C. elegans Transcription Cofactors

Project Number: <BIO-EFR-0902>

<Bioinfomatics Project for Walhout Lab Umass Medical School>
A Major Qualifying Project Report
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Bachelor of Science
in Biotechnology
by

Victor Zeng

Date: April 29th, 2010
Approved:

Prof. Elizabeth Ryder, Advisor

Table of Contents

TABLE OF CONTENTS 2

TABLE OF FIGURES 4

ABSTRACT 5

ACKNOWLEDGMENT 6

1 INTRODUCTION 7

2 BACKGROUND 9

2.1 TRANSCRIPTION 9
2.1.1 DNA 9
2.1.2 GENERAL TRANSCRIPTION FACTORS (GTFS) 11
2.1.3 REGULATORY TRANSCRIPTION FACTORS (TFS) 13
2.1.4 TRANSCRIPTION CO-FACTORS (TCFS) 14
2.2 OMICS 17
2.2.1 GENOME 17
2.2.2 OMIC DATABASES 18
2.2.3 INTERACTOME 19

3 METHODS 20

3.1 IDENTIFICATION OF TCF DOMAIN SEQUENCES 20
3.2 IDENTIFICATION OF ORTHOLOGOUS TCFS 21
3.3 EVALUATION OF THE PREDICTED TCF LIST 21
3.3.1 INTERACTOME DATABASES 21
3.3.2 INTERACTOME DATABASE PARSER 22
3.3.3 HIGH-THROUGHPUT YEAST-TWO HYBRID DATA 23
3.3.4 TCF TF CONVERGENCE 23

4 RESULTS 24

4.1 IDENTIFICATION OF PREDICTED TCFS 24
4.1.1 IDENTIFICATION OF HISTONE MODIFICATION PROTEINS VIA GENE CLASS SEARCHES 24
4.1.2 IDENTIFICATION OF HISTONE MODIFICATION PROTEIN VIA KNOWN DOMAINS 25
4.1.2 IDENTIFICATION OF NUCLEOSOME REMODELING COMPLEXES VIA TCF ORTHOLOGS 27
4.1.3 IDENTIFICATION OF HISTONE MODIFICATION INTERACTORS VIA KNOWN DOMAINS 27
4.1.4 DOMAIN BLAST RESULTS 28
4.1.5 IDENTIFICATION OF TCFS VIA ORTHOLOG SEARCHES 32
4.2 EVALUATING TCF PREDICTIONS WITH PHENOME DATA 32
4.3 EVALUATING TCF PREDICTIONS WITH INTERACTOME DATA 33
4.3 TCF TF OVERLAP 35

5 DISCUSSION 37

5.1 PREDICTED TCF DISTRIBUTION 37

5.2 PREDICTED TCF RNAI LETHALITY RATE 38
5.3 TCF TF CONVERGENCE 38
5.4 INTERACTOME EVALUATION 39

REFERENCES 40

APPENDIX A PREDICTED TCFS AND PHENOME DATA 44

APPENDIX B BLAST RESULT ALIGNMENTS 50

APPENDIX B.3 LSD-1 SPR-5 ALIGNMENT 50
APPENDIX B.2 BROMO-DOMAIN BLAST ALIGNMENT 50
APPENDIX B.3 CHROMO-DOMAIN BLAST ALIGNMENT 59
APPENDIX B.4 PLANT-HOMEO-DOMAIN BLAST ALIGNMENT 60
APPENDIX B.4 SET-DOMAIN BLAST ALIGNMENT 65
APPENDIX B.4 SET-DOMAIN BLAST ALIGNMENT 67

Table of Figures

FIGURE 1. NURF-1 TRANSCRIPT VARIANT 10
FIGURE 2. BASAL TRANSCRIPTION MACHINERY 12
FIGURE 3. TRANSCRIPTION REGULATORS 14
FIGURE 4. SET DOMAIN 26
FIGURE 5. JMJC DOMAIN 26
FIGURE 6. BROMO DOMAIN 27
FIGURE 7. CHROMO DOMAIN 27
FIGURE 8. PLANT HOMEO DOMAIN 27
FIGURE 9. SET DOMAIN BLAST OUTPUT 28
FIGURE 10. JMJC DOMAIN BLAST OUTPUT 29
FIGURE 11. BROMO DOMAIN BLAST OUTPUT 30
FIGURE 12. CHROMO DOMAIN BLAST OUTPUT 30
FIGURE 13. PLANT HOMEO DOMAIN BLAST OUTPUT 31
FIGURE 14. THE PREDICTED TCFS STATISTICS 33
FIGURE 15. TCF INTERACTOME MAPPING 34
FIGURE 16. INTERACTOME DATA ANALYSIS 35
FIGURE 17. TCF TF CONVERGENCE 36

Abstract

Transcription Cofactors (TCFs) are essential non-DNA binding gene expression regulatory proteins. 162 TCFs were predicted within *C. elegans* using literature search and BLAST. Predicted TCFs consist of mediators, TAFs, nucleosome remodeling, modification, and tail binding proteins. Using a proprietary PSI-MI2.5 parser, 98 known interactions were queried with only 9 interactions with predicted Transcription factors (TFs). 45.7% predicted TCFs shows cause embryonic lethality from RNAi phenotypes. The predicted TCFs can be experimented with predict TFs to find novel interactions.

Acknowledgment

There are many individuals that I would like to thank and acknowledge, as without their support this project would not have been possible.

I would like to thank Doctor Elizabeth Ryder, My WPI Advisor for her support and encouragement throughout my project.

Secondly, I would like to express my deepest appreciation to Doctor Marian Walhout, of the Umass Medical School for sponsoring my MQP. Her continual support and guidance has made this project a reality.

I would also like to express my gratitude to the members of the Walhout Lab for their invested interest and encouragement throughout the completion of my project.

1 Introduction

A biological system functions through the interactions of molecules such as carbohydrates, lipids, nucleic acids, and proteins. To fully understand the biological systems, biologists strive to complete the interactome, which contains all interactions of every molecule within an organism. Regulation of gene expression is an area of biological studies, and the number of molecules involved in the regulations and their interactions increase the intricacy of differential gene expression and dictate an organism's complexity (*Levine et al. 2003*). There are many protein components involved in the regulation of gene expression. General Transcription Factors (GTFs) are required for basal transcription and they are regulated by other factors in transcription. Regulatory Transcription Factors (TFs) are DNA binding regulatory factors that are not required for basal transcription. Transcription Co-Factors (TCFs) are non-DNA binding regulatory factors that are not required for basal transcription. GTFs, TFs, and TCFs are all essential for the regulation of transcription through their protein-protein interactions. A complex organism such as *Homo sapiens* has approximately 2600 predicted TFs (*Babu et al. 2004*). A simpler organism such as *Caenorhabditis elegans* (*C. elegans*) has a lower number of regulatory factors resulting in less molecular interactions, and causing it to be more feasible for systems biology studies than human. Studies of *C. elegans* molecular interactions within gene expression regulation will provide experiences and answers for future studies of *Homo sapiens*.

To understand the regulation of transcription, researchers must know which proteins are GTFs, TFs, and TCFs. Because of high level of conservation over evolution, the GTFs of *C. elegans* are indentified (*Verrijzer et al. 1995*). A DNA binding domain-based analysis generated a list of 934 *C. elegans* TFs (*Reece-Hoyes et al 2005)*. There has been research with the effort in explaining TCF functions (*Roeder 2004*), but little is known of which proteins within *C. elegans* are TCFs. In addition, there is a lack of knowledge regarding the interaction between the three types of regulatory factors of transcription.

Currently, the *C. elegans* interactome is incomplete, and some of the recorded interactions are results of computational prediction that lacks experimental support. The prediction of *C. elegans* TCFs will allow future protein-protein interaction detection with the predicted TFs and identified GTFs. The detection of these proteins' interaction can be done using high-throughput screening. The knowledge of their interaction will further the understanding of the transcription regulation network of *C. elegans*. In addition, the results of their interactions will aid the completion of the *C. elegans* interactome.

To accompany the development of the interactome, newer methods of data mining from the interactome database need to be developed. The size of the interactome database will increase with additional datasets causing searches within the interactome to be more difficult. Currently, there are multiple parties building interactome databases with different information. The current method permits the search of the interactions of only a single interactor within a database at a time. For biology studies using the interactome, biologists will need to gather all known interactions of many proteins in a time efficient manner. The creation of a program to perform batch interaction screening with all of the interactome databases will decrease the research time.

For this project, a list of predicted TCFs was determined for future detection of protein interactions, and a program was created for fast search of the interactome data. The prediction of TCFs required a comprehensive literature search for research regarding TCFs. Using the knowledge of the literature, protein families that relate to TCFs, and protein domains that show TCFs function were identified. The predicted TCFs were then gathered from the *C. elegans* genome through searches and Basic Local Alignment Search Tool (BLAST). Evaluation of the predicted TCFs then was done using the interactome data parser and the extensive *C. elegans* phenome.

2 Background

Gene expression regulations manage organisms’ reproduction, development, and responses to external stimulus. Transcription, translation, localization, and degradation are some cellular processes where countless factors regulate organisms’ gene expression. Transcription is the initial step in the central dogma, and it precedes the other cellular processes. The inhibition of factors that are involved in transcription regulation results in lethality (*Fraser et al. 2000*). This is because transcription regulation factors are key components of the gene expression regulation network. The mapping of transcription regulation factor interactions will improve the understanding of gene expression.

2.1 Transcription

Eukaryote transcription is a widely studied subject in biology due to its crucial role in the central dogma. Biologists view general transcription as a stepwise assembly line (*Dignam 1983*). General transcription consists of the initiation, elongation, and termination steps of RNA polymerization. In eukaryotes, the components of basal transcription are DNA, General Transcription Factors (GTFs), and RNA polymerases. Many different lineages of proteins that interact with the general transcription components emerged via evolution. There are additional proteins that interact with those proteins that interact with the general transcription components. Together all of the proteins produce a network of protein interactions that regulate transcription.

2.1.1 DNA

Transcription is the production of RNA using a template DNA. DNA contains multiple regions, and each region serves a critical role in the transcription process. In eukaryotes, the DNA has regions that are transcribed into RNA, promoter and enhancers that bind TFs, and many un-transcribed regions that form the tertiary structure of DNA. The interaction of proteins within these regions regulates the outcome of transcription.

The region of a gene that polymerases transcribe encodes the RNA transcripts. The transcribed RNA in eukaryotic organisms consists of sections termed exons and introns. The exons are selected and the introns are removed based on splice pattern to form the final transcript prior to translation (*Crick 1979*). This process produces transcription variants from one gene (Fig. 1). Sequences such as the start codon AUG and the stop codons of UAA, UAG, and UGA can also be found within the RNA transcript. The Open Reading Frame (ORF) of the spliced mRNA that is located between the start and the stop codons will be translated. Based on the different splice variants, a different set of start and end codon may be encouterd by the ribosome, and thus a different ORF is produced. Some of the transcribed fragments do not possess ORF, such as the DNA that encode enzymatic RNA. For those transcribed fragments that do possess ORFs, ribosomes can translate the resulting RNA transcripts into proteins (*Rosenberg et al. 1979*).

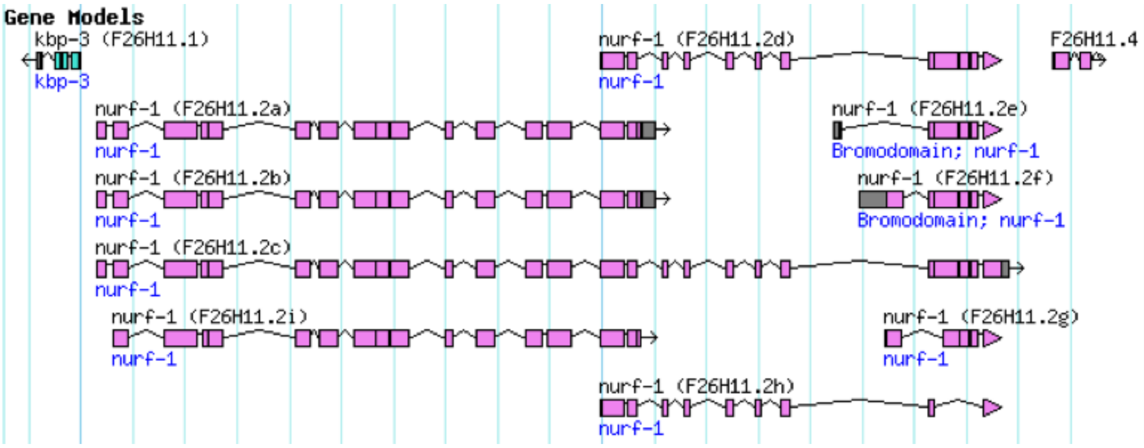


Figure 1. nurf-1 Transcript Variants

The 6 different transcript variants currently known for *C. elegans* nurf-1 are shown.

RNA polymerases are initially recruited to the promoter and transcribe in a 5'->3' fashion, thus the promoter region of DNA is found at the 5' end of the transcribed region. The TATA box is a specific DNA sequence of TATAA within the promoter. The TATA Binding Protein (TBP) associates with the TATA box and creates a base for the assembly of the transcriptional machinery (*Nakajima et al. 1988*). Biologists refer to the combination of promoters and their ORFs as genes.

When genes are not activated for transcription, histones super coil DNA (*Almer et al. 1986*). Histones are chromatin structural proteins, and form nucleosome complexes with DNA (*Laybourn et al. 1991*). Nucleosome complexes are very compacted. This mechanism prevents most factors involved in transcription from accessing the DNA, and inhibits unsystematic transcription of the compacted genes. In contrast, the promoters of activated genes are depleted of nucleosome, which allows the interaction between TFs and promoter sequence.

Enhancers are similar to the promoters because of their TF binding capability. Unlike the promoter, which must be located directly upstream of the regulated gene, the enhancer may be distal from the gene it regulates. In eukaryotes, the tertiary folding of DNA allows a distal enhancer to become extremely close to the gene it regulates. In some cases, the enhancer may exist on a completely separate chromosome as the regulated gene (*Geyer et al. 1990*). Finally, studies show some transcribed regions of DNA have TF binding affinity. For example, the murine immunoglobin Hμ core enhancer is located within the second intron of its regulated gene (*Blackwood et al. 1998*).

2.1.2 General Transcription Factors (GTFs)

Basal transcription in eukaryotes requires not only the DNA; it needs essential proteins that are termed GTFs, along with RNA polymerases. *In vitro*, GTFs are recruited to the promoter to form the Pre-Initiation Complex (PIC) with the RNA polymerases (*Rowland et al. 1994*). PIC is necessary for transcription because of its ability to recruit the RNA polymerases to genes being transcribed and aid the RNA polymerases with the down stream activity (Fig. 2).

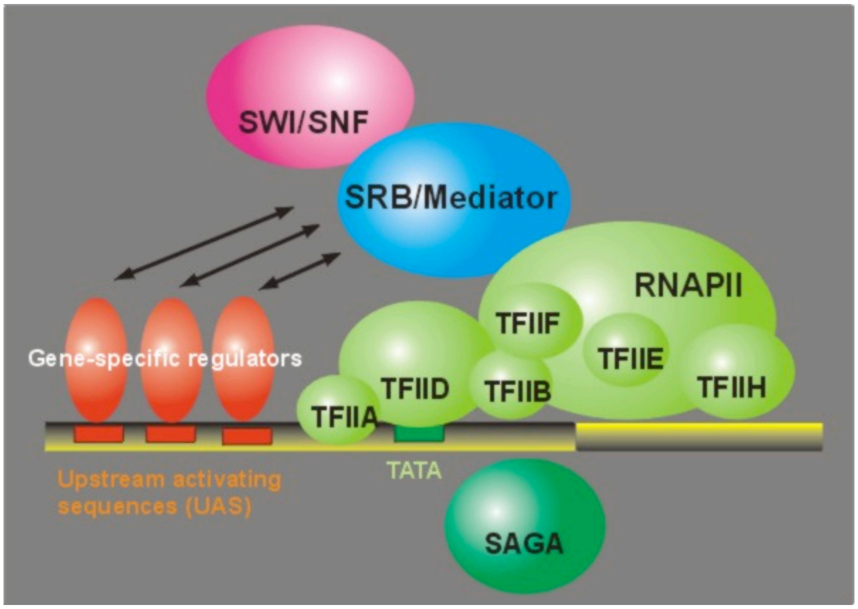


Figure 2. Basal Transcription Machinery

The basal transcription machinery is shown in light green, the mediator complex in blue, histone remodeling complexes in pink and dark green, and TFs in orange (Holstege et al. 1998).

One of the important roles of GTFs is the recruitment of the RNA polymerases. As described previously, the TATA box within the promoter is the base for the PIC for the transcription of mRNA via RNA polymerase II. TBP (TATA binding protein) is a subunit of GTF TFIID, and it is the base of a complex formed with TBP Associated Factors (TAFs) (*Lee. T. et al. 2000*). TAFs are distinguished from the GTF machinery in that they are not required for basal transcription. TFIIA could be considered either a GTF or a TAF, and it interacts with TBP in a similar manner as to TAFs. TFIIA is not required for basal transcription *in-vitro,* which initiated the debate of whether it is a GTF. TFIIA is essential *in-vivo* due to constitutive ubiquitous TFIID repressors in the nucleus (*Ozer et al. 1994*).

GTF TFIIB is another unit of the PIC that is required for basal transcription. TFIIB creates the bridge between TFIID and RNA polymerase II (*Verrijzer et al. 1995*). In addition, TFIIB is shown to have protein-protein interaction with TFs, such as the cAMP-response element binding protein (CREB) (*Tini et al. 2002*). TFIIB is regulated by CREB through protein-protein interactions, and these interactions are the key to understanding the regulation of transcription.

TFIIF binds DNA in a non-sequence specific manner, and is required for basal transcription in eukaryotes (*Robert et al. 1998*). TFIIF has protein affinity for both TFIIB and RNA polymerase II, and it is predicted to aid TFIIB in the bridging with RNA polymerase II. Though TFIIF has some structural functions in the PIC, its main purpose is to wrap DNA around the transcription complex. As the transcription complex travels down stream, DNA functions like a conveyer belt with the aid of TFIIF. Studies show phosphorylation of TFIIF may terminate transcription pauses (*Tan et al. 1995*).

Two GTFs are not involved in the recruitment of RNA polymerases to the transcription start site, but are still required for the PIC due to their duty during the elongation step of transcription. These two GTFs resemble TFIIF because phosphorylation of their c-terminal domain can also affect transcription pauses (*Kugel et al. 1998*). TFIIE is one of these GTFs, and its enzymatic function is DNA melting, which is to break the hydrogen bond of the double stranded DNA base pairing. TFIIE performs its activity through its zinc ribbon catalytic domain (*Okuda et al. 2004*). TFIIH functions as a helicase in conjunction with TFIIE. TFIIH is identified as a helicase due to its ability to use cellular energy ATP to unwind the DNA helix and separate DNA during elongation. These GTFs are essential for granting the polymerase access to the DNA while moving down stream.

2.1.3 Regulatory Transcription Factors (TFs)

TFs are DNA-binding proteins that are not required for basal transcription. TFs have functions similarity to the TBP in DNA binding and protein recruitment. TFs can both activate and repress transcription by recruiting or blocking the formation of PIC respectively. (*Roeder 1996*) Based on the DNA binding domain, the binding affinity of TFs may vary significantly. Some transcription factors have multiple DNA binding domains, which grants them more specificity for DNA interaction. There are many DNA binding domains for TFs. The majority of transcription factors have catalytic sites within their secondary structures. These secondary structures fit within the major grove of DNA allowing it to interact with the aromatic bases. (*Mitchell et al. 1989*) There are also TFs that bind to the minor grove of DNA, such as the TFs with the AT-hook domain. The AT hook domain does not bind a specific sequence but targets AT-rich regions of DNA. TFs typically bind to the enhancer and promoter regions of DNA. There have also been cases showing TFs association to heterochromatin (*Raff et al. 1994*). The DNA binding domains create an extensive network of interactions between TFs and DNA.

TFs also interact with bio-molecules other than DNA, such as proteins and lipids. The interactions with these bio-molecules regulate TFs binding with DNA. Heterodimerization and homodimerization of TFs create additional DNA specificity to the dimer, and affect DNA binding (*Helin et al. 1993*). The functions of some TFs are altered by different environments, such as hypoxia (*Zheng et al. 1998*). Nuclear hormone receptors are a group of TFs with ligands. These TF ligands interact with a variety of lipid hormones, such as estrogen, steroid, thyroid, vitamin A, and vitamin D receptors. These lipid hormones are hydrophobic and may penetrate the nuclear membrane for direct signaling to their target receptors (*Evans 1988*). TFs may also interact with non-TF proteins that result in different functions and regulations.

2.1.4 Transcription Co-Factors (TCFs)

TCFs are proteins that are involved in transcription that do not interact with DNA, and are not required for basal transcription. TCFs are typically recruited by TFs for their functions via protein-protein interactions (*Roeder 2004*). Most TCFs are found in complexes within the nucleus. TCFs are believed to have an assortment of distinct functions, including recruitment of transcription machinery, nucleosome remodeling, and histone modification (Fig. 3).

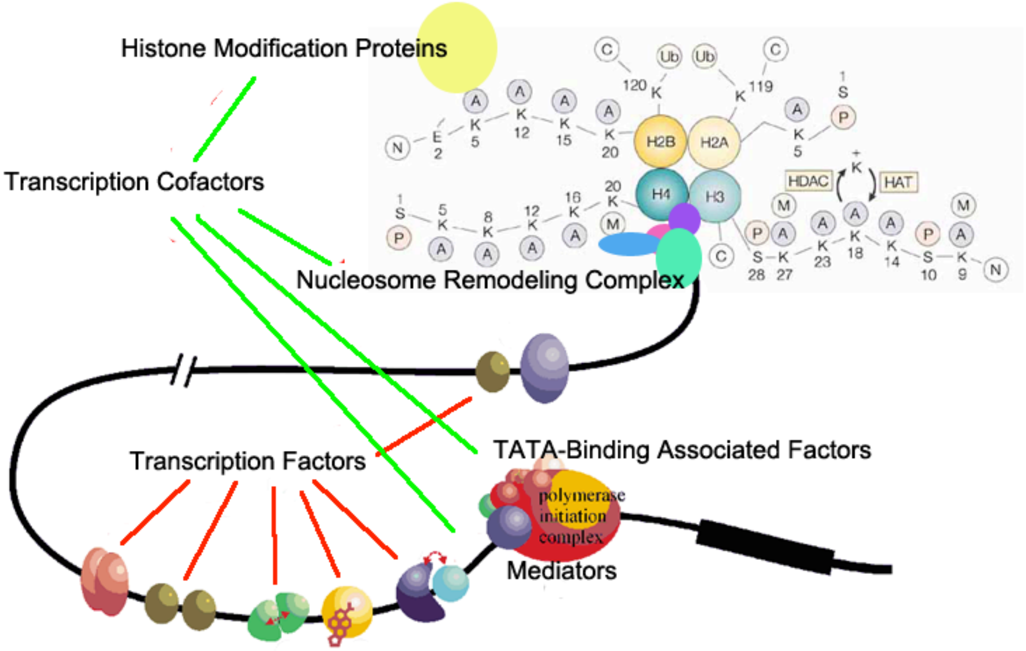


Figure 3. Transcription Regulators

The DNA is shown as a black line, and the protein-coding gene in a black box. The TCFs regulating this gene's expression in green lines, and the DNA bind TFs in red lines.

The Mediator Complex proteins are considered TCFs due to their ability to recruit GTFs to specific TFs. Mediator complex proteins resemble GTF TFIIB in ability to their bridge between TBP and polymerases. Mediator complex proteins vary based on the ligand of the TFs they bind. There are the ARC/DRIP mediators that interact with vitamin receptors, and the TRAP mediators that bind the thyroid receptors (*Rachez et al. 2001*). These mediators tend to associate with TFs activated by ligands to allow GTF and polymerase recruitment. They may also form complexes with other TCFs to recruit their enzymatic functions *(Roeder 2004)*.

TAFs of GTF TFIID are considered TCFs. TAFs associate with TBP similar to TFIIA and TFIIB. TAF proteins can recruit and mask TBP from repressors and activators. Multiple TFs such as Sp1 require TAFs for bridging and recruitment of GTFs. (*Pugh et al. 1990*) TAFII250 also shows histone modification activities, and it highly resembles GCN5, a major yeast histone acetyl-transferase protein (*Mizzen et al. 1996*). TAFs can be considered as TCFs based on their ability to recruit transcription machinery and conduct histone modification.

Nucleosome remodeling is another function of TCF that work with histone modification to allow transcription machinery access to the naked DNA. There are many studied nucleosome-remodeling complexes such as the NuRD complex, Swi/Snf, RSC (*Roeder 2004*). These complexes all have very common features, such as proteins in the ATPase and helicase families. These nucleosome-remodeling proteins resemble TFIIH in its function to modify DNA strand's conformation. DNA super-helices are flattened by nucleosome remodling complex's helicases, while ATPases provide the kinetics for the physical movement (*Sudarsanam et al. 1999*).

Histone modification is the most important role of TCFs in the process of transcription. In eukaryotes, histone octamers cause the formation of heterochromatin from euchromatin, which inhibits transcription of the compacted region. The lysine rich histone tail has very basic chemical properties, and tightly binds to acidic DNA (*Allfrey 1964*). The lysines and arginines of histone tails are very susceptable to post-translational modification. Histone AcetylTransferases (HATs) work in pairs with Histone DeACetylases (HDACs), and this reflects the high amount of changes in post-translation modification of histone that occurs in cells. ADP-ribosylation, methylation, phosphorylation, and summoylation are some other forms of histone modifications. These modifications create the histone code, which regulates transcription.

The major functions of TCFs are achieved through protein-protein interactions of either direct association, or post-translational modification. Using protein sequence consensuses, biology can predict possible protein domains for TCF activity. Some TCF domains were predicted based on their ability to interact with post-translationally modified proteins. The bromo-domain is a 110 amino acid peptide that folds to create multiple alpha helixes. It is a very important domain for some TCF due its ability to bind with acetylated lysine (*Zeng et al. 2002*). Many HAT and nucleosome remodeling proteins have the bromo-domain because it allows these proteins to localize to histone tails through recognition of histone acetylation.

The chromo-domain is 50 amino acids long and folds to create alpha helix and beta sheets that have specific protein affinity for methylated lysine. Peptide variability may cause chromo-domain proteins to associate with different methylated lysines of histone tails (*Cavalli et al. 1998; Brehm et al. 2004*). CHD-1 a highly conserved chromo-domain containing protein that binds to lysine-4 of the histone H3 tail, while the chromo-domain of Polycomb Protein binds to lysine-29 of the histone H3 tail. The chromo-domain of these two proteins only differs by 5 amino acids, thus, the selectivity may be caused by other factors. The chromo-domain's activity is essential for numerous TCF functions.

Plant Homeo Domain (PHD) finger is a cysteine rich protein domain that is approximately 50-80 amino acids long. This domain has distinct similarity to a zinc-finger, but does not have DNA binding capabilities. PHD activity is predicted to allow adhesion of protein complexes through direct association (*Aasland et al. 1995*). The PHD domain shows strong signs of self-association *in-vitro*, and it occurs in many proteins of chromatin remodling complexes. The PHD domain also interacts specifically with tri-methylated lysine.

Post-translational modification proteins exist in many other biological systems other than histones modification. Those proteins have high levels of similarity compare to the histone modification proteins, and cause the identification of a specific histone modification domain difficult. There are two domains known for their histone modification functions. The SET domain is a 130 amino acid peptide, and studies have connected it with transcription silencing and activation. The function of SET domain is methylation of histone lysine (*Dillon et al. 2005*). Different sub families of SET domain target different lysines of histone tails. SET domain is a major player in the histone code and is a TCF domain. Jmjc domain is the second post-translational modification domain, and its functions as a histone methylase (*Klose et al. 2006*). Jmjc also contains multiple subfamilies that vary in peptide sequence.

2.2 Omics

Omics is a term used in biology that originated with the creation of the Genome and Proteome. Omics is the holistic approach in annotating all molecules of organisms. Currently, with the advancement of computer technology, omic information is annotated in computer databases. System biologists utilize engineering to develop new methods to streamline experimental process to provide the vast amount of data required for omic databases. These databases are stored in servers that allow biologists throughout the world to access the knowledge via the World Wide Web. Bioinformatics has emerged as a field for the analysis of the databases while making them easier access. There has been tremendous development in the storage, mining, visualization, and computation of omic databases.

2.2.1 Genome

As the oldest of the omic databases, the genome database is developed with sophistication. In 1989, Jean Thierry-Mieg of University of Montpellier and Richard Durbin of Sanger Institute developed "A *C. elegans* DataBase" (ACeDB). ACeDB is an information system, and it is very different from traditional computer databases. (Biology Research Computer Hierarchy) Traditional filing in a database uses a family system, and this means the directory consists of parent, offspring, and siblings. In AceDB the file relationship is the user-defined, which is more suitable for storage of biological data. For example, user defined directory allows a hierarchy with gene, RNA transcript, ORF, and protein in order while allowing RNA transcript, ORF, and protein to be siblings under genes. Using user-defined directory can increase the mining speed in large biological databases.

An intelligent browsing system was also created to access the genome information generated from a variety of experimented data. Experts of bioinformatics created graphical browsers that allows biologists to navigate through the genome and conduct, data mining. www.wormbase.org is a browser of the AceDB. Using a graphical interface Wormbase can visually illustrate a specific locus of gene, whether experimentally proven or predicted. This browser has *in-silico* abilities, which are computer predictions. Using different algorithms, Wormbase can determine signature features within the genome, such as AT-rich regions, and repeated sequences. In addition, the browser has an algorithm for Genome wide sequence alignment that allows biologists to search for particular sequence patterns. Other than the DNA sequence, the developers of the browser incorporated the ability to access information of other omic databases.

2.2.2 Omic Databases

There are many omic databases other than the genome. All of the omic databases work synergistically. The information of each database can be validated and cross-referenced by the others. Biological processes involve countless different molecules. During transcription, RNA polymerases transcribe template DNA into RNA. The transcriptome is an attempt to identify all RNA transcripts produced during transcription. A recent study of the *C. elegans* transcriptome has shown the 14% of sequenced transcripts do not have a corresponding gene in the genome. (*Shin et al. 2008*) This study of transcriptome has demonstrated missing information within the genome, and provided knowledge for future improvements.

During translation, the ribosome may translate RNA transcripts into protein. The ORFeome utilizes computation to predict possible ORFs using genome and transcriptome data (*Reboul et al. 2003*). The ORFeome prediction can also be validated using proteome data. The proteome is a collection of all proteins produced in an organism under all enviornmental conditions during all developmental stages. The verification of proteome data is easier with the creation of protein mass spectroscopy technique (*Mann et al. 1993*). With the knowledge of protein sequences, the ORF can be verified, and provide information for future studies involving transcription variants. This is another example of the synergy between omic databases.

Using the genomic information, system biologists built the phenome to study the phenotypic function of genes. The phenome incorporates phenotypic data for mutated genes' alleles as well from RNA Interference (RNAi) experiments. In vivo, micro RNA and small interfering RNA act by binding mRNA (*Fire et al. 1998*). The RNA Induced Silencing Complex (RISC) dice and break down the double stranded RNA to prevent translation. Synthetic RNA can be created to perform RNAi on specific genes (*Kamath et al. 2002*). RNAi can be induced in organisms through multiple methods such as injection, feeding, and soaking during all different developmental stages (*Rual et al. 2004*). Using RNAi biologists may observe the phenotypic outcome resulting from the reduce expression of genes. Both experimental data of mutants and RNAi phenotype provide extensive information of gene functions.

Outside of the central dogma, many bio-molecules need to be included within omic database because of their involvement in gene expression. Carbohydrates and lipids are involved in organisms' metabolism, and can interact with proteins to produce differential gene expression. Carbohydrates such as glycans are involved in cellular signaling. The glycome is the annotation of carbohydrates in organisms. Lipids are also involved in cellular signaling. Molecules such as hormones and vitamins are annotated within the lipidome. The interactions of molecules within the glycome and the lipidome can be studied in conjunction with other omic databases to further the understanding of gene expression regulation.

2.2.3 Interactome

The Majority of Omics focuses on the annotation of bio-molecules. Interactome is developed to annotate the interactions between the bio-molecules of an organism. Studies predicted that human has approximately 650,000 protein interactions, which is about 3 times more than C.elegans (*Stumpf et al. 2008*). Numerous data are needed for the construction of an interactome. To perform the require number of repeated experiments, the aid of robotics automation and high-throughput system is used. High-throughput devices are designed to perform multiple experiments simultaneously. Micro array chips may be used to show more than hundreds of interaction at a time (*Bader et al. 2003*).

Many labs have been using high throughput yeast 2-hybrid system for gathering of large datasets to accomplish this goal. Yeast 2-hybrid system is able to determine binary protein interaction, and has been use by scientists for the mapping of *Saccharomyces cerevisiae* (*Ito et al. 1999*), *Drosophila melanogaster* (*Giot et al. 2003*), C. elegans (*Li et al. 2004*), and possibly in human (*Rual et al. 2005*). Protein mapping using interactome can provide a visualization of gene regulation networks.

3 Methods

The goal of this project was to create a list of *C. elegans* TCFs. There are many groups of TCF related proteins, such as histone modification and tail-binding proteins, nucleosome remodeling proteins, transcription mediators, and TAFs (Roeder 2004). Some proteins of these groups are incorporated into gene-classes that were gathered from Wormbase. Some proteins of these groups are indentified as complexes, and were gathered through literature research. A few groups have conserved protein domains, and were gathered using BLAST of *C. elegans* genome.

3.1 Identification of TCF Domain Sequences

Based on countless previous researches, approximately 800 protein domains were identified through sequence alignment of various species' proteins with identical biochemical functions. Biologists use homologs, orthologs, and paralogs of domain containing proteins to determine the consensus domain sequences of functional peptide. The sequences of TCF protein domains were gathered from a database named Simple Modular Architecture Research Tool (SMART). SMART creates protein domain consensus sequences by aligning sequences of all proteins that are known to contain the specific domain. The protein sequences are gathered from databases, such as the National Center for Biotechnological Information (NCBI). SMART has low consensus accuracy for those protein domains that possess multiple sub families. This is because the functional peptides of those protein domains differ greatly between the sub families. The sequences of these domains are gathered through literature research.

BLAST was use to identify those proteins with the desired domain within the *C. elegans* genome. Those proteins with E-values less than e-10 were considered as a predicted TCF. Through literature research, the active sites and highly conserved amino acids of each protein domain of interest were gathered. Proteins with E-values higher than e-10 were kept as predicted TCF if all active sites and conserved regions matched within the alignment resulted from BLAST.

3.2 Identification of orthologous TCFs

Mulitiple TCF complexes were determined using previous research in multiple model organisms. The sequences of proteins that were identified as predicted TCF were used in a BLAST of *C. elegans* genome. To maintain the confidence of the orthologs gathered from BLAST, only results with E-value under e-50 were kept. Using this approach, many orthologs of the previously identified proteins were found. The majority of the orthologs found were already identified as predicted *C. elegans* TCFs through literature search. A literatures have predicted the orthologs of many non-*C. elegans* TCF complex proteins within *C. elegans*, they were also incorporated into the TCF list (*Chue et al. 2006*).

3.3 Evaluation of the predicted TCF list

The silencing of those proteins that are involved in the regulation of transcription can generate many different phenotypes. The phenotypes of gene silencing experiments were gathered from Wormbase to evaluate the specific lethality phenotype of predicted TCFs. Only the phenotypes from RNAi experiment were used for this analysis. Wormbase also stores the phenotypic outcome of many mutants. For the evaluation of the predicted TCFs, mutant phenotypes were not used. Although the level of gene silencing varies amongst genes, RNAi experiments are done with the same molecular approach. Based on the alleles of a gene, a range of different changes can occur to the gene expression. Certain mutations such as point mutation of a protein's active site may result in the complete loss of function. In other mutations, the protein's functions are not altered in the same manner, which may result in a range of different phenotypes. For the comparison of large groups of genes using phenome data, utilizing only RNAi phenotypes is more precise.

3.3.1 Interactome Databases

To better our understanding of gene expression, the interaction network of transcription regulation proteins needs to be completed. To determine the need for future experiments to be performed for the identification of novel interaction, predicted TCF data within the current interactome were evaluated. All of the interaction databases including Intact, MINT, and DIP are stored as plain ASCI text files using the Proteomics Standards Initiative – Molecular Interactions level 2.5 (PSI-MI 2.5) (Hermjakob 2006).

This file format adopts the XML structure that assigns classes to information using HTML tags, and subclasses are created within HTML tags of parent classes. The classes that can be assigned to data with PSI-MI 2.5 databases are dictated by the human proteome organization.

Currently, the *C. elegans* interactome is incomplete, and the majority of the data relies on orthologous interactions using the interactome data of other species. The interactome data are also spread amongst multiple databases, causing the search for interactions to be difficult. The European Molecular Biology Laboratories' European Bioinformatics Institute (EMBL-EBI) hosts the Intact protein interaction database for multiple major model organisms. The University of Rome Tor Vergata has created a database for the annotation of protein interaction termed Molecular INTeraction (MINT). Database of Interacting Protein (DIP) was created by the University of California Los Angelas. All of these databases have visual User Interfaces (UI) that allow users to search for interaction of a particular protein over the Internet. These UI lack the ability for batch protein interaction search that is required for the predicted TCFs. By creating a third party database parser, the tedious manual search can be avoided.

3.3.2 Interactome Database Parser

Perl is a widely used coding language that is heavily utilized in database servers. In addition, Perl is a powerful text parser, and can easily manipulate the text within PSI-MI 2.5. Perl was chosen as the language to program the interactome search software. The PSI-MI 2.5 also has a very intelligent method of separating interactome information. The interactors are stored within one section of the file, while the interactions and experimental information are stored separately in their own sections of the file. This method allows all information to be recorded once with the database, and the data parser only has to iterate through a single section of the file to gather specific data based on XML class tags. For the perl based search software, each piece of information within each section of the interactome is stored within an array. An associative array was then created by the program to link arrays of interactors, interactions, and experimental data together. By inserting a txt file of the gene names of interests, the software parses through the associate array to output known interaction information.

Using the perl PSI-MI 2.5 parser, the interactions of the predicted TCFs were gathered from both Intact and MINT. These interactions are shown in figure 4.4. DIP was not used because it did not match the same PSI-MI 2.5 standard as Intact and MINT. The Inconsistency of DIP XLM class compared to Intact and MINT caused error during the parsing of the perl search software, and the resulting information from DIP was invalid. In the future, the search software can be patched and debugged to allow for the usage of DIP data.

3.3.3 High-Throughput Yeast-Two Hybrid Data

The high-throughput yeast-two hybrid protein interaction detection data from the Vidal lab was also used to determine interactions of each predicted TCF (*Li et al. 2004*). The interactions detected from this experiment are not stored within a PSI-MI 2.5 file. Because the data is stored within an excel file, the perl search software could not be used. A basic perl parser was used to find interactions of the predicted TCFs within the Vidal lab data to prevent rigorous manual search. The interactions found using Vidal lab data increased the final number of interactions gathered for the predicted TCFs.

3.3.4 TCF TF Convergence

To identify the similarity between the predicted TCFs and previously predicted TF from Walhout lab, a comparison was done between the both lists. The result may provide further insight to the functionality of TCFs. The search of the convergence of the two lists was done using perl arrarys. Information such as containing domains of those proteins that are blong to both lists was gathered from Uni-Prot.

4 Results

This project was conducted to create a predicted list of *C. elegans* TCFs (Appendix A). The prediction of *C. elegans* TCFs was done so future researchers may utilize the predicted proteins for the detection of interaction with other transcription related proteins, such as TFs, and GTFs. The prediction of *C.elegans* TCFs was carried out using data from previous TCF related research, such as the study of *C. elegans* TCF protein complexes, the study of TCFs of various eukaryotes, and the study of proteins with functions related to TCF functions. The predictions of *C. elegans* TCFs were analyzed using phenome and interactome data. In addition, to improve the mining of interactome data, a program was written to output interactions of the predicted proteins. This program was made to allow the search of multiple proteins' interactions using multiple interactome datasets at the same time.

4.1 Identification of Predicted TCFs

TCFs are non-DNA binding nuclear proteins that regulate cellular gene expression. Through literature research, TCFs were identified based on two functionalities. One is to regulate transcription through the histone code, and the other is to regulate transcription via the recruitment of transcriptional machinery. Both functionalities of TCFs involve proteins and complexes of different activities that can be categorized into sub-families of TCFs. The particular genes coding for each sub family within *C. elegans* were identified through the search of *C. elegans* gene classes, known domains, or TCF orthologs.

4.1.1 Identification of Histone Modification Proteins via Gene Class Searches

The histone code involves proteins with post-translational modification ability. There are many possible post-translational modifications of histones including acetylation, methylation, phosphorylation, sumoylation, and ADP-ribosylation. The particular gene coding for specific *C. elegans* histone modification proteins were identified through the search of *C. elegans* gene classes, and known domains. For acetylation, there is a particular *C. elegans* gene class of histone acetyl-transferase named mys, which were originally identified from histone acetylation complexes. The full name of mys is MYST, and it is the abbreviation of the 4 histone acetylation complexes, which are MOZ, Ybf2/Sas3, Sas2 and Tip60. There are two *C. elegans* gene classes of histone deacetylase named hda and hdac. The proteins of these gene classes were gathered through Wormbase. These proteins were assigned to their gene class based on their public name. The public names of *C. elegans* genes are typically based on the major mutant or RNAi phenotype, but sometime the names are based on the predicted gene function.

There are also gene classes for those proteins involving the methylation of histones. Set is a gene class of histone methyltransferase, and it is name after the SET domain. A gene class of histone demethylase is named lsd, which stands for lysine specific histone demethylase. There is only one protein of this gene class in *C. elegans* that was discovered via its homology to the human lsd protein. One histone kinase family of *C. elegans* was found during the literature search. The air gene class in *C. elegans* is based on its homology to *Drosophila* Aurora kinase and yeast Ipl protein. These proteins were included as predicted TCFs.

Both summoylation and adp-ribosylation have literatures supporting their occurrences on histones (*Realini et al. 1992*). There was no literature showing specific ADP-ribosylase activity on histones within *C. elegans*. There are two gene classes of ADP-ribosylases within *C. elegans* genome, and they are arl and arf. The proteins of these two gene classes do not have experimental support of their activity on histones, and are not included as predicted TCFs. Ubc-9 was found to cause summoylation of histones within *C. elegans* (*R. Hay 2005*). Multiple proteins of ubc gene class are also ubiquitin-conjugating enzymes that are paralog of ubc-9. These proteins do not currently have literature supporting any histone activity, and are not included as predicted TCFs.

4.1.2 Identification of Histone Modification Proteins via Known Domains

Some protein domains are reported to have histone modification ability. The SET domain (*Dillon et al. 2005*), and Jmjc domain (Klose *et al. 2006*) were also chosen because of their unique histone methyltransferase and demethylase activities. The sequences of these two domains are shown below (Figs 4, 5). Set and Jmjc domain sequence were not gathered from SMART because they have multiple subfamilies, and did not have an accurate consensus sequence on SMART. Both sequences were gathered through literature along with the conserved amino acids and active sites.

```
ARSRIAGLGLYAKVDISMGDFIIEYKGEIIRSEVCEVREI 2420
RYVAQNRGVYMFRIDEEWVIDATMAGGPARYINHSCDPNC 2460
STQILDAGSGAREKKIIITANRPISANEELTYDYQFELEG 2500
TTDKIPCLCGAPNCVKWMN
```

Figure 4. SET-Domain

The domain sequence of SET-domain gathered from set-16 is shown. The highlighted sequences are the lysine targeting amino acids (shown in yellow) and the catalytic site (shown in red) (Dillon et al. 2005).

```
(JHDM1)
FSQTPLEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGR 040
DKKFYNPHHTFPKVQNYCLMSVANCYTDFHIDFSGTSVWY 080
HVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE 120
KCHVAILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQ 160
SCKTQLRVYQVEN

(PHF2/8)
SDNNEMKEIAKPPRFVQEISMVNRLWPDVSGAEYIKLLQR 040
EEYLPEDQRPKVEQFCLAGMAGSYTDFHVDFGGSSVYYHI 080
LKGEKIFYIAAPTEQNFAAYQAHETSPDTTTWFGDIANGA 120
VKRVVIKEGQTLLIPAGWIHAVLTPVDSLVFGGNFLHLGN 160
LEMQMRVYHL

(JARID1/2)
GMCFSTFCWHTEDHWTYSVNYNHFGERKIWYGVGGEDAEK 040
FEDALKKIAPGLTGRQRDLFHHMTTAANPHLLRSLGVPIH 080
SVHQNAGEFVITFPRAYHAGFNEG

(JHDM3/JMJD2)
DAQVEEWNMNRLGTILEDTNYEIKGVNTVYLYFGMYKTTF 040
PWHAEDMDLYSINFLHFGAPKYWFAISSEHADRFERFMSQ 080
QFSYQNEYAPQCKAFLRHKTYLVTPELLRQAGIPYATMVQ 120
RPNEFIITFPRGYHMGFNLGYNLAESTNFASQRWIDYGKD 160
AVLCDC

(UTX/UTY)
KWGKQINELSKLPAFCRLIAGSNMLSHLGHQVHGMNTVKL 040
FMKVPGCRTPAHQDSNHMASININIGPGDCEWFAVPYEYW 080
GKMHKLCEKNGVDLLTGTFWPIIDDLLDAGIPVHRFTQKA 120
GDMVYVSGGAIHWVQASGWCNNISWNVAPLNFQQLSISLL 160
SYEY
```

Figure 5. Jmjc-Domain

The domain sequences of 5 sub-families of Jmjc-domain are shown. The highlighted sequences are the Fe (II) targeting amino acids (shown in yellow) and the α-ketoglutarate targeting amino acids (shown in red) (*Klose et al. 2006*).

There are many other forms of histone modification, and also many domains for acetyltransferase, kinase, ubiquitin conjugase, and adp-ribosylase. Those domains were not chosen as TCF domains because results from using a BLAST will produce many false positives. False positives are caused by proteins within *C. elegans* that possess those domains for post-translational modification and do not have direct activity on histones and the regulation of transcription.

4.1.2 Identification of Nucleosome Remodeling complexes via TCF Orthologs

The histone code modulates transcription throught the recruitment of nucleosome remodeling complexes. Many nucleosome remodeling complexes have been discovered in model organisims, such as the SWR1/SRCAP of *A. thaliana*, the ISWI/NURF of *D. melanogaster,* the NuRD/CHD of *H. sapien,* and the SWI/SNF of *S. cerevisiae*. The *C. elegans* counterparts of these complexes were found through literature research. A study that used BLAST to determine the *C. elegans* othology of each protein within each complex (*Chue et al. 2006*). TCFs found in this study were included in the predicted TCF list.

4.1.3 Identification of Histone Modification Interactors via Known Domains

There are protein domains with the ability to associate with the post-translational modification of histones discussed previously. The bromo-domain (*Zeng et al. 2002*), chromo-domain (*Cavalli et al. 1998*), and plant homeo-domain (*Aasland et al. 1995*) were chosen as TCF domains due to their specificity for binding to modified histone tails (Figs. 6,7,8). Proteins with these domains are prevalent in histone remodeling complexes, and other complexes that are involved in transcription regulation via the histone code.

```
PKRQTNQLQYLLRVVLKTLWKH------------------ 040
-------------QFAWPFQQPVDAVKLNLPDYYKIIKTP 080
MDMGTIKKRLENNYY---WNAQECIQDFNTMFTNCYIYNK 120
-----------------PGDDIVLMAEALEKLFLQKINEL 160
PT
```

Figure 6. Bromo-Domain

The domain sequence of bromo-domain was gathered from SMART. The highlighted sequences are done based on experiments of point mutations that cause loss of protein function (*Zeng et al. 2001*).

```
EYA-VEKIIDRR---------------------------- 040
--------------------------------VRKGKVEY 080
YLKWKGYPETE-NTWEPENNLD-------CQDLIQQYEAS 120
RK
```

Figure 7. Chromo-Domain

The domain sequence of chromo-domain gathered from SMART. The highlighted sequences are done based on the methyl recognition sites of the chromo-domain (*Brehm et al. 2004*).

```
FCR--VCKD------------------------------- 040
---GGELLCCD--TCP-SSYHI-HCLNPPLP--------- 080
--------------------------------------EI 120
PNGEWLCPRCT
```

Figure 8. Plant-Homeo-Domain

The domain sequence of Plant-Homeo-Domain gathered from SMART. The highlighted sequences are done based on the highly conservative cysteines residues (*Aasland et al. 1995*).

4.1.4 Domain Blast Results

A BLAST of *C.elegans* genome with the identified domain sequences (see Methods) was performed using Wormbase. The BLAST results for bromo-domain, chromo-domain, plant-homeo-domain, SET-domain and Jmjc-domain are shown in tables below. The E-value of these blast results are gathered and shown within the tables. In appendix B, the actual alignment of BLAST sequences are shown with the highlighted active sites and highly conserved amino acids.

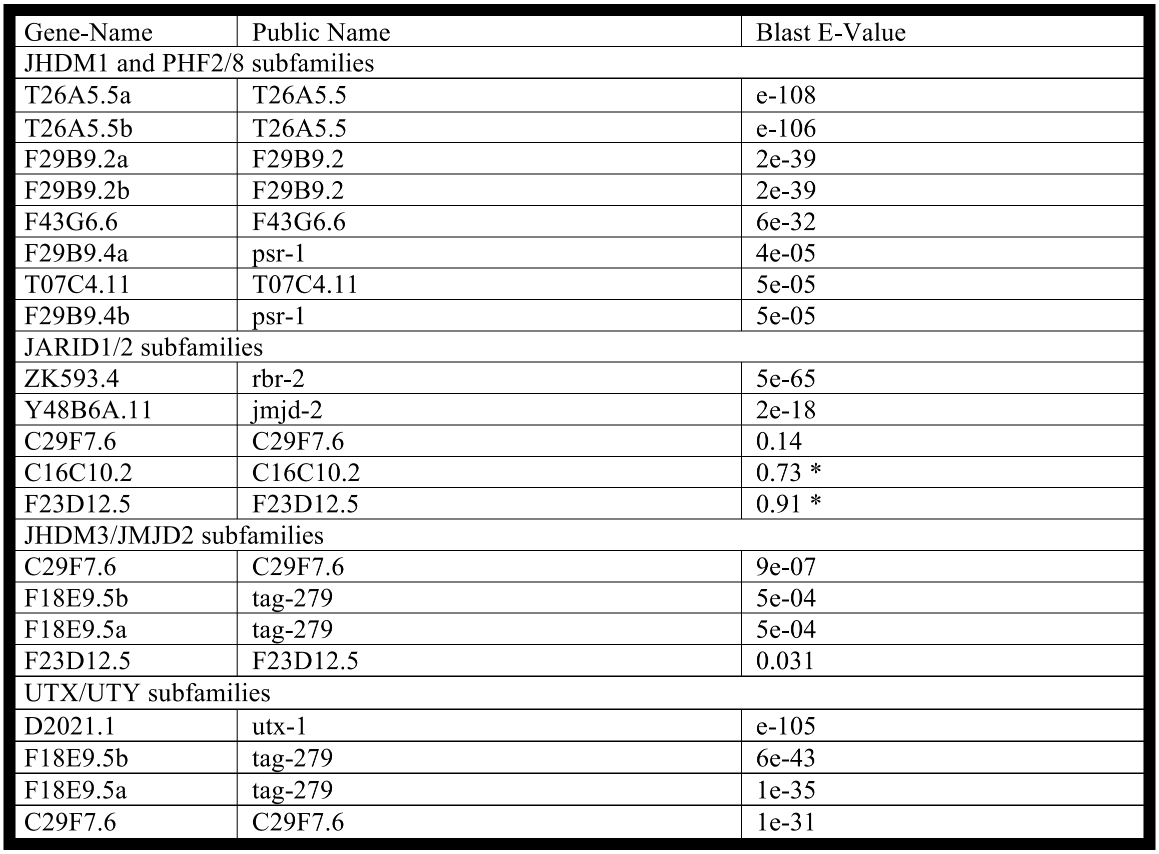
Based on the e-value of set domain BLAST result, there were very high levels of alignments for all proteins (Fig. 9). There were also high levels of alignment of the catalytic site and lysine recognition site of SET-domain (shown in Appendix B.5). Many proteins of the set gene class were in the BLAST result, and were not included within this figure.

Gene-Name	Public Name	Blast E-Value
C43E11.3a	met-1	2e-19
C43E11.3b	met-1	2e-19
Y2H9A.1	mes-4	8e-12
R05D3.11	met-2	5e-10
R06A4.7	mes-2	1e-08
T12F5.4	lin-59	6e-08

Figure 9. Set-Domain BLAST Output

The BLAST results of SET-Domain sequence excluding those proteins of the set gene class with their alignment e-value are shown.

In general, BLAST with the five domain subfamilies of the Jmjc domain identified different proteins, with some overlap (Fig.10). PHF2/8 and JHDM1 identified identical proteins, and these are listed together in the table. All genes with E-values lower than the cut-off value of e^-10 were included as predicted TCFs. Although psr-1 and T07C4.11 resulted in very high E-value from BLAST of JHDM1, the alignment has shown match for all of the important catalytic residues, and they were included as predicted TCF. Both rbr-2 and jmjd-2 were in the BLAST result of JARID1/2 and JHDM3/JMJD2 with low E-value, they were both included in the table once. The E-values of all genes other than rbr-2 and jmjd-2 resulted from the BLAST JARID1/2 and JHDM3/JMJD2 were high, the majority of them had alignment of the α-ketoglutarate targeting residue, but not the Fe (II) targeting residue (shown in Appendix B.6). These proteins were not included as TCFs. tag-279 and C29F7.6 did result with low E-value during the BLAST of UTX/UTY, and matching on each catalytic residues (shown in Appendix B.6). tag-279 and C29F7.6 were included as TCFs.



Gene-Name	Public Name	Blast E-Value
JHDM1 and PHF2/8 subfamilies		
T26A5.5a	T26A5.5	e-108
T26A5.5b	T26A5.5	e-106
F29B9.2a	F29B9.2	2e-39
F29B9.2b	F29B9.2	2e-39
F43G6.6	F43G6.6	6e-32
F29B9.4a	psr-1	4e-05
T07C4.11	T07C4.11	5e-05
F29B9.4b	psr-1	5e-05
JARID1/2 subfamilies		
ZK593.4	rbr-2	5e-65
Y48B6A.11	jmjd-2	2e-18
C29F7.6	C29F7.6	0.14
C16C10.2	C16C10.2	0.73 *
F23D12.5	F23D12.5	0.91 *
JHDM3/JMJD2 subfamilies		
C29F7.6	C29F7.6	9e-07
F18E9.5b	tag-279	5e-04
F18E9.5a	tag-279	5e-04
F23D12.5	F23D12.5	0.031
UTX/UTY subfamilies		
D2021.1	utx-1	e-105
F18E9.5b	tag-279	6e-43
F18E9.5a	tag-279	1e-35
C29F7.6	C29F7.6	1e-31

Figure 10. Jmjc-Domain BLAST Output

The BLAST results of 5 different sub families for jmjc-Domain with their alignment e-value are shown. Genes shown with an asterisk in its BLAST e-value were not included as predicted TCF.

The e-values of many bromo-domain BLAST results were high (Fig. 11). There were high levels of alignment of the sites that caused the loss of function via mutation, thus most of these proteins were included in the TCF list (shown in Appendix B.2). The EGF receptor received a very high BLAST E-value, which demonstrates the lack of alignment. In addition, only one of the 5 loss of function mutation sites is aligned for EGF receptor, and was not included as an TCF. The transcript variants of each gene received identical E-values and alignment except for the h transcript variant of nurf-1. The alignment of this variant of nurf-1 showed only 2 of the 5 loss of function mutation sites matched with the bromo-domain sequence.

Gene-Name	Public Name	Blast E-Value
F57C7.1a	Female Sterile Homeotic Protein	8e-30
F57C7.1b	Female Sterile Homeotic Protein	2e-28
Y119C1B.8a	tag-332	7e-27
Y119C1B.8b	tag-332	1e-26
F13C5.2	Bromodomain Containing Protein	8e-16
H20J04.2	H20J04.2	2e-13
R10E11.1c	cbp-1	4e-13
R10E11.1b	cbp-1	4e-13
R10E11.1a	cbp-1	4e-13
F26H11.2e	nurf-1	2e-12
F26H11.2f	nurf-1	2e-12
F26H11.2d	nurf-1	2e-12
F26H11.2g	nurf-1	2e-12
F26H11.2c	nurf-1	3e-12
Y47G6A.6	pcaf-1	2e-11
C26C6.1a	pbrm-1	4e-09
F01G4.1	psa-4	5e-08
ZK783.4	flt-1	3e-07
C01H6.7a	tag-298	9e-07
C01H6.7b	tag-298	1e-06
W04A8.7	taf-1	1e-05
F11A10.1c	lex-1	1e-04
F11A10.1b	lex-1	2e-04
F11A10.1a	lex-1	2e-04
F26H11.2h	nurf-1	0.20
C34C6.3	EGF receptor	0.40 *

Figure 11. Bromo-Domain BLAST Output

The BLAST results of Bromo-Domain with their alignment e-value are shown. Genes shown with an asterisk in its BLAST e-value were not included as predicted TCF.

The E-values of all chromo-domain BLAST results were high (Fig. 12). The high E-value resulted because chromo-domain has multiple sub-families, and the domain consensus of all sub-families gathered from SMART is different from the chromo-domain sequence found in *C. elegans*. There were high levels of congruence of the methyl recognition sites during alignment (shown in Appendix B.3), and all of the BLAST results of chromo-domain were included as TCFs.

Gene-Name	Public Name	Blast E-Value
K08H2.6	hpl-1	2e-05
ZK1236.2	cec-1	0.010
K01G5.2c	hpl-2	0.11
K01G5.2b	hpl-2	0.11
F32E10.2	Chromo-domain Containing Protein	0.24
K01G5.2a	hpl-2	0.53

Figure 12 . Chromo-Domain BLAST Output

The BLAST results of Chromo-Domain with their alignment e-value are shown.

The E-values of many plant-homeo-domain BLAST results were high (Fig. 14). Though the E-values were high, there were high levels of alignment of the highly conserved cysteine residues of Plant-Homeo-Domain (shown in Appendix B.4). All PHD BLAST results were included as predicted TCFs.

Gene-Name	Public Name	Blast E-Value
T13G8.1	chd-3	3e-17
F26F12.7	let-418	4e-12
ZK783.4	flt-1	2e-09
C44B9.4	athp-1	2e-07
T12D8.1	set-16	2e-07
ZK593.4	rbr-2	4e-07
F17A2.3	PHD-finger Protein	6e-06
Y59A8A.2	Y59A8A.2	2e-04
K09A11.5	PHD-finger Protein	2e-04
C28H8.9a	C28H8.9a	5e-04
F33E11.6b	F33E11.6b	0.003
H05L14.2	Zinc finger C3HC4 type Protein	0.010
F26H11.2i	nurf-1	0.023
F26H11.2b	nurf-1	0.023
F26H11.2a	nurf-1	0.023
F26H11.2c	nurf-1	0.027
H20J04.2	H20J04.2	0.049
F42A9.2	lin-49	0.051
C11G6.3	PHD-finger Protein	0.083
F54F2.2a	zfp-1	0.20
Y51H1A.4	ing-3	0.20

Figure 13. Plant-Homeo-Domain BLAST Output

The BLAST results of Plant-Homeo-Domain with their alignment e-value are shown.

4.1.4 Identification of PIC Recruitment TCFs via Gene Class Searches

Multiple *C.elegans* TCF complexes were determined based on previous research. These TCF complexes include mediator complexes, such as Activator Recruited Complex (ARC), Cofactor Required for SP1 (CRSP), and Thyroid Hormone Associated Proteins (TRAP) (*Rachez et al. 2001*). These mediator complexes bind TF to recruit RNA polymerase. Previously, all of the *C. elegans* proteins within these mediator complexes were renamed with the mdt prefix for their public name (*Bourbon et al. 2004*). These proteins can be gathered from Wormbase using a global search of the mdt gene class. There were 6 proteins of mdt prefix that are not included within the mdt gene class due to their previous public names. These 6 proteins were manually gathered from Wormbase. The *C.elegans* TAF complex has also been previously addressed as a TCF complex

(*Roeder 2004*). The proteins within the TAF complexes were gathered from WormBase through search of the taf gene class.

4.1.5 Identification of TCFs via Ortholog Searches

Using BLAST, many orthologs of TCFs were identified. All of the TCFs but one identified in this manner were previously identified through other methods, such as domain based search, and literature research. The one gene that was new to the TCF list was spr-5. This gene is an ortholog of lsd-1, a histone methyltransferase. spr-5 had a very high alignment with lsd-1 with an e-value of e^-147. spr-5 is very likely a paralog of lsd-1.

4.2 Evaluating TCF Predictions with Phenome Data

Due to the heavy involvement of TCFs in the gene expression regulation network, the gene silencing is predicted to produce a high lethality rate. The phenome was accessed specifically for searches of phenotype related to lethality, which are characteristics of worms that die prematurely during any stage of the life cycle. Based on a large scale RNAi experiment of *C. elegans* chromosome I 5.5% of the genes result in embryonic lethality (Kamath et al. 2003). A high percentage of embryonic lethality phenotype shown the genes is related to transcription regulation, and more likely to be TCFs. This evaluation of RNAi lethality is a low estimate of 46% embryonic lethal, because those genes without any RNAi experimental data in the phenome are considered as negative for RNAi lethality. The RNAi phenotypes gathered are shown in Appendix A.

Based on the data gathered from the phenome of each sub group of TCFs, methyltransferase have the highest distribution of the overall list, and have the lowest lethality rate. Corresponding, the histone methyl-binding domains also have a low lethality rate. The chromatin remodeling complexes all had a very high lethality rate (Fig. 14).

Categories	Count	Emb Lethal	Larval Lethal	Lethal	General Lethality	Percentage
Total	162	74 (46%)	32 (20%)	33 (20%)	86 (53%)	100%
Complex Proteins	73	45 (62%)	21 (29%)	20 (27%)	53 (73%)	45%
TAF	17	8 (47%)	1 (5.9%)	4 (24%)	9 (53%)	10%
Mediator	23	17 (74%)	6 (26%)	5 (22%)	20 (87%)	14%
SWI/SNF	10	5 (50%)	6 (60%)	5 (50%)	9 (90%)	6.2%
NuRD/CHD	3	2 (67%)	2 (67%)	2 (67%)	3 (100%)	1.9%
ISWI/NURF	4	3 (75%)	2 (50%)	2 (50%)	3 (75%)	2.5%
SWR1/SRCAP	7	5 (71%)	1 (14%)	2 (29%)	5 (71%)	4.3%
COMPASS	3	1 (33%)	1 (33%)	0 (0%)	1 (33%)	1.9%
NuA3	1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0.6%
TIP60/NuA4	4	3 (75%)	2 (50%)	0 (0%)	3 (75%)	2.5%
Histone Modification	60	19 (32%)	7 (12%)	10 (17%)	22 (37%)	37%
Acetyltransferase	6	3 (50%)	1 (17%)	1 (17%)	4 (67%)	2.5%
Deacetylase	7	3 (43%)	2 (29%)	1 (14%)	3 (43%)	4.3%
Methyltransferase	35	7 (20%)	3 (8.6%)	3 (8.6%)	9 (26%)	22%
Demethylase	11	4 (36%)	1 (9.1%)	4 (36%)	5 (45%)	6.8%
Kinase	2	2 (100%)	0 (0%)	1 (50%)	2 (100%)	1.2%
Summoylase	1	1 (100%)	0 (0%)	1 (100%)	1 (100%)	0.61%
Histone Tail Binding	32	14 (44%)	4 (13%)	4 (13%)	15 (47%)	20%
Bromo-domain	13	8 (62%)	3 (23%)	2 (15%)	9 (69%)	8.0%
Chromo-domain	4	1 (25%)	0 (0%)	0 (0%)	1 (25%)	2.5%
PHD	15	5 (33%)	1 (6.7%)	2 (13%)	5 (33%)	9.3%

Figure 14. The Predicted TCF List Statistics

The numbers and percentages of each type of TCF found in the current predicted list of TCF are shown. In addition, this table shows the number and percentage of the genes within the predicted TCF list that demonstrates the selected phenotypes of lethality using RNAi data. Emb lethal means the worm dies in the embryonic stage, larval lethal means the worm dies in the larval stage, and lethal means the worm dies prematurely not in a developmental stage. General lethality measures whether a gene have any of the three described phenotypes. Those sub families of TCFs shown in red have a lower percent of lethality rate compare to others.

4.3 Evaluating TCF Predictions with Interactome Data

A goal of this project was to determine the number of known interactions exist for the predicted TCFs. A perl based computer software was created to automatically search through multiple databases for known interactions quickly. Overall, 98 interactions were found within all the interactome databases queried with the yeast-2-hybrid data provided by the Vidal lab (Fig. 15). Only 9 of those interactions were between TCFs and TFs. Based on the theories of TCFs and TFs functionalities, many more interactions are currently unknown.

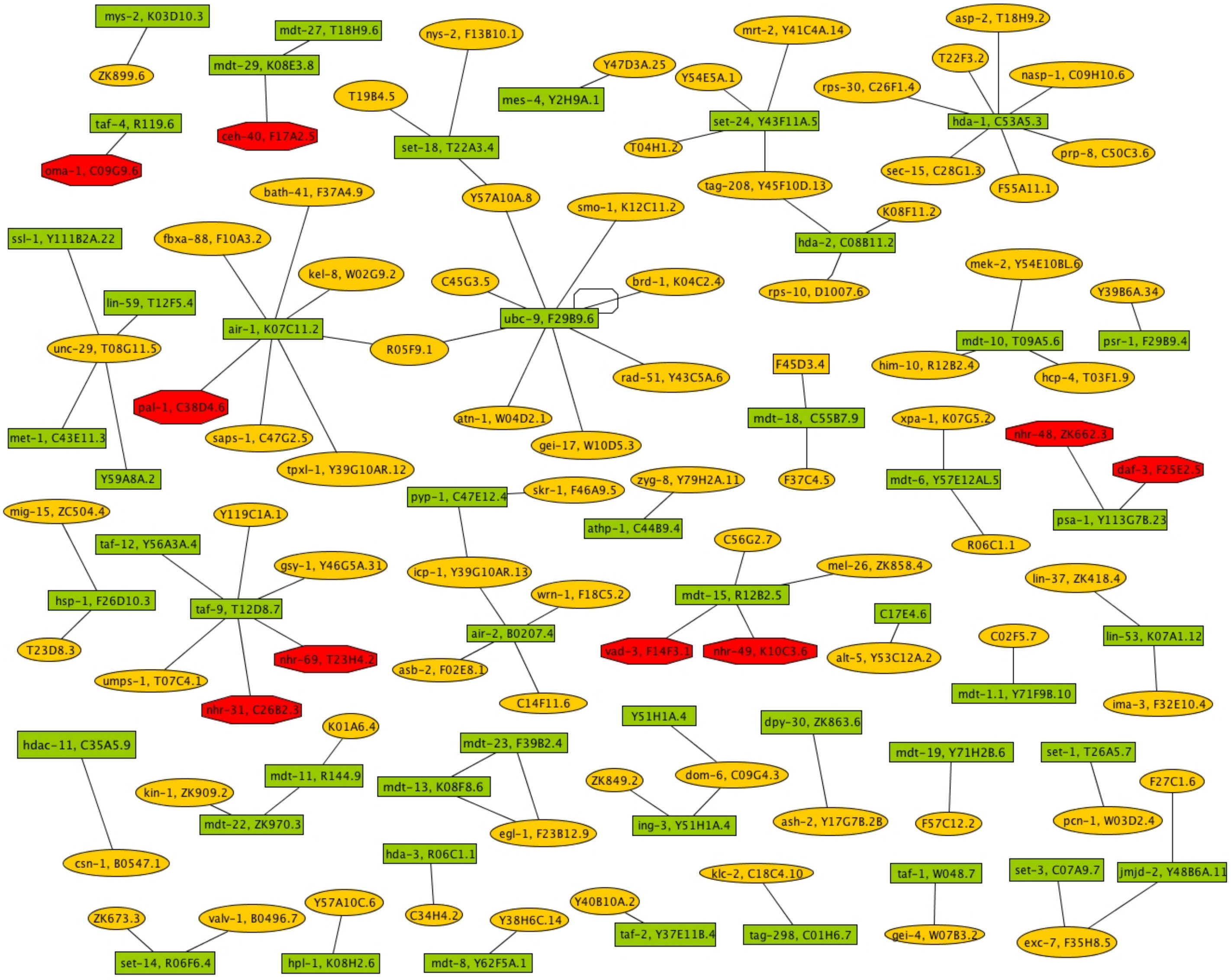


Figure 15. TCF Interactome Mapping

The known interactions of the predicted TCFs shown in green, the predicted TFs shown in red, and non-TCF-TF proteins shown in yellow.

Based on the predicted number of interactions within *C. elegans* and the current number of proteins within the proteome, there are approximately 4.8 interactions per protein (Fig. 16). Because of the functionality of TCFs, more interactions are expected, compared to other cellular proteins, which makes the 4.8 a low estimate of interactions per TCF. The data from this project has shown 98 interactions of TCFs, between the lists of 162 predicted TCFs, there are only 0.6 interactions per protein. The interactome is far from complete because 0.6 is much lower than the already low estimate of 4.8 interactions per protein.

Categories	Interactors	%	Interactions	%	Interactions per Interactors
Overall	24,202	100%	116,000	100%	4.8
Intact	2,854	11.8%	3,520	3.0%	1.2
MINT	3,678	15.2%	3,503	3.0%	1.0
Vidal Lab	2,608	10.8%	8,378	7.2%	3.2
Predicted TCFs	162	0.7%	98	<0.1%	0.6

Figure 17. Interactome Data Analysis

The number of overall protein interactors within *C.elegans* according to Sanger Institute's proteome is shown. The number of interactions is based on the predicted size of *C.elegans* interactome (*Simonis et al. 2009*). The number of interactors and interactions of Intact and MINT are determined using the interaction detection software with the available data. The number of interactors and interactions of Vidal Lab data are calculated from the provided spreadsheet. This table also shows the number of interactors and interactions within the interactome map of the predicted TCFs, and the number of TCFs that are interactors within the map.

4.3 TCF TF Overlap

The final list of predicted TCFs were compare to the predicted TF list created from a previous project. Few genes were used in both lists (Fig. 17). The predicted TF list was created using all proteins possessing a DNA binding domain, and shown to bind DNA through experimentation. TCFs are believed to be non-DNA binding gene expression regulators, so those proteins that are also within the predicted TF list have a possibility of being non-TCFs.

Gene-Name	Public Name	TF Feature	TCF Feature
F15E6.1	set-9	AT Hook	Histone Methyltransferase
C01G8.9	let-526	ARID/BRIGHT	SWI/SNF Complex
Y113G7B.23	psa-1	MYB	SWI/SNF Complex
Y71H2AM.17	Y71H2AM.17	HMG Box	SWI/SNF Complex
F37A4.8	isw-1	AT Hook, MYB	ISWI/NURF Complex
C17E4.6	C17E4.6	YL1 TF	SWR/SRCAP Complex
Y105E8A.17	ekl-4	MYB	SWR/SRCAP Complex

Figure 17. TCF TF Convergence

A group of genes that were predicted as both TCFs and TFs is shown. This figure as shows the DNA binding domain that resulted these genes to be predicted as TFs. In addition, the functional features that resulted in the TCF prediction are shown in the figure.

From looking closely at the DNA-binding domains possessed by each gene shown to be in both predicted TCF and TF lists, it was determined that they were all non-specific DNA-binding domain except for C17E4.6. C17E4.6 was gathered for the predicted TF list base on a literature that provided experimental result of DNA-binding, and no DNA specificity was specified.

5 Discussion

Through this project a list of 162 *C. elegans* proteins were predicted as TCFs. The predicted list consists of a variety of different proteins. The RNAi phenotype and distribution of each predicted TCF were analyzed. The interactome was also evaluated for possibility of future expansion. Through these analyses, many new hypotheses were made that can be tested in future research of gene expression.

5.1 Predicted TCF Distribution

To achieve a high level of confidence with the predicted TCFs, only those proteins that have previously been studied as TCF, or possess known domains that are related to TCF functions were gathered. Proteins with histone methylation function had the biggest representation within the predicted TCF list, with 35 predicted methyltransferases and 11 predicted methylases. These proteins comprise 28.4% of the predicted TCF list. Other histone modification proteins were fewer in numbers, with the most common being proteins with histone acetylation function that encompass 6.8% of the predicted TCF list. This difference may be due to the number of previous studies conducted on the different types of histone modification proteins.

The majority of methylation activity within eukaryotic cells is DNA methylation and histone methylation. Based on previous studies, there are no signs of DNA methylation within *C. elegans*, thus explaining the lack of DNA methylation proteins (Bird 2002). This finding simplified the characterization of histone methylation proteins within *C. elegans*. In contrast, other proteins that perform post-translational modification on histones (e.g. kinases, acetylases) have many homologs that have identical enzymatic functions. For example, the kinase domain of *air-1* has more than 50 proteins that have lower than e^25 alignment using Wormbase BLAST. None of those proteins other than the two Aurora kinases has literature supporting any histone modification activity. It is possible that many *air-1* homologs may modify histone proteins, but not enough experimental data are available to know; thus, none of these proteins except the Aurora kinases were included in the predicted TCF list. This complication in the histone modification protein identification process may have created the difference between the numbers of each type of histone modification proteins gathered within the TCF list.

Although certain difficulty exists during the identification of specific histone modification proteins, there are still hypotheses that can be made based on the findings within this project. There is a potential that histone methylation is the primary method of transcription regulation via histone code within *C. elegans*. However, equal numbers of proteins were found that binds histone tails modified by methylation and acetylation, which does not support the hypothesis. Future experiments can be done on all the *C. elegans* summoylase, adp-ribosylase, kinase, and acetylase to determine histone activity.

5.2 Predicted TCF RNAi Lethality Rate

There were many gene-silencing experiments performed previously on the predicted TCF proteins, and they are annotated within online databases. The RNAi data from these experiments were used for the analysis of the predicted TCFs. RNAi data comparison of each predicted TCF group showed that histone methylation proteins have a lower percentage of lethality compare to other types of histone modification. The RNAi data of the histone tail-binding domain matches the data of histone modification proteins. Histone tail lysine-acetylation binding bromo-domain proteins have a higher lethality phenotype percentage than lysine-methylation binding proteins of chromo-domain and PHD. These data suggest that methylation may regulate less essential pathways.

Overall, the RNAi phenotypes of all 162 genes showed that 46% of them are embryonic lethal and 53% of them have showed some form of lethality. Comparison to the result of 5% embryonic lethality from the genome wide RNAi of chromosome I (*Kamath et al. 2003*) shows this group of 162 genes have distinct characteristics, and are not a random selection. TCFs are believed to be centralized in the regulation net work, and their silencing may result in the silencing of the production of many other proteins. The high lethality rate of RNAi provides evidence that the predicted TCFs have regulatory functions on general transcription, and the expression of other proteins.

5.3 TCF TF Convergence

The comparison between the predicted TCFs and TFs has shown some duplication between the two lists. There are in total 7 proteins that exist in both lists, and each of these 7 proteins possesses DNA binding domains that caused them to be predicted as TFs. The domains of all 7 proteins are non-specific DNA binding domains, such as high mobility group, MYB. 6 out of the 7 proteins are a part of the predicted TCF complexes. It is possible that the non-specific DNA binding domains are utilized for the localization of TCF complexes. Experiments can be done on those predicted TCF proteins with DNA binding domain. Point mutation can be performed on the DNA binding domain to observe the outcome. If the proteins are non-active to perform their function due to the site mutation, then the non-specific DNA binding domain can be used as a method to determine more TCFs.

5.4 Interactome Evaluation

TCFs are a vital component of eukaryotic transcription regulation. There are 24,202 proteins currently recorded in the *C. elegans* proteome. Approximately 900 to 1500 of those proteins are predicted as TFs. Using the current interactome data, only 98 interactions are found with the predicted TCFs, and only 9 interactions with the predicted TFs. TCFs are theoretically predicted to be highly interactive with TFs.

The predicted completed interactome using the current proteome count and full interaction prediction estimates 4.8 interactions per protein. Using the current interactome data, only 0.6 interactions per protein were found for the highly interactive TCFs. These data illustrate that the current interactome is far from complete. The determination of novel interaction for the interactome will further our understanding of the transcription regulation. A high throughput yeast 2 hybrid screen of all predicted TCFs and TFs will likely yield a large number of novel interactions.

References

- Aasland R., Gibson T. J., and Steward F. A., "The PHD finger: implication for chromatin-mediated transcriptional regulation". *Elsevier Science* 4.965 (1995): 56-59.
- Allfrey V. G., Faulkner R., and Mirsky A. E., "Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis." *PNAS* 51.5 (1964): 786-94.
- Almer A., Hans R., Albert H., Wolfram H., “Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements”. EMBO J 5 (1986): 2689–96.
- Babu M. M., Luscombe N. M., Aravind L., Gerstein M., Teichmann S. A., "Structure and evolution of transcriptional regulatory networks". *Curr. Opin. Struct. Biol.* **14**.3 (2004): 283–91.
- Blackwood E. M., and Kadonaga J. T., "Going the distance: a current view of enhancer action". Science 281 (1998): 60-3.
- Brehm A., Tufteland K. R., Aasland R., and Becker P. B., “The many colours of chromodomains”. BioEssays 26 (2004):133-40.
- Cavalli G., and Paro R., "Chromo-domain proteins: linking chromatin structure to epigenetic regulation". *Current Opinion in Cell Biology* 10.3 (1998): 554-60.
- Crick F. "Split genes and RNA splicing". Science 204.20 (1979): 263-71. Dignam J. D., Lebovitz R. M., and Roeder R. G., "Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei". Nucleic Acids Research 11.5 (1983): 1475-89.
- Evans R. M., "The steroid and thyroid hormone receptor superfamily". *Science* 240.4854 (1988): 889-95.
- Fire A., Xu S., Montgomery M. K., “RNA as a target of double-stranded RNA-mediated genetic interference in Caenorhabditis elegans”. PNAS 95 (1998) 15501-7.
- Fraser A. G., Kamath R. S., Zipperlen P., Martinez-Campos M., Sohrmann M., and Ahringer J., “Functional genomic analysis of C. elegans chromosome I by systematic RNA interference”. Nature. 408 (2000): 325-30.
- Geyer P. K., Green M. M., and Corces V. G., "Tissue-specific transcriptional enhancer may act in trans on the gene located in the homologous chromosome: the molecular basis of tranvection in drosophila". EMBO 9.7 (1990): 2247-56.
- Giot L., Bader J. S., Brouwer C., Chaudhuri A., Kuang B., Li Y., Hao Y. L., Ooi C. E., Godwin B., Vitols E., Vijayadamodar G., Pochart P., Machineni H., Welsh M., Kong Y., Zerhusen B., Malcolm R., Varrone Z., Collis A., Minto M., Burgess S., McDaniel L., Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli E., Aanensen N., Carrolla S., Bickelhaupt E., Lazovatsky Y., DaSilva A., Zhong, C. A. Stanyon, R. L. Finley, Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach J., Knight J., Shimkets R. A., McKenna M. P., Chant J., and Rothberg J. M., “A protein interaction map of drosophila melanogaster”. Science 302.5651 (2003) 1727-36.

- Helin K., Wu C., Fattaey A. R., Lees J. A., Dynlacht B. D., Ngwu C., and Harlow E., "Heterodimerization of the transcription factors E2F-1 and DP-1 leads to cooperative trans-activation". *Genes Dev.* 7 (1993): 1850-61.
- Ito T., Tashiro K., Muta S., Ozawa R., Chiba T., Nishizawa M., Yamamoto K., Kuhara S., Sakaki Y., “Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins”. PNAS 97.3 (2000): 1143-7.
- Kamath R. S., Fraser A. G., Dong Y., Poulin G., Durbin R., Gotta M., Kanapin A. Le Bot N., Moreno S., Sohrmann M., Welchman D. P., Zipperlen P., and Ahringer J., “Systematic functional analysis of the Caenorhabditis elegans genome using RNAi”. Nature 421 (2002): 231-7.
- Kugel J. F., and Goodrich J. A., "Promoter escape limits the rate of RNA polymerase II transcription and is enhanced by TFIIE, TFIIH, and ATP on negatively supercoiled DNA". *PNAS* 95.16 (1998): 9232-37.
- Lackner M. R., Kornfeld K., Miller L. M., Horvitz R. H., and Kim S. K., "A MAP Kinase homolog, mpk-1, is involved in ras-mediated induction of vulval cell fates in Caenorhabditis elegans". *Genes Dev.* 8 (1994): 160-73.
- Laybourn P. J., and Kadonaga J. T., "Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II". Science 254 (1991): 238-45.
- Lee, T. I., and Young R. A., “Transcription of eukaryotic protein-coding genes.” *Annual Review of Genetics* 34 (2000): 77-137.
- Levine M., and Tjian R., "Transcription regulation and animal diversity". *Natue* 424 (2003): 147-52.
- Li S., Armstrong C. M., Bertin N., Ge H., Milstein S., Boxem M., Vidalain P., Han J. J., Chesneau A., Hao T., Goldberg D. S., Li N., Martinez M., Rual J., Lamesch P., Xu L., Tewari M., Wong S. L., Zhang L. V., Berriz G. F., Jacotot L., Vaglio P., Reboul J., Hirozane-Kishikawa T., Li Q., Gabel H. W., Elewa A., Baumgartner B., Rose D. J., Yu H., Bosak S., Sequerra R., Fraser A., Mango S. E., Saxton W. M., Strome S., van den Heuvel S., Piano F., Vandenhaute J., Sardet C., Gerstein M., Doucette-Stamm L., Gunsalus K. C., Harper J. W., Cusick M. E., Roth F. P., Hill D. E., and Vidal M., “A map of the interactome network of the metazoan C. elegans”. Science 303.5657 (2004) 540-3.
- Mann M., Hojrup P., and Roepstorff P., “Use of mass spectrometric molecular weight information to identify proteins in sequence databases”. Biological Mass Spectrometry 22.6 (1993): 338-45.
- Mitchell P. J., and Tjian R., "Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins". *Science* 246.4916 (1989): 371-9.
- Mizzen C. A., Yang X., Kokubo T., Brownell J. E., Bannister A. J., Owen-Hughes T., Workman J., Wang L., Berger S. L., Kouzarides T., Nakatani Y., and Allis C. D., "The TAFII250 Subunit of TFIID Has Histone Acetyltransferase Activity". *Cell* 87.7 (1996): 1261-70.
- Nakajima N., Horikoshi M., and Roeder R. G., "Factors involved in specific transcription by mammalian RNA polymerase II: purification, genetic specificity, and TATA box-promoter interaction of TFIID". Molecular and Cellular Biology 8.10 (1988): 4028-40.

- Noble D., "*The Music of Life Biology beyond the Genome"*. New York: Oxford UP, USA, (2006): 21.
- Okuda M., Tanaka A., Araill Y., Satoh M., Okamura H., Nagadoi A., Hanaoka F., Ohkuma Y., and Nishimura Y., "A Novel Zinc Finger Structure in the Large Subunit of Human General Transcription Factor TFIIE". *J.Biol.Chem* 279.49 (2004): 51395-403.
- Ozer J., Moore P. A., Bolden A. H., Lee A., Rosen C. A., and Lieberman P. M., "Molecular cloning of the small (y) subunit of human TFIIA reveals functions critical for activated transcription". *Genes Dev.* 8 (1994): 2324-35.
- Pugh F. B., and Tjian R., "Mechanism of transcriptional activation by sp1 evidence for coactivators". *Cell* 61 (1990): 1197-207.
- Rachez C., and Freeman L. P., "Mediator complexes and transcription". *Current Opinion in Cell Biology* 13 (2001): 274-80.
- Raff J. W., Kellum R., and Alberts B., "The Drosophila GAGA transcription factor is associated with specific regions of heterochromatin throughout the cell cycle". *EMBO* 13.24 (1994): 5977-83.
- Reboul J., Vaglio P., Rual J., Lamesch P., Martinez M., Armstrong C. M., Li S., Jacotot L., Bertin N., Janky R., Troy M., Hudson J. R., Hartley J. L., Brasch M. A., Vandenhaute J., Boulton S., Endress G. A., Jenna S., Chevet E., Papasotiropoulos V., Tolias P. P., Ptacek J., Snyder M., Huang R., Chance M. R., Lee H., Doucette-Stamm L., Hill D. E., and Vidal M., "C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression". Nature genetics 34 (2003): 35-41
- Reece-Hoyes J., Deplancke B., Shingles J., Grove C. A., Hope I. A., and Walhout A. J. M., "A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks". Genome Biology 6 (2005): R110.
- Robert F., Douzlech M., Forget D., Egly J., Greenblatt J., Burton Z. F., and Coulombe B., "Wrapping of Promoter DNA around the RNA Polymerase II Initiation Complex Induced by TFIIF". *Molecular Cell* 2.3 (1998): 341-51.
- Roeder R. G., "The role of general initiation factors in transcription by RNA polymerase II". TIBS 21 (1996): 327-35.
- Roeder R. G., "Transcriptional regulation and the role of diverse coactivators in animal cells". *FEBS* 579 (2004): 909-15.
- Rosenberg M., and Court D., "Regulatory sequences involved in the promotion and termination of RNA transcription". Annual Review of Genetics 13 (1979): 319-53.
- Rowland T., Baumann P., and Jackson S. P., "The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria". Science 264.5163 (1994): 1926-329.
- Rual J., Ceron J., Koreth J., Hao T., Nicot A., Hirozane-Kishikawa T., Vandenhaute J., Orkin S. H., Hill D. E., van den Heuvel S., and Vidal M., "Toward Improving Caenorhabditis elegans phenome mapping with an ORFeome based RNAi library". Genome Res. 14 (2004): 2162-8.
- Rual J., Venkatesan K., Hao T., Hirozane-Kishikawa T., Dricot A., Li N., Berriz G. F., Gibbons F. D., Dreze M., Ayivi-Guedehoussou N., Klitgord N., Simon C.,

Boxem M., Milstein S., Rosenberg J., Goldberg D. S., Zhang L. V., Wong S. L., Franklin G., Li S., Albala J. S., Lim J., Fraughton C., Liamosas E., Cevik S., Bex C., Lamesch P., Sikorski R. S., Vandenhaute J., Zoghbi H. Y., Smolyar A., Bosak S., Sequerra R., Doucette-Stamm L., Cusick M. E., Hill D. E., Roth F. P., and Vidal M., "Towards a proteome-scale map of the protein-protein interaction network". Nature 437 (2005): 1173-8.

- Shin H., Hirst M., Bainbridge M. N., Magrini V., Mardis E., Moerman D. G., Marra M. A., Baillie D. L., and Jones S. J. M., "Transcriptome analysis for Caenorhabditis elegans based on novel expressed sequence tags". BMC Biology 6.30 (2008).
- Stumpf, M. P. H., Thorne T., Silva E. D., Stewart R., An H. J., Lappe M., and Wiuf C., "Estimating the size of the human interactome". PNAS 105.19 (2008): 6959-64.
- Sudarsanam P., Cao Y., Wu L., Laurent B. C., and Winston F., "The nucleosome remodeling complex, Snf/Swi, is required for the maintenance of transcription in vivo and is partially redundant with the histone acetyltransferase, Gcn5". *Embo* 18.11 (1999): 3101-06.
- Tan S., Conaway R. C., and Conaway J. W., "Dissection of transcription factor TFIIF functional domains required for initiation and elongation". *Biochemistry* 92 (1995): 6042-46.
- Tini M., Benecke A., Um S., Torchia J., Evans R. M., and Chambon P., "Association of CBP/p300 Acetylase and Thymine DNA Glycosylase Links DNA Repair and Transcription". *Molecular Cell* 9.2 (2002): 265-77.
- University of Manitoba, "BIRCH, Biology Research Computer Hieracy", http://home.cc.umanitoba.ca/~psgendb/, 01/11/2010.
- Verrijzer P. C., Chen J., Yokomori K., and Tjian R., "Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II". *Cell* 81.7 (1995): 1115-25.
- Zeng L., and Zhou M., "Bromodomain: an acetyl-lysine binding domain". *FEBS* 513 (2002): 124-8.
- Zhang H., Smolen G. A., Palmer R., Christoforou A., Van den Heuvel S., and Haber D. A., "SUMO modification is required for in vivo Hox gene regulation by the caenorhabditis elegans polycomb group protein sop-2". Nature Genetics 36.5 (2004) 507-11.
- Zheng M., Aslund F., and Storz G., "Activation of the OxyR transcription factor by reversible disulfide bond formation". *Science* 279.5357 (1998): 1718-722.

Appendix A Predicted TCFs and Phenome Data

Predicted TCFs with any form of lethality phenotype due to RNAi are shown in red.

			embryonic lethal	larval lethal	lethal
F57C7.1	Female Sterile Homeotic Protein	bromo-domain			
Y119C1B.8	tag-332	bromo-domain	√		
F13C5.2	Bromodomain Containing Protein	bromo-domain	√	√	
H20J04.2	H20J04.2	bromo-domain			
R10E11.1	cbp-1	bromo-domain, Histone Acetyltransferase	√		
F26H11.2	nurf-1	bromo-domain, PHD, ISWI/NURF Complex			
Y47G6A.6	pcaf-1	bromo-domain, Histone Acetyltransferase			√
C26C6.1	pbrm-1	bromo-domain	√	√	
F01G4.1	psa-4	bromo-domain, SWI/SNF Complex	√		
ZK783.4	flt-1	bromo-domain			
C01H6.7	tag-298	bromo-domain	√	√	
W04A8.7	taf-1	bromo-domain, TAF	√		√
F11A10.1	lex-1	bromo-domain	√		
K08H2.6	hpl-1	chromo-domain			
ZK1236.2	cec-1	chromo-domain			
K01G5.2	hpl-2	chromo-domain			
F32E10.2	Chromo-domain Containing Protein	chromo-domain	√		
T13G8.1	chd-3	PHD	√		
F26F12.7	let-418	PHD, NuRD/CHD Complex	√		
C44B9.4	athp-1	PHD			
T12D8.1	set-16	PHD, Histone Methyltransferase	√	√	
ZK593.4	rbr-2	PHD, Histone Demethylase			
F17A2.3	PHD-finger Protein	PHD			
Y59A8A.2	Y59A8A.2	PHD	√		√
K09A11.5	PHD-finger Protein	PHD			
C28H8.9	C28H8.9a	PHD			
F33E11.6	F33E11.6b	PHD			

H05L14.2	Zince finger C3HC4 type Protein	PHD			
F42A9.2	lin-49	PHD			
C11G6.3	PHD-finger Protein	PHD			
F54F2.2	zfp-1	PHD	√		√
Y51H1A.4	ing-3	PHD			
Y37E11B.4	taf-2	TAF	√		
C11G6.1	taf-3	TAF			
R119.6	taf-4	TAF	√	√	
F30F8.8	taf-5	TAF	√		√
W09B6.2	taf-6.1	TAF			
Y37E11AL.8	taf-6.2	TAF	√		
F54F7.1	taf-7.1	TAF			
Y111B2A.16	taf-7.2	TAF			√
ZK1320.12	taf-8	TAF			
T12D8.7	taf-9	TAF	√		
K03B4.3	taf-10	TAF	√		
F48D6.1	taf-11.1	TAF			
K10D3.3	taf-11.2	TAF			
F43D9.5	taf-11.3	TAF	√		
Y56A3A.4	taf-12	TAF			
C14A4.10	taf-13	TAF			
T23C6.1	mdt-1.2	Mediator			
ZK546.13	mdt-4	Mediator	√		
Y57E12AL.5	mdt-6	Mediator	√		
Y62F5A.1	mdt-8	Mediator	√		
T09A5.6	mdt-10	Mediator	√	√	√
R144.9	mdt-11	Mediator	√	√	√
R12B2.5	mdt-15	Mediator	√		
Y113G7B.18	mdt-17	Mediator	√	√	√
C55B7.9	mdt-18	Mediator	√		
Y71H2B.6	mdt-19	Mediator	√		√
Y104H12D.1	mdt-20	Mediator			
C24H11.9	mdt-21	Mediator			
ZK970.3	mdt-22	Mediator	√		
T18H9.6	mdt-27	Mediator	√		
W01A8.1	mdt-28	Mediator			
K08E3.8	mdt-29	Mediator			√
F32H2.2	mdt-31	Mediator	√		
Y71F9B.10	sop-3, mdt-1.1	Mediator		√	
Y54E5B.3	let-49. mdt-7	Mediator	√		
F47A4.2	dpy-22, mdt-12	Mediator	√	√	

K08F8.6	let-19, mdt-13	Mediator	√		
C38C10.5	rgr-1, mdt-14	Mediator	√	√	
F39B2.4	sur-2, mdt-23	Mediator	√		
VC5.4	mys-1	Histone Acetyltransferase, TIP60/NuA4 Complex	√	√	
K03D10.3	mys-2	Histone Acetyltransferase	√		
R07B5.8	mys-3	Histone Acetyltransferase			
C34B7.4	mys-4	Histone Acetyltransferase			
C53A5.3	hda-1	Histone Deacetylase	√		
C08B11.2	hda-2	Histone Deacetylase	√	√	
Y51H1A.5	hda-3	Histone Deacetylase	√	√	√
C10E2.3	hda-4	Histone Deacetylase			
R06C1.1	hda-5	Histone Deacetylase			
F41H10.6	hdac-6	Histone Deacetylase			
C35A5.9	hdac-11	Histone Deacetylase			
T26A5.7	set-1	Histone Methyltransferase	√	√	√
C26E6.9	set-2	Histone Methyltransferase			√
C07A9.7	set-3	Histone Methyltransferase			
C32D5.5	set-4	Histone Methyltransferase	√		√
C47E8.8	set-5	Histone Methyltransferase		√	
C49F5.2	set-6	Histone Methyltransferase			
F02D10.7	set-8	Histone Methyltransferase			
F15E6.1	set-9	Histone Methyltransferase	√		
F33H2.7	set-10	Histone Methyltransferase			
F34D6.4	set-11	Histone Methyltransferase			
K09F5.5	set-12	Histone Methyltransferase			
K12H6.11	set-13	Histone Methyltransferase			
R06F6.4	set-14	Histone Methyltransferase	√	√	

R11E3.4	set-15	Histone Methyltransferase			
T21B10.5	set-17	Histone Methyltransferase			
T22A3.4	set-18	Histone Methyltransferase			
W01C8.3	set-19	Histone Methyltransferase			
W01C8.4	set-20	Histone Methyltransferase			
Y24D9A.2	set-21	Histone Methyltransferase	√		
Y32F6A.1	set-22	Histone Methyltransferase			
Y41D4B.12	set-23	Histone Methyltransferase			
Y43F11A.5	set-24	Histone Methyltransferase			
Y43F4B.3	set-25	Histone Methyltransferase	√		
Y51H4A.12	set-26	Histone Methyltransferase			
Y71H2AM.8	set-27	Histone Methyltransferase			
Y73B3B.2	set-28	Histone Methyltransferase			
Y92H12BR.6	set-29	Histone Methyltransferase			
ZC8.3	set-30	Histone Methyltransferase			
C15H11.5	set-31	Histone Methyltransferase			
C41G7.4	set-32	Histone Methyltransferase			
Y108F1.3	set-33	Histone Methyltransferase	√		
K07C11.2	air-1	Histone Kinase	√		√
B0207.4	air-2	Histone Kinase	√		
F29B9.6	ubc-9	Histone Summoylase	√		√
T08D10.2	lsd-1	Histone Demethylase	√	√	√
Y40B1B.6	spr-5	Histone Demethylase			
T26A5.5	T26A5.5	Histone Demethylase			√
F29B9.2	F29B9.2	Histone Demethylase	√		
F43G6.6	F43G6.6	Histone Demethylase			
F29B9.4	psr-1	Histone Demethylase			

T07C4.11	T07C4.11	Histone Demethylase	√		√
Y48B6A.11	jmjd-2	Histone Demethylase			
D2021.1	utx-1	Histone Demethylase	√		√
F18E9.5	tag-279	Histone Demethylase			
C29F7.6	C29F7.6	Histone Demethylase			
Y2H9A.1	mes-4	SET-domain			√
R06A4.7	mes-2	SET-domain			
C43E11.3	met-1	SET-domain			
R05D3.11	met-2	SET-domain			√
T12F5.4	lin-59	SET-domain		√	√
C18E3.2	C18E3.2	SWI/SNF Complex	√		√
F26D10.3	hsp-1	SWI/SNF Complex	√		√
C01G8.9	let-526	SWI/SNF Complex	√	√	
Y113G7B.23	psa-1	SWI/SNF Complex		√	√
R07E5.3	snfc-5	SWI/SNF Complex		√	
Y111B2A.22	ssl-1	SWI/SNF Complex		√	√
B0041.7	xnp-1	SWI/SNF Complex			
Y71H2AM.17	Y71H2AM.17	SWI/SNF Complex	√	√	
ZK1128.5	tag-246	SWI/SNF Complex		√	√
ZK616.4	ZK616.4	SWI/SNF Complex	√		
K07A1.12	lin-53	NuRD/CHD Complex	√	√	√
M04G2.1	mep-1	NuRD/CHD Complex		√	√
F37A4.8	isw-1	ISWI/NURF Complex	√	√	√
C47E12.4	pyp-1	ISWI/NURF Complex	√	√	
K07A1.11	rba-1	ISWI/NURF Complex	√		√
C08B11.6	C08B11.6	SWR1/SRCAP Complex	√		
C17E4.6	C17E4.6	SWR1/SRCAP Complex	√		
CD4.7	CD4.7	SWR1/SRCAP Complex			
M04B2.3	gfl-1	SWR1/SRCAP	√		

		Complex			
Y37D8A.9	mrg-1	SWR1/SRCAP Complex			
R08C7.3	htz-1	SWR1/SRCAP Complex	√	√	√
Y105E8A.17	ekl-4	SWR1/SRCAP Complex	√		√
C14B1.4	swd-3.1	COMPASS Complex			
ZK863.6	dpy-30	COMPASS Complex	√	√	
C46A5.9	hcf-1	COMPASS Complex			
Y53G8AR.2	Y53G8AR.2	NuA3 Complex			
Y111B2A.11	epc-1	TIP60/NuA4 Complex	√		
C47D12.1	trr-1	TIP60/NuA4 Complex	√	√	
ZK1127.3	ZK1127.3	TIP60/NuA4 Complex			

Appendix B Blast Result Alignments

Appendix B.3 Lsd-1 Spr-5 Alignment

```
>Y40B1B.6 CE20240 WBGene00005010 locus:spr-
           5#status:Confirmed#UniProt:Q9XWP6#protein_id:CAA21604.1
          Length = 770

 Score =  520 bits (1339), Expect = e-147,   Method: Composition-based stats.
 Identities = 281/650 (43%), Positives = 404/650 (62%), Gaps = 12/650 (1%)

Query: 82  DRPTEIEAAFFPEVQMSRSFSDVFLMIRNTTLSIWLASATTECTAEDVIKHLTPPYNTEI 141
           DRPT+ E AFFPE+   ++  +VFL++RN+TL+ W  +   ECTA DV  ++ PP+N+++
Sbjct: 40  DRPTDHELAFFPELWEHKTAVEVFLLLRNSTLATWQYNPLKECTALDVRNNVFPPFNSDL 99

Query: 142 HLVQNIVLFLSRFGMINIGFFFPKTELVNNM--EKKFXXXXXXXXXXXXXXXTQLLTFGFD 199
            L+QNIV +LSR G+IN G +   T++   +  +++               TQL +FGFD
Sbjct: 100 DLIQNIVHYLSRHGLINFGRYVRSTKISRFLVRDRRSVIVIGAGAAGISAATQLESFGFD 159

Query: 200 VAVVEASGLTGGRVRSLISKHGELIETGCDSLRNLDESVITTLLHQVPLNENIMSENTIV 259
           V V+EA    GGR+ S  SK GE++ETG D+LR +++S + TLLHQV   E+ + + T V
Sbjct: 160 VIVLEARNCIGGRIHSFKSKSGEIMETGGDTLRKIEDSPMATLLHQVNFEEHGVFDFTSV 219

Query: 260 FSKGKYVPVARCHVINGLYANLKAGLAHASHGPEQRGENGLYISRQQAYENYFNMIERST 319
           F +G+ +   + H+    Y +    L + +H  E R + G +ISRQQAYEN  +M ER T
Sbjct: 220 FVEGRPLNEEKIHLFLDHYKSAHGALNYQAHQCEHRDDQGSFISRQQAYENLLSMCERGT 279

Query: 320 LLSYYNFAKEKVNLNAERKHLYEVLKTNRLTALLAEQKLKNTPP-----SDELLLKSLQI 374
           L+ YYNF K    +   R+H +  +K  R+TAL+AE +LK          D +L +SL+
Sbjct: 280 LIKYYNFCKSLETVARAREHHFNQMKQLRMTALMAENQLKKMEEEGNLEQDPVLRRSLKR 339

Query: 375 DIEKAIRQFDEACERFEICEERIADLEKNPRCKQSMHP-NDFIHYNFLLGFEERLFGAQL 433
           DI  ++ +F+E  + FE  +     L ++P+ KQ MHP ++F  +NF+LGFEE L GAQL
Sbjct: 340 DIATSLEKFEEVADAFETADNHWQRLNEHPQAKQYMHPGSEFATFNFMLGFEEYLVGAQL 399

Query: 434 EKVQFSCNVNELKLKSQVARVQEGLAQVLINVANERKVKIHHNQRVIEIDTGSSDAVILK 493
           EKVQFSC+  + K     AR+ EG+A++L  ++ +RK+ I    RV++ID    + V+LK
Sbjct: 400 EKVQFSCDSMQNKENGVAARLTEGIAELLTQLSEKRKLDIRLKHRVLDIDYSGFEHVLLK 459

Query: 494 LRKPDGSVGILNADYVVSTLPIGVLKKTIIGDERAPVFRPPLPKSKFAAIRSLGNGLINK 553
           +++ +G +  + A +VVSTLPIGVLKKTII DERAP F P LP  K  AIR++G G +NK
Sbjct: 460 VQRENGDIEEMKAAFVVSTLPIGVLKKTIIADERAPTFTPSLPDKKVEAIRNIGCGSVNK 519

Query: 554 IVFVFETRFWPES--INQFAIVPDKISERAAMFTWSSLPESRTLTTHYVGENRFHDTPVT 611
            +  F+  FW  +   NQF  V   I  R +M  WSS+P S+ L T+ VGE    + P
Sbjct: 520 CILEFDRVFWTANGGRNQFVTVSPNIKTRGSMNIWSSVPGSKVLCTYIVGEEAMLELPDD 579

Query: 612 ELITKALEMLKTVF-KDCP-SPIDAYVTNWHTDELAFGTGTFMSLRTEPQHFDALKEPLK 669
            +I  A+  L+  F  +CP +PI A++T WH DELAFG+G FMSLRTE   FD + EPLK
Sbjct: 580 VIIQNAMINLQKAFGNNCPRAPISAHITRWHDDELAFGSGAFMSLRTETTSFDDVMEPLK 639

Query: 670 TRDGKPRVFFAGEHTSALEHGTLDGAFNSGLRAAADLANTCIEIPFINRS 719
           T DG  RV+FAGEHT +    T+ GA+ SG RAAAD++N  I I F++ S
Sbjct: 640 TSDGMSRVYFAGEHTCSSYTSTIQGAWMSGARAAADISNDHIGIGFVDIS 689
```

Appendix B.2 Bromo-domain BLAST Alignment

```
>F57C7.1a CE31548 WBGene00010199 female sterile homeotic protein
           (Bromodomain
           protein)#status:Partially_confirmed#UniProt:Q20947#prote
           in_id:CAA93473.3
          Length = 1209
```

```
 Score =  124 bits (312), Expect = 8e-30,   Method: Composition-based stats.
 Identities = 55/110 (50%), Positives = 77/110 (70%)

Query: 1   PKRQTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLE 60
            P R TN L ++L  V+K   KH+ +WPFQ PVDA+KL +P+Y+ I+ TPMD+ TI+KRL
Sbjct: 280 PTRHTNCLDFVLFTVVKDALKHKHSWPFQLPVDAIKLEIPEYHNIVNTPMDLRTIEKRLR 339

Query: 61  NNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKINELP 110
            N YYW A++ I+D N +F NCY +N P  D+  MA+ LEK  L ++ +LP
Sbjct: 340 NLYYWCAEDAIKDINQVFINCYSFNPPEYDVYKMAKTLEKQVLSQLTQLP 389

 Score = 62.8 bits (151), Expect = 3e-11,   Method: Composition-based stats.
 Identities = 34/87 (39%), Positives = 44/87 (50%)

Query: 24  FAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYI 83
            FA  F  PVD +KL + DY ++I  PMD+ TIKK+L+   Y   +E + D N M  NC
Sbjct: 575 FAQVFYLPVDPIKLKIYDYLEVITNPMDLQTIKKKLDFKQYAEPEEFVHDINLMVDNCCK 634

Query: 84  YNKPGDDIVLMAEALEKLFLQKINELP 110
            YN  G      A  L   F Q+    P
Sbjct: 635 YNPKGSPAHSNALELRSFFEQRWKLFP 661


>F57C7.1b CE18761 WBGene00010199 female sterile homeotic protein
            (Bromodomain
            protein)#status:Partially_confirmed#UniProt:Q20948#prote
            in_id:CAA93475.1
          Length = 1087


 Score =  119 bits (299), Expect = 2e-28,   Method: Composition-based stats.
 Identities = 52/111 (46%), Positives = 76/111 (68%)

Query: 1   PKRQTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLE 60
            P R TN L ++L  V+K   KH+ +WPFQ PVDA+KL +P+Y+ I+ TPMD+ TI+KRL
Sbjct: 280 PTRHTNCLDFVLFTVVKDALKHKHSWPFQLPVDAIKLEIPEYHNIVNTPMDLRTIEKRLR 339

Query: 61  NNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKINELPT 111
            N YYW A++ I+D NT+F NC  +N   DDI +M E +E +  + +  +P+
Sbjct: 340 NLYYWCAEDAIKDLNTLFDNCKKFNDRNDDIYIMCENIEGVVQRGLEWMPS 390


 Score = 63.5 bits (153), Expect = 2e-11,   Method: Composition-based stats.
 Identities = 34/87 (39%), Positives = 44/87 (50%)

Query: 24  FAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYI 83
            FA  F  PVD +KL + DY ++I  PMD+ TIKK+L+   Y   +E + D N M  NC
Sbjct: 575 FAQVFYLPVDPIKLKIYDYLEVITNPMDLQTIKKKLDFKQYAEPEEFVHDINLMVDNCCK 634

Query: 84  YNKPGDDIVLMAEALEKLFLQKINELP 110
            YN  G      A  L   F Q+    P
Sbjct: 635 YNPKGSPAHSNALELRSFFEQRWKLFP 661


>Y119C1B.8a CE44037 WBGene00022473 locus:tag-
            332#status:Partially_confirmed#UniProt:Q95Y80#protein_id
            :AAK39326.3
          Length = 853


 Score =  114 bits (286), Expect = 7e-27,   Method: Composition-based stats.
 Identities = 49/110 (44%), Positives = 73/110 (66%)
```

```
Query: 1   PKRQTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLE 60
           P R TN+L Y++  VLK   KH+  WPFQ+PVDAV L +P Y++ +  PMD+ TI+ RL+
Sbjct: 37  PTRHTNKLDYIMTTVLKEAGKHKHVWPFQKPVDAVALCIPLYHERVARPMDLKTIENRLK 96

Query: 61  NNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKINELP 110
           + YY  AQECI D  T+F NCY +N   DD+ +MA+ + ++  + + + P
Sbjct: 97  STYYTCAQECIDDIETVFQNCYTFNGKEDDVTIMAQNVHEVIKKSLEQAP 146


 Score = 76.3 bits (186), Expect = 3e-15,   Method: Composition-based stats.
 Identities = 34/88 (38%), Positives = 53/88 (60%)

Query: 22  HQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNC 81
            +FAWPF +PVDA +L L DY+KIIK PMD+ ++K ++E+  Y    +   D   M  NC
Sbjct: 279 QEFAWPFNEPVDAEQLGLHDYHKIIKEPMDLKSMKAKMESGAYKEPSDFEHDVRLMLRNC 338

Query: 82  YIYNKPGDDIVLMAEALEKLFLQKINEL 109
           ++YN  GD +       +++F ++  EL
Sbjct: 339 FLYNPVGDPVHSFGLRFQEVFDRRWAEL 366


>Y119C1B.8b CE33207 WBGene00022473 locus:tag-
           332#status:Partially_confirmed#UniProt:Q86S79#protein_id
           :AAO21405.1
          Length = 765


 Score =  114 bits (285), Expect = 1e-26,   Method: Composition-based stats.
 Identities = 49/110 (44%), Positives = 73/110 (66%)

Query: 1   PKRQTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLE 60
           P R TN+L Y++  VLK   KH+  WPFQ+PVDAV L +P Y++ +  PMD+ TI+ RL+
Sbjct: 37  PTRHTNKLDYIMTTVLKEAGKHKHVWPFQKPVDAVALCIPLYHERVARPMDLKTIENRLK 96

Query: 61  NNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKINELP 110
           + YY  AQECI D  T+F NCY +N   DD+ +MA+ + ++  + + + P
Sbjct: 97  STYYTCAQECIDDIETVFQNCYTFNGKEDDVTIMAQNVHEVIKKSLEQAP 146


 Score = 76.3 bits (186), Expect = 3e-15,   Method: Composition-based stats.
 Identities = 34/88 (38%), Positives = 53/88 (60%)

Query: 22  HQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNC 81
            +FAWPF +PVDA +L L DY+KIIK PMD+ ++K ++E+  Y    +   D   M  NC
Sbjct: 279 QEFAWPFNEPVDAEQLGLHDYHKIIKEPMDLKSMKAKMESGAYKEPSDFEHDVRLMLRNC 338

Query: 82  YIYNKPGDDIVLMAEALEKLFLQKINEL 109
           ++YN  GD +       +++F ++  EL
Sbjct: 339 FLYNPVGDPVHSFGLRFQEVFDRRWAEL 366


>F13C5.2 CE19384 WBGene00017423 bromodomain-containing
           protein#status:Confirmed#UniProt:O76561#protein_id:AAC64
           610.1
          Length = 374

 Score = 78.2 bits (191), Expect = 8e-16,   Method: Composition-based stats.
 Identities = 37/82 (45%), Positives = 47/82 (57%)

Query: 24  FAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYI 83
           F +PF++PVD V L L DY+++IK PMDM TI+K+L    Y  A E  +DF  M  NC
Sbjct: 137 FTFPFRKPVDVVLLGLTDYHEVIKKPMDMSTIRKKLIGEEYDTAVEFKEDFKLMINNCLT 196

Query: 84  YNKPGDDIVLMAEALEKLFLQK 105
```

```
           YN  GD +   A    K F  K
Sbjct: 197 YNNEGDPVADFALQFRKKFAAK 218


>H20J04.2 CE27187 WBGene00019217 status:Partially_confirmed UniProt:Q9
            N5L9#protein_id:AAF39888.2
          Length = 1427

 Score = 70.5 bits (171), Expect = 2e-13,   Method: Composition-based stats.
 Identities = 31/99 (31%), Positives = 57/99 (57%), Gaps = 2/99 (2%)

Query: 11   LLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQEC 70
            L+  +LK   + + +WPF QPVD+ ++  PDYY +IK PM++ T+  +++   Y    E
Sbjct: 1328 LIETLLKEAMRQECSWPFLQPVDSKEV--PDYYDVIKRPMNLRTMMNKIKQRIYNKPIEV 1385

Query: 71   IQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKINEL 109
              DF  + +NC  YN+P ++I  ++  L      +++E+
Sbjct: 1386 RNDFQLILSNCETYNEPENEIYKLSRELHDFMADRLDEI 1424


>R10E11.1c CE42151 WBGene00000366 locus:cbp-
           1#status:Partially_confirmed#UniProt:B0M0M3#protein_id:C
           AP72377.1
          Length = 2016

 Score = 68.9 bits (167), Expect = 4e-13,   Method: Composition-based stats.
 Identities = 38/104 (36%), Positives = 61/104 (58%), Gaps = 1/104 (0%)

Query: 4    QTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNY 63
           Q + +++LL V  K L K + A PF+ PVDA  LN+PDY++IIK PMD+ T+ K+L
Sbjct: 855 QEDLIKFLLPVWEK-LDKSEDAPFRVPVDAKLLNIPDYHEIIKRPMDLETVHKKLYAGQ 913

Query: 64   YWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKIN 107
           Y NA +   D   M  N ++YN+    +      L ++F+ +++
Sbjct: 914 YQNAGQFCDDIWLMLDNAWLYNRKNSKVYKYGLKLSEMFVSEMD 957


>R10E11.1b CE21117 WBGene00000366 locus:cbp-
           1#status:Partially_confirmed#UniProt:P34545#protein_id:C
           AD18875.1
          Length = 2056

 Score = 68.9 bits (167), Expect = 4e-13,   Method: Composition-based stats.
 Identities = 38/104 (36%), Positives = 61/104 (58%), Gaps = 1/104 (0%)

Query: 4    QTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNY 63
           Q + +++LL V  K L K + A PF+ PVDA  LN+PDY++IIK PMD+ T+ K+L
Sbjct: 866 QEDLIKFLLPVWEK-LDKSEDAPFRVPVDAKLLNIPDYHEIIKRPMDLETVHKKLYAGQ 924

Query: 64   YWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKIN 107
           Y NA +   D   M  N ++YN+    +      L ++F+ +++
Sbjct: 925 YQNAGQFCDDIWLMLDNAWLYNRKNSKVYKYGLKLSEMFVSEMD 968


>R10E11.1a CE28069 WBGene00000366 locus:cbp-
           1#bromodomain#status:Partially_confirmed#UniProt:P34545#
           protein_id:CAA82353.2
          Length = 2045

 Score = 68.9 bits (167), Expect = 4e-13,   Method: Composition-based stats.
 Identities = 38/104 (36%), Positives = 61/104 (58%), Gaps = 1/104 (0%)

Query: 4    QTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNY 63
           Q + +++LL V  K L K + A PF+ PVDA  LN+PDY++IIK PMD+ T+ K+L
```

```
Sbjct: 855 QEDLIKFLLPVWEK-LDKSEDAAPFRVPVDAKLLNIPDYHEIIKRPMDLETVHKKLYAGQ 913

Query: 64  YWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKIN 107
           Y NA +   D   M  N ++YN+    +      L ++F+ +++
Sbjct: 914 YQNAGQFCDDIWLMLDNAWLYNRKNSKVYKYGLKLSEMFVSEMD 957


>F26H11.2e CE15909 WBGene00009180 locus:nurf-
           1#Bromodomain#status:Confirmed#UniProt:Q6BER5#protein_id
           :CAB04198.1
          Length = 405

 Score = 67.4 bits (163), Expect = 2e-12,   Method: Composition-based stats.
 Identities = 34/86 (39%), Positives = 50/86 (58%), Gaps = 4/86 (4%)

Query: 22  HQFAWPFQQPVDAVKLN-LPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTN 80
           H+ + PF+ PVD   LN  PDY K IK PMD+ TI K++E   Y    + + D N MF N
Sbjct: 260 HRMSTPFRNPVD---LNEFPDYEKFIKKPMDLSTITKKVERTEYLYLSQFVNDVNQMFEN 316

Query: 81  CYIYNKPGDDIVLMAEALEKLFLQKI 106
              YN  G+ +   AE ++++F +K+
Sbjct: 317 AKTYNPKGNAVFKCAETMQEVFDKKL 342


>F26H11.2f CE15910 WBGene00009180 locus:nurf-
           1#Bromodomain#status:Confirmed#UniProt:Q6BER5#protein_id
           :CAB04195.1
          Length = 510

 Score = 67.0 bits (162), Expect = 2e-12,   Method: Composition-based stats.
 Identities = 34/86 (39%), Positives = 50/86 (58%), Gaps = 4/86 (4%)

Query: 22  HQFAWPFQQPVDAVKLN-LPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTN 80
           H+ + PF+ PVD   LN  PDY K IK PMD+ TI K++E   Y    + + D N MF N
Sbjct: 365 HRMSTPFRNPVD---LNEFPDYEKFIKKPMDLSTITKKVERTEYLYLSQFVNDVNQMFEN 421

Query: 81  CYIYNKPGDDIVLMAEALEKLFLQKI 106
              YN  G+ +   AE ++++F +K+
Sbjct: 422 AKTYNPKGNAVFKCAETMQEVFDKKL 447


>F26H11.2d CE42388 WBGene00009180 locus:nurf-
           1#status:Confirmed#UniProt:Q6BER5#protein_id:CAB54234.4
          Length = 808

 Score = 66.6 bits (161), Expect = 2e-12,   Method: Composition-based stats.
 Identities = 34/86 (39%), Positives = 50/86 (58%), Gaps = 4/86 (4%)

Query: 22  HQFAWPFQQPVDAVKLN-LPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTN 80
           H+ + PF+ PVD   LN  PDY K IK PMD+ TI K++E   Y    + + D N MF N
Sbjct: 663 HRMSTPFRNPVD---LNEFPDYEKFIKKPMDLSTITKKVERTEYLYLSQFVNDVNQMFEN 719

Query: 81  CYIYNKPGDDIVLMAEALEKLFLQKI 106
              YN  G+ +   AE ++++F +K+
Sbjct: 720 AKTYNPKGNAVFKCAETMQEVFDKKL 745


>F26H11.2g CE37638 WBGene00009180 locus:nurf-
           1#status:Confirmed#UniProt:Q6BER5#protein_id:CAH60782.1
          Length = 413

 Score = 66.6 bits (161), Expect = 2e-12,   Method: Composition-based stats.
 Identities = 34/86 (39%), Positives = 50/86 (58%), Gaps = 4/86 (4%)
```

```
Query: 22  HQFAWPFQQPVDAVKLN-LPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTN 80
           H+ + PF+ PVD   LN  PDY K IK PMD+ TI K++E   Y    + + D N MF N
Sbjct: 268 HRMSTPFRNPVD---LNEFPDYEKFIKKPMDLSTITKKVERTEYLYLSQFVNDVNQMFEN 324

Query: 81  CYIYNKPGDDIVLMAEALEKLFLQKI 106
              YN  G+ +   AE ++++F +K+
Sbjct: 325 AKTYNPKGNAVFKCAETMQEVFDKKL 350


>F26H11.2c CE36931 WBGene00009180 locus:nurf-
            1#status:Partially_confirmed#UniProt:Q6BER5#protein_id:CA
            H04722.1
          Length = 2266

 Score = 66.2 bits (160), Expect = 3e-12,   Method: Composition-based stats.
 Identities = 34/86 (39%), Positives = 50/86 (58%), Gaps = 4/86 (4%)

Query: 22   HQFAWPFQQPVDAVKLN-LPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTN 80
            H+ + PF+ PVD   LN  PDY K IK PMD+ TI K++E   Y    + + D N MF N
Sbjct: 2121 HRMSTPFRNPVD---LNEFPDYEKFIKKPMDLSTITKKVERTEYLYLSQFVNDVNQMFEN 2177

Query: 81   CYIYNKPGDDIVLMAEALEKLFLQKI 106
               YN  G+ +   AE ++++F +K+
Sbjct: 2178 AKTYNPKGNAVFKCAETMQEVFDKKL 2203


>Y47G6A.6 CE24372 WBGene00021636 locus:pcaf-
           1#status:Partially_confirmed#UniProt:Q9N3S7#protein_id:A
           AF60658.1
          Length = 767

 Score = 63.9 bits (154), Expect = 2e-11,   Method: Composition-based stats.
 Identities = 35/95 (36%), Positives = 53/95 (55%), Gaps = 10/95 (10%)

Query: 15  VLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDF 74
           +LK L   + AWPF  PVD  ++  P+YY  IK P+D  T++++L+   Y +    I D
Sbjct: 658 ILKKLTADKNAWPFASPVDVKEV--PEYYDHIKHPIDFKTMQEKLKRKAYTHQHLFIADL 715

Query: 75  NTMFTNCYIYNKPGDDIVLMAEALEKLFLQKINEL 109
           N +F NCY++N         AEA+   +  K+NEL
Sbjct: 716 NRLFQNCYVFNG--------AEAVYYKYGYKLNEL 742


>C26C6.1a CE30254 WBGene00007042 locus:pbrm-1 HMG (high mobility
           group) box, Bromodomain (5 domains), Zinc finger, C2H2
           type#status:Partially_confirmed#UniProt:Q18210#protein_i
           d:CAA96697.2
          Length = 1883

 Score = 55.8 bits (133), Expect = 4e-09,   Method: Composition-based stats.
 Identities = 30/76 (39%), Positives = 40/76 (52%)

Query: 36  KLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMA 95
           K   P YY +IK PMDM  IK +LEN  Y    + + DF  M +N   +N+   DI   A
Sbjct: 743 KEEFPAYYDVIKKPMDMMRIKHKLENRQYVTLLDVVSDFMLMLSNACKFNETDSDIYKEA 802

Query: 96  EALEKLFLQKINELPT 111
            +L+K  L+   EL T
Sbjct: 803 VSLQKALLEMKRELDT 818


 Score = 50.4 bits (119), Expect = 2e-07,   Method: Composition-based stats.
 Identities = 25/65 (38%), Positives = 38/65 (58%)
```

```
Query: 40  PDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALE 99
           P+YY+II+ P+DM TI+ R++ + Y      I D   MF+N   +N+P   I + A  LE
Sbjct: 570 PEYYQIIQNPIDMKTIRMRIDGHQYPQVDAMINDCRVMFSNARDFNEPRSMIHMDAIQLE 629

Query: 100 KLFLQ 104
           K  L+
Sbjct: 630 KAVLR 634


 Score = 46.2 bits (108), Expect = 4e-06,   Method: Composition-based stats.
 Identities = 24/74 (32%), Positives = 40/74 (54%), Gaps = 1/74 (1%)

Query: 36  KLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMA 95
           K + PDYY  IK P+ +  I KRL+N  Y + +  + D   M++N + YN    ++ + A
Sbjct: 374 KESYPDYYDEIKNPVSIFMINKRLKNGKY-DLKSLVADLMQMYSNAFDYNLESSEVYISA 432

Query: 96  EALEKLFLQKINEL 109
           E L+ L +    +L
Sbjct: 433 EKLKALTISTCKQL 446


 Score = 39.3 bits (90), Expect = 4e-04,   Method: Composition-based stats.
 Identities = 23/72 (31%), Positives = 35/72 (48%)

Query: 38  NLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEA 97
           + P YY+ I  P+D+ TI +   N  Y   +E   D   +F N   ++  G DI   AE
Sbjct: 226 DFPLYYEKIAKPIDLKTIAQNGVNKKYSTMKELKDDLFLLFKNAQQFSGNGSDIFKDAEQ 285

Query: 98  LEKLFLQKINEL 109
           L+ +  +KI  L
Sbjct: 286 LKTVVKEKIARL 297


 Score = 31.2 bits (69), Expect = 0.12,   Method: Composition-based stats.
 Identities = 21/70 (30%), Positives = 32/70 (45%), Gaps = 8/70 (11%)

Query: 40  PDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALE 99
           P+YY+ +K P+D+ TI+ +L+   Y    +   DF     N   Y K         E+ E
Sbjct: 79  PEYYEQVKEPIDVTTIQHKLKIPEYLTYDQFNDDFMMFIKNNLTYYKD--------ESEE 130

Query: 100 KLFLQKINEL 109
              + KI EL
Sbjct: 131 HKDMMKIQEL 140


>F01G4.1 CE05553 WBGene00004204 locus:psa-4 SNF2alpha
            like#status:Confirmed#UniProt:Q19106#protein_id:CAA92978.
            1
          Length = 1474




 Score = 52.4 bits (124), Expect = 5e-08,   Method: Composition-based stats.
 Identities = 24/71 (33%), Positives = 40/71 (56%)

Query: 39   LPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEAL 98
            LPDYY++I  PMD   I K++E   Y   +E   D N +  N   YN+ G +I + +E +
Sbjct: 1217 LPDYYQVISKPMDFDRINKKIETGRYTVMEELNDDMNLLVNNAQTYNEEGSEIYVSSETI 1276

Query: 99   EKLFLQKINEL 109
             KL+ ++ ++
Sbjct: 1277 GKLWKEQYDKF 1287
```

```
>ZK783.4 CE34152 WBGene00001470 locus:flt-
            1#status:Partially_confirmed#UniProt:Q23590#protein_id:AA
            C24421.2
          Length = 1376

 Score = 49.7 bits (117), Expect = 3e-07,   Method: Composition-based stats.
 Identities = 33/104 (31%), Positives = 53/104 (50%), Gaps = 2/104 (1%)

Query: 6    NQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYW 65
            N  + L +++L  L     A PF +PV+  KL +P Y  II  PMD+ TI+++ E   Y
Sbjct: 1262 NMNKELCQLMLDELVVQANALPFLEPVNP-KL-VPGYKMIISKPMDLKTIRQKNEKLIYE 1319

Query: 66   NAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKINEL 109
              ++  +D   MF NC  +N    +I     +L K F ++  +L
Sbjct: 1320 TPEDFAEDIELMFANCRQFNIDHSEIGRAGISLHKFFQKRWKQL 1363


>C01H6.7a CE05190 WBGene00007256 locus:tag-
           298#Bromodomain#status:Confirmed#UniProt:Q17581#protein_
           id:CAA95779.1
          Length = 636

 Score = 48.1 bits (113), Expect = 9e-07,   Method: Composition-based stats.
 Identities = 28/92 (30%), Positives = 48/92 (52%), Gaps = 6/92 (6%)

Query: 10  YLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQE 69
           ++LR +++   +  FA+P    +       PDY  IIKTPMD+ TI++ +E+  Y +
Sbjct: 158 HILRKLVEKDPEQYFAFPVTPSM------APDYRDIIKTPMDLQTIRENIEDGKYASLPA 211

Query: 70  CIQDFNTMFTNCYIYNKPGDDIVLMAEALEKL 101
              +D   + +N + YN+P     L A+ L  L
Sbjct: 212 MKEDCELIVSNAFQYNQPNTVFYLAAKRLSNL 243


>C01H6.7b CE40891 WBGene00007256 locus:tag-
           298#status:Confirmed#UniProt:A5JYT2#protein_id:CAN86573.
           1
          Length = 582

 Score = 47.4 bits (111), Expect = 1e-06,   Method: Composition-based stats.
 Identities = 28/92 (30%), Positives = 48/92 (52%), Gaps = 6/92 (6%)

Query: 10  YLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQE 69
           ++LR +++   +  FA+P    +       PDY  IIKTPMD+ TI++ +E+  Y +
Sbjct: 158 HILRKLVEKDPEQYFAFPVTPSM------APDYRDIIKTPMDLQTIRENIEDGKYASLPA 211

Query: 70  CIQDFNTMFTNCYIYNKPGDDIVLMAEALEKL 101
              +D   + +N + YN+P     L A+ L  L
Sbjct: 212 MKEDCELIVSNAFQYNQPNTVFYLAAKRLSNL 243


>W04A8.7 CE42634 WBGene00006382 locus:taf-1 transcription initiation
            factor
            TFIID#status:Partially_confirmed#UniProt:Q9XUL9#protein_i
            d:CAC14425.2
          Length = 1744

 Score = 44.3 bits (103), Expect = 1e-05,   Method: Composition-based stats.
 Identities = 25/81 (30%), Positives = 44/81 (54%), Gaps = 2/81 (2%)

Query: 28   FQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKP 87
            F  PV++ K+   DYY IIK P+ +  IKK++    Y   ++ + D   MF N  +YN
```

```
Sbjct: 1433 FVTPVNSKKV--VDYYNIIKNPISLQEIKKKISEQSYLLRKDFLDDIKLMFDNSRMYNGD 1490

Query: 88   GDDIVLMAEALEKLFLQKINE 108
             + + L A+ + +L  +++ E
Sbjct: 1491 NNILTLTAQQMLQLAGKRMIE 1511


 Score = 36.6 bits (83), Expect = 0.002,   Method: Composition-based stats.
 Identities = 31/119 (26%), Positives = 55/119 (46%), Gaps = 12/119 (10%)

Query: 1    PKRQTNQL---QYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKK 57
            P   TN L    YLL  +++ +     +  F   VD  K+  P YY  I  PMD+  +++
Sbjct: 1525 PLLDTNDLIGFSYLLGEIVQKMKNIPKSALFHTRVDPKKI--PAYYLKISDPMDLSIMEQ 1582

Query: 58   RLENNYYWNAQECIQDFNTMFTNCYIYNKP-------GDDIVLMAEALEKLFLQKINEL 109
            + ++  Y +  E ++D   ++TN  ++N           ++  MAE L K  +  + EL
Sbjct: 1583 KSKSQEYKSIDEFLKDAEKIYTNSVVFNGAESVYSLKAKEMFEMAEMLVKDQMDTLGEL 1641


>F11A10.1c CE20665 WBGene00008682 locus:lex-
           1#status:Confirmed#UniProt:P54816#protein_id:CAO82045.1
          Length = 1242

 Score = 40.8 bits (94), Expect = 1e-04,   Method: Composition-based stats.
 Identities = 35/112 (31%), Positives = 52/112 (46%), Gaps = 16/112 (14%)

Query: 1   PKRQTNQLQYLLRVVLKT----------LWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPM 50
           P R+T + +Y   V+ K           L + +    F +PVD  +    DYY+II+TP+
Sbjct: 853 PSRRTIRQKYFEHVIEKINTPPKVFDPRLMRDRRFVEFVEPVDPDEAE--DYYEIIETPI 910

Query: 51  DMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYN----KPGDDIVLMAEAL 98
            M  I ++L N  Y +A + + D   + TN   YN    K G  I  MA  L
Sbjct: 911 CMQDIMEKLNNCEYNHADKFVADLILIQTNALEYNPSTTKDGKLIRQMANTL 962


>F11A10.1b CE41384 WBGene00008682 locus:lex-
            1#status:Confirmed#UniProt:P54816#protein_id:CAO82044.1
          Length = 1289

 Score = 40.4 bits (93), Expect = 2e-04,   Method: Composition-based stats.
 Identities = 31/96 (32%), Positives = 47/96 (48%), Gaps = 6/96 (6%)

Query: 7    QLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWN 66
            Q++   +  L  L + +    F +PVD  +    DYY+II+TP+ M  I ++L N  Y +
Sbjct: 916  QMRLFFKERLTRLMRDRRFVEFVEPVDPDEAE--DYYEIIETPICMQDIMEKLNNCEYNH 973

Query: 67   AQECIQDFNTMFTNCYIYN----KPGDDIVLMAEAL 98
            A + + D   + TN   YN    K G  I  MA  L
Sbjct: 974  ADKFVADLILIQTNALEYNPSTTKDGKLIRQMANTL 1009


>F11A10.1a CE40608 WBGene00008682 locus:lex-1 TAT-binding homolog
            like#status:Confirmed#UniProt:P54816#protein_id:CAA92684.
            2
          Length = 1291

 Score = 40.4 bits (93), Expect = 2e-04,   Method: Composition-based stats.
 Identities = 31/96 (32%), Positives = 47/96 (48%), Gaps = 6/96 (6%)

Query: 7    QLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWN 66
            Q++   +  L  L + +    F +PVD  +    DYY+II+TP+ M  I ++L N  Y +
Sbjct: 918  QMRLFFKERLTRLMRDRRFVEFVEPVDPDEAE--DYYEIIETPICMQDIMEKLNNCEYNH 975

Query: 67   AQECIQDFNTMFTNCYIYN----KPGDDIVLMAEAL 98
```

```
            A + + D   + TN   YN    K G  I  MA  L
Sbjct: 976  ADKFVADLILIQTNALEYNPSTTKDGKLIRQMANTL 1011


>F26H11.2h CE42387 WBGene00009180 locus:nurf-
           1#status:Confirmed#UniProt:Q6BER5#protein_id:CAQ16138.1
          Length = 554

 Score = 30.4 bits (67), Expect = 0.20,   Method: Composition-based stats.
 Identities = 15/52 (28%), Positives = 29/52 (55%), Gaps = 1/52 (1%)

Query: 55  IKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKI 106
           +K++    Y + +Q  + D N MF N   YN  G+ +   AE ++++F +K+
Sbjct: 441 VKEQKRTEYLYLSQ-FVNDVNQMFENAKTYNPKGNAVFKCAETMQEVFDKKL 491


>C34C6.3 CE43092 WBGene00007916 EGF receptor\/notch-like
           protein#status:Partially_confirmed#UniProt:Q18424#protei
           n_id:CAA91258.3
          Length = 529

 Score = 29.3 bits (64), Expect = 0.40,   Method: Composition-based stats.
 Identities = 12/33 (36%), Positives = 20/33 (60%)

Query: 55  IKKRLENNYYWNAQECIQDFNTMFTNCYIYNKP 87
           I+KR   +Y +   Q C Q FN+   +C+ Y++P
Sbjct: 196 IEKRCFCSYGFFGQRCDQKFNSQNDHCFAYDEP 228
```

Appendix B.3 Chromo-Domain BLAST Alignment

```
>K08H2.6 CE06164 WBGene00001995 locus:hpl-1 murine modifier 2
          protein
          like#status:Confirmed#UniProt:Q21370#protein_id:CAA9415
          2.1
          Length = 184

 Score = 43.5 bits (101), Expect = 2e-05,   Method: Composition-based stats.
 Identities = 17/48 (35%), Positives = 29/48 (60%)

Query: 2  YAVEKIIDRRVRKGKVEYYLKWKGYXXXXXXXXXXXXLDCQDLIQQYE 49
          + VEK++++R+ +G  EYY+KW+G+            L C  +IQ+YE
Sbjct: 37 FVVEKVLNKRLTRGGSEYYIKWQGFPESECSWEPIENLQCDRMIQEYE 84


>ZK1236.2 CE00380 WBGene00000414 locus:cec-
          1#Nucleolin#status:Confirmed#UniProt:P34618#protein_id:
          AAA28192.1
          Length = 304





 Score = 34.7 bits (78), Expect = 0.010,   Method: Composition-based stats.
 Identities = 14/25 (56%), Positives = 19/25 (76%)

Query: 2  YAVEKIIDRRVRKGKVEYYLKWKGY 26
          Y VE I++ R +KGK E+Y+KW GY
Sbjct: 8  YTVESILEHRKKKGKSEFYIKWLGY 32


>K01G5.2c CE25038 WBGene00001996 locus:hpl-2 'chromo' (CHRromatin
          Organization MOdifier)
          domain#status:Confirmed#UniProt:Q9U3C6#protein_id:CAB54
          267.2
          Length = 303

 Score = 31.2 bits (69), Expect = 0.11,   Method: Composition-based stats.
 Identities = 13/49 (26%), Positives = 28/49 (57%), Gaps = 1/49 (2%)

Query: 2  YAVEKIIDRRVRK-GKVEYYLKWKGYXXXXXXXXXXXXLDCQDLIQQYE 49
          + VEK++D+R  K G+ E+ ++W+G+            L C +++ ++E
Sbjct: 19 FMVEKVLDKRTGKAGRDEFLIQWQGFPESDSSWEPRENLQCVEMLDEFE 67


>K01G5.2b CE25037 WBGene00001996 locus:hpl-2 'chromo' (CHRromatin
          Organization MOdifier)
          domain#status:Confirmed#UniProt:O17918#protein_id:CAB07
          243.2
          Length = 301

 Score = 31.2 bits (69), Expect = 0.11,   Method: Composition-based stats.
 Identities = 13/49 (26%), Positives = 28/49 (57%), Gaps = 1/49 (2%)

Query: 2  YAVEKIIDRRVRK-GKVEYYLKWKGYXXXXXXXXXXXXLDCQDLIQQYE 49
          + VEK++D+R  K G+ E+ ++W+G+            L C +++ ++E
Sbjct: 19 FMVEKVLDKRTGKAGRDEFLIQWQGFPESDSSWEPRENLQCVEMLDEFE 67


>F32E10.2 CE04475 WBGene00017990 chromo domain of heterochromatin
           protein#status:Confirmed#UniProt:Q19972#protein_id:AAA83
           357.1
          Length = 270

 Score = 30.0 bits (66), Expect = 0.24,   Method: Composition-based stats.
 Identities = 13/26 (50%), Positives = 18/26 (69%)

Query: 1   EYAVEKIIDRRVRKGKVEYYLKWKGY 26
           EYAVE+++  R  KG   Y ++WKGY
Sbjct: 86  EYAVERVLAHRKVKGSPLYLVQWKGY 111


>K01G5.2a CE16191 WBGene00001996 locus:hpl-2 'chromo' (CHRromatin
          Organization MOdifier)
          domain#status:Confirmed#UniProt:O17916#protein_id:CAB07
          241.1
          Length = 175

 Score = 28.9 bits (63), Expect = 0.53,   Method: Composition-based stats.
 Identities = 13/49 (26%), Positives = 28/49 (57%), Gaps = 1/49 (2%)

Query: 2  YAVEKIIDRRVRK-GKVEYYLKWKGYXXXXXXXXXXXXLDCQDLIQQYE 49
          + VEK++D+R  K G+ E+ ++W+G+            L C +++ ++E
Sbjct: 19 FMVEKVLDKRTGKAGRDEFLIQWQGFPESDSSWEPRENLQCVEMLDEFE 67
```

Appendix B.4 Plant-Homeo-domain BLAST Alignment

```
>T14G8.1 CE03657 WBGene00000482 locus:chd-3 helicase-DNA-binding
           like
           protein#status:Confirmed#UniProt:Q22516#protein_id:CAA91
           810.1
          Length = 1787

 Score = 83.2 bits (204), Expect = 3e-17,   Method: Composition-based stats.
 Identities = 28/43 (65%), Positives = 35/43 (81%)

Query: 1   FCRVCKDGGELLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRC 43
```

```
           +CR+CK+   +L CDTCPSSYH +C++PPL EIP GEW CPRC
Sbjct: 330 YCRICKETSNILLCDTCPSSYHAYCIDPPLTEIPEGEWSCPRC 372


 Score = 59.7 bits (143), Expect = 3e-10,   Method: Composition-based stats.
 Identities = 19/42 (45%), Positives = 27/42 (64%)

Query: 2   CRVCKDGGELLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRC 43
           C VC   GEL+ CDTC  +YH+ C++  + + P G+W CP C
Sbjct: 268 CEVCNQDGELMLCDTCTRAYHVACIDENMEQPPEGDWSCPHC 309


>F26F12.7 CE17716 WBGene00002637 locus:let-418 DNA
           helicase#status:Confirmed#UniProt:Q19815#protein_id:AAC2
           5894.1
          Length = 1829

 Score = 65.9 bits (159), Expect = 4e-12,   Method: Composition-based stats.
 Identities = 23/44 (52%), Positives = 32/44 (72%), Gaps = 1/44 (2%)

Query: 1   FCRVCKDGGELLCCDTCPSSYHIHCLNPPLPEIPNGE-WLCPRC 43
           FC++CK+   LL CD+C  S+H +C++PPL E+P  E W CPRC
Sbjct: 319 FCKICKETENLLLCDSCVCSFHAYCIDPPLTEVPKEETWSCPRC 362


 Score = 60.5 bits (145), Expect = 2e-10,   Method: Composition-based stats.
 Identities = 21/43 (48%), Positives = 27/43 (62%)

Query: 1   FCRVCKDGGELLCCDTCPSSYHIHCLNPPLPEIPNGE-WLCPRC 43
           +C  CK  GELL CDTCP +YH  C++  + E P G+W C  C
Sbjct: 258 YCEECKQDGELLLCDTCPRAYHTVCIDENMEEPPEGDWSCAHC 300


>ZK783.4 CE34152 WBGene00001470 locus:flt-
            1#status:Partially_confirmed#UniProt:Q23590#protein_id:AA
            C24421.2
          Length = 1376

 Score = 57.0 bits (136), Expect = 2e-09,   Method: Composition-based stats.
 Identities = 20/45 (44%), Positives = 28/45 (62%), Gaps = 2/45 (4%)

Query: 1    FCRVCK--DGGELLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRC 43
             C++CK  DG E+L CD C S  H+ C  P + ++P G+W C RC
Sbjct: 1088 LCQICKSMDGDEMLVCDGCESGCHMECFRPRMTKVPEGDWFCQRC 1132


>C44B9.4 CE30897 WBGene00008081 locus:athp-1 S.pombe hypothetical
           protein C27F7.07C
           like#status:Partially_confirmed#UniProt:Q18605#protein_i
           d:CAA97781.2
          Length = 1150

 Score = 50.4 bits (119), Expect = 2e-07,   Method: Composition-based stats.
 Identities = 16/42 (38%), Positives = 27/42 (64%)

Query: 2   CRVCKDGGELLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRC 43
           C +C  GG +LCC+ CP+S+H+ C+     ++P+  + C RC
Sbjct: 62  CGICSSGGNILCCEQCPASFHLACIGYESSDLPDDNFYCNRC 103


 Score = 33.1 bits (74), Expect = 0.029,   Method: Composition-based stats.
 Identities = 15/42 (35%), Positives = 21/42 (50%), Gaps = 1/42 (2%)

Query: 2   CRVCKDGGELLCCDTCPSSYHIHCLNPPLPEI-PNGEWLCPR 42
```

```
           C +  D   +L CD C   +H  C+ PPL  +     W+CPR
Sbjct: 224 CNLKDDWTRMLKCDFCDLIWHQKCVTPPLIHVRAYFYWMCPR 265


>T12D8.1 CE42503 WBGene00011729 locus:set-16 PHD-finger. (2
           domains), SET
           domain#status:Partially_confirmed#UniProt:O46025#protein
           _id:CAB05024.2
          Length = 2519

 Score = 50.1 bits (118), Expect = 2e-07,   Method: Composition-based stats.
 Identities = 22/46 (47%), Positives = 27/46 (58%), Gaps = 3/46 (6%)

Query: 2   CRVCKDGGE---LLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRCT 44
           C  C  GG+   LL CD C  SYHI+C+ P L +IP G W C  C+
Sbjct: 524 CEGCGTGGDEANLLLCDECDVSYHIYCMKPLLDKIPQGPWRCQWCS 569


>ZK593.4 CE35704 WBGene00004319 locus:rbr-2 Human XE169
           like#status:Partially_confirmed#UniProt:Q23541#protein_i
           d:CAA93426.2
          Length = 1477

 Score = 49.3 bits (116), Expect = 4e-07,   Method: Composition-based stats.
 Identities = 21/48 (43%), Positives = 28/48 (58%), Gaps = 5/48 (10%)

Query: 1   FCRVCKDGGE---LLCCDT--CPSSYHIHCLNPPLPEIPNGEWLCPRC 43
           FC  C +G +   LL CD   C +  H +C +P L E+P GEW CP+C
Sbjct: 321 FCVACNEGKDEDLLLLCDIDGCNNGRHTYCCDPVLDEVPEGEWRCPKC 368


 Score = 33.5 bits (75), Expect = 0.022,   Method: Composition-based stats.
 Identities = 15/39 (38%), Positives = 21/39 (53%), Gaps = 2/39 (5%)

Query: 7    DGGELLCCDTCPSSYHIHCL--NPPLPEIPNGEWLCPRC 43
            D    L C  C S +H+ C   +P L ++P G +LC RC
Sbjct: 1216 DSESTLTCIMCDSEFHVRCCEWSPFLEKLPEGCFLCVRC 1254


>F17A2.3 CE05646 WBGene00008902 PHD-
          finger.#status:Predicted#UniProt:Q19511#protein_id:CAA9
          2158.1
          Length = 463

 Score = 45.4 bits (106), Expect = 6e-06,   Method: Composition-based stats.
 Identities = 18/43 (41%), Positives = 27/43 (62%), Gaps = 1/43 (2%)

Query: 2  CRVCKDGGELLCCDTCPSSYHIHCLN-PPLPEIPNGEWLCPRC 43
          C +C DGG ++ C+TCP+S+H  CL    +PE     ++C RC
Sbjct: 30 CGMCADGGTIIWCETCPASFHAFCLGLKTIPEPEKDTFICHRC 72


>Y59A8A.2 CE44093 WBGene00013339 status:Partially_confirmed UniProt:
           Q9GRZ5#protein_id:CAC14404.2
          Length = 599

 Score = 40.4 bits (93), Expect = 2e-04,   Method: Composition-based stats.
 Identities = 19/45 (42%), Positives = 24/45 (53%), Gaps = 3/45 (6%)

Query: 2   CRVCKDGGELLCCDTCPSSYHIHCLNPPLPEIP---NGEWLCPRC 43
           CR   +  +   CD C  SYHI CL+PPL  +P   N  W+C  C
Sbjct: 518 CRKSTEQHKQTQCDECHKSYHIGCLSPPLTRLPKRNNFGWICHEC 562
```

```
>K09A11.5 CE34205 WBGene00010708 PHD-
          finger.#status:Partially_confirmed#UniProt:Q21375#prote
          in_id:CAA90618.2
          Length = 650


 Score = 40.0 bits (92), Expect = 2e-04,   Method: Composition-based stats.
 Identities = 19/43 (44%), Positives = 25/43 (58%), Gaps = 1/43 (2%)

Query: 2  CRVCKDGGELLCCDTCPSSYHIHCLNPPLPEIPNGE-WLCPRC 43
          C +C   GE+L C +CP+S+HI CL       PNG  + C RC
Sbjct: 53 CCICARRGEVLWCHSCPASFHIKCLGYDTDPQPNGTIFTCRRC 95


>C28H8.9a CE06896 WBGene00016200 status:Confirmed UniProt:Q09477 pro
           tein_id:AAA62297.3
          Length = 372

 Score = 38.9 bits (89), Expect = 5e-04,   Method: Composition-based stats.
 Identities = 16/42 (38%), Positives = 24/42 (57%)

Query: 2   CRVCKDGGELLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRC 43
           C   ++  +LL CD C   YH++CL P L + P+ E+ C  C
Sbjct: 317 CGTSENDDKLLFCDDCDRGYHLYCLTPALEKAPDDEYSCRLC 358


>F33E11.6b CE39929 WBGene00018013 status:Confirmed UniProt:Q2A950 pr
           otein_id:ABD63225.1
          Length = 447

 Score = 36.2 bits (82), Expect = 0.003,   Method: Composition-based stats.
 Identities = 18/46 (39%), Positives = 23/46 (50%), Gaps = 4/46 (8%)

Query: 2   CRVC-KDGGELLCCDTCPSSYHIHCLNPP---LPEIPNGEWLCPRC 43
           C  C K GGE++CC TC  +YH  C+  P      +   EW C  C
Sbjct: 335 CDSCEKTGGEMICCATCKIAYHPQCIEMPERMAALVKTYEWSCVDC 380


 Score = 32.3 bits (72), Expect = 0.052,   Method: Composition-based stats.
 Identities = 14/45 (31%), Positives = 24/45 (53%), Gaps = 8/45 (17%)

Query: 2   CRVC------KDGGELLCCDTCPSSYHIHCLNPPLPEIPNGEWLC 40
           CR+C      +   E++ CD C   +H +C+   L ++P G W+C
Sbjct: 380 CRLCSICNKPEKEDEIVFCDRCDRGFHTYCVG--LKKLPQGTWIC 422


>H05L14.2 CE42798 WBGene00010367 Zinc finger, C3HC4 type (RING
            finger)#status:Partially_confirmed#UniProt:O17709#protein
            _id:CAB16922.3
          Length = 2199

 Score = 34.7 bits (78), Expect = 0.010,   Method: Composition-based stats.
 Identities = 16/52 (30%), Positives = 24/52 (46%), Gaps = 18/52 (34%)

Query: 2    CRVC----KDGGELLCCDTCPSSYHIHCLNPPLPEIPNGEWL-----CPRCT 44
            C +C    ++  E + CDTC   YH HC++          WL     CP+C+
Sbjct: 2144 CLICTEIIEEAVETVTCDTCTREYHYHCIS---------RWLKINSVCPQCS 2186


>F26H11.2i CE43186 WBGene00009180 locus:nurf-
           1#status:Partially_confirmed#UniProt:B6VQ92#protein_id:C
           AR97823.1
          Length = 1619
 Score = 33.5 bits (75), Expect = 0.023,   Method: Composition-based stats.
 Identities = 13/34 (38%), Positives = 21/34 (61%), Gaps = 2/34 (5%)

Query: 2   CRVC-KDGGELLCCDTCPSSYHIHCLN-PPLPEI 33
           CRVC K  G ++ C  C +++H+ C +  P PE+
Sbjct: 350 CRVCGKSSGRVVGCTQCEAAFHVECSHLKPFPEV 383


>F26H11.2b CE35295 WBGene00009180 locus:nurf-
           1#status:Partially_confirmed#UniProt:Q6BER5#protein_id:C
           AC42289.2
          Length = 1693

 Score = 33.5 bits (75), Expect = 0.023,   Method: Composition-based stats.
 Identities = 13/34 (38%), Positives = 21/34 (61%), Gaps = 2/34 (5%)

Query: 2   CRVC-KDGGELLCCDTCPSSYHIHCLN-PPLPEI 33
           CRVC K  G ++ C  C +++H+ C +  P PE+
Sbjct: 422 CRVCGKSSGRVVGCTQCEAAFHVECSHLKPFPEV 455


>F26H11.2a CE35294 WBGene00009180 locus:nurf-
           1#status:Partially_confirmed#UniProt:Q6BER5#protein_id:C
           AB04197.2
          Length = 1691

 Score = 33.5 bits (75), Expect = 0.023,   Method: Composition-based stats.
 Identities = 13/34 (38%), Positives = 21/34 (61%), Gaps = 2/34 (5%)

Query: 2   CRVC-KDGGELLCCDTCPSSYHIHCLN-PPLPEI 33
           CRVC K  G ++ C  C +++H+ C +  P PE+
Sbjct: 422 CRVCGKSSGRVVGCTQCEAAFHVECSHLKPFPEV 455


>F26H11.2c CE36931 WBGene00009180 locus:nurf-
           1#status:Partially_confirmed#UniProt:Q6BER5#protein_id:C
           AH04722.1
          Length = 2266


 Score = 33.1 bits (74), Expect = 0.027,   Method: Composition-based stats.
 Identities = 13/34 (38%), Positives = 21/34 (61%), Gaps = 2/34 (5%)

Query: 2   CRVC-KDGGELLCCDTCPSSYHIHCLN-PPLPEI 33
           CRVC K  G ++ C  C +++H+ C +  P PE+
Sbjct: 422 CRVCGKSSGRVVGCTQCEAAFHVECSHLKPFPEV 455


>H20J04.2 CE27187 WBGene00019217 status:Partially_confirmed UniProt:Q9
            N5L9#protein_id:AAF39888.2
          Length = 1427

 Score = 32.3 bits (72), Expect = 0.049,   Method: Composition-based stats.
 Identities = 12/45 (26%), Positives = 20/45 (44%), Gaps = 3/45 (6%)

Query: 2    CRVCKDGG---ELLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRC 43
            CR C+      +L+ C  C + YH+ C    +      +W+C  C
Sbjct: 1115 CRSCRRKAAAHDLVLCSECDNCYHLKCAKLDVNSDAPADWMCTSC 1159


>F42A9.2 CE07224 WBGene00003034 locus:lin-49 zinc-finger
           protein#status:Confirmed#UniProt:Q20318#protein_id:AAB03
           164.1
          Length = 1042
 Score = 32.3 bits (72), Expect = 0.051,   Method: Composition-based stats.
 Identities = 17/47 (36%), Positives = 22/47 (46%), Gaps = 7/47 (14%)

Query: 2   CRVCKDG-----GELLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRC 43
           C +C DG      +++ CD C  S H  C    +P IP G   C RC
Sbjct: 198 CNICLDGDTSNCNQIVYCDRCNLSVHQDCYG--IPFIPEGCLECRRC 242


>C11G6.3 CE05257 WBGene00007524 PHD-
           finger.#status:Partially_confirmed#UniProt:Q17909#protei
           n_id:CAA94113.1
          Length = 385

 Score = 31.6 bits (70), Expect = 0.083,   Method: Composition-based stats.
 Identities = 16/47 (34%), Positives = 22/47 (46%), Gaps = 5/47 (10%)

Query: 2   CRVCKD----GGELLCCDTCPSSYHIHCLNPPLPEIPNGEWLCPRCT 44
           C VC      G  ++ CD C   +H HC+     E  + +W C RCT
Sbjct: 310 CPVCSVAYTVGANMIGCDQCQDWFHWHCVGLT-AEPTDSKWFCTRCT 355


>F54F2.2a CE25003 WBGene00006975 locus:zfp-
          1#status:Confirmed#UniProt:P34447#protein_id:AAK26137.1
          Length = 867

 Score = 30.4 bits (67), Expect = 0.20,   Method: Composition-based stats.
 Identities = 18/50 (36%), Positives = 23/50 (46%), Gaps = 9/50 (18%)

Query: 2  CRVCKD-----GGELLCCD--TCPSSYHIHCLNPPLPEIPNGEWLCPRCT 44
          C VC D        L+ CD   C  + H  C    + E+P GEW C +CT
Sbjct: 8  CCVCADENGWTDNPLIYCDGENCEVAVHQGCYG--IQEVPEGEWFCAKCT 55


>Y51H1A.4 CE20286 WBGene00013095 locus:ing-3 PHD-
           finger.#status:Confirmed#UniProt:Q9XWJ8#protein_id:CAA21
           665.1
          Length = 490

 Score = 29.6 bits (65), Expect = 0.28,   Method: Composition-based stats.
 Identities = 16/46 (34%), Positives = 24/46 (52%), Gaps = 6/46 (13%)

Query: 1   FCRVCKDGGELLCCDTCPSS---YHIHCLNPPLPEIPNGEWLCPRC 43
           FC   K  G+++ CD    +   +H  C+   + E P G+W CPRC
Sbjct: 432 FCNE-KSYGDMVQCDNRHCTLRWFHYPCIG--MVEPPTGKWYCPRC 474
```

Appendix B.4 SET-domain BLAST Alignment

```
>C43E11.3b CE08681 WBGene00016603 locus:met-
           1#status:Partially_confirmed#UniProt:A4LBC3#protein_id:A
           BO52817.1
          Length = 1590

 Score = 90.9 bits (224), Expect = 2e-19,   Method: Composition-based stats.
 Identities = 53/137 (38%), Positives = 75/137 (54%), Gaps = 9/137 (6%)

Query: 2387 GLGLYAKVDISMGDFIIEYKGEIIRSEVCEVREIRYVAQNRGVYMFRIDEE-WVIDATMA 2445
            G GL A  DI  G FIIEY GE++  +  E R+ +Y A  +  + +  D   + IDAT+
Sbjct: 681  GCGLRAVKDIKKGRFIIEYIGEVVERDDYEKRKTKYAADKKHKHHYLCDTGVYTIDATVY 740

Query: 2446 GGPARYINHSCDPNCSTQILDAGSGARE----KKIIITANRPISANEELTYDYQFELEGT 2501
            G P+R++NHSCDPN    I +  S  R      ++   + R I A EE+T+DYQF   G
Sbjct: 741  GNPSRFVNHSCDPNA---ICEKWSVPRTPGDVNRVGFFSKRFIKAGEEITFDYQFVNYG- 796
```

```
Query: 2502 TDKIPCLCGAPNCVKWM 2518
             D   C CG+ +C  W+
Sbjct: 797  RDAQQCFCGSASCSGWI 813


>C43E11.3a CE30503 WBGene00016603 locus:met-
           1#status:Partially_confirmed#UniProt:A4LBC2#protein_id:A
           BO52816.1
          Length = 1604


 Score = 90.9 bits (224), Expect = 2e-19,   Method: Composition-based stats.
 Identities = 53/137 (38%), Positives = 75/137 (54%), Gaps = 9/137 (6%)

Query: 2387 GLGLYAKVDISMGDFIIEYKGEIIRSEVCEVREIRYVAQNRGVYMFRIDEE-WVIDATMA 2445
            G GL A  DI  G FIIEY GE++  +  E R+ +Y A  +  + +  D   + IDAT+
Sbjct: 695  GCGLRAVKDIKKGRFIIEYIGEVVERDDYEKRKTKYAADKKHKHHYLCDTGVYTIDATVY 754

Query: 2446 GGPARYINHSCDPNCSTQILDAGSGARE----KKIIITANRPISANEELTYDYQFELEGT 2501
            G P+R++NHSCDPN    I +  S  R      ++   + R I A EE+T+DYQF   G
Sbjct: 755  GNPSRFVNHSCDPNA---ICEKWSVPRTPGDVNRVGFFSKRFIKAGEEITFDYQFVNYG- 810

Query: 2502 TDKIPCLCGAPNCVKWM 2518
             D   C CG+ +C  W+
Sbjct: 811  RDAQQCFCGSASCSGWI 827


>Y2H9A.1 CE27781 WBGene00003222 locus:mes-4 SET
           domain#status:Confirmed#UniProt:Q9NH52#protein_id:CAA162
           76.2
          Length = 898

 Score = 65.5 bits (158), Expect = 8e-12,   Method: Composition-based stats.
 Identities = 46/143 (32%), Positives = 70/143 (48%), Gaps = 15/143 (10%)

Query: 2376 DRVYLARSRIAGLGLYAKVDISMGDFIIEYKGEIIRSEVCEVREIRYVAQNRGV----YM 2431
            +++ LA +   G G++AK  I   ++I EY GEII  +  + R +  V+ +R      YM
Sbjct: 537  EKIKLAATLCKGYGVFAKGQIEKDEYICEYVGEII-DKAEKKRRLDSVSISRDFQANHYM 595

Query: 2432 FRIDEEWVIDATMAGGPARYINHSCDPNCSTQILDA------GSGAREKKIIITANRPIS 2485
              + +   +DA   G  +RYINHSCDPN ++ +              + +  I A R I
Sbjct: 596  MELHKGLTVDAARYGNISRYINHSCDPNAASFVTKVFVKKTKEGSLYDTRSYIRAIRTID 655

Query: 2486 ANEELTYDYQFELEGTTDKIP-CLCGAPNCVKWM 2518
              +E+T+ Y    E   + +P C CGA NC+  M
Sbjct: 656  DGDEITFSYNMNNE---ENLPDCECGAENCMGTM 686

>R06A4.7 CE28067 WBGene00003220 locus:mes-2 SET
            domain#status:Confirmed#UniProt:O17514#protein_id:CAB0558
            9.2
          Length = 773

 Score = 60.5 bits (145), Expect = 1e-08,   Method: Composition-based stats.
 Identities = 45/131 (34%), Positives = 64/131 (48%), Gaps = 5/131 (3%)

Query: 2370 MRREWKDRVYLARSRIAGLGLYAKVDISMGDFIIEYKGEIIRSEVCEVREIRYVAQNRGV 2429
            M R  + R Y   S+IAG GL+        +FI EY GE I  +  E R   Y  + +
Sbjct: 615  MTRMIQKRTYCGPSKIAGNGLFLLEPAEKDEFITEYTGERISDDEAERRGAIY-DRYQCS 673

Query: 2430 YMFRIDEEWVIDATMAGGPARYINH-SCDPNCSTQILDAGSGAREKKIIITANRPISANE 2488
            Y+F I+    ID+   G  AR+ NH S +P C  + +     A E +I   A R +  +E
Sbjct: 674  YIFNIETGGAIDSYKIGNLARFANHDSKNPTCYARTMVV---AGEHRIGFYAKRRLEISE 730
```

```
Query: 2489 ELTYDYQFELE 2499
            ELT+DY +   E
Sbjct: 731  ELTFDYSYSGE 741

>R05D3.11 CE42016 WBGene00019883 locus:met-
            2#status:Partially_confirmed#UniProt:P34544#protein_id:AA
            K21437.2
          Length = 1316

 Score = 59.7 bits (143), Expect = 5e-10,   Method: Composition-based stats.
 Identities = 30/79 (37%), Positives = 44/79 (55%), Gaps = 2/79 (2%)

Query: 2438 WVIDATMAGGPARYINHSCDPNCSTQ-ILDAGSGAREKKIIITANRPISANEELTYDYQF 2496
            +VIDA   G   R++NHSCDPN   Q ++      R   +     + + A +ELT+DYQ+
Sbjct: 1231 YVIDAKQRGNLGRFLNHSCDPNVHVQHVMYDTHDLRLPWVAFFTRKYVKAGDELTWDYQY 1290

Query: 2497 ELEGT-TDKIPCLCGAPNC 2514
              + T T ++ C CGA NC
Sbjct: 1291 TQDQTATTQLTCHCGAENC 1309

>T12F5.4 CE13601 WBGene00003040 locus:lin-
           59#status:Confirmed#UniProt:O44757#protein_id:AAB96746.1
          Length = 1312

 Score = 52.8 bits (125), Expect = 6e-08,   Method: Composition-based stats.
 Identities = 40/138 (28%), Positives = 58/138 (42%), Gaps = 24/138 (17%)

Query: 2367 YQKMRREWKD----RVYLARSRIAGLGLYAKVDISMGDFIIEYKGEIIRSEVCEVREIRY 2422
            Y   RR WK+    ++ ++    +   L  K+    G+F+ EY GE+I  E  + +
Sbjct: 625  YCSNRRFWKEDCGNKLCVSNGPRSKRVLKTKIARRAGEFLCEYAGEVITREQAQEK---- 680

Query: 2423 VAQNRGVYMFRIDEEWVIDATMAGGPARYINHSCDPNCSTQILDAGSGAREKKIIITANR 2482
             AQ+R   +  I     +DAT     AR+I HSC PN   ++       R     ++
Sbjct: 681  FAQDRDPRIIAIAAHLFVDATKRSNIARFIKHSCKPNSRLEVWSVNGFYRAGVFALSDLN 740

Query: 2483 PISANEELTYDYQFELEGTTDKIP----CLCGAPNC 2514
            P   N E+T D        +D +P    C CGA  C
Sbjct: 741  P---NAEITVD-------KSDLLPFDMACNCGATEC 766
```

Appendix B.4 SET-domain BLAST Alignment

(JHDM1)

```
>T26A5.5a CE32733 WBGene00020821 status:Partially_confirmed UniProt:
           Q95Q98#protein_id:AAN65291.1
          Length = 1076

 Score =  385 bits (988), Expect = e-108,   Method: Composition-based stats.
 Identities = 173/173 (100%), Positives = 173/173 (100%)

Query: 1   FSQTPLEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCLM 60
           FSQTPLEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCLM
Sbjct: 93  FSQTPLEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCLM 152

Query: 61  SVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE 120
           SVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE
Sbjct: 153 SVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE 212

Query: 121 KCHVAILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN 173
           KCHVAILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN
Sbjct: 213 KCHVAILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN 265
```

```
>T26A5.5b CE32734 WBGene00020821 status:Confirmed UniProt:Q95Q98 pro
           tein_id:AAN65292.1
          Length = 505

 Score =  381 bits (978), Expect = e-106,   Method: Composition-based stats.
 Identities = 173/173 (100%), Positives = 173/173 (100%)

Query: 1   FSQTPLEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCLM 60
           FSQTPLEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCLM
Sbjct: 93  FSQTPLEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCLM 152

Query: 61  SVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE 120
           SVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE
Sbjct: 153 SVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE 212

Query: 121 KCHVAILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN 173
           KCHVAILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN
Sbjct: 213 KCHVAILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN 265


>F29B9.2a CE09781 WBGene00017920 status:Confirmed UniProt:Q9GYI0 pro
           tein_id:AAK29799.1
          Length = 910

 Score =  157 bits (398), Expect = 2e-39,   Method: Composition-based stats.
 Identities = 74/169 (43%), Positives = 105/169 (62%), Gaps = 3/169 (1%)

Query: 6   LEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCLMSVANC 65
           ++++ K P  V++I  V   WPD     +I    R++  Y P    PKV+ +CL  +A
Sbjct: 446 MKEIAKPPRFVQEISMVNRLWPDVSGAEYIKLLQREE--YLPEDQRPKVEQFCLAGMAGS 503

Query: 66  YTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE-KCHV 124
           YTDFH+DF G+SV+YH+LKG K+F++  PTE NF  YQ    + +   +FG
Sbjct: 504 YTDFHVDFGGSSVYYHILKGEKIFYIAAPTEQNFAAYQAHETSPDTTTWFGDIANGAVKR 563

Query: 125 AILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN 173
            +++ G T+LIP+GWIHAV TP DSLVFGGNFLH  + + Q+RVY +EN
Sbjct: 564 VVIKEGQTLLIPAGWIHAVLTPVDSLVFGGNFLHLGNLEMQMRVYHLEN 612


>F29B9.2b CE27145 WBGene00017920 status:Confirmed UniProt:Q9BI67 pro
           tein_id:AAK29800.1
          Length = 897

 Score =  157 bits (397), Expect = 2e-39,   Method: Composition-based stats.
 Identities = 74/169 (43%), Positives = 105/169 (62%), Gaps = 3/169 (1%)

Query: 6   LEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCLMSVANC 65
           ++++ K P  V++I  V   WPD     +I    R++  Y P    PKV+ +CL  +A
Sbjct: 433 MKEIAKPPRFVQEISMVNRLWPDVSGAEYIKLLQREE--YLPEDQRPKVEQFCLAGMAGS 490

Query: 66  YTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSVE-KCHV 124
           YTDFH+DF G+SV+YH+LKG K+F++  PTE NF  YQ    + +   +FG
Sbjct: 491 YTDFHVDFGGSSVYYHILKGEKIFYIAAPTEQNFAAYQAHETSPDTTTWFGDIANGAVKR 550

Query: 125 AILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN 173
            +++ G T+LIP+GWIHAV TP DSLVFGGNFLH  + + Q+RVY +EN
Sbjct: 551 VVIKEGQTLLIPAGWIHAVLTPVDSLVFGGNFLHLGNLEMQMRVYHLEN 599


>F43G6.6 CE20788 WBGene00005013 status:Partially_confirmed UniProt:Q
           20367#protein_id:CAA90395.1
          Length = 548
 Score =  132 bits (333), Expect = 6e-32,   Method: Composition-based stats.
 Identities = 67/175 (38%), Positives = 101/175 (57%), Gaps = 13/175 (7%)

Query: 1   FSQTP-LEDLVKSPELVRQIDWVGNQWPDALRQRWISFNGRDKKFYNPHHTFPKVQNYCL 59
           FS  P L+++ + P  V+ I      W D   +  +S + R           PK++  C
Sbjct: 223 FSDHPELKEMARPPRFVQDISMAKRLWSDVTSKSALSDDHR-----------PKIEQICA 271

Query: 60  MSVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAFFGKSV 119
            ++AN YTDFH+DF GTSV++HV KG K+F++  PTE NF +YQ    + + + + G ++
Sbjct: 272 AAMANSYTDFHVDFGGTSVYFHVFKGEKIFYIAAPTEENFVMYQAHETSTDSSIWLGHTL 331

Query: 120 EKC-HVAILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHSQSCKTQLRVYQVEN 173
           +      +++ G T+LIP+GWIHAV T  DSL FGGNFLH  +    +RV  +EN
Sbjct: 332 KGALKRVVVKEGQTLLIPAGWIHAVLTTIDSLAFGGNFLHLGNLIMHMRVVDMEN 386


>F29B9.4a CE27146 WBGene00004205 locus:psr-
           1#status:Confirmed#UniProt:Q9GYI4#protein_id:AAF99922.2
          Length = 400

 Score = 44.3 bits (103), Expect = 4e-05,   Method: Composition-based stats.
 Identities = 40/135 (29%), Positives = 58/135 (42%), Gaps = 22/135 (16%)

Query: 37  FNGRDKKFYNPHHTFPKVQNYCLMSVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTE 96
           F+  D K   PH  F       +M  A   T  HID  GTS W  +L+G K + LIPP
Sbjct: 166 FHYADDKKRPPHRWF-------VMGPARSGTAIHIDPLGTSAWNSLLQGHKRWVLIPPIA 218

Query: 97  TNFFI---YQEFIKTVNDNAFFGKSVEK--------CHVAILE----PGDTMLIPSGWIH 141
               +     E  K  ++   + ++V K           A +E    PG+TM +PSGW H
Sbjct: 219 PRDLVKPMAHEKGKHPDEGITWFQTVYKRVRSPSWPKEYAPIECRQGPGETMFVPSGWWH 278

Query: 142 AVYTPDDSLVFGGNF 156
            V   + ++    N+
Sbjct: 279 VVINEEYTIAVTHNY 293


>T07C4.11 CE40266 WBGene00011563 status:Partially_confirmed UniProt:
           Q14V35#protein_id:CAK55173.1
          Length = 367


 Score = 43.5 bits (101), Expect = 5e-05,   Method: Composition-based stats.
 Identities = 25/111 (22%), Positives = 52/111 (46%), Gaps = 13/111 (11%)

Query: 57  YCLMSVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTETNFFIYQEFIKTVNDNAF-- 114
           +  +  +  +T  H D   +  W   + GRK ++++PP   N F       +V ++ F
Sbjct: 139 FVYIGASGSWTKLHSDVVSSHSWSANICGRKQWFMMPPGSENLFR-----SSVTESGFVD 193

Query: 115 ----FGKSVEKCHVA--ILEPGDTMLIPSGWIHAVYTPDDSLVFGGNFLHS 159
               + +  E+  V   + EPG+ + +PS W H  +  +D++    N+++S
Sbjct: 194 DIREYERLFEQAKVIKFVQEPGEIVFVPSNWYHQAHNLEDTISINHNWMNS 244


>F29B9.4b CE39926 WBGene00004205 locus:psr-
           1#status:Confirmed#UniProt:Q27GT3#protein_id:ABD63227.1
          Length = 284


 Score = 43.5 bits (101), Expect = 5e-05,   Method: Composition-based stats.
 Identities = 40/135 (29%), Positives = 58/135 (42%), Gaps = 22/135 (16%)

Query: 37  FNGRDKKFYNPHHTFPKVQNYCLMSVANCYTDFHIDFSGTSVWYHVLKGRKVFWLIPPTE 96
           F+  D K   PH  F       +M  A   T  HID  GTS W  +L+G K + LIPP
```

```
Sbjct: 50  FHYADDKKRPPHRWF-------VMGPARSGTAIHIDPLGTSAWNSLLQGHKRWVLIPPIA 102

Query: 97  TNFFI---YQEFIKTVNDNAFFGKSVEK--------CHVAILE----PGDTMLIPSGWIH 141
               +     E  K  ++   + ++V K           A +E    PG+TM +PSGW H
Sbjct: 103 PRDLVKPMAHEKGKHPDEGITWFQTVYKRVRSPSWPKEYAPIECRQGPGETMFVPSGWWH 162

Query: 142 AVYTPDDSLVFGGNF 156
            V   + ++    N+
Sbjct: 163 VVINEEYTIAVTHNY 177
```

(JARID1/2)

```
>ZK593.4 CE35704 WBGene00004319 locus:rbr-2 Human XE169
           like#status:Partially_confirmed#UniProt:Q23541#protein_i
           d:CAA93426.2
          Length = 1477


 Score =  241 bits (615), Expect = 5e-65,   Method: Composition-based stats.
 Identities = 104/104 (100%), Positives = 104/104 (100%)

Query: 1   GMCFSTFCWHTEDHWTYSVNYNHFGERKIWYGVGGEDAEKFEDALKKIAPGLTGRQRDLF 60
           GMCFSTFCWHTEDHWTYSVNYNHFGERKIWYGVGGEDAEKFEDALKKIAPGLTGRQRDLF
Sbjct: 505 GMCFSTFCWHTEDHWTYSVNYNHFGERKIWYGVGGEDAEKFEDALKKIAPGLTGRQRDLF 564

Query: 61  HHMTTAANPHLLRSLGVPIHSVHQNAGEFVITFPRAYHAGFNEG 104
           HHMTTAANPHLLRSLGVPIHSVHQNAGEFVITFPRAYHAGFNEG
Sbjct: 565 HHMTTAANPHLLRSLGVPIHSVHQNAGEFVITFPRAYHAGFNEG 608


>Y48B6A.11 CE41181 WBGene00012982 locus:jmjd-
           2#status:Partially_confirmed#UniProt:Q9U297#protein_id:C
           AB54451.2
          Length = 922


 Score = 86.7 bits (213), Expect = 2e-18,   Method: Composition-based stats.
 Identities = 44/111 (39%), Positives = 59/111 (53%), Gaps = 11/111 (9%)

Query: 1   GMCFSTFCWHTEDHWTYSVNYNHFGERKIWYGVGGEDAEKFEDALKK-------IAPGLT 53
           GM  +TF WH ED   YS+N+ HFG  K W+ +  E A++FE  + +        AP
Sbjct: 256 GMYKTTFPWHAEDMDLYSINFLHFGAPKYWFAISSEHADRFERFMSQQFSYQNEYAP--- 312

Query: 54  GRQRDLFHHMTTAANPHLLRSLGVPIHSVHQNAGEFVITFPRAYHAGFNEG 104
            + +    H T    P LLR  G+P  ++ Q   EF+ITFPR YH GFN G
Sbjct: 313 -QCKAFLRHKTYLVTPELLRQAGIPYATMVQRPNEFIITFPRGYHMGFNLG 362


>C29F7.6 CE08447 WBGene00007813 status:Partially_confirmed UniProt:O
           17619#protein_id:CAB07325.1
          Length = 732


 Score = 30.8 bits (68), Expect = 0.14,   Method: Composition-based stats.
 Identities = 26/94 (27%), Positives = 39/94 (41%), Gaps = 11/94 (11%)

Query: 6   TFCWHTEDHWTYSVNYNHFGERKIWYGVGGEDAEKFEDALKKIAPGLTGRQRDLFHHMTT 65
           T C H E+    S+N N    + IWY V  E + KFE  L K         ++L+ + +
Sbjct: 507 TTC-HIENQAIGSLNLNLGPGKCIWYAVASEHSAKFEQLLMK---------KNLWPYDSV 556

Query: 66  A-ANPHLLRSLGVPIHSVHQNAGEFVITFPRAYH 98
              N   L + G+P+    Q   + V      YH
Sbjct: 557 LWPNEEELLNWGIPVMKFIQETDDTVYVGTGTYH 590
```

```
>C16C10.2 CE01493 WBGene00007623 status:Confirmed UniProt:Q09462 p
          rotein_id:CAA86740.1
          Length = 262


 Score = 28.5 bits (62), Expect = 0.73,   Method: Composition-based stats.
 Identities = 12/37 (32%), Positives = 20/37 (54%)

Query: 26 ERKIWYGVGGEDAEKFEDALKKIAPGLTGRQRDLFHH 62
          E+K  Y +  ED +K  D +KK+      + +D +HH
Sbjct: 33 EKKKDYKLRAEDYQKKRDTIKKLKKSAMDKNQDEYHH 69


>F23D12.5 CE15893 WBGene00009089 status:Partially_confirmed UniProt:
           Q19760#protein_id:CAA94916.1
          Length = 867


 Score = 28.1 bits (61), Expect = 0.91,   Method: Composition-based stats.
 Identities = 14/38 (36%), Positives = 20/38 (52%)

Query: 10  HTEDHWTYSVNYNHFGERKIWYGVGGEDAEKFEDALKK 47
           H E+    S+N NH     +WYGV  E + + E  +KK
Sbjct: 624 HLENQALGSININHGPGDCVWYGVPMEYSGRMEVLIKK 661
```

(JHDM3/JMJD2)

```
>C29F7.6 CE08447 WBGene00007813 status:Partially_confirmed UniProt:O
           17619#protein_id:CAB07325.1
          Length = 732


 Score = 49.3 bits (116), Expect = 9e-07,   Method: Composition-based stats.
 Identities = 41/144 (28%), Positives = 67/144 (46%), Gaps = 23/144 (15%)

Query: 13  GTILEDTNYEIKGVNTVYLYFGMYKTTFPWHAEDMDLYSINFLHFGAPK-YWFAISSEHA 71
           G + E     I GVN V +YF    +    H E+  + S+N L+ G  K  W+A++SEH+
Sbjct: 480 GGLNEYIKESIGGVNEVQMYFKQPGSRTTCHIENQAIGSLN-LNLGPGKCIWYAVASEHS 538

Query: 72  DRFERFMSQQ--FSYQNEYAPQCKAFLRHKTYLVTPELLRQAGIPYATMVQRPNEFIITF 129
            +FE+ + ++  + Y +   P               E L   GIP    +Q  ++ +
Sbjct: 539 AKFEQLLMKKNLWPYDSVLWPN-------------EEELLNWGIPVMKFIQETDDTVYVG 585

Query: 130 PRGYH----MGF--NLGYNLAEST 147
              YH    +GF  N+ +N+AEST
Sbjct: 586 TGTYHWVQSIGFTGNVSWNIAEST 609


>F18E9.5b CE30958 WBGene00017571 locus:tag-
           279#status:Partially_confirmed#UniProt:Q95QK3#protein_id
           :AAM54191.1
          Length = 1061


 Score = 40.4 bits (93), Expect = 5e-04,   Method: Composition-based stats.
 Identities = 31/127 (24%), Positives = 53/127 (41%), Gaps = 15/127 (11%)

Query: 10  NRLGTILEDTNYEIKGVNTVYLYFGMYKTTFPWHAEDMDLYSINFLHFGAPKYWFAISSE 69
           NR G +L     ++ G+NTV +Y     +  P H E+  + SIN+        WFA+  E
Sbjct: 778 NREGNLLNYAGVDVLGINTVQMYAKPIGSRTPAHMENSLMASINWNRGPGTCVWFAVPYE 837

Query: 70  HADRFERFMSQQ-FSYQNE-YAPQCKAFLRHKTYLVTPELLRQAGIPYATMVQRPNEFII 127
           +  + E  + +    YQ++ Y P  K  L             + G+P     Q+ +E +
Sbjct: 838 YWGQLEFMIGEHGHKYQDQDYWPSEKELL-------------ELGVPVIKFEQKADEMVY 884
```

```
Query: 128 TFPRGYH 134
                +H
Sbjct: 885 VNTGCFH 891


>F18E9.5a CE30957 WBGene00017571 locus:tag-
           279#status:Partially_confirmed#UniProt:Q19565#protein_id
           :AAM54190.1
          Length = 1020


 Score = 40.4 bits (93), Expect = 5e-04,   Method: Composition-based stats.
 Identities = 31/127 (24%), Positives = 53/127 (41%), Gaps = 15/127 (11%)

Query: 10  NRLGTILEDTNYEIKGVNTVYLYFGMYKTTFPWHAEDMDLYSINFLHFGAPKYWFAISSE 69
           NR G +L     ++ G+NTV +Y     +  P H E+  + SIN+        WFA+  E
Sbjct: 737 NREGNLLNYAGVDVLGINTVQMYAKPIGSRTPAHMENSLMASINWNRGPGTCVWFAVPYE 796

Query: 70  HADRFERFMSQQ-FSYQNE-YAPQCKAFLRHKTYLVTPELLRQAGIPYATMVQRPNEFII 127
           +  + E  + +    YQ++ Y P  K  L             + G+P     Q+ +E +
Sbjct: 797 YWGQLEFMIGEHGHKYQDQDYWPSEKELL-------------ELGVPVIKFEQKADEMVY 843

Query: 128 TFPRGYH 134
                +H
Sbjct: 844 VNTGCFH 850


>F23D12.5 CE15893 WBGene00009089 status:Partially_confirmed UniProt:
           Q19760#protein_id:CAA94916.1
          Length = 867


 Score = 34.3 bits (77), Expect = 0.031,   Method: Composition-based stats.
 Identities = 30/143 (20%), Positives = 54/143 (37%), Gaps = 21/143 (14%)

Query: 13  GTILEDTNYEIKGVNTVYLYFGMYKTTFPWHAEDMDLYSINFLHFGAPKYWFAISSEHAD 72
           G +L      + G+N   +Y          H E+  L SIN  H      W+ +  E++
Sbjct: 594 GNLLNFAQESLAGLNKPQVYCKPPGARTTAHLENQALGSININHGPGDCVWYGVPMEYSG 653

Query: 73  RFERFMSQQF--SYQNEYAPQCKAFLRHKTYLVTPELLRQAGIPYATMVQRPNEFIITFP 130
           R E  + +     Y++ Y P             + + LR   IP    +Q+P + +
Sbjct: 654 RMEVLIKKHRLNVYKSGYWP-------------SEQELRNEKIPSQKFLQKPGDMVYVGI 700

Query: 131 RGYH------MGFNLGYNLAEST 147
             +H         N+ +N+A+ T
Sbjct: 701 GTFHWVQSNDFAINVSWNVAQPT 723
```

(UTX/UTY)

```
>D2021.1 CE01878 WBGene00017046 locus:utx-1 glucose repression
            mediator
            protein#status:Partially_confirmed#UniProt:Q09519#protein
            _id:AAB36864.1
          Length = 1168

 Score =  375 bits (962), Expect = e-105,   Method: Composition-based stats.
 Identities = 164/164 (100%), Positives = 164/164 (100%)

Query: 1    KWGKQINELSKLPAFCRLIAGSNMLSHLGHQVHGMNTVKLFMKVPGCRTPAHQDSNHMAS 60
            KWGKQINELSKLPAFCRLIAGSNMLSHLGHQVHGMNTVKLFMKVPGCRTPAHQDSNHMAS
Sbjct: 863  KWGKQINELSKLPAFCRLIAGSNMLSHLGHQVHGMNTVKLFMKVPGCRTPAHQDSNHMAS 922

Query: 61   ININIGPGDCEWFAVPYEYWGKMHKLCEKNGVDLLTGTFWPIIDDLLDAGIPVHRFTQKA 120
            ININIGPGDCEWFAVPYEYWGKMHKLCEKNGVDLLTGTFWPIIDDLLDAGIPVHRFTQKA
```

```
Sbjct: 923  ININIGPGDCEWFAVPYEYWGKMHKLCEKNGVDLLTGTFWPIIDDLLDAGIPVHRFTQKA 982

Query: 121  GDMVYYVSGGAIHWVQASGWCNNISWNVAPLNFQQLSISLLSYEY 164
             GDMVYYVSGGAIHWVQASGWCNNISWNVAPLNFQQLSISLLSYEY
Sbjct: 983  GDMVYYVSGGAIHWVQASGWCNNISWNVAPLNFQQLSISLLSYEY 1026



>F18E9.5a CE30957 WBGene00017571 locus:tag-
           279#status:Partially_confirmed#UniProt:Q19565#protein_id
           :AAM54190.1
          Length = 1020


 Score =  169 bits (428), Expect = 6e-43,   Method: Composition-based stats.
 Identities = 75/161 (46%), Positives = 110/161 (68%)

Query: 4   KQINELSKLPAFCRLIAGSNMLSHLGHQVHGMNTVKLFMKVPGCRTPAHQDSNHMASINI 63
           KQ+NE+ KLP F       N+L++ G  V G+NTV+++ K  G RTPAH +++ MASIN
Sbjct: 722 KQMNEIEKLPTFLLPNREGNLLNYAGVDVLGINTVQMYAKPIGSRTPAHMENSLMASINW 781

Query: 64  NIGPGDCEWFAVPYEYWGKMHKLCEKNGVDLLTGTFWPIIDDLLDAGIPVHRFTQKAGDM 123
           N GPG C WFAVPYEYWG++  +  ++G       +WP   +LL+ G+PV +F QKA +M
Sbjct: 782 NRGPGTCVWFAVPYEYWGQLEFMIGEHGHKYQDQDYWPSEKELLELGVPVIKFEQKADEM 841

Query: 124 VYVSGGAIHWVQASGWCNNISWNVAPLNFQQLSISLLSYEY 164
           VYV+ G  HWVQ++ +C N+SWNV   NF QL+ S++++++
Sbjct: 842 VYVNTGCFHWVQSNSFCINVSWNVGQPNFTQLATSIVAHDH 882


>F18E9.5b CE30958 WBGene00017571 locus:tag-
           279#status:Partially_confirmed#UniProt:Q95QK3#protein_id
           :AAM54191.1
          Length = 1061


 Score =  169 bits (428), Expect = 6e-43,   Method: Composition-based stats.
 Identities = 75/161 (46%), Positives = 110/161 (68%)

Query: 4   KQINELSKLPAFCRLIAGSNMLSHLGHQVHGMNTVKLFMKVPGCRTPAHQDSNHMASINI 63
           KQ+NE+ KLP F       N+L++ G  V G+NTV+++ K  G RTPAH +++ MASIN
Sbjct: 763 KQMNEIEKLPTFLLPNREGNLLNYAGVDVLGINTVQMYAKPIGSRTPAHMENSLMASINW 822

Query: 64  NIGPGDCEWFAVPYEYWGKMHKLCEKNGVDLLTGTFWPIIDDLLDAGIPVHRFTQKAGDM 123
           N GPG C WFAVPYEYWG++  +  ++G       +WP   +LL+ G+PV +F QKA +M
Sbjct: 823 NRGPGTCVWFAVPYEYWGQLEFMIGEHGHKYQDQDYWPSEKELLELGVPVIKFEQKADEM 882

Query: 124 VYVSGGAIHWVQASGWCNNISWNVAPLNFQQLSISLLSYEY 164
           VYV+ G  HWVQ++ +C N+SWNV   NF QL+ S++++++
Sbjct: 883 VYVNTGCFHWVQSNSFCINVSWNVGQPNFTQLATSIVAHDH 923


>F23D12.5 CE15893 WBGene00009089 status:Partially_confirmed UniProt:
           Q19760#protein_id:CAA94916.1
          Length = 867



 Score =  145 bits (365), Expect = 1e-35,   Method: Composition-based stats.
 Identities = 64/164 (39%), Positives = 102/164 (62%)

Query: 1   KWGKQINELSKLPAFCRLIAGSNMLSHLGHQVHGMNTVKLFMKVPGCRTPAHQDSNHMAS 60
```

```
           ++ +Q++E+ KLP   R     N+L+     + G+N  +++ K PG RT AH ++  + S
Sbjct: 573 RFKEQLDEIKKLPDCLRPDGAGNLLNFAQESLAGLNKPQVYCKPPGARTTAHLENQALGS 632

Query: 61  ININIGPGDCEWFAVPYEYWGKMHKLCEKNGVDLLTGTFWPIIDDLLDAGIPVHRFTQKA 120
           ININ GPGDC W+ VP EY G+M  L +K+ +++    +WP   +L +  IP  +F QK
Sbjct: 633 INPNHGPGDCVWYGVPMEYSGRMEVLIKKHRLNVYKSGYWPSEQELRNEKIPSQKFLQKP 692

Query: 121 GDMVYVSGGAIHWVQASGWCNNISWNVAPLNFQQLSISLLSYEY 164
           GDMVYV  G  HWVQ++ +  N+SWNVA   F QL+ +++ +++
Sbjct: 693 GDMVYVGIGTFHWVQSNDFAINVSWNVAQPTFNQLAAAMVIHDH 736


>C29F7.6 CE08447 WBGene00007813 status:Partially_confirmed UniProt:O
           17619#protein_id:CAB07325.1
          Length = 732


 Score =  132 bits (331), Expect = 1e-31,   Method: Composition-based stats.
 Identities = 61/164 (37%), Positives = 95/164 (57%), Gaps = 2/164 (1%)

Query: 1   KWGKQINELSKLPAFCRLIAGSNMLSHLGHQVHGMNTVKLFMKVPGCRTPAHQDSNHMAS 60
           K+   I EL KLP F +   G N   ++   + G+N V+++ K PG RT  H ++  + S
Sbjct: 461 KFQPLIQELDKLPNFLKTKGGLN--EYIKESIGGVNEVQMYFKQPGSRTTCHIENQAIGS 518

Query: 61  ININIGPGDCEWFAVPYEYWGKMHKLCEKNGVDLLTGTFWPIIDDLLDAGIPVHRFTQKA 120
           +N+N+GPG C W+AV  E+  K  +L  K  +       WP  ++LL+ GIPV +F Q+
Sbjct: 519 LNLNLGPGKCIWYAVASEHSAKFEQLLMKKNLWPYDSVLWPNEEELLNWGIPVMKFIQET 578

Query: 121 GDMVYVSGGAIHWVQASGWCNNISWNVAPLNFQQLSISLLSYEY 164
            D VYV  G  HWVQ+ G+  N+SWN+A   F Q +++ L +++
Sbjct: 579 DDTVYVGTGTYHWVQSIGFTGNVSWNIAESTFDQFAMAALVHDH 622
```