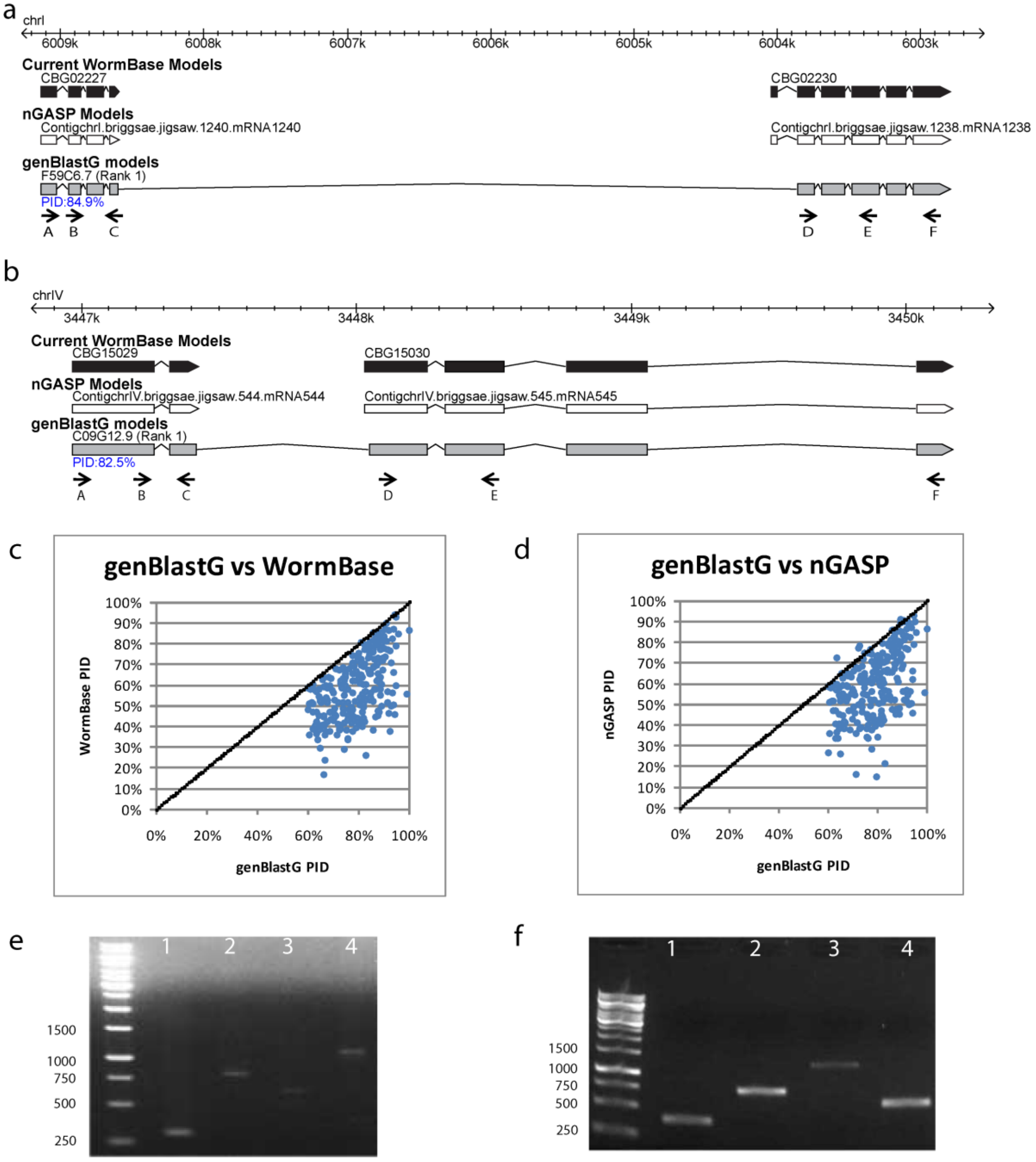


Figure 4-10 Representative gene model merge cases. (a) CBG02227 and CBG02230 are merged into F59C6.7 with improved PID. This gene model merge is also supported by the nGASP annotation. (b) CBG15029 and CBG15030 are merged into C09G12.9 with improved PID. This gene model merge was not fixed by nGASP. (c) Scatter plot showing PID comparison between genBlastG predictions and WormBase gene models. (d) Scatter plot showing PID comparison between genBlastG predictions and nGASP gene models. (e) Merging of two genes CBG02227 and CBG02230 shown in (a) is experimentally validated. The number in prentices indicates expected band size. Lane 1: A→C (303bp); Lane 2: D→F (806bp); Lane 3: B→E (632bp); Lane 4: A→F (1121bp). All four lanes show bands of expected sizes, supporting the merge. (f) Merging of two genes CBG15029 and CBG15030 shown in (b) is experimentally validated. Lane 1: A→C (383bp); Lane 2: D→F (744bp); Lane 3: A→F (1220bp); Lane 4: B→E (555bp). All four lanes show bands of expected sizes, supporting the merge. The numbers beside panel (e) and (f) indicate ladder sizes in bp. The data used to generated this figure was collected by J. Chu, R. She, and J. Wang.

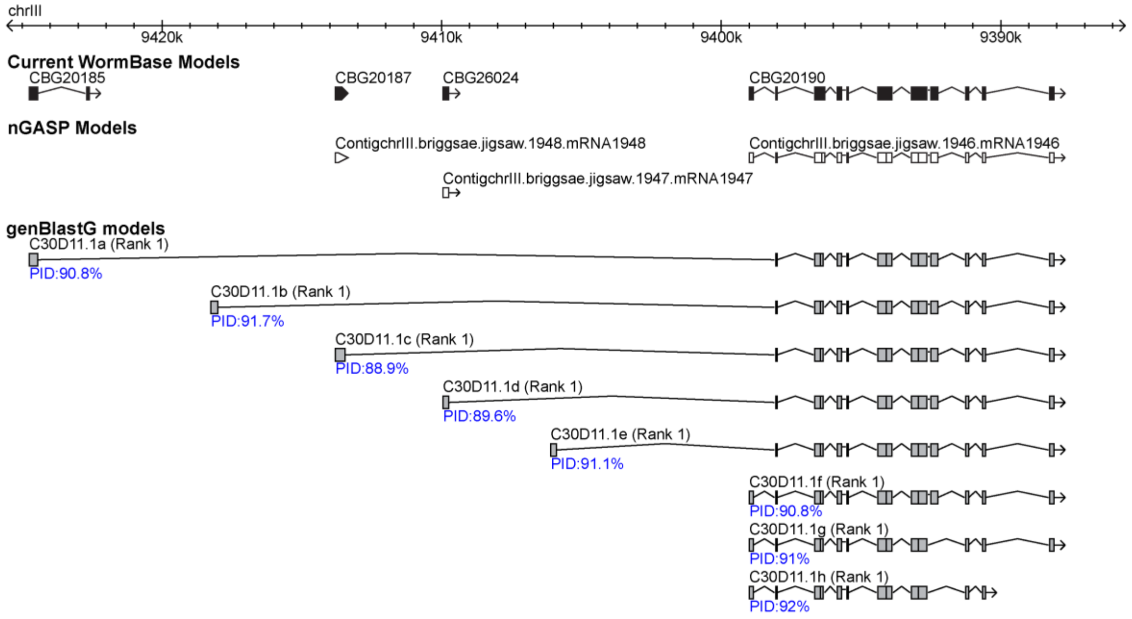


Figure 4-11 Some WormBase gene models of *C. briggsae* are in fact different isoforms of a same gene. Gene models CBG20185, CBG20187, CBG206024, and CBG20190 in fact represent components of one single gene. The genBlastG models represent putative isoforms of one *C. briggsae* gene. The data used to generate this figure was collected by J. Chu.

4.5.3 Gene model trimming/extension suggested by genBlastG.

The 5' and 3' ends of many current WormBase gene models are defective, according to the homology to their corresponding *C. elegans* genes. In these cases, their ends need to be trimmed or extended (Figure 4-12a and b). We found 3,825 such cases with genBlastG average PID at 78% and WormBase average PID at 70% (p < 0.0001). Out of 3,825 cases, 3,544 nGASP models show different start and end positions. The average PID for these nGASP models is 69% (p < 0.0001). Overall, we also see genBlastG models are more similar to *C. elegans* genes in comparison to WormBase models and nGASP models with 1,032 cases showing at least 10% PID improvement (Figure 4-12c and d). In some cases, changing the start or the end positions leads to a change of reading frames and thus creating an entirely different protein sequence. For example in Figure 4-12b, the exon frames for F43G9.13 are 2, 3, 3, 1, and 3 and the exon frames for CBG12505 are 1, 1, 2, and 1. The change in reading frame produces an entirely different protein sequence that dramatically improved similarity to C. elegans query.

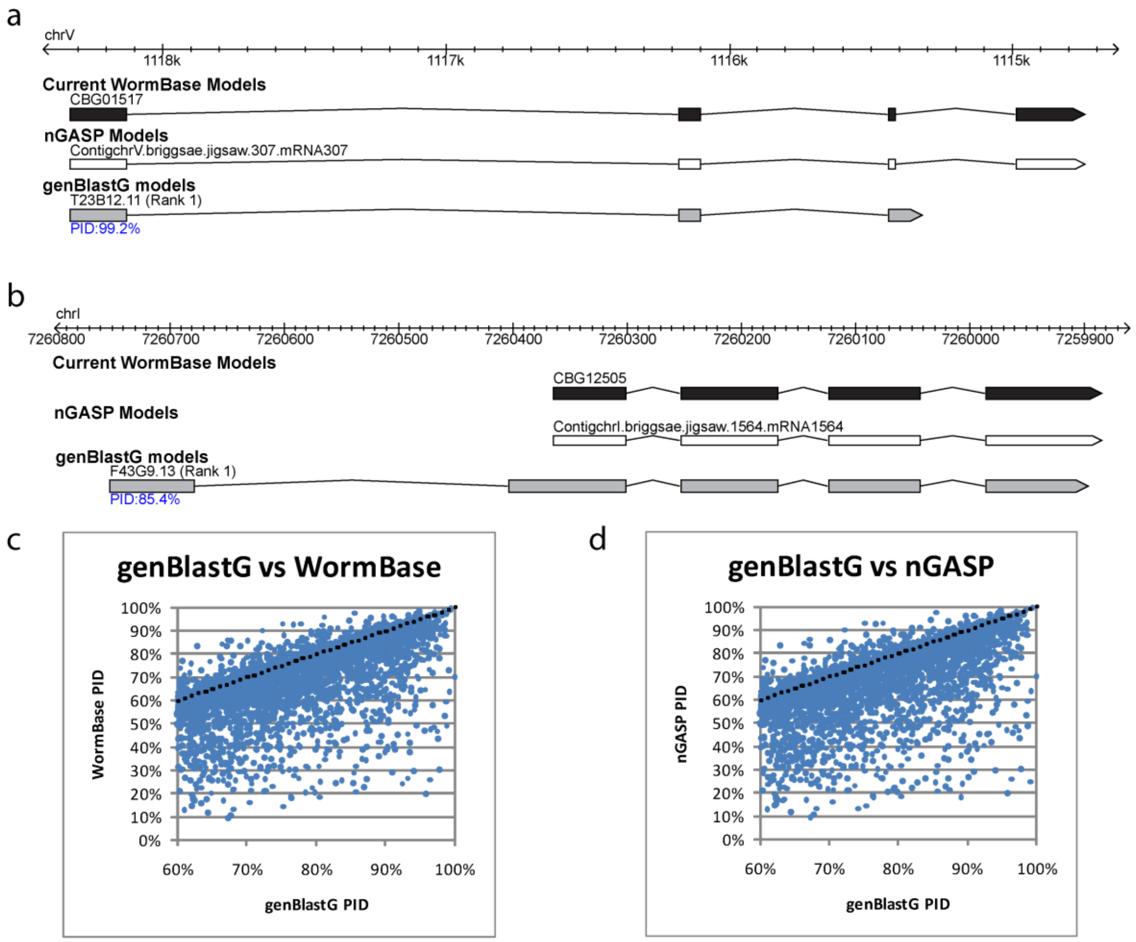


Figure 4-12 Representative cases of gene model trimming/extension. (a) The gene model for CBG01517 is predicted by genBlastG to be shorter based on its similarity to the *C. elegans* gene T23B12.11. (b) The extra predicted exon indicates CBG12505 should be extended based on the similarity to F43G9.13. Both examples were not fixed by nGASP. (c) Scatter plot showing PID comparison between genBlastG and WormBase. (d) Scatter plot showing PID comparison between genBlastG and nGASP. The data used to generate this figure was collected by J. Chu and R. She.

4.5.4 genBlastG reveals *de novo* gene formation in *C. briggsae*.

genBlastG has been designed to identify gene models that extend beyond homologous genomic regions to identify appropriate start and/or stop codons. In contrast, GeneWise chooses to predict incomplete gene models that end where homologous regions end, without attempting to predict start and/or stop codons. In our analysis, we have found special cases of gene models extensions at the 3' ends. Comparative examination of the ends of the gene models indicates gene models predicted by GeneWise do not end at canonical stop codons while genBlastG models do. Our experimental validation of two such cases suggests that these longer models are most likely true (Figure 4-13). For example, Y105E8B.6 is 216 bp in *C. elegans* (query) while genBlastG prediction and experimental validation show the *C. briggsae* transcript size to be about 300 bp. These longer models likely resulted from a mutation that removed stop codon in *C. briggsae*, or from a mutation that generated a new stop codon in *C. elegans*.

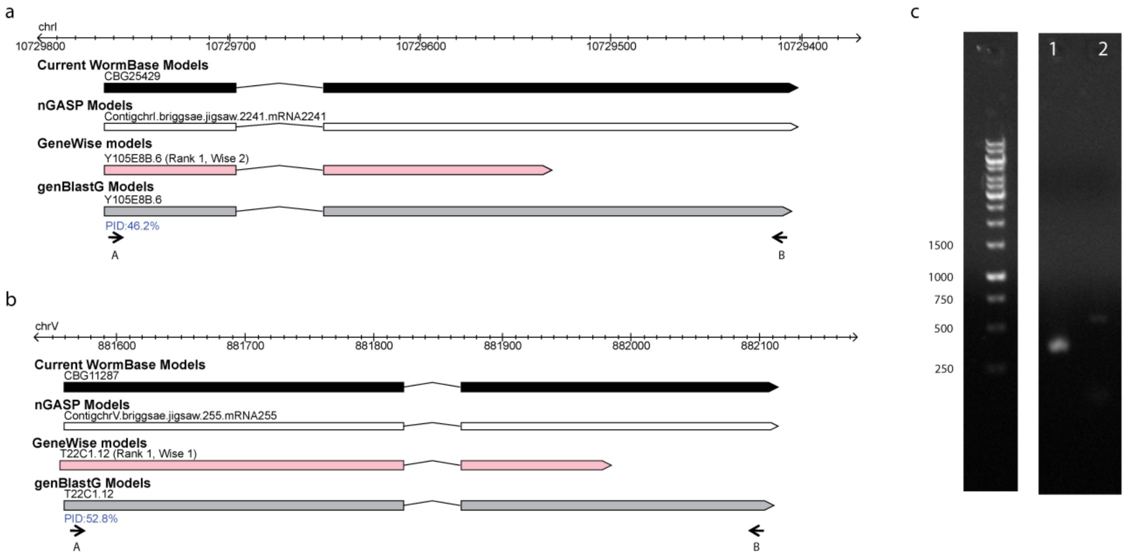


Figure 4-13 Representative novel genes in *C. briggsae* generated by the mutation of the stop codon. (a and b) Two examples of *de novo* gene formation likely risen from mutated stop codon. In both examples, GeneWise predictions will terminate even though no stop codons are found. genBlastG attempted to extend the gene models to stop codons. The arrows and labels indicate primers used for PCR validation. (c) Both gene models are experimentally validated, suggesting that these two C. briggsae gene models recruited non-coding sequences as coding exons. Lane 1: Y105E8B.6 (301bp), Lane 2: T22C1.12 (482bp). The data used to generated this figure was collected by J. Chu and J. Wang.

4.5.5 Internal exon differences suggested by genBlastG.

Internal exon differences are cases where the exons in between are either in different length, missing, or have extra exon(s). We found 4,594 cases where genBlastG has different internal exons from current WormBase *C. briggsae* models and of those, 692 cases showing at least 10% PID improvement and 3,876 cases showing differences within 10%. genBlastG has an average PID of 79% and WormBase 74% ($p < 0.0001$). Of the 4,594 cases, 4,489 nGASP models still contain differences in internal exons with an average PID of 73% ($p < 0.0001$). Figure 4-14 shows an example where an extra exon is predicted by genBlastG and thereby improved the model PID with the *C. elegans* query. Even over all cases, we find that genBlastG models are mostly equally similar or more similar to the *C. elegans* query sequence (Figure 4-14b and c). Figure 4-14d shows a ~400bp band representing the expected size of the revised gene model. We further verified two other cases (CBG16922 and CBG06025) and found both to be as what genBlastG predicted.

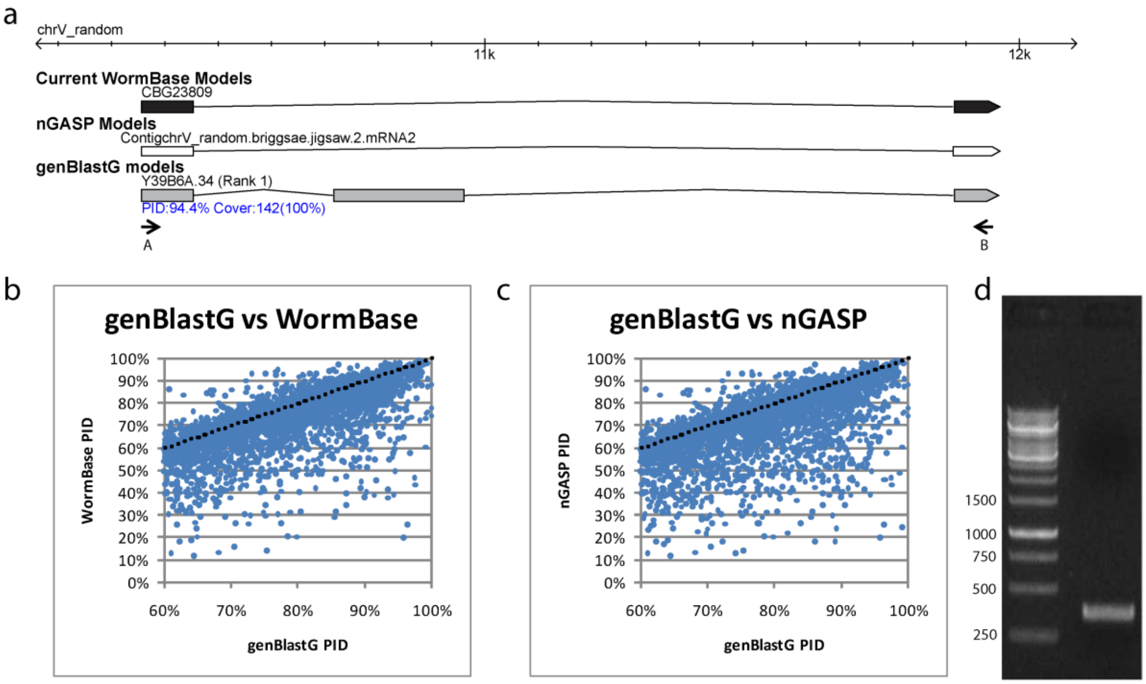


Figure 4-14 Representative case of internal exon changes. (a) An example of internal exon difference where genBlastG predicts an extra exon and improves PID. (b) Scatter plot showing PID comparison between genBlastG and WormBase. (c) Scatter plot showing PID comparison genBlastG and nGASP. (d) The addition of the exon shown in (a) is experimentally validated. The band is amplified using primer A and B to produce an expected band size of 364bp. The numbers beside panel (d) indicate ladder sizes in bp. The data used to generate this figure was collected by J. Chu, R. She, and J. Wang.

4.5.6 Missed gene models predicted by genBlastG.

We also observed 85 genBlastG gene models that do not overlap with any current WormBase gene models in the *C. briggsae* genome, suggesting that these are the gene models that were entirely missed in the current WormBase annotation. Among them, nine have also been found by nGASP independently. All of these models show 60% or higher PID with their *C. elegans* queries. Among the 85 cases, the shortest gene model is 105 bp and the longest gene model is 1,407 bp. On average, the models show 72% PID and 375bp in length. Figure 4-15 shows four examples with each case validated using PCR amplification from a cDNA library. Based on homology-based gene prediction, our results here suggest that there are many models are still missing in the current WormBase annotation.

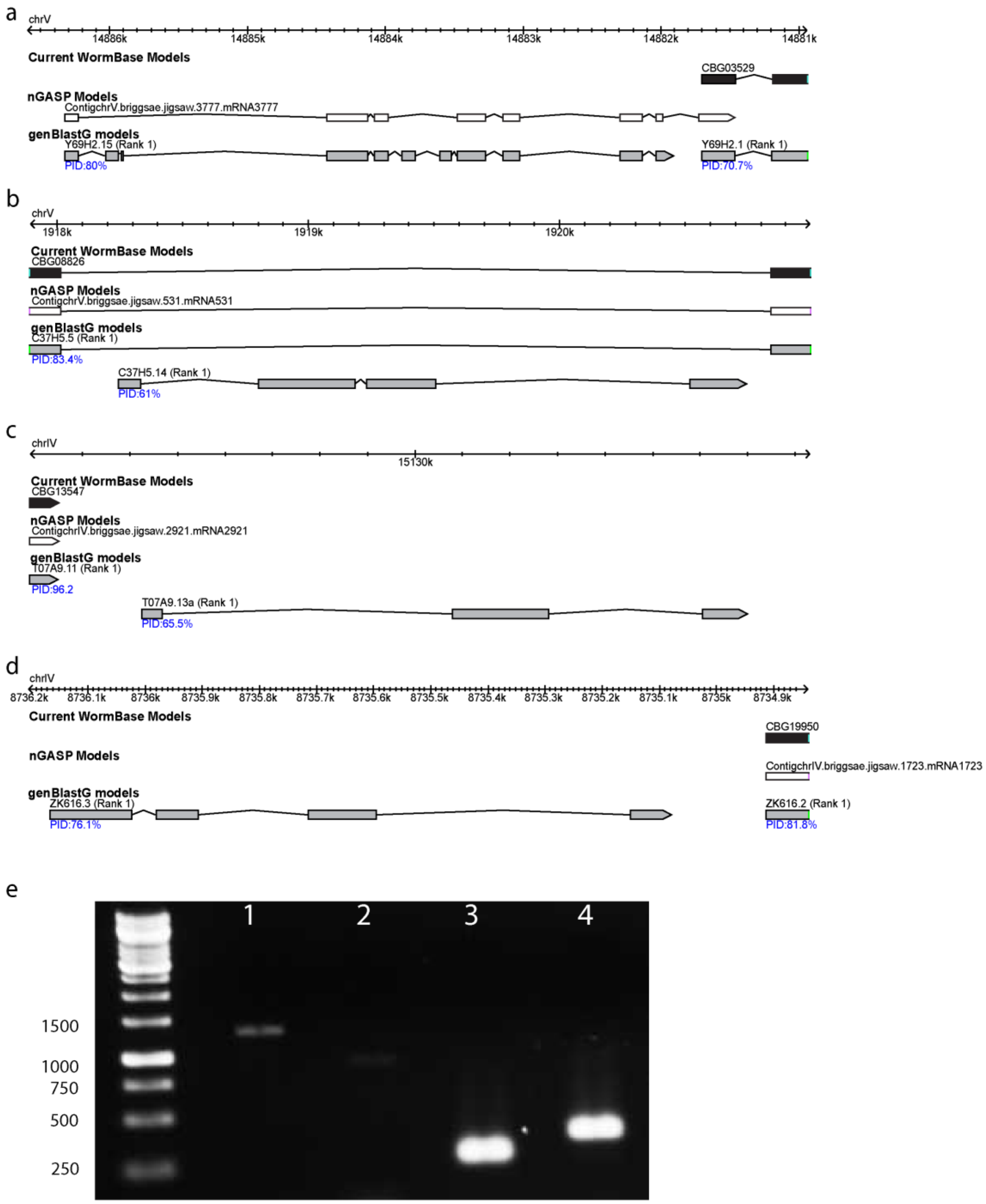


Figure 4-15 Examples of missed gene models predicted by genBlastG. (a-d) Gene models predicted using the query Y69H2.15, C37H5.14, T07A9.13, and ZK616.3. (e) All four missed gene models are experimentally validated. The lanes and the expected band size are as follows: Lane 1: Y69H2.15 (1,279bp), Lane 2: C37H5.14 (910bp), T07A9.13 (292bp), and Lane 4: ZK6.6.3 (356bp). The data used to generate this figure was collected by J. Chu and J. Wang.

4.5.7 Predicting genes in the human genome

Although genBlastG was originally developed to annotate gene models in the newly sequenced *Caenorhabditis* species including, it can be applied to annotate homologous genes in other species. We have applied it to predict genes in the human genome using human proteins as queries. The human genome and protein sequences are downloaded from the ENSEMBL database (http://www.ensembl.org/) (Hubbard *et al.* 2009). To compare the performance of genBlastG and GeneWise in predicting genes in human genome, we randomly selected 75 peptide sequences as queries and run both genBlastG and GeneWise to predict gene models in the human genome. About two-thirds of these 75 peptides range between 250 aa and 750 aa in size. Five peptides are longer than 1,000 aa. For genBlastG, the runtime of all 75 queries are all below 2 seconds per query, with the average time of less than 1 second. In contrast, runtime of GeneWise range from tens of seconds to thousands of seconds, depending on the query length, with the average runtime of 457 seconds. The result resembles that obtained from experiments on worm genomes. Similar to the experiments on the *C. briggsae* genome, accuracy comparison between genBlastG and GeneWise in human genome is also based on the alignment PID between the predicted model and the corresponding query. On average, PIDs of GeneWise models are slightly lower (95% vs. 93%, p = 0.03, paired-Student t-Test).

4.6 Conclusion

In this section of my thesis, I described a novel gene prediction tool, genBlastG, which exploits homology information contained in BLAST HSP alignments to identify gene structures. Our experiments show that genBlastG achieved better accuracy and speed than GeneWise and Exonerate, which are two popular homology-based gene prediction programs. genBlastG, GeneWise, and Exonerate are homology-based programs that predict genes based on protein alignments. The speed of genBlastG is largely due to its simpler model, with most of the alignment information taken directly from HSPs that are already obtained by BLAST, whereas GeneWise and Exonerate needs to align the protein using complex HMM models and build everything from scratch. On the other hand, there is still plenty of room for improvements, since our current method only utilizes the homology evidences from BLAST HSPs. We believe that its performance can be further improved by integrating more evidences, such as more sophisticated splice site detection models, for example, by using base pair and codon composition information in detecting splice sites.

Using genBlastG, we have demonstrated that many *C. briggsae* gene models in the current WormBase may be defective. We have also demonstrated that at least 1,805 gene models predicted using genBlastG can be used to replace the current *C. briggsae* gene models in WormBase. These gene models show much higher PID to their homologous *C. elegans* genes. Experimental validation from some of these genBlastG gene models suggests these 1,805 gene models are likely true. Many additional gene models in *C. briggsae* could be revised based on genBlastG predictions after more careful examination. In addition to predicted novel gene models, we found cases where the gene length is extended in *C. briggsae* in comparison to its *C. elegans* ortholog. Two cases that were selected for experimental validation show that the extended gene models are real (Figure 4-13). These extended gene models likely arose due to mutated stop codons. Similar phenomenon was observed in some mammals where novel genes were formed by incorporating non-coding DNA and became longer in humans in comparison to their orthologs in chimpanzee, gorilla, gibbon, and macaqueas (Knowles and McLysaght 2009).

In summary, we have demonstrated that genBlastG can be used to identify high-quality gene models based on homology between two *Caenorhabditis* species. More than one thousand gene models can be revised. We anticipate that genBlastG will also be useful as a homology-based gene prediction program for genes in other genomes.

5: IDENTIFICATION OF DAF-19 REGULATED GENES WITH DIVERGENT X-BOX MOTIFS USING COMPARATIVE GENOMICS

Note regarding contributions:

In this study, I performed all bioinformatics analyses including comparative genomics search and X-box motif detection. I also performed most of the molecular biology experiments. I designed all promoter::mCherry fusion constructs. I also performed all screening for Mos integrants, genetic crosses, and expression pattern analysis. M Tarailo-Graovac assisted with the initial Mos integrant screening and genetic crossing. J Wang and J Trinh assisted in vector cloning and *E. coli* transformation. D Tu performed all microinjection procedures. B Uyar assisted with gene prediction and annotation in *Caenorhabditis* species. D Zhang analyzed transcriptome data. DL Baillie and N Chen conceived the study.

5.1 Abstract

RFX transcription factors (TFs) play important roles in cilia biogenesis and maintenance. Many of their target genes are associated with disease conditions. RFX TFs bind to X-box motifs. However, the consensus X-box motif (a 14 bp consensus) generated based on validated instances may not adequately represent all functional X-box motifs since “typical” X-box motifs are not found in many human ciliary genes. We hypothesize that some functional X-box motifs are divergent from these validated X-box motifs and that these motifs are missed in consensus-based computational searches. To test this hypothesis, we compared the gene sets between *C. elegans*, *C. briggsae*, *C. remanei*, and *C. brenneri* to find genes with conserved X-box motif residing 500 bp upstream of the start codon in all orthologous genes except in *C. elegans*. We identified 10 *C. elegans* genes that satisfy this criterion. One of these genes is F25B4.2 that contains two putative divergent X-box motifs: a distal and a promixal motif. We observed F25B4.2 expression in ciliated neurons that is driven by the proximal motif but repressed by the distal motif. Our data suggest that two divergent X-box motifs cooperate to regulate the expression of F25B4.2 in location and intensity. This is the first report to discover a potential repressive X-box motif in *C. elegans*. We postulate that regulation via two X-box motifs may be a general regulatory mechanism used by RFX transcription factors. Our identifications of divergent X-box motifs will also improve our understanding on RFX/DAF-19-mediated regulation in *C. elegans* and in other organisms including humans.

5.2 Introduction

Regulatory Factor X (RFX) is an evolutionarily conserved DNA binding protein family that has been identified in organisms ranging from single cellular eukaryotes, including the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe* to humans (Emery *et al.* 1996). All RFX transcription factors (TFs) contain a single DNA binding domain, which is very well conserved, showing about 40% identity between yeast, nematodes, and mammals and perfect identity in nucleotide positions that are in direct contact with RFX DNA binding domains (DBD) (Gajiwala *et al.* 2000; Chu *et al.* 2010). In humans as well as all other mammals, seven RFX transcription factors (TF) have been uncovered, including RFX6 and RFX7 that were recently identified in our laboratory (Aftab *et al.* 2008). In mammals, RFX1, 2, and 3 were found to function in ciliated cells of the kidney (Haycraft *et al.* 2001; Boito *et al.* 2005) and brain (Ma *et al.* 2006). Mutations in these genes and their target genes have been associated with an expanding array of devastating human disease conditions, including polycystic kidney disease (Haycraft *et al.* 2003; Praetorius and Spring 2005) and Bardet-Biedl syndrome (Badano *et al.* 2006). In addition to their role in regulating genes in kidney and brain, RFX genes have been found to play important roles in other human tissues as well. RFX5 has been demonstrated to be important in regulating Major histocompatibility complex class II (MHC II) gene expression in the immune system (Garvie *et al.* 2007). RFX6, which we recently identified in the human genome, is almost exclusively expressed in the human pancreatic islets (Aftab *et al.* 2008). Subsequent studies from other groups have further confirmed the role of RFX6 in human pancreas (Smith *et al.* 2010; Soyer *et al.* 2010).

Accumulating evidence suggests that RFX genes regulate the transcription of ciliary genes in metazoans, albeit analysis on the molecular evolutionary relationship between RFX TFs and ciliary genes showed that RFX TFs and ciliary genes evolved independently before the establishment of metazoans (Chu *et al.* 2010; Piasecki *et al.* 2010). The first critical evidence linking RFX TFs and the ciliary genes was reported by Swoboda and colleagues (Swoboda *et al.* 2000). The authors cloned *daf-19* in the nematode *Caenorhabditis elegans* and found that it is the first and only RFX gene in *C. elegans*. They showed that in the absence of functional DAF-19, ciliated neurons in *C. elegans* lost their cilia and displayed chemosensory defects (Che), dye filling defect (Dyf), and constitutive dauer formation (Daf-c) (Swoboda *et al.* 2000). Furthermore, they demonstrated that DAF-19 regulates the expression of ciliary genes, including *che-2*, *osm-1*, *osm-6* and many Bardet-Biedl Syndrome (BBS) genes through binding to a DNA element called the X-box motif, which was first identified as binding site for human RFX5 (Emery *et al.* 1996; Swoboda *et al.* 2000). Later, it was discovered that many ciliary genes in the fruit fly *Drosophila melanogaster* are also regulated by RFX genes, including *CG15161*, the homolog of *dyf-6* (Laurencon *et al.* 2007).

X-box motif, the binding motif of RFX DBDs, has been found to be highly conserved as well. Many validated instances of X-box motifs in yeast, *C. elegans*, and humans are 14 bp in size. Because of their large size, X-box motifs have been used as a ciliary gene indicator in genomics and bioinformatics projects. Efimenko and colleagues searched the *C. elegans* promoter regions (defined here as 1000 bp genomic region upstream of the start codon) for candidate X-box motifs that resemble an “average consensus X-box motif” and identified 730 potential DAF-19 target genes in *C. elegans* (Efimenko *et al.* 2005). Independently, Blacque and colleagues identified 53 putative DAF-19 target genes in *C. elegans* through searching for the presence of putative X-box motifs in promoter regions (defined as 1500 bp genomic region upstream of the start codon) and comparing relative gene expression in four different tissues (Blacque *et al.* 2005). Taking advantage of the availability of two newly sequenced genomic sequences in *Caenorhabditis* genus, *C. briggsae* and *C. remanei*, our laboratory searched for X-box motifs in the promoter regions (defined as 2000 bp genomic region upstream of the start codon) of orthologous genes in all three species and predicted 93 candidate DAF-19 regulated genes (Chen *et al.* 2006), including *dyf-5*. The putative X-box motifs identified in these three studies all show resemblance to known X-box motifs.

An important caveat of these projects is that these homology-based searches of X-box motifs bias against the discovery of more diverged X-box motifs. Since these studies identified *bona fide* DAF-19 regulated genes by relying heavily on the known cases (consensus) of X-box motifs, which are mostly 14 bp long, they could also miss genuine X-box motifs that may show significant differences from that consensus. For example, the X-box motif found in the promoter of *nph-1* has 15 bp, which was missed in all three projects

(Winkelbauer *et al.* 2005). In fact, it was found that X-box motifs can vary in length because X-box motifs can contain a variable sequence in the middle spacer region (Emery *et al.* 1996; Lubelsky *et al.* 2005). These motifs cannot be found in consensus based searches. This situation can be more severe in humans. For example, the HMM profile built for searching the *C. elegans* genome, which retrieves many ciliary genes in *C. elegans*, fails to identify X-box motifs in most known ciliary genes in the human genome despite the high conservation of RFX DBDs. One possibility is that these ciliary genes are not regulated by RFX transcription factors in humans, however, we hypothesize that there exist many ciliary genes whose X-box motifs are different from consensus motifs. The goal of this chapter is to test this hypothesis and identify X-box motifs that are divergent from known cases of X-box motifs, by taking advantage of the recent availability of the genome sequences of four *Caenorhabditis* species: *C. elegans*, *C. briggsae*, *C. remanei*, and *C. brenneri*. These genomes were chosen as a model system for this study because *C. elegans* has been used effectively to identify and characterize functionally ciliary genes, and because all of these organisms have comparatively compact genomes. The availability of these four genomes allows us to look for putative X-box motifs that are conserved in three species but not in the fourth, which may suggest potential divergent forms of X-box motif. Some ciliary genes in *C. elegans* contain X-box motifs in their promoters and have readily identifiable orthologs in humans (Emery *et al.* 1996; Swoboda *et al.* 2000; Chu *et al.* 2010), which makes this study useful for understanding the function of RFX TFs in humans.

5.3 Methods and Materials

5.3.1 Strains used

Worm strains are maintained using standard procedures in 20^{o}C unless otherwise noted (Brenner 1974). The following strains were used in this study: DR86 *daf-19(m86)*, EG5003 *unc-119(ed3) III; cxTi10882 IV*, JT204 *daf-12(sa204)*, JT6924 *daf-12(sa204); daf-19(m86)*. Strains generated in this study are listed in Table 5-1.

Table 5-1 Strains generated in this chapter listing the strain name and the allele. The strains were generated by J. Chu and D. Tu.

Strain Name	Allele	Description
JNC20	*unc-119*(ed3); dotSi1 [prF25B4.2::mCherry, *Cb-unc-119*(+)] IV	F25B4.2 Wild type promoter
JNC21	*unc-119*(ed3); dotSi2 [prF25B4.2::mCherry del(-149), *Cb-unc-119*(+)] IV	Proximal deletion
JNC22	*unc-119*(ed3); dotSi3 [prF25B4.2::mCherry del(-199), *Cb-unc-119*(+)] IV	Distal deletion
JNC29	*unc-119*(ed3); dotSi10 [prF25B4.2::mCherry del(-140) del(-190), *Cb-unc-119*(+)] IV	Double deletion
JNC33	*daf-19*(m86); dotSi2	Proximal deletion in *daf-19* background
JNC34	*daf-19*(m86); dotSi10	Double deletion in *daf-19* background
JNC36	*daf-19*(m86); dotSi1	Wild type promoter in *daf-19* background
JNC37	*daf-19*(m86); dotSi3	Distal deletion in *daf-19* background
JNC23	*unc-119*(ed3); dotSi4 [prM04C9.5::mCherry, *Cb-unc-119*(+)] IV	*dyf-5* wild type promoter

JNC31	*unc-119*(ed3); dotSi4 [prM04C9.5::mCherry replace -285 to -271 →gtcctcacaagtaac, *Cb-unc-119*(+)] IV	*dyf-5* promoter replaced with distal motif
JNC35	*unc-119*(ed3); dotSi4 [prM04C9.5::mCherry replace -285 to -271 →gtctccaatggcaac, *Cb-unc-119*(+)] IV	*dyf-5* promoter replaced with proximal motif

5.3.2 Genomic data and gene model improvement

Genomic DNA data for all four *Caenorhabditis* species were obtained from WS204 version of WormBase (ftp.wormbase.org). The gene set for *C. elegans* was also obtained from the WS204 version. The gene set for *C. briggsae*, *C. remanei*, and *C. brenneri* were obtained by running genBlastG. genBlastG is a homology based gene finder based on genBlastA that we developed previously to look for homologous gene regions (see Chapter 4). For principles behind genBlastA, see (She *et al.* 2009) and Chapter 3. 20,173 *C. elegans* proteins were used as input for genBlastG. These sequences represent the longest alternative transcript if more than one exists. genBlastG returns 264,411 gene models for *C. briggsae*, 319,750 for *C. remanei*, and 425,947 for *C. brenneri*. Many of these gene models are overlapping and redundant due to multiple genes with highly similar sequences (such as gene families or tandem gene duplications) in *C. elegans*. A filtering procedure was used so that each genomic region would contain only one gene model with the highest sequence percent identity (PID) to the query. The filtering procedure was carried out as follows: (1) All the predictions are sorted by PID in decreasing order. (2) For each two overlapping model, if the overlapping region is greater than 5% of the length for either gene, then the model with higher PID is kept and the model with lower PID is filtered out. (3) To ensure the quality of the gene set, we only kept gene models that show PID >= 40% to the query. The filtering procedure resulted in 16,577 gene models for *C. briggsae*, 18,426 for *C. remanei*, and 23,473 for *C. brenneri*. In the last step, we combined these gene models with the current WormBase models to generate a hybrid gene set. In the hybrid set, genBlastG's predicted models replace corresponding WormBase gene models if genBlastG's prediction shows at least 2% improvement in PID. The final gene models were uploaded in GFF3 format to a MYSQL server and visualized on Generic Genome Browser (Stein *et al.* 2002).

5.3.3 X-box motif search and comparative genomics

We generated a HMM profile based on 31 validated X-box motifs (Table 5-2). The X-box motifs were first aligned by ClustalW (Larkin *et al.* 2007). The resulting alignment file was used to generate the HMM profile by hmmb. The profile was used to probe the entire genome by hmmfs. hmmb and hmmfs are both part of the HMMER suite (Durbin *et al.* 1998) (http://hmmer.janelia.org) . Mapping of orthologous relationships between genes in *C. elegans* and the other three species were generated by Inparanoid (Remm *et al.* 2001).

We also generated position weight matrix (PWM) based on the 31 validated motifs for 6 bp from the right and 6 bp from the left. These PWMs are used by TFMscan with the p-value parameter (-p) setting to 5 (Liefooghe *et al.* 2006).

Table 5-2 Validated X-box motifs. An X-box motif containing gene is considered validated when the expression of the gene is dependent on the X-box motif (mutagenesis study) or on DAF-19 (DAF-19 knock out study). Expression ratio is based on comparative transcriptomics between JT204 and JT6924. JT204 is a control strain with wild-type *daf-19* and JT6924 is a *daf-19* mutant strain. Value > 1 means higher expression in JT6924, value < 1 means lower expression in JT6924 (see Methods). Genes that do not have any mapped reads could not be calculated and thus labeled as "n/a". X-box distance was collected by published references. HMMER Score was collected by J. Chu. Expression Ratio data was collected by D. Zhang.

Gene name	Sequence name	Distance from ATG	HMMER Score	X-box sequence	Expression Ratio	Reference
che-13	F59C6.7	-74	5.33	GTTGCTATAGCAAC	0.08	(Haycraft *et al.* 2003; Efimenko *et al.* 2005)
xbx-1	F02D8.3	-79	7.73	GTTTCCATGGTAAC	0.26	(Swoboda *et al.* 2000; Schafer *et al.* 2003; Efimenko *et al.* 2005)
xbx-2	D1009.5	-77	7.98	GTTGCCATGACAAC	0.33	(Blacque *et al.* 2005; Efimenko *et al.* 2005)
xbx-3	M04D8.6	-97	5.72	GTTGTCTTGGCAAC	n/a	(Efimenko *et al.* 2005)
xbx-4	C23H5.3	-82	7.98	GTTGCCATGACAAC	n/a	(Efimenko *et al.* 2005)
xbx-5	T24A11.2	-121	6.84	GTCTCCATGACAAC	0.38	(Efimenko *et al.* 2005)
xbx-6	F40F9.1	-151	7.53	GTTTCCATGGAAAC	0.71	(Efimenko *et al.* 2005)
xbx-7	R148.1	-69	4.56	GTCACCATAGGAAC	0.29	(Efimenko *et al.* 2005)
	ZK328.7	-89	6.51	GTTACCATGGCAAT	0.00	(Blacque *et al.* 2005)
bbs-9	C48B6.8	-81	7.53	GTTTCCATGACAAC	0.35	(Blacque *et al.* 2005)
che-11	C27A7.4	-85	7.04	ATCTCCATGGCAAC	1.84	(Efimenko *et al.* 2005)
odr-4	Y102E9.1	-200	4.09	ATCGTCATGGTAAC	0.56	(Efimenko *et al.* 2005)
osm-5	Y41G9A.1	-115	6.92	GTTACTATGGCAAC	0.55	(Haycraft *et al.* 2001; Qin *et al.* 2001; Efimenko *et al.* 2005)
nhr-44	T19A5.4	-76	6.91	GTCTTCATGGCAAC	0.51	(Efimenko *et al.* 2005)
nph-1	M28.7	-77	5.57	GTTGCCAGGGGCAAC	0.47	(Winkelbauer *et al.* 2005)
nph-4	R13H4.1	-168	5.93	ATTTCCATGACAAC	2.20	(Winkelbauer *et al.* 2005)
nud-1	F53A2.4	-263	3.81	GTATCCATGGGAAC	1.02	(Efimenko *et al.* 2005)
dyf-2	ZK520.3	-140	5.84	GTTACCAAGGCAAC	0.18	(Efimenko *et al.* 2006)
osm-6	R31.3	-100	6.13	GTTACCATAGTAAC	0.27	(Collet *et al.* 1998; Swoboda *et al.* 2000;

						Efimenko *et al.* 2005)
dyf-3	C04C3.5	-88	4.22	GTTTCTATGGGAAC	0.00	(Murayama *et al.* 2005; Ou *et al.* 2005)
	Y110A7A.20	-60	4.26	GTCTCTATAGCAAC	0.18	(Blacque *et al.* 2005)
che-2	F38G1.1	-117	7.05	GTTGTCATGGTGAC	0.26	(Fujiwara *et al.* 1999; Swoboda *et al.* 2000; Efimenko *et al.* 2005)
osm-1	T27B1.1	-86	5.27	GCTACCATGGCAAC	0.22	(Perkins *et al.* 1986; Swoboda *et al.* 2000; Efimenko *et al.* 2005; Bell *et al.* 2006)
bbs-1	Y105E8A.5	-99	5.45	GTTCCCATAGCAAC	0.16	(Ansley *et al.* 2003; Blacque *et al.* 2004; Efimenko *et al.* 2005)
bbs-2	F20D12.3	-94	6.31	GTATCCATGGCAAC	0.12	(Ansley *et al.* 2003; Blacque *et al.* 2004; Efimenko *et al.* 2005)
bbs-5	R01H10.6	-65	8.64	GTCTCCATGGCAAC	0.41	(Li *et al.* 2004; Efimenko *et al.* 2005)
bbs-7	Y75B8A.12	-94	6.92	GTTGCCATAGTAAC	0.00	(Ansley *et al.* 2003; Blacque *et al.* 2004; Efimenko *et al.* 2005)
bbs-8	T25F10.5	-84	4.22	GTACCCATGGCAAC	0.53	(Ansley *et al.* 2003; Blacque *et al.* 2004; Efimenko *et al.* 2005)
tub-1	F10B5.4	-183	5.25	ATCTCCATGACAAC	0.15	(Efimenko *et al.* 2005; Mak *et al.* 2006)
che-12	B0024.8	-767		ATCAGCTTGAAAAC	2.83	(Bacaj *et al.* 2008)
dyf-5	M04C9.5	-285	5.95	GTTACCATAGAAAC	0.32	(Chen *et al.* 2006; Burghoorn *et al.* 2007)

5.3.4 Transcriptome data

JT204 and JT6924 worms were grown on 25 small plates until the plate were crowded with gravid adults. The animals were collected by washing with M9. Embryos were harvested by bleaching (mixing 1 ml of household bleach and 0.25 ml of 10M NaOH to 3.75 ml of worms with M9). We washed with M9 four times to remove excess bleach and transferred the embryos to 250 ml liquid culture. Liquid culture method was adapted from http://info.med.yale.edu/mbb/koelle/protocols/protocol_liquid_culture.html. The embryos were allowed to hatch and grow to gravid adult stage. Those adults were collected and bleached again to collect the embryos. We staged the embryos to 2-fold/3-fold stage by swirling gently in S-media for up to 8 hours in 20^{o}C. We then extracted RNA from the two strains using Trizol/Chloroform (Invitrogen #15596-026) and precipitated by isopropanol. The RNA was washed with 75% ethanol and dissolved in DEPC water. The sample was treated with DNaseI (Invitrogen #18047-019) to remove contaminating genomic DNA. RNA was purified again by phenol/Chloroform extraction. Sample quantity was measured by Nanodrop (Thermo Fisher #ND-1000). Sequencing of the RNA was done by BC Genome Sciences Centre using the Illumina Solexa technology. The pair-end sequencing reads were mapped using MAQ to a virtual transcriptome based on WormBase WS204. The MAQ alignment parameter ‘n’ was set to 3 and ‘a’ was set to 700 for JT204 and 400 for JT6924. We used read depth as a way to measure the level of expression. Read depth was calculated as follows: a read segment, defined by the length that spans from one read to the other read of the same pair, is calculated for each read pair. For cases where only a single read maps, then the segment is defined as the length of the read. The expression level of a gene is defined as the maximum depth of the read segments. To calculate the differential expression level, we first normalized all the read count to that of JT204. The ratio is the result of the following equation: (Depth in JT6924 x Normalization factor) / Depth in JT204.

5.3.5 Construct generation and cloning

Deletion constructs were made by standard site-directed mutagenesis and PCR stitching method (Hobert 2002). Briefly, primers were designed to contain the particular deletion. See below for a list of primers used and their sequences. A left fragment was amplified using Primer A and Primer DeletionR (either distal or proximal). A right fragment was amplified using Primer DeletionF (either distal or proximal) and Primer B. The left and right fragment was stitched together using Primer A* and B. The mCherry was amplified using Primer C and D from pCFJ190 (A generous gift from E.M. Jorgensen). The final stitching between the promoter fragment and mCherry was done using Primer A* and D*. Primer A* contains SbfI site and Primer D* contains SpeI site. The construct and the plasmid pCFJ178 were cut using the respective restriction enzymes in 37^{o}C for 2.5 hours. The construct was ligated into the linearized plasmid overnight at room temperature. The final ligation reaction was transformed into DH5α cells by electroporation. The transformants were plated onto LB-Ampicillin plates. Living colonies were picked to grow in a 5ml of LB broth and the DNA was extracted using Qiagen Mini-prep kit (Cat#:27104). All the primers used in this study are listed in Table 5-3.

.

Table 5-3 Primers used in this chapter. The primers were designed by J. Chu.

Primer name	Primer sequence
F25B4.2_A	CAAAATTACCTATCGCACTACGTT
F25B4.2_A*	CCTGCAGGCCTGCAGGAAGCTGAAACGTCGGAGATAATAC
F25B4.2_B	TATCTTCTTCACCCTTTGAGACCATCATCCACGATTAATCTGAAACTCA
M04C9.5_A	CCTGCAGGCCTGCAGGAATTGAATTAGCCGCGGAGC
M04C9.5_B	TATCTTCTTCACCCTTTGAGACCATGGCTTCTTGCCCTTATATTTTCC
mCherry_C	ATGGTCTCAAAGGGTGAAGA
mCherry_D	GGCCTCTTCGCTATTACGC
mCherry_D*	ACGACGGCCAGTGAATTATCACTAGTACTAGT
F25B4.2_deletionF_distal	CACTTTTCAATTCGAAATGTCATGGGCGTTG
F25B4.2_deletionR_distal	CCATGACATTTCGAATTGAAAAGTGTCGAAATTCTTAGAG
F25B4.2_deletionF_proximal	GGCGCCACTGAAACCCGCATTTTAAACTCCAT
F25B4.2_deletionR_proximal	CGGGTTTCAGTGGCGCCGTGGCGACA
M04C9.5_replaceF_distal	GTCCTCACAAGTAACTGTCTGTTACACCCTTTTCTC
M04C9.5_replaceR_distal	GTTACTTGTGAGGACCAAGAGCAAACGGCGGAG
M04C9.5_replaceF_proximal	GTCTCCAATGGCAACTGTCTGTTACACCCTTTTCTC
M04C9.5_replaceR_proximal	GTTGCCATTGGAGACCAAGAGCAAACGGCGGAG
ChrIV-R	TGTTTACTAGACCGGGGCTC
mCherry-genoF	AAAACCGCACACAAAATACC
178-genoF	TCCCCATTTCACCAGAGAAC

5.3.6 Mos Single Copy Insertion (MosSCI)

DNA purified from the transformation was used directly for injection. The injection mix for MosSCI was made as suggested from the literature (Frokjaer-Jensen *et al.* 2008): pJL43.1 (50ng/ul), purified plasmid (50ng/ul), pGH8 (10ng/ul), pCFJ90 (2.5ng/ul), pCFJ104 (5ng/ul). The mix was injected into EG5003 worms. Worms that move and show none of the mCherry markers were individually plated. To confirm for insertion, we performed PCR with primers ChrIV-F, mCherry-genoF, and 178-genoR to genotype individual mothers. A worm with homozygous insertion would have a single band at around 2.2kb; a worm with no insertion would have a single band at around 4kb; a worm with heterozygous insertion would have both bands.

5.3.7 Dye-filling assay

The methods for dye-filling was adapted from Worm Atlas (Altun and Hall 2005). Briefly, we washed one plate of mixed population using 1ml of M9 buffer. Then, we collected worms by centrifugation at 1500rpm for 1 minute and removed supernatant. Then, we resuspended the worms in 1ml of M9 buffer mixed with 5ul of 2mg/ml DiO (Molecular Probes, Cat#:D275). To allow the worms to take up the dye, we covered the tubes in tin foil and slowly shake at room temperature for 2 hours. The worms were spun down again and transferred to a fresh seeded plate to allow the dye to pass through the gut. The worms were washed and spun as before just prior to transferring worms to the glass slide.

5.3.8 Genetic crosses

We obtained males for each strain containing the Mos insert by heat shocking 30 L4 hermaphrodites in 33^{o}C for 4 hours. We crossed four males with Mos insert to two *daf-19* (DR86) L4 hermaphrodites. Fifteen hermaphrodite F1s were selected randomly and individually plated. The genotype of the *daf-19* gene in these F1s was confirmed by Tetra-ARM PCR (Ye *et al.* 2001). To find homozygous Mos insertion and homozygous *daf-19* mutation, we individually plated 200 F2s and screened for dauer phenotype as 85% of *daf-19* worms enter the dauer stage even in favourable condition (Perkins *et al.* 1986). Candidates were screened and confirmed by genotyping.

5.3.9 Microscopy visualization

Worms were immobilized using sodium azide on 3% agarose pad. Images were captured under Zeiss spinning disc confocal microscope (Zeiss Axio Observer.Z1) equipped with Hamamatsu ImagEM camera. Image capture and visualization were performed using Volocity software (www.improvision.com).

5.4 Results

5.4.1 Comparative genomics search for putative DAF-19 regulated genes

Comparing different genomes to find similarities and differences relies heavily on the reliability of genome annotations. While the *C. elegans* genome was well curated by WormBase curators and the entire *C. elegans* community during the last dozen years since it was sequenced (*C. elegans* Sequencing Consortium 1998), the genome of *C. briggsae* (Stein *et al.* 2003), *C. remanei* and *C. brenneri* are not as well annotated. To ensure that the gene sets are of comparable quality for this project, we first revised gene sets for *C. briggsae*, *C. remanei,* and *C. brenneri* using genBlastG, a homology-based gene prediction program genBlastG recently developed in our laboratory (She *et al*., submitted to Bioinformatics). We revised gene models so that they have maximized protein sequence identity to their corresponding *C. elegans* homologs. For each gene in the three species, we have attempted to predict one isoform, using the longest *C. elegans* homolog as query. Altogether, we used 20,173 *C. elegans* protein sequences (version WS204) as queries. The revised *C. briggsae* gene set has 23,299 gene models, which is comparable to that of *C. elegans*. In contrast, *C. remanei* has 31,830 gene models and *C. brenneri* has 35,071 gene models, because of some heterozygosity contained in the sequenced genome (Barriere *et al.* 2009).

With the revised gene sets, we then identified putative X-box motifs using the same procedure described previously (Chen *et al.* 2006). First, an HMM profile for X-box motifs was generated using 31 X-box motifs, which were validated in previous studies (Table 5-2). An X-box motif is considered validated if it is proven to be essential for cilia-specific gene expression such that their absence or mutation lead to loss of cilia-specific gene expression, or the expression of the X-box motif-containing gene is DAF-19 dependent. Genome-wide X-box search using HMMER predicted 5,332, 6,667, 6,920, and 10,651 putative motifs in *C. elegans*, *C. briggsae*, *C. remanei*, and *C. brenneri*, respectively. To maximize the ability to search for divergent X-box motifs, all candidate X-box motifs that have HMM score > 0 are considered in the following analysis.

Divergent X-box motifs in *C. elegans* are defined as putative X-box motifs that were not predicted in *C. elegans,* but high-score X-box motifs (HMM score > 3.0) are found within 500 bp upstream of their orthologs in *C. briggsae*, *C. remanei*, and *C. brenneri*. The 500 bp search window was chosen because it defines a stringent criterion to find high-quality X-box motifs. Almost all (30/31) validated X-box motifs reside within 500 bp (Table 5-2). By applying these criteria, we identified 10 promoter regions that harbour candidate divergent X-box motifs (Table 5-4).

Table 5-4 Putative divergent X-box motifs identified using TFMscan. Putative X-box sequences are shown with the first and last 6 bp while the middle spacer regions are only indicated with its length in bp. Positions are listed as distance from the translational start site. Expression ratio is based on comparative transcriptomics between JT204 and JT6924. JT204 is a control strain with wild-type daf-19 and JT6924 is a daf-19 mutant strain. Value > 1 means higher expression in JT6924, value < 1 means lower expression in JT6924 (see Methods). Genes that do not have any mapped reads could not be calculated and thus labeled as “n/a”. Putative X-box sequence and distance was collected by J. Chu. Expression Ratio data was collected by D. Zhang.

Sequence name	Expression Ratio	Description	Putative X-box sequence	X-box position
C46A5.8	2.63	Uncharacterized gene	gttctc-(57)-gaaaac	-296
F25B4.2	0.77	Homologous to Pellino	gtcctc-(3)-agtaac gtctcc-(3)-ggcaac	-199 -149
F39H12.2	3.06	Uncharacterized gene	atttcc-(0)-agtaac atttccc-(35)-gaaaac gttacc-(29)-gacaac	-281 -255 -91
F48B9.8	1.51	Uncharacterized gene	n/a	n/a
M04D8.7	2.29	Uncharacterized gene	gttact-(55)-gaaaac	-195
R05H10.5	0.50	Homologous to glutathione peroxidase	gtatcc-(43)-gaaaac	-216
T27E7.3	n/a	Uncharacterized gene	gttttc-(32)-gaaaac	-358
T27E7.4	n/a	Uncharacterized gene	gttact-(24)-agaaac	-463
T27E7.5	n/a	Uncharacterized gene	n/a	n/a
T27E7.9	n/a	Uncharacterized gene	n/a	n/a

5.4.2 Analysis of potential divergent X-box motifs

Unable to identify X-box motifs in these 10 promoter regions using HMMER could be explained by several reasons. First, these 10 genes may not be regulated by DAF-19 due to loss of X-box motifs in evolution. Second, X-box motifs for DAF-19 may exist in these 10 promoters but they are located outside of these 500 bp windows. Third, the similarity between these X-box motifs and the typical X-box motifs is low. Fourth, X-box motifs in these promoters may have a larger space between two half sites that can be as large as 60 bp (Emery *et al.* 1996; Lubelsky *et al.* 2005). We searched two halves of X-box motifs separately with no regard to the sequence or the length of the spacer region. We applied TFMscan that is based on position weighted matrix (PWM) (Liefooghe *et al.*

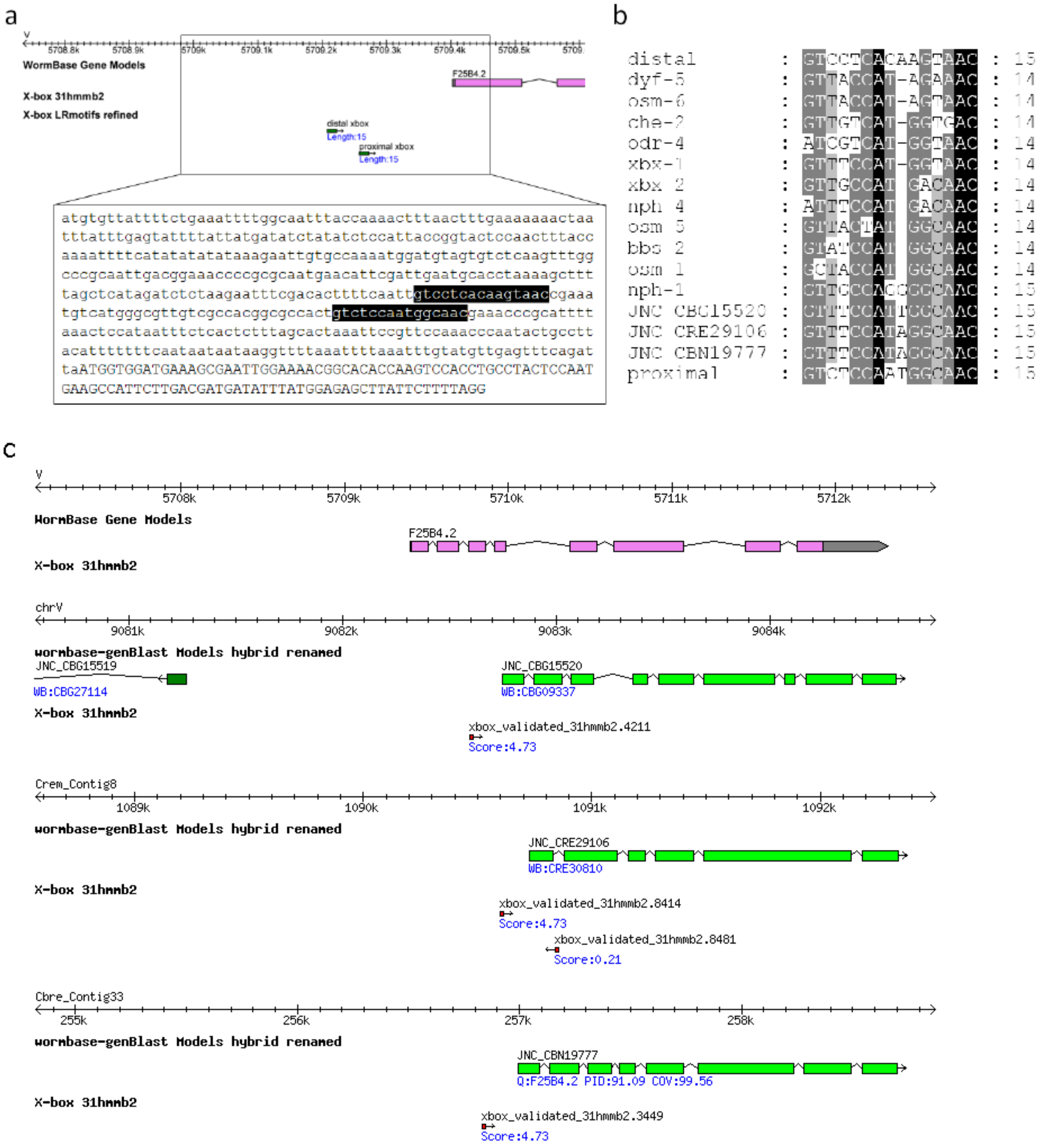
2006). To predict putative divergent X-box motifs, we pair together a left half and a right half given that there is no other half sites predicted in between. With this search strategy, we allow putative X-box motifs to have a greater flexibility in terms of sequence and length. Of the 10 genes, we found seven contained putative divergent X-box motifs within the 500 bp upstream sequence (Table 5-4). The length flexibility in the spacer region allowed us to find two half sites that are further apart.

None of these 10 genes except F25B4.2 has expression information available in public databases (McKay *et al.* 2003; Hunt-Newbury *et al.* 2007). GFP reporter strain revealed neuronal expression for F25B4.2, which is also supported by SAGE data (McKay *et al.* 2003). In order to gain insight into the dependency of gene expression of these genes on DAF-19, we examined the difference at the transcription level for all genes in *C. elegans*, including these 10 genes. We prepared cDNA libraries for two strains: a control strain (JT204) and a *daf-19* mutant strain (JT6924) and sequenced these two libraries using the Illumina Solexa Genome Analyzer. We decided to use these two strains in particular due to a mutation in the *daf-12* gene (sa204) in both of these two strains. A *daf-12* mutant suppresses the Daf-c phenotype of *daf-19* and allows propagation of *daf-19* worms in large quantities (see Methods and Materials). The difference in transcription is calculated as a ratio of sequencing read depth between JT6924 and JT204 (see Methods). We expect genes that are positively regulated by DAF-19 to show lower transcript level in JT6924 (lower ratio) and genes that are negatively regulated by DAF-19 to show higher transcript level in

JT6924 (higher ratio). As expected, the average ratio of the 31 validated X-box motif regulated genes is 0.52 while the average ratio for genes that are not regulated by X-box motifs/DAF-19 remain relatively unchanged. Of the 10 genes we identified, four showed ratio higher than 1.5 while F25B4.2 and R05H10.5 showed ratio lower than 0.75. We did not detect expression for the remaining four genes (Table 5-4). Next, we will examine whether putative divergent X-box motifs in F25B4.2 are functional.

5.4.3 F25B4.2 is a conserved gene that harbors X-box motifs in 4 *Caenorhabditis* species but not in *C. elegans*

We hypothesized that F25B4.2 is regulated by DAF-19 through binding to these two putative divergent X-box motifs. The two putative X-box motifs are located at 199-bp and 149-bp upstream of the start codon (Figure 5-1a, Table 5-4). For convenience, we named the X-box motif located at -199 the distal motif and the X-box motif located at -149 the proximal motif. Among the two, the proximal motif displays higher conservation especially in the last 6 nucleotides where it is identical to many known X-box motifs (Figure 5-1b). However, these two motifs differ from the consensus at the 3rd nucleotide (consensus = T) and at the 8th nucleotide (consensus = T). We believe the reason that these two elements were missed by HMMER is due to the degeneracy at these two positions. In contrast, HMMER identified a 15 bp X-box motif in each of the three orthologous upstream regions. The 15 bp X-box motif identified in three other species was found because these two positions are conserved (Figure 5-1b and c).

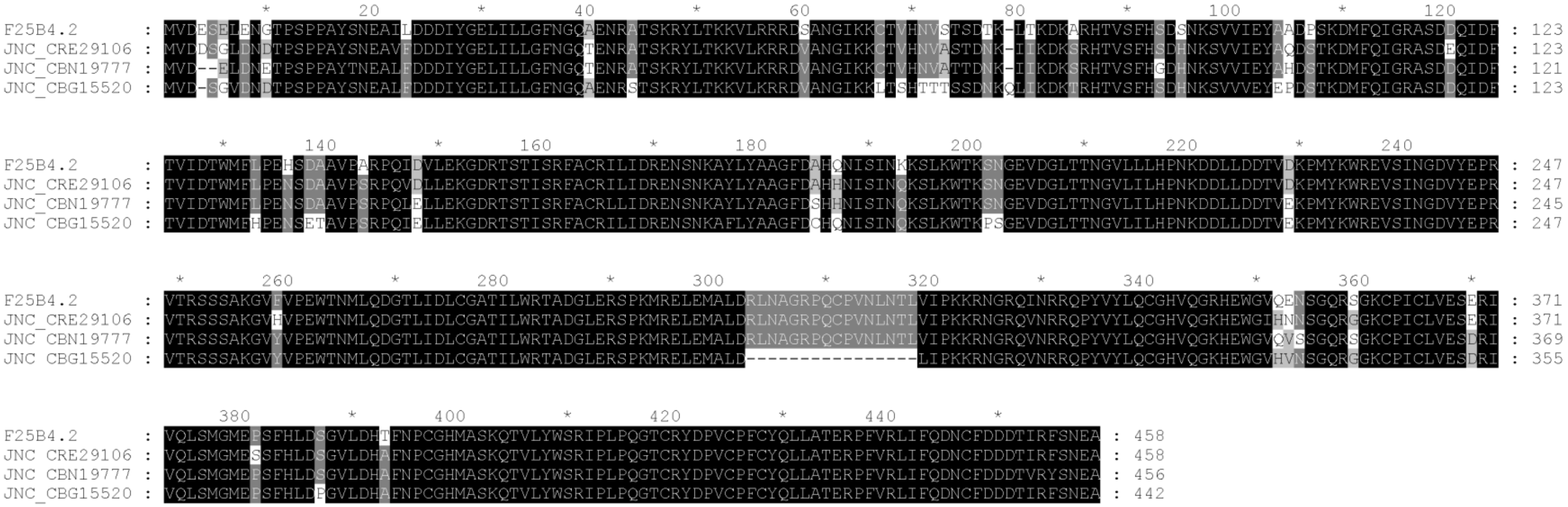


```
atgtgttattttctgaaattttggcaatttaccaaaactttaactttgaaaaaaactaa
tttatttgagtattttattatgatatctatatctccattaccggtactccaactttacc
aaaattttcatatatatatataaagaattgtgccaaaatggatgtagtgtctcaagtttgg
cccgcaattgacggaaaccccgcgcaatgaacattcgattgaatgcacctaaaagcttt
tagctcatagatctctaagaatttcgacacttttcaattgtcctcacaagtaaccgaaa
tgtcatgggcgttgtcgccacggcgccactgtctccaatggcaacgaaacccgcatttt
aaactccataatttctcactctttagcactaaattccgttccaaacccaatactgcctt
acatttttttcaataataataaggttttaaattttaaatttgtatgttgagtttcagat
taATGGTGGATGAAAGCGAATTGGAAAACGGCACACCAAGTCCACCTGCCTACTCCAAT
GAAGCCATTCTTGACGATGATATTTATGGAGAGCTTATTCTTTTAGG
```

```
distal        : GTCCTCACAAGTAAC : 15
dyf-5         : GTTACCAT-AGAAAC : 14
osm-6         : GTTACCAT-AGTAAC : 14
che-2         : GTTGTCAT-GGTGAC : 14
odr-4         : ATCGTCAT-GGTAAC : 14
xbx-1         : GTTTCCAT-GGTAAC : 14
xbx-2         : GTTGCCAT-GACAAC : 14
nph-4         : ATTTCCAT-GACAAC : 14
osm-5         : GTTACTAT-GGCAAC : 14
bbs-2         : GTATCCAT-GGCAAC : 14
osm-1         : GCTACCAT-GGCAAC : 14
nph-1         : GTTGCCACGGGCAAC : 15
JNC_CBG15520  : GTTTCCATTGGCAAC : 15
JNC_CRE29106  : GTTTCCATAGGCAAC : 15
JNC_CBN19777  : GTTTCCATAGGCAAC : 15
proximal      : GTCTCCAATGGCAAC : 15
```

Figure 5-1 (a) The location and sequence of the putative divergent X-box motifs upstream of F25B4.2. (b) The alignment of the putative proximal and distal motifs to known X-box motifs. Also include in the alignment are the putative X-box motifs in the orthologs of F25B4.2 in three other *Caenorhabditis* species. (c) Generic Genome Browser view of orthologous regions in four *Caenorhabdits* species. Every species contains a clearly identified X-box motif in the upstream region except in *C. elegans*. The ID on top of the gene models indicates the species it is from: CBG = *C. briggsae*, CRE = *C. remanei*, CBN = *C. brenneri*. The data used to generate this figure was collected by J. Chu and B. Uyar.

F25B4.2 is conserved in four sequenced *Caenorhabdits* species with more than 80% identity at the protein level (Figure 5-2). The presence of X-box in upstream region of a gene usually suggests regulation by DAF-19 (Swoboda *et al.* 2000). F25B4.2 protein sequence shows about 40% identity to human Pellino gene family. Pellino proteins are E3 ligases known to participate in balancing inflammatory response (Butler *et al.* 2007). Pellino proteins interact with IRAK and mediate NFkB nuclear translocation to promote activation of pro-inflammatory genes (Rich *et al.* 2000; Strelow *et al.* 2003). Pellino1 is also suggested to play a part in TGF-β pathways to promote anti-inflammatory response preventing hyperactivation of inflammatory response (Choi *et al.* 2006; Chang *et al.* 2009). However, Pellino is not currently known to have any role in cilia development or cilia maintenance in human or in any other organisms.



```
                         *        20         *        40         *        60         *        80         *       100         *       120
F25B4.2      : MVDESELENGTPSPPAYSNEAILDDDIYGELILLGFNGQAENRATSKRYLTKKVLRRRDSANGIKKCTVHNVSTSDTK-LTKDKARHTVSFHSDSNKSVVIEYAADPSKDMFQIGRASDDQIDF : 123
JNC_CRE29106 : MVDDSGLDNDTPSPPAYSNEAVFDDDIYGELILLGFNGQTENRATSKRYLTKKVLKRRDVANGIKKCTVHNVASTDNK-IIKDKSRHTVSFHSDHNKSVVIEYAQDSTKDMFQIGRASDEQIDF : 123
JNC_CBN19777 : MVD--ELDNETPSPPAYTNEALFDDDIYGELILLGFNGQTENRATSKRYLTKKVLKRRDVANGIKKCTVHNVATTDNK-IIKDKSRHTVSFHGDHNKSVVIEYAHDSTKDMFQIGRASDDQIDF : 121
JNC_CBG15520 : MVD-SGVDNDTPSPPAYSNEALFDDDIYGELILLGFNGQAENRSTSKRYLTKKVLKRRDVANGIKKLTSHTTTSSDNKQLIKDKTRHTVSFHSDHNKSVVVEYEPDSTKDMFQIGRASDDQIDF : 123

                     *       140         *       160         *       180         *       200         *       220         *       240
F25B4.2      : TVIDTWMFLPEHSDAAVPARPQIDVLEKGDRTSTISRFACRILIDRENSNKAYLYAAGFDAHQNISINKKSLKWTKSNGEVDGLTTNGVLLLHPNKDDLLDDTVDKPMYKWREVSINGDVYEPR : 247
JNC_CRE29106 : TVIDTWMFLPENSDAAVPSRPQVDLLEKGDRTSTISRFACRILIDRENSNKAYLYAAGFDAHHNISINQKSLKWTKSNGEVDGLTTNGVLILHPNKDDLLDDTVDKPMYKWREVSINGDVYEPR : 247
JNC_CBN19777 : TVIDTWMFLPENSDAAVPSRPQLELLEKGDRTSTISRFACRLLIDRENSNKAYLYAAGFDSHHNISINQKSLKWTKSNGEVDGLTTNGVLILHPNKDDLLDDTVEKPMYKWREVSINGDVYEPR : 245
JNC_CBG15520 : TVIDTWMFHPENSETAVPSRPQIELLEKGDRTSTISRFACRILIDRENSNKAFLYAAGFDCHQNISINQKSLKWTKPSGEVDGLTTNGVLILHPNKDDLLDDTVEKPMYKWREVSINGDVYEPR : 247

                 *       260         *       280         *       300         *       320         *       340         *       360         *
F25B4.2      : VTRSSSAKGVFVPEWTNMLQDGTLIDLCGATILWRTADGLERSPKMRELEMALDRLNAGRPQCPVNLNTLVIPKKRNGRQINRRQPYVYLQCGHVQGRHEWGVQENSGQRSGKCPICLVESERI : 371
JNC_CRE29106 : VTRSSSAKGVHVPEWTNMLQDGTLIDLCGATILWRTADGLERSPKMRELEMALDRLNAGRPQCPVNLNTLVIPKKRNGRQVNRRQPYVYLQCGHVQGKHEWGIHNNSGQRGGKCPICLVESERI : 371
JNC_CBN19777 : VTRSSSAKGVYVPEWTNMLQDGTLIDLCGATILWRTADGLERSPKMRELEMALDRLNAGRPQCPVNLNTLVIPKKRNGRQVNRRQPYVYLQCGHVQGKHEWGVQVSSGQRSGKCPICLVESDRI : 369
JNC_CBG15520 : VTRSSSAKGVYVPEWTNMLQDGTLIDLCGATILWRTADGLERSPKMRELEMALD----------------LIPKKRNGRQVNRRQPYVYLQCGHVQGKHEWGVHVNSGQRGGKCPICLVESDRI : 355

                     380         *       400         *       420         *       440         *
F25B4.2      : VQLSMGMEPSFHLDSGVLDHTFNPCGHMASKQTVLYWSRIPLPQGTCRYDPVCPFCYQLLATERPFVRLIFQDNCFDDDTIRFSNEA : 458
JNC_CRE29106 : VQLSMGMESSFHLDSGVLDHAFNPCGHMASKQTVLYWSRIPLPQGTCRYDPVCPFCYQLLATERPFVRLIFQDNCFDDDTIRFSNEA : 458
JNC_CBN19777 : VQLSMGMEPSFHLDSGVLDHAFNPCGHMASKQTVLYWSRIPLPQGTCRYDPVCPFCYQLLATERPFVRLIFQDNCFDDDTVRYSNEA : 456
JNC_CBG15520 : VQLSMGMEPSFHLDPGVLDHAFNPCGHMASKQTVLYWSRIPLPQGTCRYDPVCPFCYQLLATERPFVRLIFQDNCFDDDTIRFSNEA : 442
```

Figure 5-2 The alignment of F25B4.2 with its orthologs in the other four *Caenorhabditis* species. CBG = *C. briggsae*, CRE = *C. remanei*, CBN = *C. brenneri*. The sequences were aligned using ClustalW (Larkin *et al.* 2007) and visualized using GeneDoc (Nicholas *et al.* 1997). The data used to generate this figure was collected by J. Chu and B. Uyar.

5.4.4 F25B4.2 is expressed in ciliated neurons in a DAF-19 dependent manner

We examined whether the promoter of F25B4.2 drives expression in ciliated neurons and whether its expression is dependent on DAF-19. We constructed a *C. elegans* strain carrying a single copy mCherry transgene driven by a 3-Kb genomic DNA sequence upstream of F25B4.2. The F25B4.2 promoter::mCherry fusion construct was stably integrated into chromosome IV at the Mos site cxTi10882 using the Mos Single Copy Insertion (MosSCI) method (Frokjaer-Jensen *et al.* 2008). This Mos element is located in an intergenic region with the flanking genes pointing towards each other. Hence this location is not likely to have functional elements disrupted after reporter gene insertion. Insertion at the Mos site is confirmed by genotyping (Figure 5-3).

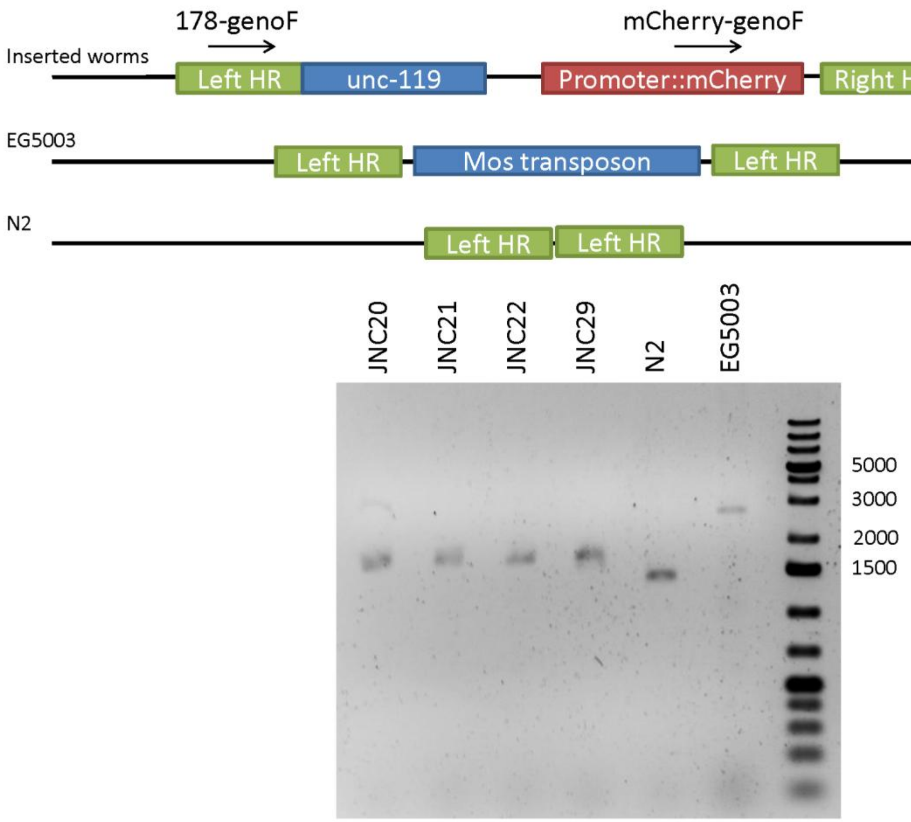


Figure 5-3 Genotyping of stably integrated strains. The insertion site on chromosome IV is depicted by the diagram on the top while an agarose gel showing the genotyping results on the bottom. The primers used for genotyping are also indicated by the arrows. Primer mCherry-genoF can only hybridize to inserted worms and not EG5003 and N2. The expected band sizes for inserted worms are 8,312bp from 178-genoF→ChrIV-R and 1,564bp from mCherry-genoF→ChrIV-R. The expected band size for EG5003 is 2,700bp (Mos1 is about 1,280bp (Benjamin and Kleckner 1992; van Luenen *et al.* 1994; Lampe *et al.* 1996)). The expected band size for N2 is 1,420bp from 178-genoF→ChrIV-R. The gel image shows homozygous insertion for JNC20, 21, 22, and 29 as well as EG5003 and N2 as controls. The number on the right hand side indicates the ladder positions. The data used to generate this figure was collected by J. Chu.

Observation of mCherry signals indicates that F25B4.2 is expressed in ciliated neurons (Figure 5-4). Dye-filling method with DiO in *C. elegans* allows 6 pairs of amphid neurons and 2 pairs of phasmid neurons to be filled with dye. Detailed analysis of F25B4.2 expression using dye-filling shows that F25B4.2 drives gene expression in ciliated neurons, including ASK, ADL, ASI, ASH, ASJ, PHA, and PHB neurons (outlined by white dash lines in Figure 5-4). Expression in AWB was not found. Additional expression was also observed in muscle cells during larval stages but not in adults. Similar expression pattern for this gene was observed previously in *C. elegans* injected with extra-chromosomal array that contained GFP reporter driven by the same putative promoter sequence (Hunt-Newbury *et al.* 2007). To confirm whether the expression pattern indicated by mCherry is dependent on DAF-19, we crossed the strain with the mCherry reporter construct to a *daf-19* mutant strain (*m86*). We found that the expression in ciliated neurons both in the head and tail was abolished (Figure 5-4), suggesting that F25B4.2 in *C. elegans* is regulated by DAF-19. This is especially evident in the cells that dye-fill (outlined by white dash lines).

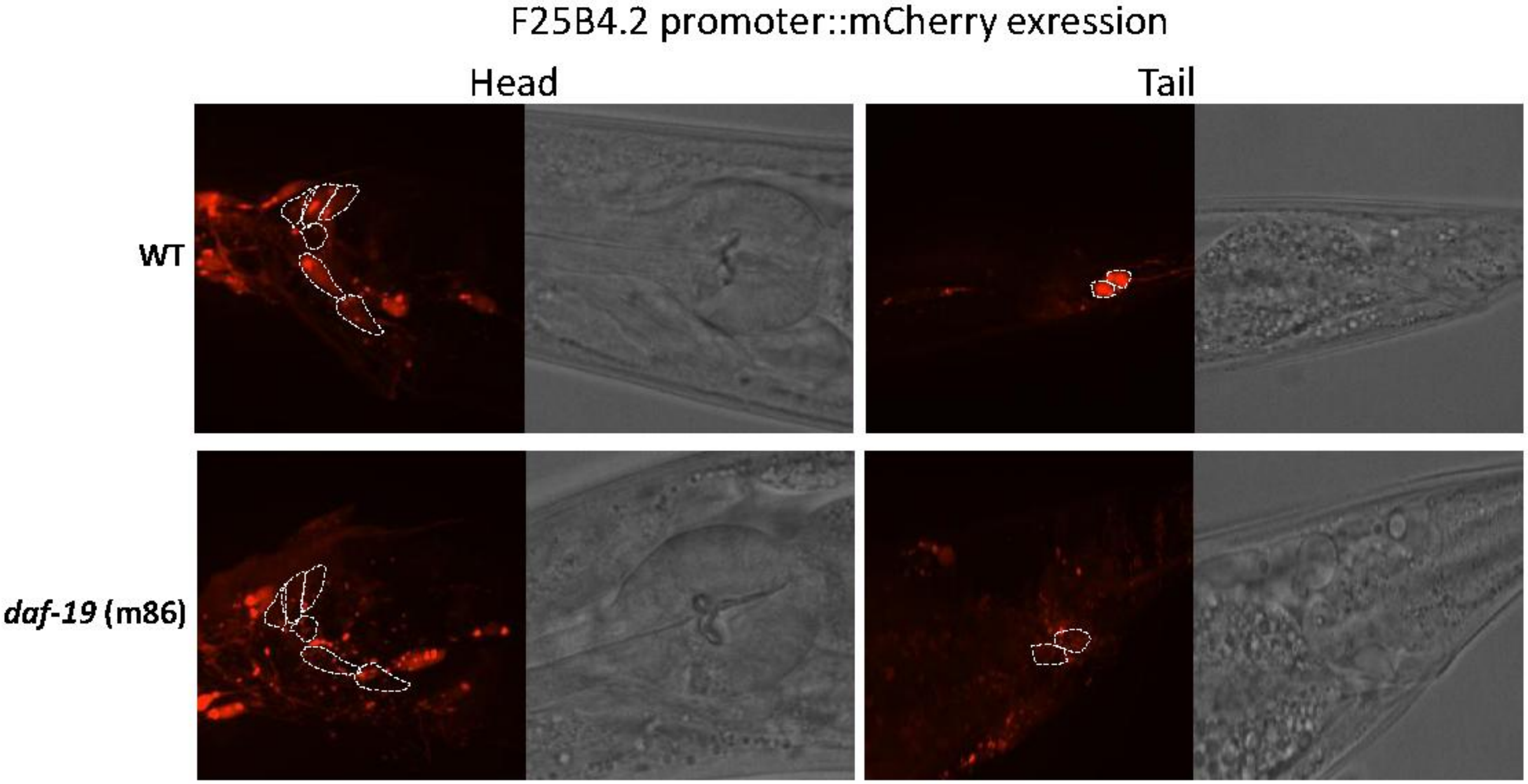


Figure 5-4 The head and tail expression patterns of F25B4.2 3kb upstream region fused to mCherry in either WT strain or *daf-19(m86)* strain. White dashed lines outline the ciliated neurons that dye fill. Neurons that dye fill in the head include ASK, ADL, ASI, AWB, ASH, and ASJ; neurons that dye fill in the tail include PHA and PHB. Because *daf-19* worms do not dye fill, the white outlines are the supposed location of these neurons. The expressions in these neurons are abolished in *daf-19(m86)* background. Exposure time = 3 seconds. The data used to generate this figure was collected by J. Chu and D. Tu.

5.4.5 Deletion analysis of putative divergent X-box motifs in F25B4.2

To test whether these two motifs are functional, we engineered three additional promoter fusion constructs with 1) only the proximal motif removed, 2) only the distal motif removed, and 3) both the proximal and distal motifs removed. If these motifs are functional, we would expect the expression pattern in ciliated neurons to be abolished. These constructs were injected and integrated using the MosSCI method (Frokjaer-Jensen *et al.* 2008). In the strain carrying the proximal deletion construct (JNC21), we observed that many amphid neurons as well as phasmid neurons lost mCherry expression (compare Figure 5-5a and b; i and j). Using dye-filling with DiO, we observed specifically that ASK, ASI, and ASJ neurons no longer show expression while ADL and ASH neurons retained expression. In the strain carrying the distal deletion construct (JNC22), we were surprised that it did not abolish any expression but instead enhanced expression (compare Figure 5-5a and c; i and k). By reducing the exposure time from 3 seconds to 800 milliseconds (about 4 fold), we were able to capture the expression intensity at a comparable level to that of JNC20 (the strain carrying the wild type promoter). In the strain carrying construct with both motifs removed (JNC29), we observed similar pattern and intensity as JNC21 where many ciliated neurons no longer show mCherry expression (Figure 5-5). Again, dye-filing with DiO reveals that ASK, ASI, ASJ neurons do not show mCherry expression anymore while ADL and ASH neurons retained expression. Taken together, our results suggest proximal motif but not the distal motif is responsible for driving F25B4.2 expression in ciliated neurons. However the distal motif may have a regulatory (repressive) role in modulating the expression level of this gene.

In order to show whether these motifs function together with DAF-19, we have crossed JNC21, JNC22, and JNC29 to a *daf-19*-deficient strain, *daf-19(m86)*. If the putative X-box motifs are functional binding sites for DAF-19, we expect these constructs in *daf-19* mutant background would show similar pattern to what was observed in Figure 5-4 where many ciliated neurons no longer show mCherry expression in *daf-19* mutant worms. As expected, we observed nearly identical expression pattern across all constructs in *daf-19* mutant strain where many ciliated neurons in both the head and tail have abolished expression in ciliated neurons (compare Figure 5-5e to h; m to p). The difference is especially striking for the distal deletion construct where an elevated expression level in wild type background dropped to very low expression in *daf-19* mutants (Figure 5-5c and g). The observation here further suggests that proximal motif is the main driving force for expression by interacting with DAF-19. We did observe one exception in the tail of JNC22 and JNC29 where a single cell is expressing mCherry in the *daf-19* mutant background (Figure 5-5o and p). This single cell is expressed towards the right side of the worm and slightly posterior to where PHA and PHB should be. Based on the cell positioning, it is most likely PQR neuron.

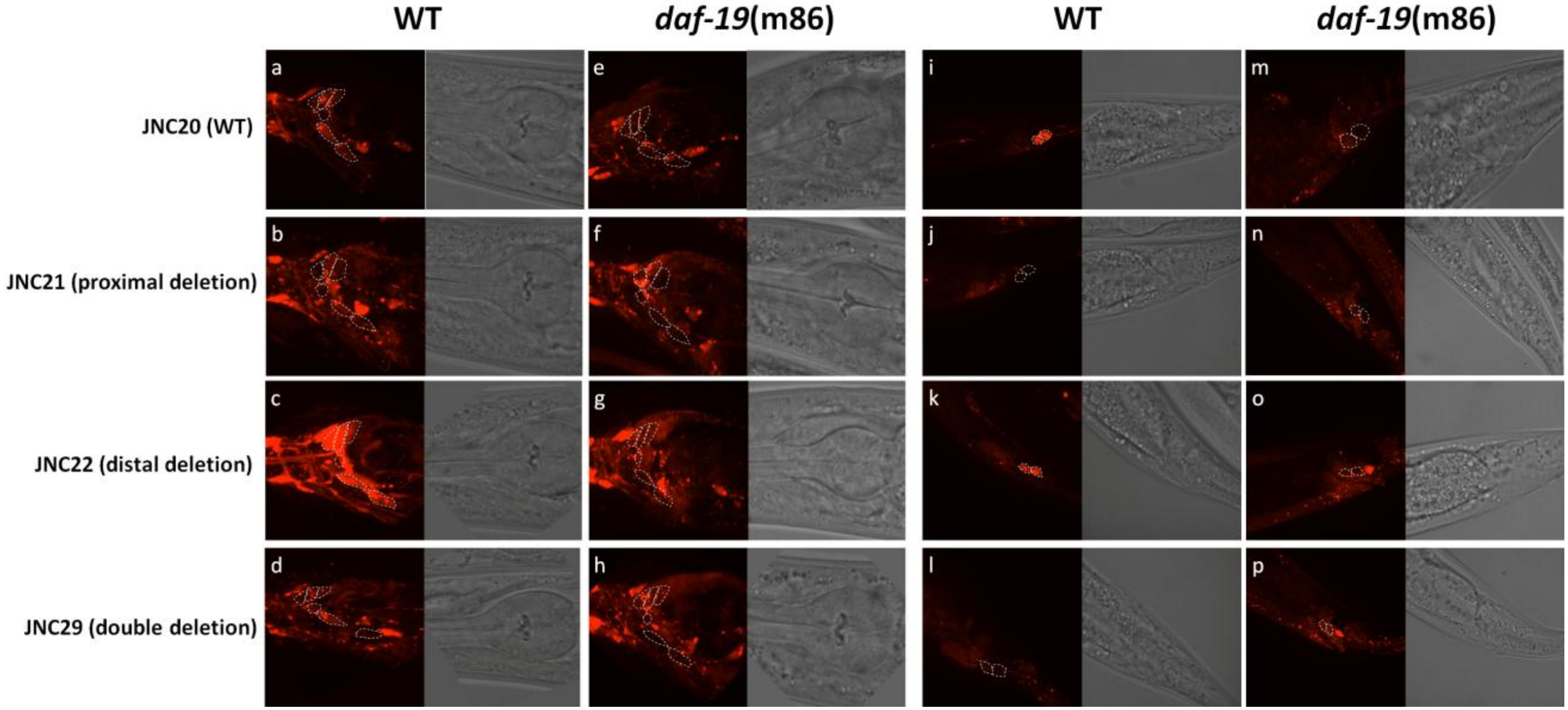


Figure 5-5 The expression of different deletion constructs in either wild type or *daf-19(m86)* backgrounds. Proximal deletion construct removes the 15bp putative X-box at -140; distal deletion construct removes the 15bp putative X-box at -190; double deletion construct removes both putative X-box motifs. Other than that, all sequences remain the same. White dashed lines outline the ciliated neurons that dye fill. The outlines for strains in *daf-19* background are supposed locations. Panels (a-g) show expression in the head and panels (i-p) show expression in the tail. Exposure time = 3 seconds. The data used to generate this figure was collected by J. Chu and D. Tu.

5.4.6 Functional analysis of proximal and distal motifs function

To further demonstrate the function of the two motifs, we used the putative divergent X-box motifs from F25B4.2 and replaced the endogenous X-box motif in the promoter of another DAF-19 regulated gene. If the motif is functional, we would expect similar ciliated neuron expression as the wild type. We chose the promoter of *dyf-5* for this experiment because *dyf-5* was identified previously to express exclusively in ciliated neurons in a DAF-19 dependent manner (Chen *et al.* 2006; Burghoorn *et al.* 2007). We replaced the endogenous X-box motif in *dyf-5* promoter region with either the proximal motif or the distal motif. Confirming what we observed before, proximal element is able to drive *dyf-5* gene expression in ciliated neurons just like the wild type promoter (Figure 5-6). On the other hand, *dyf-5* promoter replaced with distal element can only show very poor level of expression (Figure 5-6). Our results demonstrated that proximal motif is indeed an X-box motif and is able to drive gene expression through DAF-19. On the other hand, the distal motif is likely an X-box motif that plays a repressive role.

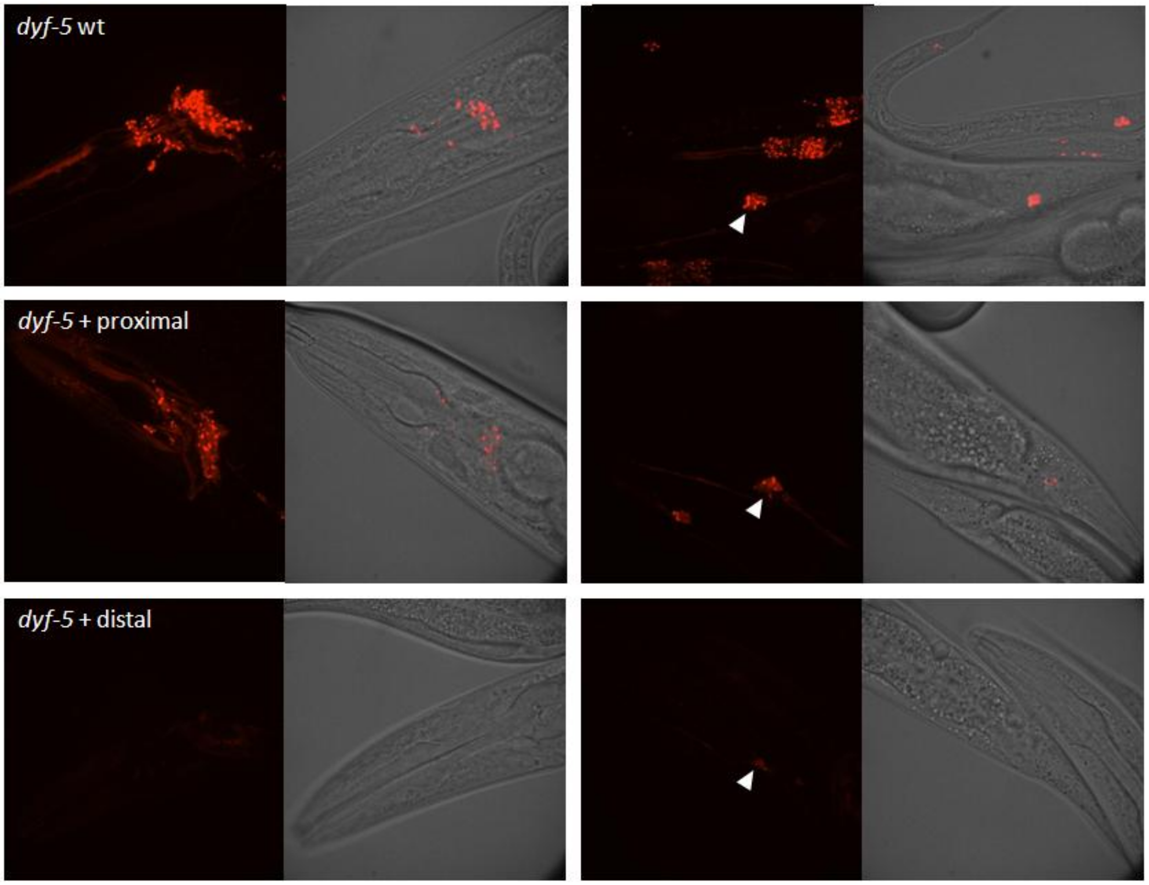


Figure 5-6 The expression pattern driven by *dyf-5* promoter replacing the endogenous X-box motif with either the proximal motif or the distal motif. Proximal motif is able to drive normal expression while distal motif is unable to. The white arrows show the location of PHA and PHB neurons. Exposure time = 3 seconds. The data used to generate this figure was collected by J. Chu and D. Tu.

5.4.7 DAF-19 target genes with two X-box motifs

Could having two X-box motifs provide an alternative regulatory mechanism for DAF-19? If this is true, we should expect to find other DAF-19 target genes with more than one X-box motifs. To examine this hypothesis, we searched the promoter region (500 bp upstream) of all 31 validated X-box motif-containing genes and see if other X-box motifs could be found in the vicinity. We used TFMscan again to search two separate half sites within the 500 bp region and identified left and right combinations without any other predictions in between. As a result, we found three genes (*osm-5*, *nph-4*, and *tub-1*) contain multiple highly probable X-box motifs within the promoter regions (Table 5-5).

Table 5-5 Known target genes with multiple X-box motifs.

Gene name	Sequence name	Position	X-box motif sequence
osm-5	Y41G9A.1	-183 -270	atctccatgacaac gtcgtcttggagac
nph-4	R13H4.1	-55 -168 -489	attgcctagaaac atttccatgacaac gtttccagaaaggaac
tub-1	F10B5.4	-67 -115	ggtgccatggcaac gttactatggcaac

5.5 Discussion

Ciliopathy is an emerging human genetic disorder caused by malformation of cilia that leads to many clinical hallmarks including obesity, polydactyly, and retinal degeneration. Swoboda and colleagues made the first link in *C. elegans* between the RFX transcription factor DAF-19 and cilia development (Swoboda *et al.* 2000). In the 10 years that followed this discovery, studies in *Chlamydomonas reinhardtii* and *C. elegans* have greatly benefited further ciliopathy research in mammals. *C. elegans*, in particular, has been instrumental in identifying the molecular nature of human BBS3, 5, 7, and 8 (Blacque *et al.* 2004; Fan *et al.* 2004; Li *et al.* 2004) by looking at the target genes of DAF-19. However, even with our current advances, the known 14 human BBS genes only constitute 25% to 50% of the ciliopathy cases (Yang *et al.* 2008). Therefore the search for additional target genes is needed. To achieve this, we need to gain a deep understanding of the molecular evolution and diversity of functional X-box motifs.

We have shown that our current comparative genomics approach is highly sensitive for DAF-19 target genes by searching for both canonical X-box motifs as well as X-box motifs that are more divergent from the known consensus. Previous studies have suggested that X-box motifs can be flexible in length (Emery *et al.* 1996; Lubelsky *et al.* 2005). Studies using RFX1 have shown that RFX DBD is able to bind to a single half site as a monomer (Siegrist *et al.* 1993; Emery *et al.* 1996). Crystal structure of RFX1 DBD with binding DNA also showed that the major protein-DNA interaction is in the “winged” part of the helix-winged DBD, which interacts with G^9, A^{12}, A^{13}, and C^{14} (Gajiwala *et al.* 2000). Lubelsky further showed using EMSA and ChIP that RFX1 binds a well conserved left half and right half separately (Lubelsky *et al.* 2005). Given the evidences from RFX1 studies, we postulate that the putative divergent X-box motifs we found are also functional since they do have a well conserved left half or right half. We demonstrated in this manuscript that the two 15 bp putative X-box motifs upstream of F25B4.2 are functional with proximal motif being the crucial motif for driving ciliated neuron expression. Our work has illustrated that comparative genomics is an effective approach for discovering divergent X-box motifs that are different from the consensus motif. This approach can be used to identify additional instances of X-box motifs, which will in turn improve our understanding on RFX/DAF-19-mediated regulation in *C. elegans* and in other organisms including humans.

It is important to be able to identify different instances of X-box motifs, and especially the divergent ones, because it is these motifs that may have different regulatory roles. For example, the PHA-4 binding site in *C. elegans* was shown to be functional when the binding sequence was altered, which also changed the temporal expression pattern of its target genes (Gaudet and Mango 2002). In a similar way, we observed that removing the distal motif cause the expression intensity to increase significantly in all expressing cells. This suggests that distal motif might have a repressive role via DAF-19 so that the combination of distal motif and proximal motif gives the right level of expression. There are known cases where RFX plays a repressive role in mammals. For example, RFX1 represses Id2 gene during cell growth arrest but activates the gene after serum induction (Wang *et al.* 2007); RFX3 represses MAP1A in non-neuronal cells (Nakayama *et al.* 2003); and RFX5 is able to repress a collagen gene COL1A2 (Sengupta *et al.* 2002).

The proximal motif identified in this study can be seen as a “strong” motif that has higher sequence conservation and drives gene expression while the distal motif can be seen as a “weak” motif that do not drive gene expression as well but may function in expression level regulation. The combination of strong and weak motifs may be a general expression level regulatory mechanism used by RFX transcription factors. In addition to F25B4.2, we also found *osm-5*, *nph-4*, and *tub-1* to have additional putative X-box motifs within 500 bp upstream. A similar phenomenon was observed among ribonucleotide reductase genes in *S. cerevisae* where RNR2, RNR3, and RNR4 are regulated by the yeast RFX gene via a strong X-box motif and a weak X-box motif (Huang *et al.* 1998). Yeast RFX negatively regulate the expression of ribonucleotide reductase genes. Removing the weak X-box motifs only show slight expression increase (1.4-1.7 fold) and removing the strong X-box motif increase the expression level by 5-fold (Huang *et al.* 1998). However, simultaneous removal of all motifs elevates the expression level by 17-fold (Huang *et al.* 1998). The strong and weak X-box motifs in yeast may work synergistically; however, our results here suggest X-box motifs in F25B4.2 work antagonistically. We postulate that distal motif work like a “sink” that binds DAF-19 but do not provide transcriptional enhancement. Other transcription factors that use cooperative binding include PurR in *Bacillus subtilis* where two binding motifs are required for high affinity binding (Bera *et al.* 2003).

Proximal and distal motif may not be the only elements at work in regulating F25B4.2. Expressions in muscle as well as many other head neurons are independent of X-box motifs or DAF-19. This suggests other transcriptional regulators also play a role in regulating F25B4.2. We also note that ADL and ASH neurons retained reporter expression in JNC21 (proximal deletion strain) and in JNC29 (double deletion strain), but show no expression in any *daf-19* mutant background strains. This expression pattern suggests possible additional DAF-19 binding sites within the 500 bp promoter region.

This project represents an important step towards identifying the entire collection of functional X-box motifs in *C. elegans*, which in turn may help identify functional X-box motifs as well as RFX target genes in humans.

6: GENERAL CONCLUSION

Since the completion of the human genome, focuses has been shifted to understanding what the functional elements are and how they function. One of the biggest challenges today is to understand gene regulation. Turning on or off a gene at the right place at the right time is critical for proper development. Thus, studying gene regulation was a major component in the ENCODE and modENCODE project where an international consortium aims to identify functional elements especially in the non-coding regions. Figuring out mechanisms of gene regulation is a key step in understanding developmental biology, gene structure and gene organization, and even evolution. Accumulating reports are hypothesizing transcriptional regulation and regulatory elements might play a major role in evolution (King 2004; Wray 2007). For instance, polymorphism in cis-regulatory elements of the KITLG locus has caused changes in pigmentation in sticklebacks and humans humans (Miller *et al.* 2007).

In my thesis, I used RFX gene family as a model to study the evolution of transcriptional regulation. RFX transcription factors are identified as master regulators of genes functioning in intraflagellar transport (IFT), a process for cilia biogenesis and maintenance. Mutations in RFX3 or its target genes causes sever defects in cilia structure, which leads to a variety of disease conditions labelled as ciliopaties. The medical relevance and the regulatory function of IFT made RFX an excellent model system. I searched and compared 153 species for the presence of RFX and ciliary genes and found that RFX is widely found in all metazoans, some fungi, and one choanoflagellate; but RFX is not found in plants/algea, protists or any prokaryotes. In my search, I identified two additional RFX genes, RFX6 and RFX7, as well as nine RFX genes in fishes. Identification of RFX6 and its exclusive expression pattern in the pancrease led to the understanding of its role in insulin production.

My data further suggested convergent evolution of RFX genes and ciliary genes. Ciliary genes are found in many eukaryotic species, ranging from algae to humans. Yet, outside of metazoans, RFX and ciliary genes do not co-exist. The two systems likely evolved independently but converged in Opisthokonts. However, why did RFX and ciliary genes persist together in metazoan evolution but not in fungi? A possible answer is that RFX acquired regulatory function on ciliary genes just prior to the establishment of metazoans but this event did not happen in fungi. This idea is supported by studies in yeast, which showed that yRFX functions in DNA repair.

Intrestingly, RFX and ciliary genes can both be found in choanoflagellates, a group of species proposed to be the closest single cellular relative to metazoans. Many genes, including adhesion proteins and transcription factors, that are thought to be metazoans specific were found in the choanoflagellate *M. brevicollis*. I believe the meeting of RFX and ciliary genes in choanoflaggelates is not a conincedent and likely to have a major impact in metazoan evolution. With an increasing number of chanoflagellate species being sequenced, RFX transcription factor family will be an excellent model for studying early transcription factor evolution in detail. More importantly, *M. brevicollis* is being developed as a model organism, which will provide the opportunity to investigate whether RFX does in fact regulate ciliary genes in choanoflagellates. The results of such future endeavour will have a dramatic impact in the way we understand evolution.

In my thesis, I have presented RFX as an excellent system for studying transcriptional regulation. The seemingly simple system of RFX binding X-box motifs to drive IFT gene transcription turned out to have many complexities. Even in *C. elegans* that only has a single RFX genes (DAF-19) yet produces multiple alternative transcripts with isoform specific expression patterns. Efimenko and colleagues have shown that DAF-19 do not regulate genes equally: while some target genes are expressed in all ciliated neurons, some only expression in a subset of ciliated neurons. How does DAF-19/RFX vary in their regulatory function? One hypothesis is that DAF-19 relies on different co-factors, similar to RFX5, to regulate genes in a tissue specific manner. Future investigations using co-IP or yeast-2-hybrid methods will be valuable in tackling this hypothesis. Another explanation, which I have shown in my thesis, is the different configuration of X-box motifs in promoter regions. Different X-box configurations in promoters have now been observed: promoter with a single strong X-box motif, promoter with a single weak X-box motif, and promoter with a strong and weak X-box motif each.

Promoter with a single strong X-box motif. This is the most common configuration observed in *C. elegans*. Genes with a single strong X-box motif depend on X-box motifs and RFX to expressed in most, if not all, of the ciliated neurons. Removing X-box motifs will completely abolish expression.

Promoter with a single weak X-box motif. Efimenko *et al*. identified a number of genes that contain X-box motifs with higher sequence variation from the consensus (weak X-box motifs). These genes are only expressed in a subset of ciliated neurons. Some genes (*xbx-6*, *nhr-44*) are also expressed in other tissue types. DAF-19 is not the only TF that promoter differential expression pattern. PHA-4 binding site in *C. elegans* was shown to be functional when the binding sequence was altered, which also changed the temporal expression pattern of its target genes (Gaudet and Mango 2002).

Promoter with a strong and weak X-box motif. F25B4.2 was identified with a strong X-box motif (proximal motif) that drives ciliated neuron expression effectively and a weak X-box motif (distal motif) that drives ciliated neuron expression very poorly. However, the weak X-box motif is important to function together with the strong X-box motif to drive proper expression level. Cooperation of regulatory motifs have been observed in both prokaryotes and eukaryotes. For instance, PurR in *Bacillus subtilis* binds to two PurBox motifs for higher affinity (Bera *et al.* 2003). *Drosophila* gap gene hunchback (hb) is regulated by multiple bicoid (bcd) binding sites (Driever and Nusslein-Volhard 1989). Multiple NFkB binding sites work synergistically to regulate US3 gene of human cytomegalovirus (Chan *et al.* 1996). More recently, RFX in other systems was discovered to exhibit similar characteristic. *S. cerevisiae* RFX negatively regulates many ribonucleotide reductase genes (e.g. RNR2, RNR3, and RNR4) through a combination of strong X-box motifs and weak X-box motifs (Huang *et al.* 1998). Human RFX1 represses MAP1A in non-neuronal cells by binding to two X-box motifs in the first exon (Nakayama *et al.* 2003). All of the above examples shows synergistic effects of having multiple binding sites of a particular transcription factor. However, the cooperation of X-box motifs in F25B4.2 seems to be antagonistic where the weak motif dampens the activity of the strong motif. My findings represent a novel way for RFX to regulate genes. Combination of strong and weak motifs, either synergistically or antagonistically, is likely a general mechanism for RFX to achieve greater dynamic range in its regulatory repertoire. Due to the high homology of RFX DBD across species, including humans, the lessons learned in *C. elegans* are directly applicable to understanding RFX regulatory mechanisms in humans.

APPENDICES

Appendix A: Full list of putative RFX genes in 154 species

ID	Specie name	Chrom	Chrom start	Chrom end	Putative DBD seq
S1	*Acyrthosiphon pisum*	SCAFFOLD15692	528	815	VDWLMNNYEKAEGVSLLRSTIYDNYLTHCSETKFDPLNAPSFGKLIRSVFTLLTVFLFIRGNSKYHYYGIRIKP
S2	*Acyrthosiphon pisum*	SCAFFOLD10511	25720	28907	TVNWLMENYEMAEGVSLPRSTLYNHYLTHCSETKIDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRIKA
S3	*Acyrthosiphon pisum*	SCAFFOLD10540	63941	82271	TVDWLMDNYEKAEGVSLLRSTIYDNYLTHCSETKFDPLNAPSFGKLIRSVFTGLQTRRLGTRGNSKYHYYGIRIK
S4	*Aedes aegypti*	supercont1.68	2491115	2522813	WLVDNYENAEGVSLPRSTLYNHYMRHCNEHKLDAVNAASFGKLIRSVFTGLRTRRLGTRGNSKYHYYGIRIKP
S5	*Aedes aegypti*	supercont1.910	93090	125072	WVRSHLEHDPNVSIPKQEVYDDYTAFCERIDIKPLSTADFGKVMKQVFPGIRPRRLGTRGHSRYCYAAMRK
R6	*Anolis carolinensis*	gi|126570215|gb|DS229241.1|	2770532	2771768	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
R7	*Anolis carolinensis*	gi|126570583|gb|DS229122.1|	2336017	2337385	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
R8	*Anolis carolinensis*	gi|126570369|gb|DS229181.1|	1446403	1453189	LEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKVIRQQFPQLTTRRLGTRGQSK
R9	*Anolis carolinensis*	gi|126569319|gb|DS230137.1|	94992	95360	ACTWIQNHLEEYPDTCLPKQDVYDAYKRYCDNLCCRSLSAANFGKIMREIFPNIKARRLGGRGQSKY
R10	*Anolis carolinensis*	gi|126569937|gb|DS229519.1|	484177	486759	LEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSK
S11	*Anopheles gambiae*	2R	46615186	46615410	VSWLMENYETADGVSLPRSTLYNHYMWHCNENKLDAVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S12	*Anopheles gambiae*	2R	24319508	24319833	WVRSHLEHDPNVSIPKQEVYDDYRAYCARINIKPLSTADFGKVMKQVFPGIRPRRLGTRGHSRYCYAAMRK
S13	*Apis mellifera*	GroupUn.167	35877	36227	WIKTHLEEDPDVSLPKQEVYDEYKMYCMRNSMKPLSTADFGKVMKQVYPRVRPRRLGTRGNSRYCYAGMRKR
U14	*Aspergillus clavatus*	supercontig_1.76	204541	204732	LKENCRKSSGSVRRDRVYCCYAEKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHY
U15	*Aspergillus flavus*	supercontig_2.12	185796	185987	LKENCRKSSGSVRRDRVYCCYAEKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHY
U16	*Aspergillus fumigatus*	supercontig_null.5	1462039	1462230	LKENCRKSSGSVRRDRVYCCYAEKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHY
U17	*Aspergillus nidulans*	supercontig 1.2	2779655	2779867	LKENCRKSTGSVRRDRVYCCYAEKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHYVDLTVIE
U18	*Aspergillus niger*	chr_5_1	725052	725243	LKENCRKSSGSVRRDRVYCCYAEKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHY
U19	*Aspergillus oryzae*	supercontig_1.1	1724987	1725178	LKENCRKSSGSVRRDRVYCCYAEKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHY
U20	*Aspergillus terreus*	supercontig 1.14	1166283	1166474	LKENCRKSSGSVRRDRVYCCYAEKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHY
S21	*Bombyx mori*	nscaf3055	1405987	1423135	VQWLLDHYETADGVSLPRSSLYAHYLRHCTSHRLEPVNAASFGKLIRSVFVGLRTRRLGTRGNSKYHYYGIRAKP
S22	*Bombyx mori*	nscaf2888	357836	368482	TWIQTHLEVDPDVSLPKQDVYDEYIAHCMSSNMKPLSTADFGKVMKQVYPSVRPRRLGTRGNSRCEVRK
M23	*Bos taurus*	7	9925855	9933480	TVQWLLENYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M24	*Bos taurus*	7	16919250	16961506	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M25	*Bos taurus*	8	43460999	43562649	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M26	*Bos taurus*	5	75052404	75195661	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M27	*Bos taurus*	3	20975655	20980557	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M28	*Bos taurus*	9	35415208	35467208	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M29	*Bos taurus*	10	55128499	55168076	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
N30	*Brugia malayi*	Bmal_supercontig14687	47035	54079	TIQWLINNYEPADGTSLPRCTLYSHYIKHCNENKLEPVNAASFGKLIRSVFHGLRTRRLGTRGNSKYHYYGIRIKP
N31	*Caenorhabditis brenneri*	Cbre_Contig399	47418	47700	VNWLFDNYEIAEGSLPRCQLYDHYRKHCEEHRMDPVNAASFGKLIRSVFQNLKTRRLGTRGNSKYHYYGIRMKE
N32	*Caenorhabditis brenneri*	Cbre_Contig120	184629	184903	VNWLFDNYEIAEGSLPRCQLYDHYRKHCEEHRMDPVNAASFGKLIRSVFQNLKTRRLGTRGNSKYHYYGIRMKE
N33	*Caenorhabditis briggsae*	chrII	922251	922556	IGWLFENYEIAEGSLPRCQLYDHYRKHCEEHRMDPVNAASFGKLIRSVFQNLKTRRLGTRGNSKYHYYGIKMKD
N34	*Caenorhabditis elegans*	II	10157446	10169268	TVNWLFENYEIGEGSLPRCELYDHYKKHCAEHRMDPVNAASFGKLIRSVFHNLKTRRLGTRGNSKYHYYGIRLKD
N35	*Caenorhabditis japonica*	Cjap_Contig1772	11957	12230	TVKWLLDNYETADGSLPRCQLYDHYRKHCSEHRMDAVNAASFGKLIRSVFLNLKTRRLGTRGNSKYHYYGIKIKE
N36	*Caenorhabditis remanei*	Crem_Contig211	11914	13255	TVNWLFDNYEIAEGSLPRCQLYDHYRKHCAEHRMDPVNAASFGKLIRSVFQNLKTRRLGTRGNSKYHYYGIRMKE
F37	*Callorhinchus milii*	AAVX01023624.1	15	1674	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
F38	*Callorhinchus milii*	AAVX01302999.1	255	434	LQWLLDNYETAEGVSLPRSSLYNHYLRHCQEQKLDPVNAASFGKLIRSVFMGLRTRRLGT
F39	*Callorhinchus milii*	AAVX01026212.1	106	285	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEQKLDPVNAASFGKLIRSIFMGLRTRRLGT
F40	*Callorhinchus milii*	AAVX01161766.1	412	999	AFNWIRNHLEDHPDTSLPKQEVYNERGYCDNLGYHSLSAADFGKIMKNVFPNMKARRLGTRGKSKYPF
U41	*Candida albicans*	supercontig_1.2	1301142	1301372	MVWLLNSCDLAPTAVIPRNRIYARYVQVCADNNLAPVSPASFGKLVKILYPNITTRRLGMRGQSKYHYCGIKLTGDE
U42	*Candida guilliermondii*	supercontig_1.6	326189	326401	WLLGVCEVSSTAVVPRNRVYARYVQSCANFGLVPITPTNMGKLVKLMFPGLRIRRLGVRGRSKYHYNGIRL
U43	*Candida lusitaniae*	supercontig_1.7	225658	225870	WIQRSCEHAPAAVVPRTRVYARYVQRCADLALHPLAPALFGRLVRVAYPNLTIRRLGVRGKSKYHYCGVRL
U44	*Candida parapsilosis*	contig_1.135	92664	92885	MVWLLTSCEVSPTAVIPRNRIYARYVQICADNSLSPLSPASFGKLVRILYPTITTRRLGMRGQLKYHYCGIRLK
U45	*Candida tropicalis*	supercontig_3.1	1372009	1372239	MVWLLNSCELSPTAVIPRNRIYARYVQVCADNSLAPVSPASFGKLVKILYPNITTRRLGMRGQSKYHYCGIKLNGDE
M46	*Canis familiaris*	20	51497894	51519585	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M47	*Canis familiaris*	20	57014252	57054444	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP

M48	*Canis familiaris*	1	94845887	95002161	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M49	*Canis familiaris*	10	35119527	35257965	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M50	*Canis familiaris*	17	63473796	63477569	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M51	*Canis familiaris*	1	60291581	60346011	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M52	*Canis familiaris*	30	24303138	24347785	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
M53	*Cavia porcellus*	scaffold_42	12840138	12855859	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M54	*Cavia porcellus*	scaffold_250	470795	494743	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M55	*Cavia porcellus*	scaffold_21	16445737	16564764	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRNESPMTYFQEKASP
M56	*Cavia porcellus*	scaffold_171	147020	296813	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKSDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M57	*Cavia porcellus*	scaffold_2	12375363	12379772	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M58	*Cavia porcellus*	scaffold_1	72390699	72449966	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKERLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M59	*Cavia porcellus*	scaffold_23	32154333	32186858	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
U60	*Chaetomium globosum*	supercontig_1.5	2835815	2836006	GKGSVPRGRVYANYASRCATERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFQLRE
I61	*Ciona intestinalis*	5q	5499071	5511694	TVHWLMDHFENSEGVSLPRALMYNHYLLHCQEQQLDPVNAASFGKLVRSVFIGLRTRRLGTRGNSKYHYYGIRIK
I62	*Ciona intestinalis*	4q	1678121	1679000	TTEWLTKNFEENSLTSVPRSIMFDEYQKFCRDSNTKPFNQAVFGKIVRACFPNLTTRRLGVRGQSRYHYAGLSVK
I63	*Ciona intestinalis*	4q	1712625	1713510	TTEWLTKNFEENSLTSVPRSIMFDEYQKFCRDSNTKPFNQAVFGKIVRACFPNLTTRRLGVRGQSRYHYAGLSVK
I64	*Ciona intestinalis*	12q	3596830	3597452	LEKNYVICDGVCLARCILYSHYLDFCNKSNIEPACAATFGKTIRHKFPLLTTRRLGTRGHSK
I65	*Ciona savignyi*	reftig_60	567930	575867	VHWLLEHFENSEGVSLPRALMYNHYLLHCQDQHLDPVNAASFGKLVRSVFIGLRTRRLGTRGNSKYHYYGIRIKA
I66	*Ciona savignyi*	reftig_9	3427758	3428417	LEENYMICDGICLARCILYNHYLDFCNKSNIEPACAATFQTIRHKFPLLTTRRLGTRGHSK
U67	*Coccidioides immitis*	supercontig_2.3	788844	789026	WLRENCRKSTGSVRRDRVYCCYADKCGTERVSVLNPASFGKLTRRLGVRGESKYHYVDLSI
U68	*Coccidioides posadasii*	supercontig_1.1	3409417	3409599	WLRENCRKSTGSVRRDRVYCCYADKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHYVDLSI
U69	*Cochliobolus heterostrophus*	scaffold_7	232298	234013	AMLWLQSVCRVAKTSVPRNRVYSKYAERCGTDRVIPLNPASFGKLVRVIFPGIQTRRLGVRGESKYHYVDLEL
U70	*Coprinus cinereus*	supercontig_2.5	1584577	1584762	WLTANYATYPDGNVPRQGLYFSYRRVCDQYGIPHINTATLGKAIRLCFPTIKTRRLGVRGNSKYHYCGIR
S71	*Culex quinquefasciatus*	supercont3.57	430472	439517	TVAWLVENYENAEGVSLPRSTLYNHYMRHCNEHKLDAVNAASFGKLIRSVFTGLRTRRLGTRGNSKYHYYGIRIKP
S72	*Culex quinquefasciatus*	supercont3.119	136315	137757	NWVRSHLEHDPNVSIPKQEVYEDYIAFCERIDIKPLSTADFGKVMKQVFPGIRPRRLGTRGHSRYCYAAMRK
F73	*Danio rerio*	3	16190097	16238314	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
F74	*Danio rerio*	8	17838466	17903457	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEQKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
F75	*Danio rerio*	1	50190919	50191098	VQWLMDNYETAEGVSLPRCTLYCHYLLHCQQTKLEPVNAASFGKLIRSVFMGLRTRRLGT
F76	*Danio rerio*	10	110739	135581	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEQKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
F77	*Danio rerio*	18	18936788	18976167	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKLDSQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
F78	*Danio rerio*	19	5172385	5175954	CNWIRSHLEEHPDTCLPKQDVYETYRKHCDNLQHRPLSAANFGKIIRDIFPNIKARRLGGRGHGIRRKT
F79	*Danio rerio*	20	43625190	43652681	TLQWLEDNYIVCEGVCLPRCILYAHYLDFCRKEKLDPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
F80	*Danio rerio*	Zv7_NA1148	36204	36494	AFNWIRNHLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGMRGKSKYPF
F81	*Danio rerio*	7	28955594	28957779	AFSWIRNHLEEHPETSLPKQEVYDEYKSYCDSLGYHALSAADFGKIMKNVFPNMKARRLGMRGKSKYP
M82	*Dasypus novemcinctus*	GeneScaffold_944	2835	27766	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M83	*Dasypus novemcinctus*	GeneScaffold_798	39166	142641	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M84	*Dasypus novemcinctus*	scaffold_5812	27914	28120	LNIHEAAEGVSLPGSTLYNQKHYLXDKLDPINAASFGKLIKSSFMGVCIRRLGIRRNSXYCNYGICIKP
M85	*Dasypus novemcinctus*	GeneScaffold_1956	2952	138989	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M86	*Dasypus novemcinctus*	GeneScaffold_6862	35121	106217	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKXXXXXXXXXXXTIRQKFPLLTTRRLGTRGHSKYHYYGIGIPE
M87	*Dasypus novemcinctus*	GeneScaffold_6646	31033	146200	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNIFPNMKARRLGTRGKSKYLECGLRKKA
U88	*Debaryomyces hansenii*	contig_1.7	921459	921677	MVWLLNSCESSATAVVPRNRIYARYVQICADNSLKPLSPASFGKLVRILFPNLTTRRLGMRGQSKYHYCGIKL
M89	*Dipodomys ordii*	GeneScaffold_6195	74027	92444	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M90	*Dipodomys ordii*	GeneScaffold_704	8463	109040	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M91	*Dipodomys ordii*	GeneScaffold_1755	17906	164955	TLQWLEESYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M92	*Dipodomys ordii*	GeneScaffold_3536	6746	12167	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRVQSKYCYSGIRRKT
M93	*Dipodomys ordii*	GeneScaffold_5466	50250	101276	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M94	*Dipodomys ordii*	GeneScaffold_5830	10741	23523	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSK
S95	*Drosophila ananassae*	scaffold_13340	915815	924627	TIKWLSRNYETADGVSLPRSTLYNHYMQHCSEQKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S96	*Drosophila ananassae*	scaffold_13266	18499904	18505287	LSWLGATYERAGSLRVEQAELYRIYLSHCQKSKLSVVNHKQFPRLVRLIFVGVIVRQLDGTELPGSYYVGIRMR
S97	*Drosophila ananassae*	scaffold_13340	13503014	13507218	NWVRSHLEHDAKVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S98	*Drosophila erecta*	scaffold_4770	15418429	15427892	TIKWLSRNYETADGVSLPRSTLYNHYMQHCSEHKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S99	*Drosophila erecta*	scaffold_4929	20014930	20020991	LAWLGATYERANDLRVEQAELYRIYLSHCQKAKLSVVNHMQFPRLVRLIFVGVIVRHLDGTELPGTYYVGIRMR
S100	*Drosophila erecta*	scaffold_4770	2523790	2528061	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK

S101	*Drosophila grimshawi*	scaffold_14906	5323969	5337621	TIKWLSRNYETADGVSLPRSTLYNHYMQHCNEQKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S102	*Drosophila grimshawi*	scaffold_15112	165166	170758	LAWLGATYERAHDHRVEQQELYTIYLSHCQKAKHSVVNRMQFPRLVRLIFVGPAVRQLDGSDLPGTHYVGIRMR
S103	*Drosophila grimshawi*	scaffold_14906	12498019	12502575	NWVRSHLEHDAQVSIPKQDVYNDYIVYCERLNIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S104	*Drosophila melanogaster*	3R	6188885	6198708	TIKWLSRNYETADGVSLPRSTLYNHYMQHCSEHKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S105	*Drosophila melanogaster*	2R	2524194	2529791	LAWLGATYERANDLRVEQAELYRIYLSHCQKAKLSVVNHMQFPRLVRLIFVGVIVRHLDGIELPGTYYVGIRMR
S106	*Drosophila melanogaster*	3R	2277571	2281743	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S107	*Drosophila mojavensis*	scaffold_6540	4957327	4971281	TIKWLSRNYETADGVSLPRSTLYNHYMQHCNEQKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S108	*Drosophila mojavensis*	scaffold_6496	24946890	24952398	LAWLGATYERAHDYRVEQQELYTIYLSHCQKAKHSVVNRVQFPRLVRLIFVGPAVRQMDGTELPGTHYVGIRMR
S109	*Drosophila mojavensis*	scaffold_6540	14223712	14228062	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S110	*Drosophila persimilis*	scaffold_34	169039	173282	LAWLGATYERAGAFRLEQQELYRIYLSHCQKAKLSVVNHMQFPRLVRLIFVGVIVRHLDGTELPGTYYVGIRTR
S111	*Drosophila persimilis*	scaffold_3	2843930	2848291	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S112	*Drosophila pseudoobscura*	2	7268071	7278147	TIKWLSRNYETADGVSLPRSTLYNHYMQHCNEQKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S113	*Drosophila pseudoobscura*	3	18225993	18231492	LAWLGATYERAGAFRLEQQELYRIYLSHCQKAKLSVVNHMQFPRLVRLIFVGVIVRHLDGTELPGTYYVGIRTR
S114	*Drosophila pseudoobscura*	2	20099132	20103344	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S115	*Drosophila sechellia*	scaffold_0	15743429	15752974	TIKWLSRNYETADGVSLPRSTLYNHYMQHCSEHKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S116	*Drosophila sechellia*	scaffold_1	187778	193430	LAWLGATYERANDLRVEQAELYRIYLSHCQKAKLSVVNHMQFPRLVRLIFVGVIVRHLDGIELPGTYYVGIRMR
S117	*Drosophila sechellia*	scaffold_6	2364297	2368469	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S118	*Drosophila simulans*	3R	15194928	15249245	TIKWLSRNYETADGVSLPRSTLYNHYMQHCSEHKLEPVNAASFGKLIRSLFSGLRTRRLGTRGKSKYHYYGI
S119	*Drosophila simulans*	2R	1347483	1353107	LAWLGATYERANDLRVEQAELYRIYLSHCQKAKLSVVNHMQFPRLVRLIFVGVIVRHLDGIELPGTYYVGIRMR
S120	*Drosophila simulans*	3R	2304459	2308605	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S121	*Drosophila virilis*	scaffold_13047	954683	966910	TIKWLSRNYETADGVSLPRSTLYNHYMQHCNEQKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S122	*Drosophila virilis*	scaffold_12875	5139624	5145541	LAWLGATYERAGNYRVEQQELYTIYLSHCQKAKLSVVNRLQFPRLVRLIFVGPAVRQMDGTDLPGTHYVGIRMR
S123	*Drosophila virilis*	scaffold_13047	17440606	17445026	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S124	*Drosophila willistoni*	scf2_1100000004902	107805	121072	TIKWLSRNYETADGVSLPRSTLYNHYMQHCNEQKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S125	*Drosophila willistoni*	scf2_1100000004512	1122060	1127300	LAWLGATYERAHDCRVDQQELYRIYLSHCQKTKLSVVNHVQFPRLVRLIFVGVIVRQLDGTELPGTHYVGIKMR
S126	*Drosophila willistoni*	scf2_1100000004943	3540318	3544731	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
S127	*Drosophila willistoni*	scf2_1100000004511	6034048	6035799	AFPPIRNDNVPKPRLLLDKYQSYDDVMEFLDVVGYHPVSLVDVGRSYENLKTIVISNSDGRRGKNVFMDAGLHAR
S128	*Drosophila yakuba*	3R	10216429	10226471	TIKWLSRNYETADGVSLPRSTLYNHYMQHCSEHKLEPVNAASFGKLIRSVFSGLRTRRLGTRGNSKYHYYGIRIKP
S129	*Drosophila yakuba*	2L	15270081	15275972	LAWLGATYERANDLRVEQAELYRIYLSHCQKAKLSVVNHMQFPRLVRLIFVGVIVRHLDGTELPGTYYVGIRMR
S130	*Drosophila yakuba*	3R	18198136	18202392	NWVRSHLEHDAQVSIPKQDVYNDYIAYCERLSIKPLSTADFGKVMKQVFPGVRPRRLGTRGNSRYCYAAMRK
M131	*Echinops telfairi*	GeneScaffold_3517	4434	21873	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M132	*Echinops telfairi*	scaffold_174732	5576	7486	LQWLLDNYETAEGVSLPRSSVYSHYLRHCQDHKLDPVNAASFGKLIRCVFMGLRTRRLGTRGNSKCHYYGIRLK
M133	*Echinops telfairi*	GeneScaffold_1025	608	114908	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTXXXXXXHYYGIRLKP
M134	*Echinops telfairi*	GeneScaffold_859	365	130308	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGT
M135	*Echinops telfairi*	GeneScaffold_2217	91	152248	TLQWLEENYEIAKGVCIPRSALYMHYLGFCEKNDTXXXXXXXXXXIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M136	*Echinops telfairi*	GeneScaffold_4516	158553	162286	AYKWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M137	*Echinops telfairi*	GeneScaffold_7797	4331	138015	TLQWLDENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKXXXXXXXXXXXXXXXXXXXXXYHYYGIGIKE
M138	*Echinops telfairi*	GeneScaffold_7531	540	18343	FSRIGNTLEEHPESSLPKQEVDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYPLGGLRKKA
M139	*Equus caballus*	Un0116	68373	95382	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M140	*Equus caballus*	7	3650858	3690354	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M141	*Equus caballus*	23	24949918	25095620	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M142	*Equus caballus*	28	29570191	29729790	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M143	*Equus caballus*	5	45922019	45926788	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M144	*Equus caballus*	10	65440510	65498175	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSNYHLGIVNIR
M145	*Equus caballus*	1	1.35E+08	1.35E+08	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYPFDGLRKKA
M146	*Erinaceus europaeus*	GeneScaffold_920	8165	170722	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M147	*Erinaceus europaeus*	GeneScaffold_2248	1064	202739	LEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHY
M148	*Erinaceus europaeus*	scaffold_344002	44149	47774	RNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M149	*Erinaceus europaeus*	GeneScaffold_7901	46211	129633	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M150	*Erinaceus europaeus*	GeneScaffold_7653	4934	22951	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYPLPGLRKKA
M151	*Felis catus*	scaffold_158686	64452	74488	LVDNFCICEGSVPRCLMYEIYVETCGHNTQTQVNPATFGKVVRLVFPDLGTRRLGTRGSARYHF
M152	*Felis catus*	GeneScaffold_5267	178544	208726	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTR
M153	*Felis catus*	GeneScaffold_5000	126702	180956	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M154	*Felis catus*	GeneScaffold_495	17919	180453	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTR

M155	*Felis catus*	GeneScaffold_103	173581	331167	LEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M156	*Felis catus*	GeneScaffold_2523	274955	278768	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M157	*Felis catus*	GeneScaffold_4221	43114	67821	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
U158	*Fusarium graminearum*	supercontig_3.4	3458625	3458816	AMLWINSVCSSGKGSVPRGRVYANYASKCANERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFTL
U159	*Fusarium oxysporum*	supercontig_2.2	178165	178362	WIHGVCERGKGSVPRGRVYANYASRCATERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFTLKE
U160	*Fusarium verticillioides*	supercontig_3.10	184472	184669	WIHGVCERGKGSVPRGRVYANYASRCATERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFTLKE
B161	*Gallus gallus*	28	1175469	1228576	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
B162	*Gallus gallus*	Z	27469709	27499519	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
B163	*Gallus gallus*	1	55755290	55828747	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
B164	*Gallus gallus*	25	1894548	1896996	ACNWIRNHLEEHADTCLPKQDVYDAYRQYCDNLCCRPLSAANFGKIIREIFPNIKARRLGGRGQSKYCYSGIRRKT
B165	*Gallus gallus*	3	66310220	66344069	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
B166	*Gallus gallus*	10	8826902	8835953	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYPL
B167	*Gallus gallus*	1	1.38E+08	1.38E+08	IADNFYLCEGTIPRWLLYEMYMENFSSNDNDKVNSATFGKVQLVFPGLGTRRLGTRGSARY
F168	*Gasterosteus aculeatus*	groupXI	11622284	11631967	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
F169	*Gasterosteus aculeatus*	groupIX	15059848	15063665	TVQWLCDNYEGAEGVSLPRCTLYYHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKS
F170	*Gasterosteus aculeatus*	groupVIII	8542911	8565618	HLQWLLDNYETAEGVSLPRCSLYNHYLRHCQEQKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRVKP
F171	*Gasterosteus aculeatus*	groupXIV	6061641	6067515	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEQKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
F172	*Gasterosteus aculeatus*	groupXIX	15070736	15085469	TLEWLEENYEIAEGVCIPRSALYMHYLDFSEKQDTQPVNAASFGKIIRQQFPALTTRRLGTRGQSKYHYYGIAVKE
F173	*Gasterosteus aculeatus*	groupX	15264797	15265090	CNWIRSHLEEHCDTCLPKQDVYETYRRHCENLQHRPLSAANFGKIIRDIFPNIKARRLGGRGQSKY
F174	*Gasterosteus aculeatus*	groupXVIII	3056044	3064717	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
F175	*Gasterosteus aculeatus*	groupXIX	7462024	7468520	AFNWIRNHLEEHPETSLPKQEVYDEYKSYCDNLGYNPLSAADFGKIMKNVFPNMKARRLGMRGKSKYCYSGLRKKA
F176	*Gasterosteus aculeatus*	groupII	11510431	11517374	AINWIRHHLEEYPETSLPKQEVYDEYKSFCDNLNYHPLSAADFGKIMKNVFPNMKARRLGMRGKSKYCYSGLRKR
M177	*Gorilla gorilla*	GeneScaffold_3729	11964	47714	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M178	*Gorilla gorilla*	GeneScaffold_1090	42251	42433	LQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTR
M179	*Gorilla gorilla*	GeneScaffold_2315	4572	137417	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M180	*Gorilla gorilla*	scaffold_25868	2076	6245	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M181	*Homo sapiens*	19	13933353	13978097	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M182	*Homo sapiens*	19	5944175	6061554	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M183	*Homo sapiens*	9	3214649	3515983	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M184	*Homo sapiens*	12	1.06E+08	1.06E+08	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M185	*Homo sapiens*	1	1.50E+08	1.50E+08	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M186	*Homo sapiens*	6	1.17E+08	1.17E+08	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M187	*Homo sapiens*	15	54170024	54322775	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
I188	*Hydra magnipapillata*	gi|196137731|gb|EQ252937.1|	97730	98074	VQWLIENYETADGVSLPRSTLYSHYLRHCSESKIDAVNAASFGKLIRSVFLGLKTRRLGRGNSKYHYYGIRVKP
I189	*Hydra magnipapillata*	gi|196137066|gb|EQ253602.1|	31911	35330	LHENYDMLEGISLRRIALHTHYLDFCNSTNVTPVHAASLGNVIRSTFPELKTRRLGTRGKSK
I190	*Hydra magnipapillata*	gi|196130311|gb|EQ260357.1|	114854	121336	TLEWLDENFTHYPGVCLPRCIMYAHYLTFCQENQLHQMCAATFGKIIRQKFPELTTRRLGTRGNSKYHYYGVAIKE
U191	*Laccaria bicolori*	scaffold_4	1793299	1796249	WLTANYAPYPDGNVPRQGLYFSYRRVCDQYGIPHINTATLGKAIRLCFPTIKTRRLGVRGNSKYHYCGIR
U192	*Lodderomyces elongisporus*	supercontig_1.4	403484	403702	MIWLLNSCEISPTAVIPRNRIYARYVQVCADYGLSPLSPASFGKLVKILYPNITTRRLGMRGQSKYHYCGIKL
M193	*Loxodonta africana*	GeneScaffold_3023	17323	27194	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M194	*Loxodonta africana*	GeneScaffold_924	692	47465	LQWLLDNYETAEGVSLPKSSLYNHYLQHCQEHKLDPVNAASFGKLIRSVFAGLRTRRLGT
M195	*Loxodonta africana*	GeneScaffold_776	22219	177107	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGT
M196	*Loxodonta africana*	GeneScaffold_6006	64902	155127	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHS
M197	*Loxodonta africana*	GeneScaffold_6394	32525	59372	AFSWIRNTLEEHPETSLPKQEVYDEYXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXYCYSGLRKKA
M198	*Macaca mulatta*	19	13655404	13685077	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M199	*Macaca mulatta*	19	5896888	5955543	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M200	*Macaca mulatta*	15	73743838	74012735	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M201	*Macaca mulatta*	11	1.08E+08	1.08E+08	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M202	*Macaca mulatta*	1	1.30E+08	1.30E+08	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M203	*Macaca mulatta*	4	1.47E+08	1.47E+08	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M204	*Macaca mulatta*	7	34402269	34498255	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYPLPGLRKKA
U205	*Magnaporthe grisea*	contig 2.768	27190	27381	AMLWIAQVCSKGKSSVPRGRVYANYASKCASERVTVLNPASFGKLVRVIFPKLKTRRLGVRGESKYHY
M206	*Microcebus murinus*	GeneScaffold_1823	17253	75427	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M207	*Microcebus murinus*	GeneScaffold_4252	101767	127432	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP

M208	*Microcebus murinus*	GeneScaffold_1997	291138	445815	LEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M209	*Microcebus murinus*	scaffold_21694	7742	8148	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKY
M210	*Microcebus murinus*	GeneScaffold_4112	18301	68830	LEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M211	*Monodelphis domestica*	7	64903620	64919074	LVDNFCICEGSVPRCLMYEIYVETCGQNAQNQVNPATFGLVRLVFPDLGTRRLGTRGSARY
M212	*Monodelphis domestica*	3	4.46E+08	4.46E+08	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M213	*Monodelphis domestica*	3	4.44E+08	4.44E+08	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFVGLRTRRLGTRGNSKYHYYGIRLKP
M214	*Monodelphis domestica*	6	1.65E+08	1.65E+08	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M215	*Monodelphis domestica*	8	86170675	86271025	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M216	*Monodelphis domestica*	2	4.97E+08	4.97E+08	AYRWIRNHLEEHTATCLPKQDVYDAYRRYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M217	*Monodelphis domestica*	2	3.85E+08	3.85E+08	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M218	*Monodelphis domestica*	1	1.65E+08	1.65E+08	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYLEHGLRKKA
P219	*Monosiga brevicollis*	scaffold_39	61843	67196	TVVWLHENFEACDDTSLGREPLFAHYIEHCKTLNQEPVNQASFGKLIRSVFPNLKTRRLGTRGNSKYHYYGIRLKE
P220	*Monosiga brevicollis*	scaffold_3	1262536	1267190	WIHEHYELKEAACVLRSSLYENYVKFCELTSQEPTNAANFGKIIRQQFPQLKTRRLGTRGQSKYHYYGLRLK
P221	*Monosiga ovata*	est	0	0	TVVWLHEHFEAAEGSLGRSTLYQHYCDHCTLHHYDPVNQASFGKLIRSVFPNLKTRRLGTRGNSKYHYYGIRLRD
P222	*Monosiga ovata*	est	0	0	TVVWLHEHFEAAEGSLGRSTLYQHYCDHCTLHHYDPVNQASFGKLIRSVFPNLKTRRLGTRGNSKYHYYGIRLRD
P223	*Monosiga ovata*	est	0	0	TVVWLHEHFEAAEGSLGRSTLYQHYCDHCTLHHYDSVNQASFGKLIRSVFPNLKTRRLGTRGNSKYHYYGIRLRD
M224	*Mus musculus*	8	86590765	86620901	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M225	*Mus musculus*	17	56915323	56970436	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M226	*Mus musculus*	19	27842635	28085630	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M227	*Mus musculus*	10	84218793	84369281	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M228	*Mus musculus*	3	94757997	94763616	AYRWIRNHLEEHMDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M229	*Mus musculus*	10	51397616	51450235	TLQWLEDNYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M230	*Mus musculus*	9	72380047	72470744	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
M231	*Myotis lucifugus*	GeneScaffold_5509	121957	144888	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M232	*Myotis lucifugus*	GeneScaffold_741	5092	80631	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M233	*Myotis lucifugus*	GeneScaffold_638	110885	306488	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGT
M234	*Myotis lucifugus*	GeneScaffold_2579	327789	531122	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKXXXXXXXGQTGKRFRARGSKYHYYGIAVKE
M235	*Myotis lucifugus*	GeneScaffold_3990	258073	261889	ACKWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M236	*Myotis lucifugus*	GeneScaffold_5342	8092	66184	LEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKXXXXXXXXXXXXXXXXXXXXYHYYGIGIKE
S237	*Nasonia vitripennis*	SCAFFOLD39	204137	209297	VQWLLENYETADGVSLPRSTLYNHYLRHCSDNKLDPVNAASFGKLIRSVFLGLRTRRLGTRGNSKYHYYGIRVKP
S238	*Nasonia vitripennis*	SCAFFOLD3	4812482	4812775	WIKTHLEEDPEVSLPKQEVYDEYKIFCSKNSMKPLSTADFGKVMKQVYPRVRPRRLGTRGNSRYCYAGMRKR
I239	*Nematostella vectensis*	scaffold_43	647360	650810	VQWLLENYETAEGVSLPRSTLYNHYLTHCQTHKLDPVNAASFGKLIRSVFLGLRTRRLGTRGNSKYHYYGIRIKP
I240	*Nematostella vectensis*	scaffold_194	181810	187624	LNENYEVADGVSLPRSALYSHYLDFCEKNSLSPVNAASFGKIIRHTFPNLKTRRLGTRGQSK
I241	*Nematostella vectensis*	scaffold_11	402747	410796	LDENYVMCEGVCLPRCILYAHYLDFCRRHKIEAACAATFGKTIRQKFPQLTTRRLGTRGHSKYHYYGIGIKE
I242	*Nematostella vectensis*	scaffold_100	77629	82568	LGENYELKEGMCLPRCVMYTHYLDFCKKNKLNPAGPATFGKIIRQKFPKLTTRRLGTRGQSKYHYYGIQVSE
I243	*Nematostella vectensis*	scaffold_71	335476	337458	AFHWIRCHLEECDNSSLPKHEVYDEYKAYCESMSARTLSAPDFGKIIKCVFPRVKARRLGTRGNSKYCYSGIQRK
U244	*Neosartorya fischeri*	supercontig_null.570	1309849	1310040	LKENCRKSSGSVRRDRVYCCYAEKCGTERVSVLNPASFGKLVRIIFPNVQTRRLGVRGESKYHY
U245	*Neurospora crassa*	contig 7.31	114999	115190	GKGSVPRGRVYANYASRCATERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFQLRE
M246	*Ochotona princeps*	scaffold_36243	748	4694	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M247	*Ochotona princeps*	scaffold_15344	1066	4994	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKPDPVNAASFRKLIRSIFMGLRTRRLSTRGNSKYHYYGIRVKP
M248	*Ochotona princeps*	GeneScaffold_422	783698	909859	TLQWLEENYEIAEGVCIPRSALYMHYLDFCDKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M249	*Ochotona princeps*	scaffold_53	799871	823100	LEENYEIAEGVCIPRSALYMHYLDFCDKNDTQPVNAASFGKVIRQQFPQLTTRRLGTRGQSK
M250	*Ochotona princeps*	GeneScaffold_4477	8990	23390	ALSWIRNTLEEHPETSLPKQEVYDEYXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXYCYSGLRKKA
M251	*Ochotona princeps*	scaffold_94928	3701	4101	ACRWIRNHLEEHADTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKY
M252	*Ornithorhynchus anatinus*	Contig10457	8984	25871	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTR
M253	*Ornithorhynchus anatinus*	Ultra497	472440	566943	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M254	*Ornithorhynchus anatinus*	X5	2258647	2374266	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M255	*Ornithorhynchus anatinus*	Ultra443	4586082	4709763	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M256	*Ornithorhynchus anatinus*	Contig784	186929	211415	LEENYIVCEGVCLPRCILYAHYLDFCRKEQLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIRE
M257	*Ornithorhynchus anatinus*	Ultra366	700986	723064	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
M258	*Oryctolagus cuniculus*	GeneScaffold_820	2921	28655	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M259	*Oryctolagus cuniculus*	scaffold_213687	47838	48029	LEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKVSPAPQAQSAPPRQGIXAQVLNH
M260	*Oryctolagus cuniculus*	scaffold_203023	14666	18088	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M261	*Oryctolagus cuniculus*	GeneScaffold_6139	3602	75116	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE

F262	*Oryzias latipes*	8	13635578	13644098	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
F263	*Oryzias latipes*	1	31369411	31377554	TVQWLCENYEGAEGVSLPRCTLYYHYLLHCQEQKLEPVNAASFGKLIRSVFVGLRTRRLGTRGNSKYHYYGLRIKS
F264	*Oryzias latipes*	4	15861411	15879280	HLQWLLDNYETAEGVSLPRCSLYNHYLRHCQEQKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRVKP
F265	*Oryzias latipes*	12	12587127	12592775	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEQKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
F266	*Oryzias latipes*	6	16561767	16577755	TLEWLEENYEIAEGVCIPRSALYMHYLDFSEKHDTQPVNAASFGKIIRQQFPALTTRRLGTRGQSKYHYYGIAVKE
F267	*Oryzias latipes*	11	3297678	3301889	CNWIRSHLEEHSDTCLPKQDVYEAYKRYCKNLRHRPLSAAIFGKIIRDIFPNIKARRLGGRGQSKYCYSGIRRKT
F268	*Oryzias latipes*	scaffold914	20398	39026	TLQWLEENYMVCEGVCLPRCILYAHYLDFCRKERLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
F269	*Oryzias latipes*	6	10233133	10241039	AFNWIRNHLEEHPETSLPKQEVYDEYKSYCDNLGYNPLSAADFGKIMKNVFPNMKARRLGMRGKSKYCYSGLRKKA
F270	*Oryzias latipes*	3	20272309	20281059	AFSWIRDHLEEYPETSLPKQEVYDEYKSFCDNLNYHPLSAADFGKMMKNVFPNMKARRLGMRGKSKYCYSGLRKK
M271	*Otolemur garnettii*	GeneScaffold_4390	117868	172244	LQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGT
M272	*Otolemur garnettii*	GeneScaffold_538	151816	269855	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M273	*Otolemur garnettii*	GeneScaffold_1329	20533	179128	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M274	*Otolemur garnettii*	GeneScaffold_1333	232758	286284	LEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M275	*Pan troglodytes*	19	14361897	14409186	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M276	*Pan troglodytes*	19	6101889	6159032	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M277	*Pan troglodytes*	9	3260630	3434900	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M278	*Pan troglodytes*	12	1.08E+08	1.08E+08	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M279	*Pan troglodytes*	1	1.30E+08	1.30E+08	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M280	*Pan troglodytes*	6	1.19E+08	1.19E+08	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M281	*Pan troglodytes*	15	53518246	53669813	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
S282	*Pediculus humanus*	1.10E+12	29062	34487	TVQWLLENYECFEGVSLPRSTMYAHYLRHCSEHKLDPVNAASFGKLIRSVFLGLRTRRLGTRGNSKYHYYGIRIK
S283	*Pediculus humanus*	1.10E+12	7944	10738	TLLWLGKNYELAEGICIPRNTLYSHYVHFCQTNSMSPLNSASFGKIIRQAFPSLTTRRLGTRGQSQYHYCGIAIKD
S284	*Pediculus humanus*	1.10E+12	989439	992782	WIKTHLEEDSEISIPKQDVYDQYLKYCENVTMKPLSTADFGKVMKQVYPGVRPRRLGTRGNSRYCYSGMR
U285	*Phanerochaete chrysosporium*	scaffold_2	2333749	2336797	WLTANYAPYPDGNVPRQGLYFSYRRVCDQYGIPHINTATLGKAIRLCFPTIKTRRLGVRGNSKYHYCGIR
U286	*Phycomyces blakesleeanus*	scaffold_13	882371	882574	NYEYEEHNVPRSGMYDHYKNQCDSQGIEPVNSATFGKLIRTVFPGIKTRRLGTRGQSKYHYCNIRLR
U287	*Pichia stipitis*	chr_6.1	1145282	1148418	MVWLLNSCESLPTAVVPRNRIYARYVQVCADNSLTPLSPASFGKLVRILFPNLTTRRLGMRGQSKYHYCGIKL
M288	*Pongo pygmaeus*	19	14077169	14108688	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M289	*Pongo pygmaeus*	19	6010451	6067810	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M290	*Pongo pygmaeus*	2b	1.94E+08	1.94E+08	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M291	*Pongo pygmaeus*	9	59652296	59805834	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M292	*Pongo pygmaeus*	12	1.08E+08	1.09E+08	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M293	*Pongo pygmaeus*	1	1.00E+08	1.00E+08	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M294	*Pongo pygmaeus*	6	1.19E+08	1.19E+08	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M295	*Pongo pygmaeus*	15	52894754	52935112	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
N296	*Pristionchus pacificus*	Ppa_Contig40	507439	507829	VNWLKANYEKADGSSLPRCTLYQHYIRHCKSMGIEPVNAASFGKLIRSIFDGLKTRRLGTRGNSK
M297	*Procavia capensis*	scaffold_3992	9436	34818	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M298	*Procavia capensis*	GeneScaffold_6923	111816	152438	HLQWLLDNYETAEGVSLPRSSLYSHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M299	*Procavia capensis*	scaffold_54300	2979	4329	LQWLLDNYETAEGVSLPKSSLYNHYLQHCQQHRLDPVNAASFGKLICSVFTGLRTRRLGTRGNSKCHYYGIRLKP
M300	*Procavia capensis*	GeneScaffold_803	8724	127905	HLQWLLDNYETAEGVSLPSKTLYTYYLRHCQEHKXXXXXXXXXXXXXXXXXXXXXXXXXXXGNSKYHYYGIRVKP
M301	*Procavia capensis*	GeneScaffold_1926	34670	154161	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQS
M302	*Procavia capensis*	GeneScaffold_3901	55775	59583	ACRWIRNHLEEHTDTCLPKQSVYDAYRKYCENLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M303	*Procavia capensis*	GeneScaffold_6516	9203	42275	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
M304	*Pteropus vampyrus*	GeneScaffold_1492	103298	125343	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M305	*Pteropus vampyrus*	scaffold_1981	55592	61228	LVDNFCICEGYSVPRCLMYEIYVETCGQNAQNQVNPATFGKVVRLVFPDLGTRRLGTRGSARYHF
M306	*Pteropus vampyrus*	GeneScaffold_488	8877	43122	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M307	*Pteropus vampyrus*	GeneScaffold_421	100212	230264	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGT
M308	*Pteropus vampyrus*	GeneScaffold_1644	288061	407783	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M309	*Pteropus vampyrus*	GeneScaffold_1943	213738	217390	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M310	*Pteropus vampyrus*	GeneScaffold_3519	16555	61819	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M311	*Pteropus vampyrus*	GeneScaffold_3374	52162	72814	AFFWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
U312	*Pyrenophora tritici-repentis*	supercontig_1.1	3520522	3520707	AMLWLKCVCRIAKTSVPRNRVYSKYAERCGTDRVIPLNPASFGKLVRVIFPGIQTRRLGVRGESKYHYVDLEL
M313	*Rattus norvegicus*	19	25745412	25776701	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M314	*Rattus norvegicus*	1	2.31E+08	2.32E+08	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M315	*Rattus norvegicus*	7	20977074	21115616	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE

M316	*Rattus norvegicus*	2	1.90E+08	1.90E+08	AYRWIRNHLEEHMDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M317	*Rattus norvegicus*	20	30335024	30390585	TLQWLEDNYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M318	*Rattus norvegicus*	8	77176638	77203848	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYSECGLRKKA
U319	*Rhizopus oryzae*	supercontig 3.12	777218	777436	NRVRDNYQERDHNVPRRNMYEHYKAHCIARHLVPVNSATFGKLIRIVFPELKTRRLGVRGQSKYHYCGIRVR
U320	*Saccharomyces castellii*	gi|30987890|gb|AACF01000217.1|	2336	2557	ALLWLMKNCESKHDSFVPRGRIFAQYASSCAQNNLKPLSQASLGKLIRTVFPDLTTRRLGMRGQSKYHYCGLRL
U321	*Saccharomyces cerevisiae*	XII	507799	510234	ALLWLMKNCKSQHDSYVPRGKIFAQYASSCSQNNLKPLSQASLGKLIRTVFPDLTTRRLGMRGQSKYHYCGLKLTVNE
U322	*Saccharomyces kluyveri*	Contig0.26	41770	42000	ALIWLMNNCIPDGDSYVPRGRIFAQYASSCAQNSLKPLSQASLGKLIRSLFPNLTTRRLGMRGQSKYHYCGLKLVNN
U323	*Saccharomyces kudriavzevii*	gi|77694828|gb|AACI02001081.1|	163	396	ALLWLMKNCRSQHDSYVPRGKIFAQYASSCSQNNLKPLSQASLGKLIRTVFPDLTTRRLGMRGQSKYHYCGLKLTINE
U324	*Saccharomyces mikatae*	contig_1338	136	369	ALLWLMKNCKSQHDSYVPRGKIFAQYASSCSQNNLKPLSQASLGKLIRTVFPDLTTRRLGMRGQSKYHYCGLKLAANE
U325	*Saccharomyces paradoxus*	contig_136	15371	15604	ALLWLMKNCKSQHDSYVPRGKIFAQYASSCSQNNLKPLSQASLGKLIRTVFPDLTTRRLGMRGQSKYHYCGLKLTANE
U326	*Schizosaccharomyces japonicus*	supercontig_1.7	20861	21076	WLKRNCEAQDAAVQRNHIYAQYVDSCNALRTKPLNPASFGKLVRLLFPAIKTRRLGTRGHSKYHYCGIRLR
U327	*Schizosaccharomyces octosporus*	supercontig_2.2	1080905	1081129	WLKQSCEDQEDAAVQRNQIYAQYVDACNVYHVKTLSSASFGKLVRMLFPTIKTRRLGTRGHSKYHYCGLKLRGHE
U328	*Schizosaccharomyces pombe*	chromosome 1	3148986	3149204	ICWLKRACEEQQDAAVQRNQIYAHYVEICNSLHIKPLNSASFGKLVRLLFPSIKTRRLGMRGHSKYHYCGIKL
M329	*Sorex araneus*	GeneScaffold_6602	19779	45566	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQDQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M330	*Sorex araneus*	GeneScaffold_3735	64110	67690	AYKWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M331	*Sorex araneus*	GeneScaffold_6411	12610	74593	LEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M332	*Sorex araneus*	GeneScaffold_6214	15030	89004	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYPLPGLRKKA
M333	*Spermophilus tridecemlineatus*	GeneScaffold_6073	35774	213633	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M334	*Spermophilus tridecemlineatus*	GeneScaffold_1485	71999	222784	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M335	*Spermophilus tridecemlineatus*	GeneScaffold_5823	70454	75274	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPNIKARRLGGRGQSKYCYSGIRRKT
M336	*Spermophilus tridecemlineatus*	GeneScaffold_1492	317633	427818	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M337	*Spermophilus tridecemlineatus*	GeneScaffold_5074	5912	84540	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYPLGGLRKKA
I338	*Strongylocentrotus purpuratus*	gb|DS006113|	98605	99758	VQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSVFLGLRTRRLGTRGNSKYHYYGIRIK
I339	*Strongylocentrotus purpuratus*	gb|DS007398|	3766	4213	LHANYMLSEGVCIPRSALYVHYLDFCCRSVIVPINAASFGKVIRQQFPQITTRRLGTRGQSK
I340	*Strongylocentrotus purpuratus*	gb|DS014603|	52790	53695	LSDNYERSDGVCVPRCVLYTHYLDFCKKHDFSPSSAATFGVIRQKFPKLTTRRLGTRGQSK
I341	*Strongylocentrotus purpuratus*	gb|DS003348|	5408	6061	LMQNYEASQGYSLPRCLIYEHYLDFCQRNVLQPVNAASFGKVIRQVFPDIRTRRLGTRGQSK
I342	*Strongylocentrotus purpuratus*	gb|DS001470|	101616	106423	LEENYCICEGVCLPRCILYSHYLDFCRKETLDPACAATFGKTIRQKFPNLTTRRLGTRGHSK
I343	*Strongylocentrotus purpuratus*	gb|DS014005|	91208	94031	NWVRSHIEESPDTSLPKQEVYEEYFCENSGHRPLSTADFGKIIKGVFPAVQARRLGTRGNSRY
B344	*Taeniopygia guttata*	Chr28	2171475	2175341	LQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
B345	*Taeniopygia guttata*	ChrZ	64577900	64578685	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
B346	*Taeniopygia guttata*	Chr1A	53670175	53681344	LEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKVIRQQFPQLTTRRLGTRGQSK
B347	*Taeniopygia guttata*	Chr3	65876523	65877908	LEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGTIRQKFPLLTTRRLGTRGHSK
B348	*Taeniopygia guttata*	Chr10	7434504	7440662	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYPL
F349	*Takifugu rubripes*	scaffold_141	666977	674324	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
F350	*Takifugu rubripes*	scaffold_189	361201	364280	TIQWLCDNYEGAEGVSLPRCTLYYHYLLHCQEHKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKS
F351	*Takifugu rubripes*	scaffold_332	231401	239952	HLQWLLDNYETAEGVSLPRCSLYNHYLRHCQEQKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRVKP
F352	*Takifugu rubripes*	scaffold_84	711067	716942	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEQKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
F353	*Takifugu rubripes*	scaffold_105	398092	409361	TLEWLEENYEMAEGVCIPRSALYMHYLDFSEKHDTQPVNAASFGKIIRQQFPALTTRRLGTRGQSKYHYYGIAVKE
F354	*Takifugu rubripes*	scaffold_274	21221	22516	CNWIRSHLEEHSDTCLPKQDVYETYRRYCENLQYRPLSAANFGKIIRDIFPNIKARRLGGRGQSKYCYSGIRRKT
F355	*Takifugu rubripes*	scaffold_3348	8214	8491	CNWIRSHLEEHSDTCLPKQDVYETYRRYCENLQYRPLSAANFGKIIRDIFPNIKARRLGGRGQSKY
F356	*Takifugu rubripes*	scaffold_111	651744	657177	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKENLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
F357	*Takifugu rubripes*	scaffold_198	178368	182328	AFNWIRNHLEEHQETSLPKQEVYDEYKSYCDNLGYNPLSAADFGKIMKNVFPTMKARRLGMRGKSKYCYSGLRKKA
F358	*Takifugu rubripes*	scaffold_14	1446440	1452582	AFSWIHNHLEEYPETSLPKQEVYDEYKSFCDNLNYHPLSAADFGKMMKNVFPNMKARRLGMRGKSKYPSLCLRKR
M359	*Tarsius syrichta*	scaffold_475280	579	797	IQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRWVPQWQLWGLR
M360	*Tarsius syrichta*	GeneScaffold_863	31131	181914	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M361	*Tarsius syrichta*	GeneScaffold_2153	15102	136629	TLQWLEENYEAEGVCIPRSALYMHYLFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M362	*Tarsius syrichta*	GeneScaffold_7469	16161	67174	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M363	*Tarsius syrichta*	GeneScaffold_7244	12076	37511	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKK
F364	*Tetraodon nigroviridis*	3	9006510	9013322	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA

F365	*Tetraodon nigroviridis*	18	2475022	2478752	IQWLCDNYEGAEGVSLPRCTLYYHYLLHCQEHKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKS
F366	*Tetraodon nigroviridis*	1	19961846	19969283	HLQWLLDNYETAEGVSLPRCSLYNHYLRHCQEQKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRVKP
F367	*Tetraodon nigroviridis*	4	2336075	2342320	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEQKLDPVNAASFGKLIRSIFMGLRTRRLGTRGNSKYHYYGIRVKP
F368	*Tetraodon nigroviridis*	13	5078530	5088777	TLEWLEENYEMAEGVCIPRSALYMHYLDFSEKHDTQPVNAASFGKIIRQQFPALTTRRLGTRGQSKYHYYGIAVKE
F369	*Tetraodon nigroviridis*	Un_random	16005587	16008196	CNWIRSHLEEHSDTCLPKQDVYETYRRYCENLQYRPLSAANFGKIIRDIFPNIKARRLGGRGQSKYCYSGIRRKT
F370	*Tetraodon nigroviridis*	14	9205929	9210771	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGETIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
F371	*Tetraodon nigroviridis*	13	11757434	11762001	AFNWIRNHLEEHQETSLPKQEVYDEYKSYCDNLGYNPLSAADFGKIMKNVFPTMKARRLGMRGKSKYCYSGLRKKA
F372	*Tetraodon nigroviridis*	5	5349575	5354831	FSWIHNHLEEYPETSLPKQEVYDEYKSFCDNLNYHPLSAADFGKMMKNVFPNMKARRLGMRGKSKYCYSGLRKR
S373	*Tribolium castaneum*	ChLG9	18947924	18954115	TVQWLLENYETAEGVSLPRSTLYAHYLRHCAENKLEPVNAASFGKLIRSVFLGLRTRRLGTRGNSKYHYYGIRVK
S374	*Tribolium castaneum*	ChLG9	19573524	19573745	VQWLLENYETAEGVSLPRSTLYAHYLRHCAENKLEPVNAASFGKLIRSVFLGLRTRRLGTRGNSKYHYYGIRVK
S375	*Tribolium castaneum*	ChLG7	13141888	13147479	SWIKTHLEEDAALSLPKQEVYEEYTVYCTQNQIKSLSQADFGKVMKQVYPKVRARRLGTRGNSRYCYSGLRR
U376	*Trichoderma atroviride*	scaffold_16	792679	795000	AMLWINSVCSKGKGSVPRGRVYANYASRCATERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFSLVEDQ
U377	*Trichoderma reesei*	scaffold_28	223052	225370	AMLWINSVCSKGKGSVPRGRVYANYASRCATERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFSLAEDQ
U378	*Trichoderma virens*	scaffold_4	2243840	2246454	AMLWINSVCSKGKGSVPRGRVYANYASRCATERITVLNPASFGKLVRVLFPGLKTRRLGVRGESKYHYVNFSLAEDQ
M379	*Tupaia belangeri*	GeneScaffold_5255	99701	197472	VQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M380	*Tupaia belangeri*	scaffold_102650	435	1652	LQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKVDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M381	*Tupaia belangeri*	GeneScaffold_559	88790	265783	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLGT
M382	*Tupaia belangeri*	GeneScaffold_1367	2850	246604	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M383	*Tupaia belangeri*	GeneScaffold_5612	129500	133758	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M384	*Tupaia belangeri*	GeneScaffold_1372	239521	334917	TLQWLEENYIVCEGVCFTRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M385	*Tupaia belangeri*	GeneScaffold_4908	3933	57132	AFSWIRNTLEEHPETSLPKQEVYDEYXXXXXXXXXXXXXXXXXXXXMKKRVPNQKARRLGTRGKSNY
M386	*Tursiops truncatus*	GeneScaffold_3219	82156	106574	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
M387	*Tursiops truncatus*	GeneScaffold_2831	96172	142311	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQEHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
M388	*Tursiops truncatus*	GeneScaffold_338	68999	224750	HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIKSIFMGLRTRRLGTRGNSKYHYYGIRVKP
M389	*Tursiops truncatus*	GeneScaffold_2471	258952	408379	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLGTRGQSKYHYYGIAVKE
M390	*Tursiops truncatus*	GeneScaffold_1690	343119	347140	AYRWIRNHLEEHTDTCLPKQSVYDAYRKYCESLACCRPLSTANFGKIIREIFPDIKARRLGGRGQSKYCYSGIRRKT
M391	*Tursiops truncatus*	GeneScaffold_809	177766	234338	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M392	*Tursiops truncatus*	scaffold_104484	20198	82549	AFSWIRNTLEEHTETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
U393	*Ustilago maydis*	supercontig 1.23	147920	148129	WLTCNYTLKPSISIPRTILHESYRRACDALGLEPLQAASFGKVLRSQFPDVVQRRLGGRGKTRFHYCG
M394	*Vicugna pacos*	GeneScaffold_361	161905	314950	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRERLTIGNSKYHYYGIRVKP
M395	*Vicugna pacos*	GeneScaffold_871	5828	121968	TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKXXXXXXXXXXXXXXXXXXXXXXYHYYGIAVKE
M396	*Vicugna pacos*	GeneScaffold_3293	347939	405234	TLQWLEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSKYHYYGIGIKE
M397	*Vicugna pacos*	GeneScaffold_1605	748609	789350	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
R398	*Xenopus tropicalis*	scaffold_649	58260	80834	TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGLRIKA
R399	*Xenopus tropicalis*	scaffold_386	698760	712622	HLQWLLDNYETAEGVSLPRSSLYNHYLRHCQDHKLDPVNAASFGKLIRSVFMGLRTRRLGTRGNSKYHYYGIRLKP
R400	*Xenopus tropicalis*	scaffold_86	2782908	2814293	LQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRL
R401	*Xenopus tropicalis*	scaffold_2034	8474	10680	LEEHTDTCLPKQDVYDAYKRYCDNLHGRPLSVANFGKIIREIFPNIKARRLGGRGQYTYCYSGLRRKS
R402	*Xenopus tropicalis*	scaffold_358	314109	315527	LEENYIVCEGVCLPRCILYAHYLDFCRKEKLEPACAATFGKTIRQKFPLLTTRRLGTRGHSK
R403	*Xenopus tropicalis*	scaffold_589	218319	271092	AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLGTRGKSKYCYSGLRKKA
V404	*Streptococcus phage P9*				FLEDECELGEDFKVPVRDVYPAYKFYCQDSGYKPLARNSFTQRMNELNFENKNAKMGGKTVRCWIGFRIK
V405	*Bovine papular stomatitis virus*				LVNRLHALNTEKIEQIKDAYARYLQDVAEGRIVPMSPADEADAVESLLSNLTNLNVREINEY
S406	*Nasonia vitripennis*	SCAFFOLD12	2957747	2988555	LGWLKATFELSPGVKIEQEELYKKYLGCCTKIGRRGVIAPHFPRCVRSVFGGIGPNPIKGENTGTLYYEGIRVR
S407	*Pediculus humanus*	########	72457	82288	LAWLRATFELSAGGKVEQQDLYKRYVESCNKMGRKGIIASHFPRFVRSVFGG
S408	*Tribolium castaneum*	ChLG4	3995087	4004743	LAWLRATYEPCVNGKVDHQELYKQYLNSCSKGRRGVISPLHFPRCVRSVFGGTPNPMKPSSANEPQYYEGIKVR
P409	*Monosiga brevicollis*	scaffold_19	676281	682156	TALRRAEMYRSYVAVCQNENRVPLCTEVFGKLIMECFPGTKCVTGPDLGRTIY-YTALSSKA
P410	*Monosiga brevicollis*	scaffold_9	869071	869268	RWITDYLQGPTATYVPKQTIYEAYKS-----ATPRANITSVFWKDMHQLFGDKLLERRAGSGGSEGKY
S411	*Drosophila melanogaster*				ALVTLIKTFKISANAVCPRNIVYLKYVENCKEHQISPICNAAFGKLVKIFHPDIKTRRLGVRGSSRYNYCGLELIKN
U412	*Allomyces macrogynus*	supercontig_1.23	343227	343439	LHRNFEAAEEYNMPRQDVYDQYKLYCDTMSVPPVSSPMFGKIVKIMFPELKTRRLGTRGQSRYHYCGIRVK

Appendix B: List of genome sources

SPECIE	VERSION	DATABASE
Acyrthosiphon pisum	1	HGSC
Aedes aegypti	AaegL1	ENSEMBL
Allomyces macrogynus	1	Broad Institute
Anolis carolinensis	1	Broad Institute
Anopheles gambiae	AgamP3	ENSEMBL
Apis mellifera	4	HGSC
Arabidopsis thaliana	8	tair
Aspergillus clavatus	1	Broad Institute
Aspergillus flavus	2	Broad Institute
Aspergillus fumigatus	1	Broad Institute
Aspergillus nidulans	1	Broad Institute
Aspergillus niger	3	Broad Institute
Aspergillus oryzae	1	Broad Institute
Aspergillus terreus	1	Broad Institute
Batrachochytrium dendrobatidis	1	Broad Institute
Bombyx mori	2	silkDB
Bos Taurus	4	ENSEMBL
Botrytis cinerea	1	Broad Institute
Brugia malayi	WS185	WormBase
Caenorhabditis brenneri	WS198	WormBase
Caenorhabditis briggsae	WS198	WormBase
Caenorhabditis elegans	WS190	ENSEMBL
Caenorhabditis japonica	WS198	WormBase
Caenorhabditis remanei	WS198	WormBase
Callorhinchus milii		http://esharkgenome.imcb.a-star.edu.sg/
Candida albicans	1	Broad Institute
Candida guilliermondii	1	Broad Institute
Candida lusitaniae	1	Broad Institute
Candida parapsilosis	1	Broad Institute
Candida tropicalis	3	Broad Institute
Canis familiaris	2	ENSEMBL
Cavia porcellus	3	ENSEMBL
Chaetomium globosum	1	Broad Institute
Chlamydomonas reinhardtii	3.1	JGI
Ciona intestinalis	2	ENSEMBL
Ciona savignyi	2	ENSEMBL
Coccidioides immitis	2	Broad Institute
Coccidioides posadasii	1	Broad Institute
Cochliobolus heterostrophus	1	JGI
Coprinus cinereus	2	Broad Institute
Cryptococcus neoformans	1	Broad Institute
Cryptosporidium hominis		www.hominis.mic.vcu.edu
Culex quinquefasciatus	3	vectorbase

Cyanidioschyzon merolae	2007	merolae.biol.s.u-tokyo.ac.jp
Danio rerio	7	ENSEMBL
Dasypus novemcinctus	"May 2005"	ENSEMBL
Debaryomyces hansenii	1	Broad Institute
Dictyostelium discoideum	"Dec 2008"	dictyBase
Dipodomys ordii	1	ENSEMBL
Drosophila ananassae	1.3	flybase
Drosophila erecta	1.3	flybase
Drosophila grimshawi	1.3	flybase
Drosophila melanogaster	5.13	ENSEMBL
Drosophila mojavensis	1.3	flybase
Drosophila persimilis	1.3	flybase
Drosophila pseudoobscura	2.3	flybase
Drosophila sechellia	1.3	flybase
Drosophila simulans	1.3	flybase
Drosophila virilis	1.2	flybase
Drosophila willistoni	1.3	flybase
Drosophila yakuba	1.3	flybase
Echinops telfairi	"July 2005"	ENSEMBL
Entamoeba histolytica	"October 2005"	Sanger
Equus caballus	2	ENSEMBL
Erinaceus europaeus	1	ENSEMBL
Felis catus	"March 2006"	ENSEMBL
Fusarium graminearum	3	Broad Institute
Fusarium oxysporum	2	Broad Institute
Fusarium verticillioides	3	Broad Institute
Gallus gallus	2	ENSEMBL
Gasterosteus aculeatus	1	ENSEMBL
Giardia lamblia	1.1	GiardiaDB
Gorilla gorilla	1	ENSEMBL
Homo sapien	NCBI 36	ENSEMBL
Hydra magnipapillata		NCBI
Laccaria bicolori	"Mar 2006"	MycorWeb
Leishmania infantum	3	Sanger
Leishmania major	5.2	Sanger
Lodderomyces elongisporus	1	Broad Institute
Loxodonta africana	BROAD E1	ENSEMBL
Macaca mulatta	1	ENSEMBL
Magnaporthe grisea	2	Broad Institute
Microcebus murinus	1	ENSEMBL
Monodelphis domestica	5	ENSEMBL
Monosiga brevicollis	1	JGI
Monosiga ovata	est	NCBI
Mus musculus	NCBI m37	ENSEMBL
Myotis lucifugus	1	ENSEMBL
Nasonia vitripennis	1	HGSC
Nematostella vectensis	1	JGI
Neosartorya fischeri	"Nov 2008"	Broad Institute
Neurospora crassa	7	Broad Institute
Ochotona princeps	2	ENSEMBL
Ornithorhynchus anatinus	5	ENSEMBL
Oryctolagus cuniculus	"May 2005"	ENSEMBL
Oryza sativa	4	RAP-DB

Oryzias latipes	1	ENSEMBL
Ostreococcus lucimarinus	2	JGI
Ostreococcus tauri	2	JGI
Otolemur garnettii	1	ENSEMBL
Pan troglodytes	2.1	ENSEMBL
Paramecium tetraurelia	2.1	genoscope
Pediculus humanus	PhumU1	TIGR
Phaeodactylum tricornutum	2	JGI
Phanerochaete chrysosporium	1	JGI
Phycomyces blakesleeanus	1.1	JGI
Physarum polycephalum	3.1	WUSTL
Physcomitrella patens	1.1	JGI
Phytophthora ramorum	1.1	JGI
Phytophthora sojae	1.1	JGI
Pichia stipitis	2	JGI
Plasmodium falciparum	5.5	plasmodb
Plasmodium yoelii	5.5	plasmodb
Pongo pygmaeus	2	ENSEMBL
Populus trichocarpa	1	JGI
Pristionchus pacificus	WS197	WormBase
Procavia capensis	1	ENSEMBL
Pteropus vampyrus	1	ENSEMBL
Pyrenophora tritici-repentis	"Apr 2008"	Broad Institute
Rattus norvegicus	3.4	ENSEMBL
Rhizopus oryzae	3	Broad Institute
Saccharomyces bayanus	"May 2003"	Broad Institute
Saccharomyces castellii		NCBI
Saccharomyces cerevisiae	1.01	ENSEMBL
Saccharomyces kluyveri	2	WUSTL
Saccharomyces kudriavzevii		NCBI
Saccharomyces mikatae	"May 2003"	Broad Institute
Saccharomyces paradoxus	"May 2003"	Broad Institute
Schizosaccharomyces japonicus	1	Broad Institute
Schizosaccharomyces octosporus	2	Broad Institute
Schizosaccharomyces pombe	"Jan 2009"	Sanger
Sclerotinia sclerotiorum	1	Broad Institute
Sorex araneus	1	ENSEMBL
Spermophilus tridecemlineatus	1	ENSEMBL
Stagonospora nodorum	1	Broad Institute
Strongylocentrotus purpuratus	2.1	HGSC
Taeniopygia guttata	3.2.4	genome.wustl.edu
Takifugu rubripes	4	ENSEMBL
Tarsius syrichta	1	ENSEMBL
Tetrahymena thermophila	"Nov 2006"	TIGR
Tetraodon nigroviridis	8	ENSEMBL
Thalassiosira pseudonana	3	JGI
Theileria annulata	"Sep 2004"	Sanger
Tribolium castaneum	3	beetlebase
Trichoderma atroviride	1	JGI
Trichoderma reesei	2	JGI
Trichoderma virens	1	JGI
Trypanosoma brucei	4	Sanger
Tupaia belangeri	1	ENSEMBL
Tursiops truncatus	1	ENSEMBL

Ustilago maydis	1	Broad Institute
Vicugna pacos	1	ENSEMBL
Xenopus tropicalis	4.1	ENSEMBL

REFERENCE LIST

Abedin, M. and N. King (2008). "The premetazoan ancestry of cadherins." Science **319**(5865): 946-8.

Aftab, S., L. Semenec, et al. (2008). "Identification and characterization of novel human tissue-specific RFX transcription factors." BMC Evol Biol **8**: 226.

Ait-Lounis, A., D. Baas, et al. (2007). "Novel function of the ciliogenic transcription factor RFX3 in development of the endocrine pancreas." Diabetes **56**(4): 950-9.

Aldrich, H. C. (1968). "The development of flagella in swarm cells of the myxomycete Physarum flavicomum." J Gen Microbiol **50**(2): 217-22.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altun, Z. F. and D. H. Hall. (2005). "Anatomical methods for C. elegans research." WormAtlas, from http://www.wormatlas.org/methods.htm.

Ansley, S. J., J. L. Badano, et al. (2003). "Basal body dysfunction is a likely cause of pleiotropic Bardet-Biedl syndrome." Nature **425**(6958): 628-33.

Ashique, A. M., Y. Choe, et al. (2009). "The Rfx4 transcription factor modulates Shh signaling by regional control of ciliogenesis." Sci Signal **2**(95): ra70.

Baas, D., A. Meiniel, et al. (2006). "A deficiency in RFX3 causes hydrocephalus associated with abnormal differentiation of ependymal cells." Eur J Neurosci **24**(4): 1020-30.

Bacaj, T., Y. Lu, et al. (2008). "The conserved proteins CHE-12 and DYF-11 are required for sensory cilium function in Caenorhabditis elegans." Genetics **178**(2): 989-1002.

Badano, J. L., N. Mitsuma, et al. (2006). "The Ciliopathies: An Emerging Class of Human Genetic Disorders." Annu Rev Genomics Hum Genet **7**: 125-148.

Balbo, P. B. and A. Bohm (2007). "Mechanism of poly(A) polymerase: structure of the enzyme-MgATP-RNA ternary complex and kinetic analysis." Structure **15**(9): 1117-31.

Barriere, A., S. P. Yang, et al. (2009). "Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes." Genome Res **19**(3): 470-80.

Beales, P. L. (2005). "Lifting the lid on Pandora's box: the Bardet-Biedl syndrome." Curr Opin Genet Dev **15**(3): 315-23.

Beaudoing, E., S. Freier, et al. (2000). "Patterns of variant polyadenylation signal usage in human genes." Genome Res **10**(7): 1001-10.

Bell, L. R., S. Stone, et al. (2006). "The molecular identities of the Caenorhabditis elegans intraflagellar transport genes dyf-6, daf-10 and osm-1." Genetics **173**(3): 1275-86.

Benjamin, H. W. and N. Kleckner (1992). "Excision of Tn10 from the donor site during transposition occurs by flush double-strand cleavages at the transposon termini." Proc Natl Acad Sci U S A **89**(10): 4648-52.

Bentley, D. R. (2006). "Whole-genome re-sequencing." Curr Opin Genet Dev **16**(6): 545-52.

Bera, A. K., J. Zhu, et al. (2003). "Functional dissection of the Bacillus subtilis pur operator site." J Bacteriol **185**(14): 4099-109.

Beresford, G. W. and J. M. Boss (2001). "CIITA coordinates multiple histone acetylation modifications at the HLA-DRA promoter." Nat Immunol **2**(7): 652-7.

Birney, E., M. Clamp, et al. (2004). "GeneWise and Genomewise." Genome Res **14**(5): 988-95.

Birney, E. and R. Durbin (2000). "Using GeneWise in the Drosophila annotation experiment." Genome Res **10**(4): 547-8.

Blacque, O. E. and M. R. Leroux (2006). "Bardet-Biedl syndrome: an emerging pathomechanism of intracellular transport." Cell Mol Life Sci **63**(18): 2145-61.

Blacque, O. E., E. A. Perens, et al. (2005). "Functional genomics of the cilium, a sensory organelle." Curr Biol **15**(10): 935-41.

Blacque, O. E., M. J. Reardon, et al. (2004). "Loss of C. elegans BBS-7 and BBS-8 protein function results in cilia defects and compromised intraflagellar transport." Genes Dev **18**(13): 1630-42.

Boito, R., M. Menniti, et al. (2005). "RFX-1, a putative alpha Adducin interacting protein in a human kidney library." FEBS Lett **579**(28): 6439-43.

Bonnafe, E., M. Touka, et al. (2004). "The transcription factor RFX3 directs nodal cilium development and left-right asymmetry specification." Mol Cell Biol **24**(10): 4417-27.

Borukhov, S. and E. Nudler (2003). "RNA polymerase holoenzyme: structure, function and biological implications." Curr Opin Microbiol **6**(2): 93-100.

Boss, J. M. and P. E. Jensen (2003). "Transcriptional regulation of the MHC class II antigen presentation pathway." Curr Opin Immunol **15**(1): 105-11.

Boss, J. M. and J. L. Strominger (1986). "Regulation of a transfected human class II major histocompatibility complex gene in human fibroblasts." Proc Natl Acad Sci U S A **83**(23): 9139-43.

Breathnach, R. and P. Chambon (1981). "Organization and expression of eucaryotic split genes coding for proteins." Annu Rev Biochem **50**: 349-83.

Brenner, S. (1974). "The genetics of Caenorhabditis elegans." Genetics **77**(1): 71-94.

Brent, M. R. (2005). "Genome annotation past, present, and future: how to define an ORF at each locus." Genome Res **15**(12): 1777-86.

Briggs, L. J., J. A. Davidge, et al. (2004). "More than one way to build a flagellum: comparative genomics of parasitic protozoa." Curr Biol **14**(15): R611-2.

Burghoorn, J., M. P. Dekkers, et al. (2007). "Mutation of the MAP kinase DYF-5 affects docking and undocking of kinesin-2 motors and reduces their speed in the cilia of Caenorhabditis elegans." Proc Natl Acad Sci U S A **104**(17): 7157-62.

Burset, M. and R. Guigo (1996). "Evaluation of gene structure prediction programs." Genomics **34**(3): 353-67.

Burset, M., I. A. Seledtsov, et al. (2000). "Analysis of canonical and non-canonical splice sites in mammalian genomes." Nucleic Acids Res **28**(21): 4364-75.

Butler, M. P., J. A. Hanly, et al. (2007). "Kinase-active interleukin-1 receptor-associated kinases promote polyubiquitination and degradation of the Pellino family: direct evidence for PELLINO proteins being ubiquitin-protein isopeptide ligases." J Biol Chem **282**(41): 29729-37.

C. elegans Sequencing Consortium (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**(5396): 2012-8.

Cai, X., C. H. Hagedorn, et al. (2004). "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs." RNA **10**(12): 1957-66.

Chan, Y. J., W. P. Tseng, et al. (1996). "Two distinct upstream regulatory domains containing multicopy cellular transcription factor binding sites provide basal repression and inducible enhancer characteristics to the immediate-early IES (US3) promoter from human cytomegalovirus." J Virol **70**(8): 5312-28.

Chang, M., W. Jin, et al. (2009). "Peli1 facilitates TRIF-dependent Toll-like receptor signaling and proinflammatory cytokine production." Nat Immunol **10**(10): 1089-95.

Chen, N., T. W. Harris, et al. (2005). "WormBase: a comprehensive data resource for Caenorhabditis biology and genomics." Nucleic Acids Res **33**(Database issue): D383-9.

Chen, N., A. Mah, et al. (2006). "Identification of ciliary and ciliopathy genes in Caenorhabditis elegans through comparative genomics." Genome Biol **7**(12): R126.

Choi, K. C., Y. S. Lee, et al. (2006). "Smad6 negatively regulates interleukin 1-receptor-Toll-like receptor signaling through direct interaction with the adaptor Pellino-1." Nat Immunol **7**(10): 1057-65.

Chu, J. S., D. L. Baillie, et al. (2010). "Convergent evolution of RFX transcription factors and ciliary genes predated the origin of metazoans." BMC Evol Biol **10**: 130.

Coghlan, A., T. J. Fiedler, et al. (2008). "nGASP--the nematode genome annotation assessment project." BMC Bioinformatics **9**: 549.

Coghlan, A. and K. H. Wolfe (2002). "Fourfold faster rate of genome rearrangement in nematodes than in Drosophila." Genome Res **12**(6): 857-67.

Collet, J., C. A. Spike, et al. (1998). "Analysis of osm-6, a gene that affects sensory cilium structure and sensory neuron function in Caenorhabditis elegans." Genetics **148**(1): 187-200.

Conway-Morris, S. (2003). "The Cambrian "explosion" of metazoans and molecular biology: would Darwin be satisfied?" Int J Dev Biol **47**(7-8): 505-15.

Costa, R. H., D. R. Grayson, et al. (1989). "Multiple hepatocyte-enriched nuclear factors function in the regulation of transthyretin and alpha 1-antitrypsin genes." Mol Cell Biol **9**(4): 1415-25.

Cracraft, J. and M. J. Donoghue (2004). Assembling the tree of life. New York, Oxford University Press.

Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-3.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-63.

Cui, X., T. Vinar, et al. (2007). "Homology search for genes." Bioinformatics **23**(13): i97-103.

Cutter, A. D. (2008). "Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate." Mol Biol Evol **25**(4): 778-86.

Davila Lopez, M. and T. Samuelsson (2008). "Early evolution of histone mRNA 3' end processing." RNA **14**(1): 1-10.

Dawson, S. C., M. S. Sagolla, et al. (2007). "Kinesin-13 regulates flagellar, interphase, and mitotic microtubule dynamics in Giardia intestinalis." Eukaryot Cell **6**(12): 2354-64.

Degnan, B. M., M. Vervoort, et al. (2009). "Early evolution of metazoan transcription factors." Curr Opin Genet Dev **19**(6): 591-9.

Deutsch, M. and M. Long (1999). "Intron-exon structures of eukaryotic model organisms." Nucleic Acids Res **27**(15): 3219-28.

Dorn, A., B. Durand, et al. (1987). "Conserved major histocompatibility complex class II boxes--X and Y--are transcriptional control elements and specifically bind nuclear proteins." Proc Natl Acad Sci U S A **84**(17): 6249-53.

Driever, W. and C. Nusslein-Volhard (1989). "The bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo." Nature **337**(6203): 138-43.

Dubruille, R., A. Laurencon, et al. (2002). "Drosophila regulatory factor X is necessary for ciliated sensory neuron differentiation." Development **129**(23): 5487-98.

Durand, B., C. Vandaele, et al. (2000). "Cloning and characterization of dRFX, the Drosophila member of the RFX family of transcription factors." Gene **246**(1-2): 285-93.

Durbin, R., S. Eddy, et al. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge University Press.

Efimenko, E., O. E. Blacque, et al. (2006). "Caenorhabditis elegans DYF-2, an orthologue of human WDR19, is a component of the intraflagellar transport machinery in sensory cilia." Mol Biol Cell **17**(11): 4801-11.

Efimenko, E., K. Bubb, et al. (2005). "Analysis of xbx genes in C. elegans." Development **132**(8): 1923-34.

El Zein, L., A. Ait-Lounis, et al. (2009). "RFX3 governs growth and beating efficiency of motile cilia in mouse and controls the expression of genes involved in human ciliopathies." J Cell Sci **122**(Pt 17): 3180-9.

Elhasid, R. and A. Etzioni (1996). "Major histocompatibility complex class II deficiency: a clinical review." Blood Rev **10**(4): 242-8.

Emery, P., B. Durand, et al. (1996). "RFX proteins, a novel family of DNA binding proteins conserved in the eukaryotic kingdom." Nucleic Acids Res **24**(5): 803-7.

Emery, P., M. Strubin, et al. (1996). "A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity." Mol Cell Biol **16**(8): 4486-94.

Fan, Y., M. A. Esmail, et al. (2004). "Mutations in a member of the Ras superfamily of small GTP-binding proteins causes Bardet-Biedl syndrome." Nat Genet **36**(9): 989-93.

Feng, C., W. Xu, et al. (2009). "Knockout of the regulatory factor X1 gene leads to early embryonic lethality." Biochem Biophys Res Commun **386**(4): 715-7.

Florea, L., G. Hartzell, et al. (1998). "A computer program for aligning a cDNA sequence with a genomic DNA sequence." Genome Res **8**(9): 967-74.

Frokjaer-Jensen, C., M. W. Davis, et al. (2008). "Single-copy insertion of transgenes in Caenorhabditis elegans." Nat Genet **40**(11): 1375-83.

Fujiwara, M., T. Ishihara, et al. (1999). "A novel WD40 protein, CHE-2, acts cell-autonomously in the formation of C. elegans sensory cilia." Development **126**(21): 4839-48.

Gajiwala, K. S. and S. K. Burley (2000). "Winged helix proteins." Curr Opin Struct Biol **10**(1): 110-6.

Gajiwala, K. S., H. Chen, et al. (2000). "Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding." Nature **403**(6772): 916-21.

Garvie, C. W. and J. M. Boss (2008). "Assembly of the RFX complex on the MHCII promoter: role of RFXAP and RFXB in relieving autoinhibition of RFX5." Biochim Biophys Acta **1779**(12): 797-804.

Garvie, C. W., J. R. Stagno, et al. (2007). "Characterization of the RFX complex and the RFX5(L66A) mutant: implications for the regulation of MHC class II gene expression." Biochemistry **46**(6): 1597-611.

Gaudet, J. and S. E. Mango (2002). "Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4." Science **295**(5556): 821-5.

Ginger, M. L., N. Portman, et al. (2008). "Swimming with protists: perception, motility and flagellum assembly." Nat Rev Microbiol **6**(11): 838-50.

Hardison, R. C. (2003). "Comparative genomics." PLoS Biol **1**(2): E58.

Haycraft, C. J., J. C. Schafer, et al. (2003). "Identification of CHE-13, a novel intraflagellar transport protein required for cilia formation." Exp Cell Res **284**(2): 251-63.

Haycraft, C. J., P. Swoboda, et al. (2001). "The C. elegans homolog of the murine cystic kidney disease gene Tg737 functions in a ciliogenic pathway and is disrupted in osm-5 mutant worms." Development **128**(9): 1493-505.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-9.

Hillier, L. W., A. Coulson, et al. (2005). "Genomics in C. elegans: so many genes, such a little worm." Genome Res **15**(12): 1651-60.

Hobert, O. (2002). "PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans." Biotechniques **32**(4): 728-30.

Horvath, G. C., M. K. Kistler, et al. (2009). "RFX2 is a candidate downstream amplifier of A-MYB regulation in mouse spermatogenesis." BMC Dev Biol **9**: 63.

Horvath, G. C., W. S. Kistler, et al. (2004). "RFX2 is a potential transcriptional regulatory factor for histone H1t and other genes expressed during the meiotic phase of spermatogenesis." Biol Reprod **71**(5): 1551-9.

Huang, M., Z. Zhou, et al. (1998). "The DNA replication and damage checkpoint pathways induce transcription by inhibition of the Crt1 repressor." Cell **94**(5): 595-605.

Hubbard, T. J., B. L. Aken, et al. (2009). "Ensembl 2009." Nucleic Acids Res **37**(Database issue): D690-7.

Hunt-Newbury, R., R. Viveiros, et al. (2007). "High-throughput in vivo analysis of gene expression in Caenorhabditis elegans." PLoS Biol **5**(9): e237.

Itoh, T., T. Tanaka, et al. (2007). "Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana." Genome Res **17**(2): 175-83.

Iwama, A., J. Pan, et al. (1999). "Dimeric RFX proteins contribute to the activity and lineage specificity of the interleukin-5 receptor alpha promoter through activation and repression domains." Mol Cell Biol **19**(6): 3940-50.

Jabrane-Ferrat, N., N. Nekrep, et al. (2002). "Major histocompatibility complex class II transcriptional platform: assembly of nuclear factor Y and regulatory factor X (RFX) on DNA requires RFX5 dimers." Mol Cell Biol **22**(15): 5616-25.

James, T. Y., F. Kauff, et al. (2006). "Reconstructing the early evolution of Fungi using a six-gene phylogeny." Nature **443**(7113): 818-22.

Kara, C. J. and L. H. Glimcher (1991). "In vivo footprinting of MHC class II genes: bare promoters in the bare lymphocyte syndrome." Science **252**(5006): 709-12.

Katan-Khaykovich, Y. and Y. Shaul (2001). "Nuclear import and DNA-binding activity of RFX1. Evidence for an autoinhibitory mechanism." Eur J Biochem **268**(10): 3108-16.

Katan-Khaykovich, Y., I. Spiegel, et al. (1999). "The dimerization/repression domain of RFX1 is related to a conserved region of its yeast homologues Crt1 and Sak1: a new function for an ancient motif." J Mol Biol **294**(1): 121-37.

Katan, Y., R. Agami, et al. (1997). "The transcriptional activation and repression domains of RFX1, a context-dependent regulator, can mutually neutralize their activities." Nucleic Acids Res **25**(18): 3621-8.

Kellis, M., N. Patterson, et al. (2004). "Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery." J Comput Biol **11**(2-3): 319-55.

Kelly, A. and J. Trowsdale (1985). "Complete nucleotide sequence of a functional HLA-DP beta gene and the region between the DP beta 1 and DP alpha 1 genes: comparison of the 5' ends of HLA class II genes." Nucleic Acids Res **13**(5): 1607-21.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-64.

King, N. (2004). "The unicellular ancestry of animal development." Dev Cell **7**(3): 313-25.

King, N., M. J. Westbrook, et al. (2008). "The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans." Nature **451**(7180): 783-8.

Kistler, W. S., G. C. Horvath, et al. (2009). "Differential expression of Rfx1-4 during mouse spermatogenesis." Gene Expr Patterns **9**(7): 515-9.

Knowles, D. G. and A. McLysaght (2009). "Recent de novo origin of human protein-coding genes." Genome Res **19**(10): 1752-9.

Korf, I., P. Flicek, et al. (2001). "Integrating genomic homology into gene structure prediction." Bioinformatics **17 Suppl 1**: S140-8.

Kumar, S., M. Nei, et al. (2008). "MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences." Brief Bioinform **9**(4): 299-306.

Lai, E., K. L. Clark, et al. (1993). "Hepatocyte nuclear factor 3/fork head or "winged helix" proteins: a family of transcription factors of diverse biologic function." Proc Natl Acad Sci U S A **90**(22): 10421-3.

Lampe, D. J., M. E. Churchill, et al. (1996). "A purified mariner transposase is sufficient to mediate transposition in vitro." EMBO J **15**(19): 5470-9.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-8.

Laurencon, A., R. Dubruille, et al. (2007). "Identification of novel regulatory factor X (RFX) target genes by comparative genomics in Drosophila species." Genome Biol **8**(9): R195.

Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." EMBO J **23**(20): 4051-60.

Leitch, C. C., N. A. Zaghloul, et al. (2008). "Hypomorphic mutations in syndromic encephalocele genes are associated with Bardet-Biedl syndrome." Nat Genet **40**(4): 443-8.

Levine, M. and R. Tjian (2003). "Transcription regulation and animal diversity." Nature **424**(6945): 147-51.

Li, J. B., J. M. Gerdes, et al. (2004). "Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene." Cell **117**(4): 541-52.

Liefooghe, A., H. Touzet, et al. (2006). "Large scale matching for Position Weight Matrices." Lecture Notes in Computer Science **4009**: 401-412.

Liefooghe, A., H. Touzet, et al. (2006). Large scale matching for Position Weight Matrices. Lecture Notes in Computer Science, Springer Verlag. **4009:** 401-412.

Lipman, D. J. and W. R. Pearson (1985). "Rapid and sensitive protein similarity searches." Science **227**(4693): 1435-41.

Liu, M., B. H. Lee, et al. (1999). "Involvement of RFX1 protein in the regulation of the human proliferating cell nuclear antigen promoter." J Biol Chem **274**(22): 15433-9.

Lopez, R., V. Silventoinen, et al. (2003). "WU-Blast2 server at the European Bioinformatics Institute." Nucleic Acids Res **31**(13): 3795-8.

Lubelsky, Y., N. Reuven, et al. (2005). "Autorepression of rfx1 gene expression: functional conservation from yeast to humans in response to DNA replication arrest." Mol Cell Biol **25**(23): 10665-73.

Ma, B., J. Tromp, et al. (2002). "PatternHunter: faster and more sensitive homology search." Bioinformatics **18**(3): 440-5.

Ma, K., S. Zheng, et al. (2006). "The transcription factor regulatory factor X1 increases the expression of neuronal glutamate transporter type 3." J Biol Chem **281**(30): 21250-5.

Maddison, D. R. and K.-S. Schulz. (2007). "The Tree of Life Web Project." from http://tolweb.org.

Mak, H. Y., L. S. Nelson, et al. (2006). "Polygenic control of Caenorhabditis elegans fat storage." Nat Genet **38**(3): 363-8.

Manber, U. (1989). Introduction to algorithms: A creative approach. Reading, MA., Addison-Wesley.

Manning, G., S. L. Young, et al. (2008). "The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan." Proc Natl Acad Sci U S A **105**(28): 9674-9.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-80.

Marshall, W. F. (2008). "The cell biological basis of ciliary disease." J Cell Biol **180**(1): 17-21.

Mathis, D. J., C. O. Benoist, et al. (1983). "The murine E alpha immune response gene." Cell **32**(3): 745-54.

McKay, S. J., R. Johnsen, et al. (2003). "Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans." Cold Spring Harb Symp Quant Biol **68**: 159-69.

Meyer, I. M. and R. Durbin (2004). "Gene structure conservation aids similarity based gene prediction." Nucleic Acids Res **32**(2): 776-83.

Miller, C. T., S. Beleza, et al. (2007). "cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans." Cell **131**(6): 1179-89.

Mitchell, D. R. (2007). "The evolution of eukaryotic cilia and flagella as motile and sensory organelles." Adv Exp Med Biol **607**: 130-40.

Miwa, K., C. Doyle, et al. (1987). "Sequence-specific interactions of nuclear factors with conserved sequences of human class II major histocompatibility complex genes." Proc Natl Acad Sci U S A **84**(14): 4939-43.

Morotomi-Yano, K., K. Yano, et al. (2002). "Human regulatory factor X 4 (RFX4) is a testis-specific dimeric DNA-binding protein that cooperates with other human RFX members." J Biol Chem **277**(1): 836-42.

Mudhasani, R. and J. D. Fontes (2005). "Multiple interactions between BRG1 and MHC class II promoter binding proteins." Mol Immunol **42**(6): 673-82.

Murayama, T., Y. Toh, et al. (2005). "The dyf-3 gene encodes a novel protein required for sensory cilium formation in Caenorhabditis elegans." J Mol Biol **346**(3): 677-87.

Nakayama, A., H. Murakami, et al. (2003). "Role for RFX transcription factors in non-neuronal cell-specific inactivation of the microtubule-associated protein MAP1A promoter." J Biol Chem **278**(1): 233-40.

Nechaev, S. and K. Adelman (2011). "Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation." Biochim Biophys Acta **1809**(1): 34-45.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-53.

Nicholas, K. B., H. B. J. Nicholas, et al. (1997). "GeneDoc: Analysis and Visualization of Genetic Variation." EMBNET.NEWS **4**(14).

O'Sullivan, D. M., D. Larhammar, et al. (1986). "Structure of the human Ia-associated invariant (gamma)-chain gene: identification of 5' sequences shared with major histocompatibility complex class II genes." Proc Natl Acad Sci U S A **83**(12): 4484-8.

Otsuki, K., Y. Hayashi, et al. (2004). "Characterization of dRFX2, a novel RFX family protein in Drosophila." Nucleic Acids Res **32**(18): 5636-48.

Ou, G., M. Koga, et al. (2007). "Sensory ciliogenesis in Caenorhabditis elegans: assignment of IFT components into distinct modules based on transport and phenotypic profiles." Mol Biol Cell **18**(5): 1554-69.

Ou, G., H. Qin, et al. (2005). "The PKD protein qilin undergoes intraflagellar transport." Curr Biol **15**(11): R410-1.

Pachter, L., M. Alexandersson, et al. (2001). Applications of generalized pair hidden Markov models to alignment and gene finding problems. Annual Conference on Research in Computational Molecular Biology.

Pan, J. (2008). "Cilia and ciliopathies: from Chlamydomonas and beyond." Sci China C Life Sci **51**(6): 479-86.

Parra, G., P. Agarwal, et al. (2003). "Comparative gene prediction in human and mouse." Genome Res **13**(1): 108-17.

Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proc Natl Acad Sci U S A **85**(8): 2444-8.

Pedersen, L. B. and J. L. Rosenbaum (2008). "Intraflagellar transport (IFT) role in ciliary assembly, resorption and signalling." Curr Top Dev Biol **85**: 23-61.

Pedersen, L. B., I. R. Veland, et al. (2008). "Assembly of primary cilia." Dev Dyn **237**(8): 1993-2006.

Perkins, L. A., E. M. Hedgecock, et al. (1986). "Mutant sensory cilia in the nematode Caenorhabditis elegans." Dev Biol **117**(2): 456-87.

Piasecki, B. P., J. Burghoorn, et al. (2010). "Regulatory Factor X (RFX)-mediated transcriptional rewiring of ciliary genes in animals." Proc Natl Acad Sci U S A **107**(29): 12969-74.

Piskurich, J. F., M. W. Linhoff, et al. (1999). "Two distinct gamma interferon-inducible promoters of the major histocompatibility complex class II transactivator gene are differentially regulated by STAT1, interferon regulatory factor 1, and transforming growth factor beta." Mol Cell Biol **19**(1): 431-40.

Piskurich, J. F., Y. Wang, et al. (1998). "Identification of distinct regions of 5' flanking DNA that mediate constitutive, IFN-gamma, STAT1, and TGF-beta-regulated expression of the class II transactivator gene." J Immunol **160**(1): 233-40.

Praetorius, H. A. and K. R. Spring (2005). "A physiological view of the primary cilium." Annu Rev Physiol **67**: 515-29.

Qin, H., J. L. Rosenbaum, et al. (2001). "An autosomal recessive polycystic kidney disease gene homolog is involved in intraflagellar transport in C. elegans ciliated sensory neurons." Curr Biol **11**(6): 457-61.

Reinhold, W., L. Emens, et al. (1995). "The myc intron-binding polypeptide associates with RFX1 in vivo and binds to the major histocompatibility complex class II promoter region, to the hepatitis B virus enhancer, and to regulatory regions of several distinct viral genes." Mol Cell Biol **15**(6): 3041-8.

Reith, W., E. Barras, et al. (1989). "Cloning of the major histocompatibility complex class II promoter binding protein affected in a hereditary defect in class II gene regulation." Proc Natl Acad Sci U S A **86**(11): 4200-4.

Reith, W., S. Satola, et al. (1988). "Congenital immunodeficiency with a regulatory defect in MHC class II gene expression lacks a specific HLA-DR promoter binding protein, RF-X." Cell **53**(6): 897-906.

Reith, W., C. Ucla, et al. (1994). "RFX1, a transactivator of hepatitis B virus enhancer I, belongs to a novel family of homodimeric and heterodimeric DNA-binding proteins." Mol Cell Biol **14**(2): 1230-44.

Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-52.

Rich, T., R. L. Allen, et al. (2000). "Pellino-related sequences from Caenorhabditis elegans and Homo sapiens." Immunogenetics **52**(1-2): 145-9.

Robertson, H. M. and J. H. Thomas (2006). "The putative chemoreceptor families of C. elegans." WormBook: 1-12.

Roeder, R. G. (2003). "The eukaryotic transcriptional machinery: complexities and mechanisms unforeseen." Nat Med **9**(10): 1239-44.

Roeder, R. G. (2005). "Transcriptional regulation and the role of diverse coactivators in animal cells." FEBS Lett **579**(4): 909-15.

Roeder, R. G. and W. J. Rutter (1970). "Specific nucleolar and nucleoplasmic RNA polymerases." Proc Natl Acad Sci U S A **65**(3): 675-82.

Ruiz-Trillo, I., G. Burger, et al. (2007). "The origins of multicellularity: a multi-taxon genome initiative." Trends Genet **23**(3): 113-8.

Saito, H., R. A. Maki, et al. (1983). "Complete primary structures of the E beta chain and gene of the mouse major histocompatibility complex." Proc Natl Acad Sci U S A **80**(18): 5520-4.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-25.

Satir, P., D. R. Mitchell, et al. (2008). "How did the cilium evolve?" Curr Top Dev Biol **85**: 63-82.

Schafer, J. C., C. J. Haycraft, et al. (2003). "XBX-1 encodes a dynein light intermediate chain required for retrograde intraflagellar transport and cilia assembly in Caenorhabditis elegans." Mol Biol Cell **14**(5): 2057-70.

Sengupta, P. K., J. Fargo, et al. (2002). "The RFX family interacts at the collagen (COL1A2) start site and represses transcription." J Biol Chem **277**(28): 24926-37.

Senti, G. and P. Swoboda (2008). "Distinct isoforms of the RFX transcription factor DAF-19 regulate ciliogenesis and maintenance of synaptic activity." Mol Biol Cell **19**(12): 5517-28.

She, R., J. S. Chu, et al. (2009). "GenBlastA: enabling BLAST to identify homologous gene sequences." Genome Res **19**(1): 143-9.

Sherman, P. A., P. V. Basta, et al. (1987). "Upstream DNA sequences required for tissue-specific expression of the HLA-DR alpha gene." Proc Natl Acad Sci U S A **84**(12): 4254-8.

Siegrist, C. A., B. Durand, et al. (1993). "RFX1 is identical to enhancer factor C and functions as a transactivator of the hepatitis B virus enhancer." Mol Cell Biol **13**(10): 6375-84.

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

Sloan, J. H. and J. M. Boss (1988). "Conserved upstream sequences of human class II major histocompatibility genes enhance expression of class II genes in wild-type but not mutant B-cell lines." Proc Natl Acad Sci U S A **85**(21): 8186-90.

Smith, S. B., H. Q. Qu, et al. (2010). "Rfx6 directs islet formation and insulin production in mice and humans." Nature **463**(7282): 775-80.

Sonnhammer, E. L. and R. Durbin (1994). "An expert system for processing sequence homology data." Proc Int Conf Intell Syst Mol Biol **2**: 363-8.

Soyer, J., L. Flasse, et al. (2010). "Rfx6 is an Ngn3-dependent winged helix transcription factor required for pancreatic islet cell development." Development **137**(2): 203-12.

Steimle, V., B. Durand, et al. (1995). "A novel DNA-binding regulatory factor is mutated in primary MHC class II deficiency (bare lymphocyte syndrome)." Genes Dev **9**(9): 1021-32.

Stein, L. D., Z. Bao, et al. (2003). "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics." PLoS Biol **1**(2): E45.

Stein, L. D., C. Mungall, et al. (2002). "The generic genome browser: a building block for a model organism system database." Genome Res **12**(10): 1599-610.

Stiernagle, T. (2006). "Maintenance of C. elegans." WormBook: 1-11.

Stoetzel, C., J. Muller, et al. (2007). "Identification of a novel BBS gene (BBS12) highlights the major role of a vertebrate-specific branch of chaperonin-related proteins in Bardet-Biedl syndrome." Am J Hum Genet **80**(1): 1-11.

Strelow, A., C. Kollewe, et al. (2003). "Characterization of Pellino2, a substrate of IRAK1 and IRAK4." FEBS Lett **547**(1-3): 157-61.

Suyama, M., D. Torrents, et al. (2004). "BLAST2GENE: a comprehensive conversion of BLAST output into independent genes and gene fragments." Bioinformatics **20**(12): 1968-70.

Swoboda, P., H. T. Adler, et al. (2000). "The RFX-type transcription factor DAF-19 regulates sensory neuron cilium formation in C. elegans." Mol Cell **5**(3): 411-21.

Tamura, K., J. Dudley, et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol **24**(8): 1596-9.

Thomas, M. C. and C. M. Chiang (2006). "The general transcription machinery and general cofactors." Crit Rev Biochem Mol Biol **41**(3): 105-78.

Toscani, A., R. V. Mettus, et al. (1997). "Arrest of spermatogenesis and defective breast development in mice lacking A-myb." Nature **386**(6626): 713-7.

Traylor-Knowles, N., U. Hansen, et al. (2010). "The evolutionary diversification of LSF and Grainyhead transcription factors preceded the radiation of basal animal lineages." BMC Evol Biol **10**: 101.

van Eggermond, M. C., I. Tezcan, et al. (2008). "Transcriptional silencing of RFXAP in MHC class II-deficiency." Mol Immunol **45**(10): 2920-8.

van Luenen, H. G., S. D. Colloms, et al. (1994). "The mechanism of transposition of Tc3 in C. elegans." Cell **79**(2): 293-301.

Vandaele, C., M. Coulon-Bublex, et al. (2001). "Drosophila regulatory factor X is an embryonic type I sensory neuron marker also expressed in spermatids and in the brain of Drosophila." Mech Dev **103**(1-2): 159-62.

VanWert, J. M., S. A. Wolfe, et al. (2008). "Binding of RFX2 and NF-Y to the testis-specific histone H1t promoter may be required for transcriptional activation in primary spermatocytes." J Cell Biochem **104**(3): 1087-101.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Wakasugi, M. and J. Ohta (1973). "Studies on the amoebo-flagellate transformation in Physarum polycephalum." Journal of Plant Research **86**(4): 299-308.

Wang, A. H., S. Gregoire, et al. (2005). "Identification of the ankyrin repeat proteins ANKRA and RFXANK as novel partners of class IIa histone deacetylases." J Biol Chem **280**(32): 29117-27.

Wang, J., H. T. Schwartz, et al. (2010). "Functional Specialization of Sensory Cilia by an RFX Transcription Factor Isoform." Genetics.

Wang, K. R., T. Nemoto, et al. (2007). "RFX1 mediates the serum-induced immediate early response of Id2 gene expression." J Biol Chem **282**(36): 26167-77.

Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Weigel, D., G. Jurgens, et al. (1989). "The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the Drosophila embryo." Cell **57**(4): 645-58.

Weinmann, R., H. J. Raskas, et al. (1974). "Role of DNA-dependent RNA polymerases II and III in transcription of the adenovirus genome late in productive infection." Proc Natl Acad Sci U S A **71**(9): 3426-39.

Weinmann, R. and R. G. Roeder (1974). "Role of DNA-dependent RNA polymerase 3 in the transcription of the tRNA and 5S RNA genes." Proc Natl Acad Sci U S A **71**(5): 1790-4.

Wheatley, D. N., A. M. Wang, et al. (1996). "Expression of primary cilia in mammalian cells." Cell Biol Int **20**(1): 73-81.

Wickstead, B. and K. Gull (2007). "Dyneins across eukaryotes: a comparative genomic analysis." Traffic **8**(12): 1708-21.

Winkelbauer, M. E., J. C. Schafer, et al. (2005). "The C. elegans homologs of nephrocystin-1 and nephrocystin-4 are cilia transition zone proteins involved in chemosensory perception." J Cell Sci **118**(Pt 23): 5575-87.

Wolfe, S. A., J. van Wert, et al. (2006). "Transcription factor RFX2 is abundant in rat testis and enriched in nuclei of primary spermatocytes where it appears to be required for transcription of the testis-specific histone H1t gene." J Cell Biochem **99**(3): 735-46.

Wolfe, S. A., J. M. van Wert, et al. (1995). "Expression of the testis-specific histone H1t gene: evidence for involvement of multiple cis-acting promoter elements." Biochemistry **34**(38): 12461-9.

Wolfe, S. A., D. C. Wilkerson, et al. (2004). "Regulatory factor X2 (RFX2) binds to the H1t/TE1 promoter element and activates transcription of the testis-specific histone H1t gene." J Cell Biochem **91**(2): 375-83.

Wray, G. A. (2007). "The evolutionary significance of cis-regulatory mutations." Nat Rev Genet **8**(3): 206-16.

Wright, M., A. Moisand, et al. (1979). "The structure of the flagellar apparatus of the swarm cells ofPhysarum polycephalum." Protoplasma **100**(3): 231-50.

Xu, Y., P. K. Sengupta, et al. (2006). "Regulatory factor for X-box family proteins differentially interact with histone deacetylases to repress collagen alpha2(I) gene (COL1A2) expression." J Biol Chem **281**(14): 9260-70.

Yang, Z., Y. Yang, et al. (2008). "A novel mutation in BBS7 gene causes Bardet-Biedl syndrome in a Chinese family." Mol Vis **14**: 2304-8.

Ye, S., S. Dhillon, et al. (2001). "An efficient procedure for genotyping single nucleotide polymorphisms." Nucleic Acids Res **29**(17): E88-8.

Zhang, D., G. J. Harry, et al. (2008). "G-protein pathway suppressor 2 (GPS2) interacts with the regulatory factor X4 variant 3 (RFX4_v3) and functions as a transcriptional co-activator." J Biol Chem **283**(13): 8580-90.

Zhang, H. (2003). "Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm." Bioinformatics **19**(11): 1391-6.

Zhang, M. Q. (2002). "Computational prediction of eukaryotic protein-coding genes." Nat Rev Genet **3**(9): 698-709.

Zhang, Z. and J. C. Reese (2005). "Molecular genetic analysis of the yeast repressor Rfx1/Crt1 reveals a novel two-step regulatory mechanism." Mol Cell Biol **25**(17): 7399-411.

Zheng, N., E. Fraenkel, et al. (1999). "Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP." Genes Dev **13**(6): 666-74.