2022

EVOLUTIONARY CONSERVATION OF THE DREAM SUBCOMPLEX MUVB

Spencer Snider
Michigan Technological University, sjsnider@mtu.edu

EVOLUTIONARY CONSERVATION OF THE DREAM SUBCOMPLEX MUVB

By

Spencer J. Snider

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Biological Sciences

MICHIGAN TECHNOLOGICAL UNIVERSITY

2022

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Biological Sciences.

Department of Biological Sciences

Thesis Advisor:	*Paul D. Goetsch*
Committee Member:	*Thomas Werner*
Committee Member:	*Zhiying Shan*
Department Chair:	*Chandrashekhar P. Joshi*

Table of Contents

List of Figures

Author Contribution Statement

Jillian Kuizenga during the Summer and Fall of 2021 was a big help in getting this study put together. While the start was rough with just learning how to set up and dealing with programing issues together, the end result was more than I thought would be possible in such a short time.

Acknowledgements

I would like to express my deepest thanks to my advisor Dr. Paul Goetsch for these past two years. Time has flown by quickly and it's been a wonderful experience being part of your lab. Thank you for all the knowledge, guidance, patience, and giving me the opportunity to work on this project.

Thank you to my committee members, Dr. Thomas Werner and Dr. Zhiying Shan. I greatly appreciate you for giving your time and expertise to contributing to this project.

I'd also like to thank my family for supporting me throughout my time here.

Abstract

As the publications of annotated genomes from species representing most domains of life continue to grow exponentially, we are gaining more insight into how proteins, cellular pathways, and protein complexes evolved. We are interested in understanding how each protein in the 8-subunit transcriptional repressor complex called DREAM interacts with each other. DREAM is comprised of 3 main components: an E2F-DP transcription factor heterodimer, a pocket protein, and the highly conserved 5-subunit subcomplex called MuvB. We hypothesize that the mechanism of DREAM's formation on chromatin dictates how DREAM functions to turn off target gene expression. Unfortunately, many interaction surfaces remain unknown, especially those that are involved in the formation of the MuvB subcomplex. While protein conservation has been performed on some DREAM subunits, primarily the E2Fs and pocket proteins, an extensive analysis of the MuvB subcomplex itself had yet to be conducted. Not only is the MuvB subcomplex understudied evolutionarily, but the studies that have evaluated MuvB conservation primarily utilize model organisms. Here, we developed a pipeline to perform protein conservation analysis of 4 of the 5 subunits of the MuvB subcomplex and the pocket protein. Our analysis makes full use of the protein sequences uploaded to Uniprot to expand to hundreds of sequences for analysis. By identifying each protein subunit in most annotated genomes, we developed the broadest model for conservation of MuvB and the pocket protein, including individual domains within each protein. We determined that the conservation of known MuvB interactions is observed outside of the animal kingdom, extending into the plant kingdom. We identified a novel conserved region within the MuvB subunit LIN37. Additionally, our protein conservation model revealed how the unique LxCxE motif in the Nematoda LIN52, known to be the MuvB interaction interface with the pocket protein, diverged from the more broadly conserved LxSxExL motif observed in humans. Interestingly, a similar LxCxE motif was observed in LIN54 homologs in plants but did not identify a corresponding LIN52 homolog, suggesting that the LIN54 DNA-binding subunit may have originally served to link the MuvB subcomplex with the pocket protein. Altogether, our findings serve to expand our understanding of the evolution of the MuvB subcomplex and how the complex may assemble.

1 Introduction

The DREAM complex is an evolutionarily conserved protein complex that represses the transcription of cell cycle genes [Pilkinton, 2007] [Fischer, 2017]. In humans, the MuvB subcomplex is a core component of the DREAM complex that acts as both a repressor and activator of cell cycle genes, depending on the stage of the cell cycle [Pilkinton, 2007]. We are interested in how the DREAM complex subunits forms on chromatin, as we hypothesize that the mechanism of its formation defines how the complex transcriptionally represses target genes. We identified that the MuvB complex mediates DREAM's repressive activity, but little is known as to how the MuvB complex forms [Goetsch, 2017]. In this study, we employed protein conservation analysis to help identify interaction interfaces between MuvB subunits. While protein conservation has been performed in other components of the DREAM complex, most specifically the E2F transcription factor family and pocket protein family of proteins, the MuvB subcomplex has not been evaluated in significant detail [Liban, 2017]. However, many important aspects of how the DREAM complex forms and functions on chromatin as well as analysis of its evolution has been described, as outlined below.

1.1 The DREAM Complex

The evolutionary conserved transcriptional repressor complex known as DREAM (Dp, Rb-like, E2F, and MuvB) is composed of three core components: an E2F-DP heterodimer (E2F4/5-DP1/2/3), a pocket protein (p130/p107), and the 5-subunit MuvB subcomplex (LIN9, LIN37, LIN52, RBAP48, and LIN54) [Korenjak, 2004] [Harrison, 2006] [Litovchick, 2007] (Figure 1). DREAM assembles during quiescence, also called G0, forming an 8-subunit protein complex that represses the transcription of cell cycle genes [Pilkinton, 2007]. The complex binds directly to chromatin through the E2F-DP transcription factor heterodimer and the MuvB subunit LIN54, recognizing site-specific DNA sequence motifs called the cell cycle-dependent element (CDE) and the cell cycle genes homology region (CHR), respectively. The pocket protein (known as LIN-35 in *Caenorhabditis elegans*) acts as a scaffold bridging MuvB and E2F-DP together to mediate transcriptional repression of cell cycle genes [Müller, 2010]. The current model is that DREAM occupancy on chromatin positions a nucleosome at the transcriptional start site of target genes and thus interfering with transcriptional initiation [Asthana, 2022]

In human cells, the current model for DREAM assembly starts with DYRK1A phosphorylation of LIN52, which mediates the interaction between MuvB and the pocket protein [Guiley, 2015]. The interaction of the pocket protein to MuvB and E2F-DP mediates in repression of DREAM target genes in *C. elegans* [Goetsch, 2017]. However, the direct interaction of the pocket protein to MuvB alone is not required for MuvB to repress DREAM target genes in C. *elegans* [Goetsch, 2019]. How DREAM assembly results in the repression of cell cycle genes and if the same mechanisms apply across all eukaryotes is not fully understood.

Previous studies have mapped out the conservation of the DREAM components (E2Fs and the pocket protein family) and found that homologs can be found throughout the animal kingdom and into Amoebozoa [Liban, 2017]. However, an extensive conservation analysis into the MuvB complex has yet to be conducted. Here we conduct an extensive analysis into MuvB subunits as well as re-analyze the pocket protein as the pocket protein interacts with MuvB. Knowledge of the MuvB subcomplex's conservation will help us better understand the evolution of the DREAM complex.

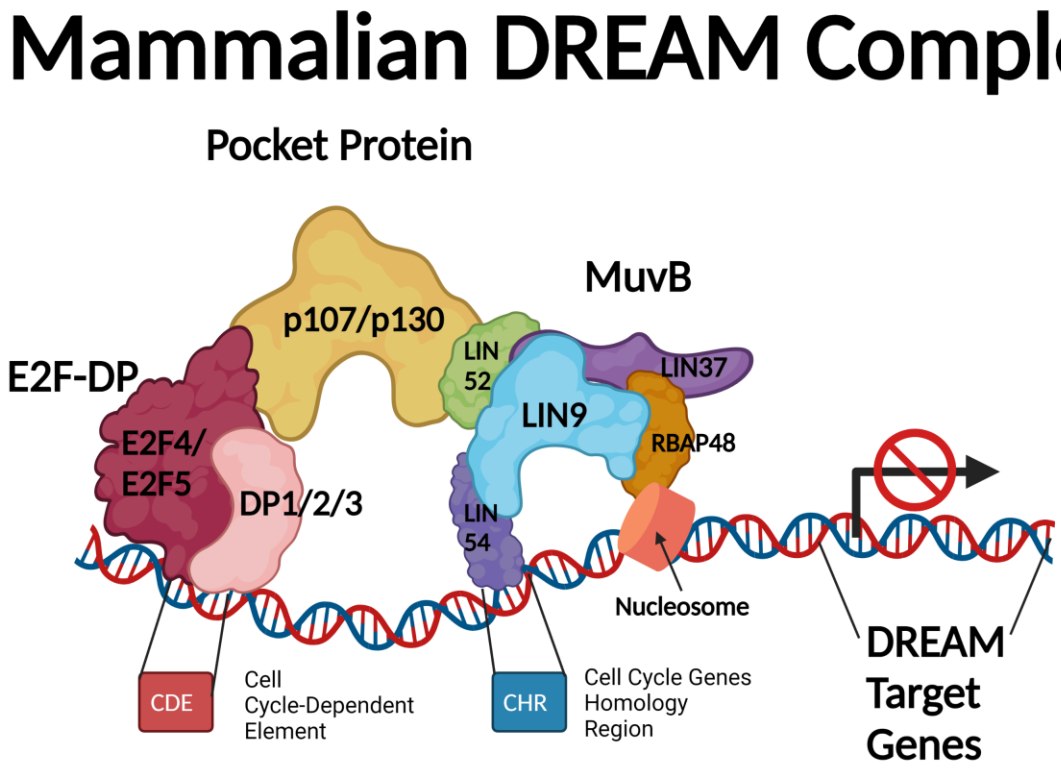

Figure 1 The Mammalian DREAM Complex repressing target genes

Model of DREAM complex repression of target genes in mammalian cells. The E2F-DP heterodimer contains either E2F4 or E2F5 and either DP1, DP2, or DP3. The pocket protein, either p107 or p130, binds to both the E2F-DP heterodimer and the MuvB subunit LIN52. The MuvB subcomplex contains LIN9, LIN37, LIN52, LIN54, and RBAP48. The E2F-DP heterodimer binds to the DNA motif called the cell cycle-dependent element (CDE) site while LIN54 binds to DNA motif called the cell cycle genes homology region (CHR) site. RBAP48 recruits a nucleosome downstream of the transcription start site of DREAM target genes. Transcription of DREAM target genes is repressed while DREAM is assembled and bound to the chromatin.

Created with BioRender.com

1.2 The MuvB Subcomplex

MuvB is a unique 5-protein complex that acts as both a repressor and activator of cell cycle genes depending on the cell cycle phase [Pilkinton, 2007] [Fischer, 2017]. When assembled in DREAM, the MuvB subcomplex functions as a repressor of cell cycle gene during G0 and G1 [Litovchick, 2007]. As a cell enters the cell cycle, cyclin dependent kinases (CDK) phosphorylate the pocket protein, leading to DREAM disassembly. However, MuvB continues to occupy the CHR site and during S-phase, the B-Myb transcription factor interacts with MuvB, forming the Myb-MuvB (MMB) transcriptional activator complex [Osterloh, 2007]. Interestingly, MMB is not wholly conserved, as *C. elegans* does not contain a B-Myb ortholog [Vorster, 2020]. At this point in the cell cycle, the function of MuvB switches from being a repressor to and activator of cell cycle genes [Osterloh, 2007]. During G2 of the cell cycle, MMB disassembles and once again MuvB remains on chromatin, supporting the FOXM1 transcription factor to cell cycle gene promoters [Sadasivam, 2012]. Interestingly, FOXM1 orthologs have not been identified in both *C. elegans* and *Drosophila* [Fischer, 2017].

In *C. elegans*, MuvB is comprised of the subunits LIN-9, LIN-37, LIN-52, LIN-53 (homolog of RBAP48), and LIN-54 [Harrison, 2006]. We know that many of the subunits are essential for MuvB function, as removal of Lin9, Lin52, Rbap48, or Lin54 are lethal in mice [Reichert, 2010] [Matsuo, 2012] [Forristal, 2014] [Miao, 2020]. The exception is the loss of Lin37, which in mouse cells causes loss of DREAM repression function but not MMB activation function [Mages, 2017]. The phenotypes observed in mammalian cells is likely due to the dominant effect of MuvB's function in activating cell cycle genes when present in MMB, as the loss of function in LIN-9, LIN-52, or LIN-54 only confers sterility in *C. elegans* [Harrison, 2006]. Structural studies have identified interaction sites between LIN52 and LIN9. LIN54 and the CHR site, and recently between LIN9, LIN37, and the histone-binding protein RBAP48 [Jiang, 2007] [Marceau, 2016] [Guiley, 2018] [Asthana, 2022]. Although pieces of the model for MuvB's assembly have been described, there remains important interaction sites between subunits that remain unknown.

1.2.1 LIN9

Much of the structure of LIN-9 is unknown, but LIN-9 is thought to be a core structural component in MuvB [Fischer, 2017]. It is known in humans that LIN9 is required for the regulation of G2/M genes [Osterloh, 2007]. LIN9 is required for embryonic stem cells to regulate the cell cycle, and the loss of LIN9 results in an increased distribution of cells in G2 and S phase [Esterlechner, 2013]. The loss of mice Lin9 results in embryonic lethality and when knocked out of mice, lethality occurs within 7 days and is associated with a substantial loss of the intestinal epithelium [Reichert 2010]. The loss of LIN9 in human cell cultures results in DREAM target genes becoming upregulated [Litovchick, 2007]. In *Drosophila*, mutations of Mip130 (homolog of LIN9) results in defects in olfactory receptors [Sam, 2012]. In *C. elegans* loss of LIN-9 results in a high-temperature arrest phenotype where they do not develop past the first larval stage [Petrella, 2011]. LIN9 interacts with LIN52 and B-Myb during later stages in the cell cycle [Guiley, 2018]. The

known interactions of LIN9 include all other MuvB components (LIN37, LIN52, LIN53, LIN54) as well as B-Myb. This suggests that LIN9 plays a critical role in being a "core" for the rest of the MuvB proteins to bind to in animals. However, in plants, it was observed that the homolog of LIN9 (ALY) does not have an apparent interaction with the plant homolog of LIN37 [Lang, 2021]. How LIN9 interacts with LIN54 though is still unknown.

1.2.2 LIN37

The function of LIN37 in MuvB appears to primarily support DREAM's transcriptional repressive activity. The loss of LIN37 alone in human cells results in MuvB losing its function to repress cell cycle genes during G0/G1, but MuvB retained its function in MMB [Mages, 2017]. As a result, loss of LIN37 causes mammalian cells to be unable to undergo cell cycle arrest [Mages, 2017] [Uxa, 2019]. Recently, it was also observed that LIN37 is required to prevent DNA end resection (the process of modifying dsDNA to change a 5` end to a 3` end that's single-stranded) in cells during G0 [Chen, 2021]. Moreover, in *C. elegans,* a LIN-37 protein null strain is the only MuvB subunit that remains fertile, even though it presents other common DREAM loss-of-function phenotypes [Harrison, 2006]. LIN37 interacts with both LIN9 and RBAP48 (LIN53 in *C. elegans*) through a CRAW domain [Asthana, 2022]. In plants it was observed that the LIN37 has an interaction with all other MuvB homologs except for the LIN9 homolog [Lang, 2021]. No interaction other than LIN9 and RBAP48 in LIN37 has been detected in animals.

1.2.3 LIN52

LIN52 in animals has been observed to have two functions that mediate DREAM or MMB assembly. The first is to mediate MuvB assembly into DREAM through its LxSxExL motif in mammals. DYRK1A phosphorylation of LIN52 S28 is required for the LxSxExL motif to bind to the pocket protein [Guiley, 2015]. However, in *C. elegans*, LIN-52 contains a stronger binding LxCxE motif but lacks a phosphorylation site for a DYRK1A-like enzyme to phosphorylate the region [Guiley, 2015]. *C. elegans* also does not appear to contain a DYRK1A homolog required for phosphorylation of LIN-52 [Litovchik, 2011]. What is interesting of *C. elegans* containing an LxCxE motif in LIN-52 is that the human papillomavirus (HPV) protein E7 outcompetes LIN52 with an LxCxE interaction motif, effectively sequestering the pocket protein and deactivating DREAM's repression of cell cycle genes [Guiley, 2015]. The second function is to bind to LIN9 as well as allow for B-Myb to bind during S-phase when DREAM is disassembled from MuvB [Guiley. 2018]. However, in plants, the homolog for LIN52 does not contain a similar LxCxE or LxSxExL motif that mediates MuvB association with the pocket protein [Lang, 2021]. Instead of the LIN52 homolog, the pocket-protein association motif was observed in the plant LIN54 homolog [Lang, 2021]. Like with *C. elegans* and the HPV E7 protein, the observed pocket protein-association motif in the plant LIN54 homolog is the LxCxE motif, but with a potential phosphorylation site [Lang, 2021]. The implications of these observed evolutionary changes of what MuvB subunit interacts with the pocket protein and with what strength of binding motif used across species has yet to be explored.

1.2.4 LIN54

MuvB binds to chromatin through LIN54 which contains a highly conserved DNA binding domain (DBD) that contains two CXC domains [Jiang, 2007]. In animals, LIN54 binds at specific DNA motifs called cell cycle genes homology regions (CHRs) which most commonly have the DNA consensus motif TTYRR [Marceau, 2016]. CHRs are located near the transcription start sites of cell cycle genes in G0 and G1 [Marceau, 2016]. A DNA motif like a CHR site has been identified in *Tetrahymena thermophila*, and the LIN54 homolog known as Anqa1 is known to bind to the CHR-like motif [Nabeel-Shah, 2021]. As discussed above, the plant LIN54 homolog also contains the LxCxE pocket protein-association interaction region [Lang, 2021]. However, little is known how both animal LIN54 and plant LIN54 interacts with other members of the MuvB subcomplex.

1.2.5 RBAP48

RBAP48 in humans (LIN-53 in *C. elegans*) is a histone binding protein [Zhang, 1998]. RBAP48 is highly conserved and is found in other chromatin binding complexes outside of DREAM, such as NuRD [Zhang, 1999]. While RBAP48 is known to be a part of the MuvB subcomplex, a recent discovery revealed that RBAP48 interacts with both LIN9 and LIN37 to bind to nucleosomes located downstream from CHR sites and near the transcription start sites of cell cycle genes [Asthana, 2022]. Structural analysis shows that RBAP48 contains multiple highly conserved WD-40 domains [Murzina, 2008]. WD40 repeats are short motifs of about 40 residues that end with a tryptophan-aspartic acid sequence (WD) [Neer, 1994]. Due to RBAP48 containing several WD-40 domains which are highly conserved and found in other protein complexes, protein conservation would be difficult to perform which is why we did not analyze RBAP48 in this study.

1.3 Pocket Protein

While the pocket protein, in humans p107 or p130, is not part of the MuvB subcomplex, MuvB does interact with the pocket protein to assemble into DREAM [Litovchick, 2007]. While phosphorylation of LIN52 of the LxSxExL region is required for DREAM assembly, hyperphosphorylation of the pocket protein leads to the disassembly of DREAM [Guiley, 2015]. Related to the pocket proteins is the well characterized retinoblastoma protein (pRb) which diverged from the ancestral pocket protein around the time of the emergence of sharks (*Callorhinchus milii*) [Liban, 2017]. The 3 proteins, pRb, p130, and p107, are members of the pocket protein family, each of which bind to E2F transcription factor proteins [Cobrinik, 2005]. Unlike pRb, p107 and p130 contain a binding site that mediates LIN52 association [Guiley, 2015]. Since pRb does not contain a similar interface, it does not bind to MuvB nor does it associate in DREAM [Litovchick, 2007]. Mice that have deficiencies in p107/p130 die shortly after birth due to skeletal defects in the skull [Forristal, 2014]. Elimination of the pocket protein LIN-35 in *C. elegans* leads to the disruption of MuvB chromatin occupancy and thus loss of DREAM target gene repression [Goetsch, 2017]. Disruption of only LIN-52 association with LIN-35, however, leads to a small decrease in transcriptional repression of some DREAM target genes and that DREAM is still able to assemble on chromatin [Goetsch,

2019]. Altogether, molecular analyses in *C. elegans* indicate that the pocket protein acts as a scaffold to stabilize MuvB function on chromatin and suppress MuvB's ability to switch to a transcriptional activator with B-Myb.

1.4 Protein Evolution

Looking at the broader picture of protein evolution, essential proteins are often tied to protein complexes [Hart, 2007]. In general, if a protein is essential for cellular or organismal health, then evolutionary pressure will preserve its amino acid sequence. Surprisingly, larger protein complexes are more likely to be essential and the subunits are more likely to have more co-complex interactions [Wang, 2009]. Additionally, the distribution of essential and non-essential proteins is not random within protein complexes but tends to cluster together as either all essential or all non-essential in protein complexes [Ryan, 2013]. When viewing from a disease perspective of protein complexes, proteins that belong to the same complex tend to show the same disease phenotype when defective [Oti, 2006] [Fraser, 2007]. In the case of the DREAM complex, because we know that the MuvB subcomplex is essential, we expect that MuvB is likely highly conserved in eukaryotes.

1.5 Study Hypothesis

The goal of this analysis was to explore the conservation of the DREAM complex, more specifically MuvB subunits and the pocket protein, to better understand how the complex has evolved. We hypothesize that the MuvB complex represents the ancestral form of the DREAM complex. Therefore, we expect that the core interaction surfaces within the MuvB complex will be the most conserved characteristic observed. To address our hypothesis, we performed protein sequence alignment of MuvB subunits mined from as many species as could be identified. We evaluated the conservation beyond the animal kingdom and asked whether homologs could be identified in different kingdoms. While sequence alignment of the MuvB subunits and the pocket protein has been done numerous times, previous published observations focused primarily on model organisms. By evaluating hundreds of homologous sequences, we performed a more detailed analysis aimed to identify conserved regions both known from structural studies as well as previously uncharacterized regions of conservation. We evaluated 4 of the 5 MuvB subunits and regions of importance within each subunit to identify conservation at the phylum level. We performed our focused analysis of the following subunits: LIN52, to evaluate the conservation of the known LIN52-pocket protein interaction site, LIN54, to evaluate the known MuvB-DNA interface, and LIN9 and LIN37, to evaluate for novel interaction interfaces with other MuvB subunits. Ultimately, this study will lay the groundwork for CRISPR/Cas9-mediated targeted mutagenesis in *C. elegans* to evaluate the importance of this study's findings.

2 Methods

Here we describe our pipeline for protein conservation analysis. Figure 2 outlines the pipeline that was developed and explains the flow.

2.1 Programs

To perform protein conservation analyses, we used the following programs:

ClustalO 1.2.4: ClustalO is a multiple sequence alignment software that uses HMM profiles to align multiple sequence with a high accuracy and high speed as well as generate percent identities [Sievers, 2013].

HMMER v 3.1b2: HMMER is a software program that makes use of Hidden Markov Models (HMM) to search against a database to find homologous proteins. By generating a profile HMM from a multiple sequence alignment of a set of proteins, the program assigns a score to each sequence hit as well as calculating E-values of scores through an algorithm [Mistry, 2016].

Jalview 2.11.2.0: Jalview is a bioinformatics software that can be used to view and edit MSAs. The software package also contains several modules to perform bioinformatic analysis. Here we used Jalview to view results of MSAs and generate consensus sequences [Waterhouse, 2009].

2.2 Initial Setup

We began the analysis by creating a directory with the following directories inside named “Alignments”, “Downloads”, “Databases”, “hmmsearchoutput”, “Initalproteins”, “HMMS”, “HITS”, “Sorted”, and “FinalIDs”.

We developed a total of 3 programs for the analysis. Program 1 and Program 2 were placed in the initial directory while the Perl script was to be placed in hmmsearchoutput directory so that the pipeline can run without issues.

2.3 Database Setup

To analyze from a wide selection of proteins across all forms of life, we created a database for the analysis that contains all annotated protein sequences of Uniprot. We went to https://beta.uniprot.org/help/downloads and used the command “wget” to download the .gz files of both the Reviewed (Swiss-Prot) and Unreviewed (TrEMBL) into the Databases directory and using the command “gunzip” to unzip the contents of the files. Once both files were unzipped, we used the command “cat” to combine both files into a single fasta file with the name uniprot_complete.fasta. The name of each file needs to match exactly for the analysis to function.

2.4 Creating a set of Initial Proteins for Analysis

We created set of 5-10 homolog proteins that were well annotated and diverse across the tree of life were collected. The proteins were collected in a .txt file named Initial{nameofyourprotein}.txt and placed in the Initalproteins directory.

2.5 First Program

Once the initial setup is done, the first program was run in the command line. The program starts by first asking for the name of the protein to be analyzed. Here, the following process took place automatically. After directly inputting the name of the protein, which needs to match exactly to the name given to the initial set of proteins from the previous step, a multi-sequence alignment (MSA) was performed on the initial set of proteins we had collected earlier. We used default settings in all instances, where ClustalO was used in this analysis. Next, the HHM model was produced using hmmbuild followed by searching against our created database using hmmsearch. Here we set the E-value to 1e-5 and set the parameter tblout. After searching against our database, the generated output file is processed in our developed perl script to filter out duplicate species and a set of commands to only select those that scored the lowest e-value in the analysis. Then a file containing the IDs is generated and placed within the "FinalIDs" directory.

2.6 Downloading Proteins found on Uniprot

After the first program was finished running and we have obtained the file in the "FinalIDs: directory, we uploaded the file to https://www.uniprot.org/uploadlists/ under "Choose File" with the options From:UniProtKB AC/IC To:UniProtKB selected. From there under "Download", the download all box was selected, Format set to FASTA (canonical), and uncompressed box also checked and a fasta file containing all the found protein sequences was downloaded. The file we downloaded was renamed {nameofprotein}All.fasta and placed in Downloads/{nameofprotein}/.

Next, we needed to create files that contained the protein sequences of only specific phylum or clades. To start. we downloaded a spreadsheet to better analyze the categories the IDs are part of. We did this by going under "Download", setting the Format to Excel and added the following columns: Taxonomic lineage (PHYLUM), Taxonomic lineage (CLASS), Taxonomic lineage (ORDER), Taxonomic lineage (all). The spreadsheet was then downloaded and from there we sorted IDs were sorted by the lineage by hand. We copied the IDs and uploaded them to Uniprot and downloaded them in the same way as the initial set of IDs. The naming format of the file downloaded we used was{nameofprotein}{nameofgroup}.fasta and the file place in the Downloads/{nameofprotein}/. The naming convention is important for the rest of the analysis below to function properly.

2.7 Creating a Segment to Analyze

Before running the second program, we created a fasta file containing a segment of a protein was created. The fasta file we made was placed in Alignments/{nameofprotein}/{nameofdomain}/ directory and named {nameofprotein}{nameofdomain}.fasta. Like before, the placement and naming convention is crucial for the analysis to function properly.

2.8 Second Program

Once we downloaded and placed all files in the correct directories, we ran the second program in the command line. The program asked for 3 inputs: name of the protein, name of the group, and the name of the domain which we inputted for each analysis. This generated a MSA of the segment created in the previous step with the sequences of the group which we can use in creating consensuses.

2.9 Creating a Consensus

We created a consensus of what the protein sequence was for each phylum or clade. To do this, we open the created fasta file generated in the previous step in Jalview. From there. we identified the region where the segment aligns with the rest of the sequences. It was possible that there is no alignment or a very poor alignment indicating that the region is not conserved within the group. In this case, we would not continue the analysis for those phylum or clades. If the segment did align, then we performed the following steps to identify the consensus sequence. First, we highlighted the region and select “hide all but selected region”. From there, we hid the segment sequence so that the segment is not calculated for the consensus. We obtained the consensus sequence by right clicking and selecting copy the consensus in the box that generates a consensus sequence in the program automatically and pasting the consensus in a .txt file with the name {nameofprotein}{nameofdomain}.txt. This was done so we can edit the name of the consensus to >{nameofgroup}. We repeated the analysis for each phylum or clade in the analysis. We also added the segment itself to the top line of the .txt file as well.

Once we generated the consensus for each phylum or clade and transferred the contents to the .txt file, we converted the .txt file into a .fasta file and perform an alignment using ClustalO and analyzed the results in Jalview.

We also generated percent identity matrixes by repeating the pervious ClustalO with the additional argument “--percent-id”.

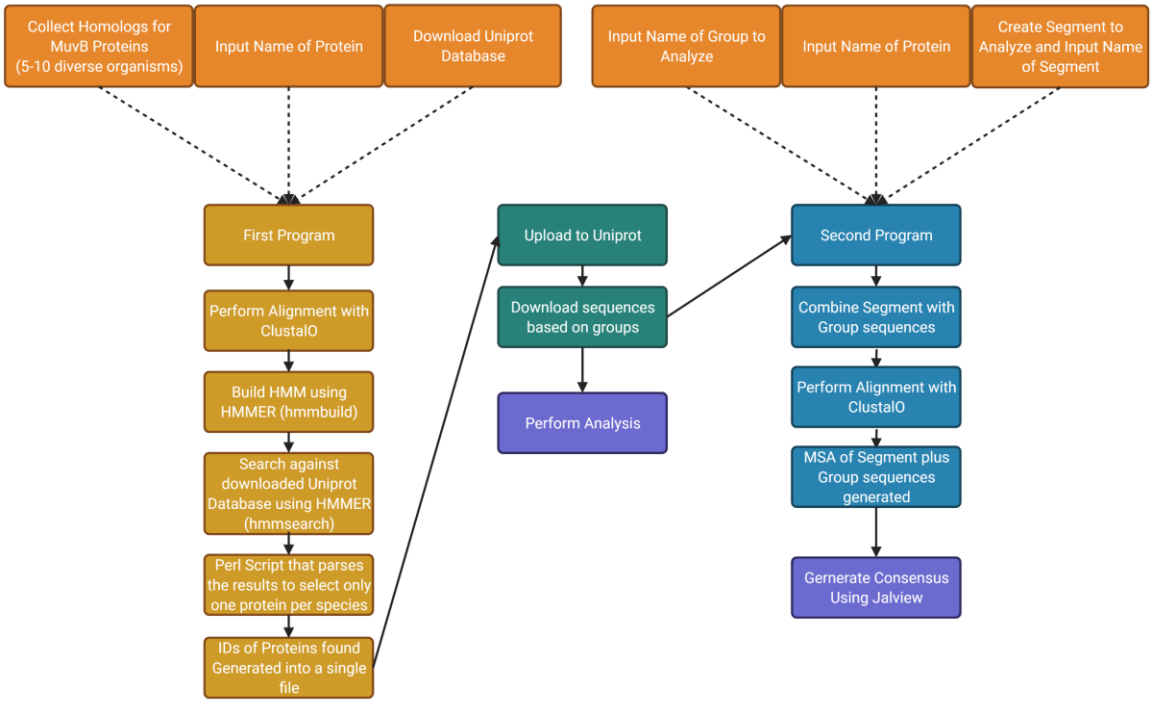

Figure 2 Pipeline of Protein Conservation Analysis

Outline of the pipeline used for the analysis performed. The flow starts at the top left. The orange boxes represent inputs or files needed to run each program. Yellow boxes represent the first program's flow. Green represent the Uniprot analysis that we did between Program 1 and Program 2. Blue boxes represent the second program's flow. Purple boxes represent analysis that we performed to generate our results used in our study.

Created with BioRender.com

3 Results

Our initial hypothesis for our conservation analysis is that the MuvB subcomplex represents the ancestral cell cycle repressor. Therefore, we expect that we will observe broader conservation in MuvB subunits across animal phyla and potentially between animal, plant, and fungi, as compared to the pocket protein. For our first analysis, we analyzed the conservation of the full-length proteins to describe how conserved each subunit is across the forms of life with annotated genomic information. Next, we were interested in whether our conservation analysis could reveal novel interaction sites between MuvB subunits. For our second analysis, we limited our conservation analysis to specific regions within each subunit. As a test case, we analyzed known interaction sites of the subunits before assessing for new regions of conservation. Our analysis lays the groundwork for future genetic analyses assessing DREAM and MuvB formation and function using the *Caenorhabditis elegans* model system.

3.1 Analysis of each MuvB subunit and Pocket Protein as a whole

To test our hypothesis that MuvB is the ancestral DREAM regulatory complex, we performed our protein conservation analysis pipeline on full-length human LIN9, LIN37, LIN52, LIN54 and the pocket proteins p130 and p107. We performed HMMER analysis to identify likely homologs across sequenced species and separated the sequences into phylum or clades using Uniprot's taxonomy classification. For each protein in the analysis, we recorded the total number of unique species with putative homologs in each phylum or clade (Figure 3). If multiple homologs for any subunit were identified in a species, then only the best match was added to the totals to avoid a single species over-representing a particular phylum in our analysis. The annotated proto-animal or protists kingdom categories, which are not proper taxonomic categories, were designated in our analysis for clarity of where each group is located phylogenetically and what they comprise of (Figure 2). We used Uniprot designations as the phylum for our analysis, this might cause some conflicts with what is proposed in the taxonomic field. However, the general structure of species relationships with each other is maintained in our analysis.

Chordata contained the most identified MuvB or pocket protein homologs of all represented phyla, as expected because Chordata genomes are a prime target for sequencing. Surprisingly, we observed that LIN37 homologs account for the fewest identified (Figure 3). Only 258 LIN37 homologs were detected, which accounts for half of the total observed in our analysis of the other MuvB subunits and the pocket protein (Figure 2). We suspect that our finding indicates a problem with LIN37 annotation in published Chordata genomes, as complete loss of LIN37 is unlikely due to its role in the repressive ability of MuvB in mammalian systems [Mages, 2017]. Our observation may also reflect that with the default HMMER settings are insufficient to identify putative homologs because the full-length protein itself is not as conserved as other MuvB subunits. However, in contrast, Arthropods are another phylum that is very well represented but doesn't display a similar drop out of identified putative LIN37 homologs

(Figure 3). Further investigation into why LIN37 drops out of Chordata only will need to be conducted to determine what is driving the inconsistency that we observed.

Overall, protein homologs of all tested full-length proteins were identified consistently across the animal kingdom, except for two phyla, Orthonectida and Tardigrada (Figure 3). In Orthonectida, represented by one species, only putative LIN9, LIN54, and pocket protein homologs were identified. In Tardigrada, represented by two species, only putative LIN54 and pocket protein homologs were identified. However, with only a few species representing both phyla, we expect that our result is due to incomplete annotation of the genomes.

Species representing in the proto-animal kingdom section contains only LIN9, LIN54, and the pocket protein (Figure 3). Phylogenetically located between fungi and animals, these organisms represent an important outgroup that provide explanations for how animals evolved. Placing each organism into a group was difficult because no species in the proto-animal kingdom have a designated phylum, so we grouped by class. Unfortunately, only a total of 5 species was observed in our analysis. Our analysis does give insight into what are potentially the most important subunits in the MuvB subcomplex. Perhaps as more genomes are sequenced and annotated, a new analysis of this kingdom of organisms will gain insight into the key evolutionary divergences of the MuvB subcomplex in animals

Like with the proto-animal kingdom, protists are not a true kingdom but is used to help clarify where these organisms are in relation other kingdoms. When we analyzed the protist kingdom, we detected the subunits LIN9, LIN37, LIN54, and the pocket protein (Figure 3). However, only one species had LIN37, and each phylum had varying representation of each MuvB subunit. The most consistent clades were Oomycote, Ciliophoran, and Stramenopiles (Figure 3). We noted that LIN54 and the pocket protein appear to be the most constant appearing subunits (Figure 3). Given the varying degree of representation of the kingdom, it's difficult to make conclusions about the MuvB subcomplex evolution at this stage until we further investigate each protein on a deeper level.

When we reviewed the amoeboa kingdom, we observed very few hits (7 in total for LIN9, 1 for LIN52, 9 for LIN54, 9 for the pocket protein, and none for LIN37) (Figure 3). The phylum that was most represented was Evosea in comparison to Tubulinea and Discosea. The lack of species detected in amoeboa does present problems in a deeper analysis at this stage. It is possible that all MuvB subunits are present but can't be detected due to inconsistent genomic annotation in the amoeboa kingdom.

Finally, the plant kingdom contains sequences for all tested subunits except LIN52 (Figure 3). Notably, LIN37 was observed in fewer species compared to the other subunits, with 20 total putative LIN37 homologs identified. All LIN37 homologs were identified within the phylum Streptophyta. Much like Chordata and Arthropoda, Streptophyta contains a much higher species count than the other phyla in the plant kingdom (Figure 3). Because of LIN37 being notably sparsely represented and LIN52

being absent, one thought is that LIN37 and LIN52 might be later evolved proteins. Another possibility is that the homology of these subunits is too far removed to be detected in our analysis or not identified in our analysis because of incomplete genomic annotations.

In fungi, no MuvB subunits where detected. However, we did identify the pocket protein in a few fungi species. It should be noted that the kingdom of fungi is between proto-animals and protists but because only the pocket protein was found, we opted to separate the table for clarity (Figure 3). The appearance of the pocket protein might indicate that a protein complex like MuvB exists in fungi, but the sequence identify of the subunits is highly diverged. Further study will need to be done to locate a MuvB-like subcomplex in fungi and by extension the DREAM complex.

Overall, these data indicate that our initial hypothesis that MuvB represents the ancestral complex is not supported, as the pocket protein remains as well conserved as portions of the MuvB complex. The most highly conserved subunit in the MuvB complex is the LIN54 DNA-binding protein, which may reflect that the DNA-binding domain in LIN54 diverges more slowly over evolutionary time. This is in comparison to the other MuvB subunits that do not contain highly defined protein domains like a DNA-binding domain. We can infer from our analysis that the LIN9 and LIN54 subunits may represent the ancestral MuvB proteins. It is possible that the ancestral MuvB contained these subunits plus the histone binding protein (RBPA48) to create a core regulatory complex with later subunits being added over time. It is possible that the DREAM complex is also an ancient construct as the pocket protein does appear consistently as well. Whether the function of DREAM is consistent throughout evolution is a question that requires further research.

MuvB and Pocket Protein Conservation Analysis

Kingdom	Clade	Phylum	Total Unique	LIN9	LIN37	LIN52	LIN54	Pocket Protein
Animals	Deuterostc	Chordata	604	526	258	540	558	589
		Echinodermata	2	2	2	2	2	2
	Protostomi	Annelida	3	3	3	3	3	3
		Arthropoda	250	220	210	208	216	229
		Brachiopoda	1	1	1	1	1	1
		Bryozoa	1	1	1	1	1	1
		Mollusca	12	12	10	8	11	12
		Nematoda	77	75	64	65	76	76
		Orthonectida	1	1	0	0	1	1
		Platyhelminthes	30	30	27	25	29	30
		Rotifera	10	10	2	9	9	10
		Tardigrada	2	0	0	0	2	2
	Early Anima	Cnidaria	10	7	7	7	8	10
		Placozoa	2	2	2	2	2	2
		Porifera	1	1	1	1	1	1
Proto-Animals		Choanoflagellata*	2	2	0	0	2	2
		Filasterea*	1	1	0	0	1	1
		Rotosphaerida*	1	1	0	0	1	0
		Apusozoa*	1	1	0	0	1	1
Protists	SAR	Oomycota	35	21	0	0	34	35
		Alevolata*	3	3	0	0	0	0
		Cercozoa	3	0	0	0	3	0
		Endomyxa	1	0	0	0	1	0
		Ciliophora	13	3	0	0	10	10
		Bacillariophyta	21	7	0	0	17	10
		Stramenopiles*	23	5	1	0	19	8
	Excavata	Heterolobosea*	2	2	0	0	2	2
		Parabasalia	1	0	0	0	1	0
	Other	Cryptophyceae*	7	1	0	0	7	4
		Haptista*	8	5	0	0	7	7
Amoeboa		Evosea	6	5	0	0	6	6
		Tubulinea	1	0	0	0	1	1
		Discosea	3	2	0	1	2	2
Plants		Streptophyta	232	206	20	0	212	219
		Rhodophyta	10	3	0	0	10	9
		Chlorophyta	47	32	0	0	38	36
		Prasinodermophyta	2	2	0	0	1	1
Fungi		Chytridiomycota	6	0	0	0	0	6
		Cryptomycota	2	0	0	0	0	2
		Mucoromycota	2	0	0	0	0	2
		Ichthyosporea*	1	0	0	0	0	1

Figure 3 Table of our MuvB and Pocket Protein Conservation Analysis

Here we set out to determine the total amount of each MuvB subunit plus the pocket protein could be identified with our protein conservation analysis. The results were sorted by either phylum or closest clade as determined by the Uniport taxonomy. Each phylum or clade was then grouped into different kingdoms. While protists are no longer a wildly used kingdom and proto-animals is not a technical kingdom, these terms are used to describe each of the phylum or clades in relation to one another. Colors in the table are used to separate the kingdoms for better visualization. Red indicating animals, red-orange indicating proto-animals, yellow indicating protests, blue indicating amoeboa, green indicating plants, and purple indicating fungi. A gap is place between plants and fungi for clarity. An asterisk (*) indicates groups chosen that are not a phylum, which include either a class, order, or clade. Another note is that the clade description “early animals” is used to differentiate phylum that evolutionally appear before Deuterostomia and Protostomia.

3.2 Percent Identity of Regions of Interests

We next set out to address the question as to whether the core interaction surfaces within the MuvB complex and between the MuvB complex and the pocket protein were the most conserved characteristics observed in our protein conservation analysis. To address this question, we evaluated 7 known interaction sites, identified 2 potentially new interaction sites, and evaluated a recently discovered *Arabidopsis thaliana* interaction site. To gain insight into how well each region of interest remained conserved across species and phyla, we evaluated each interaction based on percent identity score compared to the human sequence (Figures 4, 5, 6, 7, and 8).

The known interaction sites observed in LIN9 include LIN9 93-129, which mediates the LIN9-LIN37 and LIN9-RBAP48 associations [Asthana, 2022]), LIN9 163-214, which also mediates LIN9-RBAP48 association and has an indirect association of LIN37 [Asthana, 2022], and LIN9 338-412, which mediates the LIN9-LIN52 and LIN9-B-Myb associations [Guiley, 2018] (Figure 4). We observed that each region is similarly conserved across phyla, except that the LIN9-LIN52 interaction site shows markedly lower conservation in comparison to the LIN9-LIN37 interaction sites. A more detailed analysis will be described in section 3.3.1 and sections 3.3.2. Overall, our findings suggest that the MuvB subcomplex maintains conservation in key interaction regions past the animal kingdom, including interaction regions in subunits that we could not detect homologs past animals.

The known interaction site observed in LIN37 includes LIN37 95-126, which mediates the LIN37-LIN9 and LIN37-RBAP48 associations [Asthana, 2022] (Figure 5). We also observed two potentially new interaction regions in LIN37 1-43 and LIN37 203-246. We observed that the known interaction site shows more robust conservation compared to the two potentially new interaction regions. A more detailed analysis will be described in section 3.3.1 and 3.4. Overall, our findings suggest that both LIN37 1-43 and LIN37 203-246 are regions that contain unknown interaction sites that maintain conservation throughout animals with LIN37 203-246 also maintaining conservation into plants as well.

The known interaction sites observed in LIN52 include LIN52 17-45, which mediates the LIN52-pocket protein association [Guiley, 2015]), and LIN52 68-113, which mediates LIN52-LIN9 and LIN-52-B-Myb associations [Guiley, 2018] (Figure 6). We also evaluated the pocket protein-LIN52 interaction surface (Figure 7). We observed that the pocket protein interaction surface appears to be conserved in not only the animal kingdom but also within many phyla in proto-animals, protists, amoeboa, plants, and surprisingly fungi. A more detailed analysis will be described in section 3.3.2 and 3.3.3. Overall, our findings suggest that the MuvB subcomplex outside of animals has a similar mechanism to interact with the pocket protein and this extends not only into plants but also fungi which is surprising as fungi do not have appear to have any MuvB subunit homologs.

Finally, the *Arabidopsis thaliana* LIN54 homolog TCX5 LxCxE site is believed to bind to the pocket protein in plants (Figure 8). We observed that, LIN54 homologs containing LxCxE appear in other phyla outside of *A. thaliana* (Streptophyta) and extends into the kingdoms of protist, amoeboa, and proto animals. A more detailed analysis will be described in section 3.3.2 and 3.3.3. Overall, our findings suggest that MuvB interacts with the pocket protein through mediation of LIN54 homologs' LxCxE motif until the emergence of animals where the role shifts to LIN52.

LIN9 Percent Identity By Region

Kingdom	Clade	Phylum	LIN9 94-130	LIN9 163-214	LIN9 338-412
Animals	Deurterostomia	Chordata	94.44	98.04	96
		Echinodermata	77.14	72	61.33
	Protostomia	Annelida	72.22	72.55	36
		Arthropoda	77.78	76.47	56
		Brachiopoda	75.68	75.51	62.67
		Bryozoa	65.71	70	55.41
		Mollusca	81.08	75.51	65.33
		Nematoda	50	48	41.33
		Orthonectida	36.67	36.73	24.64
		Platyhelminthes	62.86	58	42.67
		Rotifera	38.89	39.58	36
		Tardigrada	-	-	-
	Early Animals	Cnidaria	55.56	76.47	69.33
		Placozoa	75	70.83	-
		Porifera	40	42.86	38.67
Proto-Animals		Choanoflagellata*	32.14	34	-
		Filasterea*	36.11	43.75	25.33
		Rotosphaerida*	33.33	30.43	23.08
		Apusozoa*	25.71	37.5	-
Protists	SAR	Oomycota	20.69	38.46	-
		Alevolata*	-	-	-
		Cercozoa	-	-	-
		Endomyxa	-	-	-
		Ciliophora	26.09	38.78	18.18
		Bacillariophyta	34.29	34.69	15.28
		Stramenopiles*	23.53	45.95	21.43
	Excavata	Heterolobosea*	26.47	33.33	22.67
		Parabasalia	-	-	-
	Other	Cryptophyceae*	20	29.41	-
		Haptista*	41.18	40	-
Amoeboa		Evosea	63.64	44.19	28
		Tubulinea	-	-	-
		Discosea	37.14	39.13	31.08
Plants		Streptophyta	32.43	37.5	25
		Rhodophyta	31.43	47.5	-
		Chlorophyta	27.27	45.16	25.33
		Prasinodermophyta	20.69	36.17	22.64

Figure 4 LIN9 Percent Identity Table based of off Human LIN9

Percent identity table for each of the LIN9 regions that we examined. Each phylum's consensus for each LIN9 region was compared to the human LIN9 region being analyzed. The results were sorted by either phylum or closest clade as determined by the Uniport taxonomy. Each phylum or clade was then grouped into different kingdoms. While protists are no longer a wildly used kingdom and proto animals is not a technical kingdom, these terms are used to describe each of the phylum or clades in relation to one another. Colors in the table are used to separate the kingdoms for better visualization. Red indicating animals, red-orange indicating proto-animals, yellow indicating protests, blue indicating amoeboa, green indicating plants, and purple indicating fungi. A gap is place between plants and fungi for clarity. An asterisk (*) indicates groups chosen that are not a phylum, which include either a class, order, or clade. Another note is that the clade description "early animals" is used to differentiate phylum that evolutionally appear before Deuterostomia and Protostomia

LIN37 Percent Identity By Region

Kingdom	Clade	Phylum	LIN37 1-43	LIN37 95-126	LIN37 203-246
Animals	Deurterostomia	Chordata	86.05	81.25	100
		Echinodermata	28.21	71.88	44.19
	Protostomia	Annelida	19.05	75	45.45
		Arthropoda	30.95	84.38	29.55
		Brachiopoda	28.57	84.38	41.86
		Bryozoa	21.43	78.12	21.74
		Mollusca	30.23	78.12	51.16
		Nematoda	15.38	47.83	20.93
		Orthonectida	-	-	-
		Platyhelminthes	32.56	65.62	34.88
		Rotifera	0	31.25	23.26
		Tardigrada	-	-	-
	Early Animals	Cnidaria	25	65.62	38.64
		Placozoa	18.18	71.88	32.56
		Porifera	25.81	31.25	29.63
Proto-Animals		Choanoflagellata*	-	-	-
		Filasterea*	-	-	-
		Rotosphaerida*	-	-	-
		Apusozoa*	-	-	-
Protists	SAR	Oomycota	-	-	-
		Alevolata*	-	-	-
		Cercozoa	-	-	-
		Endomyxa	-	-	-
		Ciliophora	-	-	-
		Bacillariophyta	-	-	-
		Stramenopiles*	16.67	34.38	23.26
	Excavata	Heterolobosea*	-	-	-
		Parabasalia	-	-	-
	Other	Cryptophyceae*	-	-	-
		Haptista*	-	-	-
Amoeboa		Evosea	-	-	-
		Tubulinea	-	-	-
		Discosea	-	-	-
Plants		Streptophyta	13.04	42.86	25.58
		Rhodophyta	-	-	-
		Chlorophyta	-	-	-
		Prasinodermophyta	-	-	-

Figure 5 LIN37 Percent Identity Table based of off Human LIN37

Percent identity table for each of the LIN37 regions that we examined. Each phylum's consensus for each LIN37 region was compared to the human LIN37 region being analyzed. The results were sorted by either phylum or closest clade as determined by the Uniport taxonomy. Each phylum or clade was then grouped into different kingdoms. While protists are no longer a wildly used kingdom and proto-animals is not a technical kingdom, these terms are used to describe each of the phylum or clades in relation to one another. Colors in the table are used to separate the kingdoms for better visualization. Red indicating animals, red-orange indicating proto-animals, yellow indicating protests, blue indicating amoeboa, green indicating plants, and purple indicating fungi. A gap is place between plants and fungi for clarity. An asterisk (*) indicates groups chosen that are not a phylum, which include either a class, order, or clade. Another note is that the clade description "early animals" is used to differentiate phylum that evolutionally appear before Deuterostomia and Protostomia.

LIN52 Percent Identity By Region

Kingdom	Clade	Phylum	LIN52 17-45	LIN52 68-113
Animals	Deurterostomia	Chordata	93.1	84.31
		Echinodermata	71.43	77.08
	Protostomia	Annelida	85.71	68.75
		Arthropoda	72.41	54.17
		Brachiopoda	85.71	86.27
		Bryozoa	73.68	72.92
		Mollusca	80.95	82.35
		Nematoda	38.1	43.75
		Orthonectida	-	-
		Platyhelminthes	67.86	59.57
		Rotifera	-	39.58
		Tardigrada	-	-
	Early Animals	Cnidaria	80.95	84.31
		Placozoa	61.9	46.81
		Porifera	40	50
Proto-Animals		Choanoflagellata*	-	-
		Filasterea*	-	-
		Rotosphaerida*	-	-
		Apusozoa*	-	-
Protists	SAR	Oomycota	-	-
		Alevolata*	-	-
		Cercozoa	-	-
		Endomyxa	-	-
		Ciliophora	-	-
		Bacillariophyta	-	-
		Stramenopiles*	-	-
	Excavata	Heterolobosea*	-	-
		Parabasalia	-	-
	Other	Cryptophyceae*	-	-
		Haptista*	-	-
Amoeboa		Evosea	-	-
		Tubulinea	-	-
		Discosea	47.62	37.5
Plants		Streptophyta	-	-
		Rhodophyta	-	-
		Chlorophyta	-	-
		Prasinodermophyta	-	-

Figure 6 LIN52 Percent Identity Table based of off Human LIN52

Percent identity table for each of the LIN52 regions that we examined. Each phylum's consensus for each LIN52 region was compared to the human LIN52 region being analyzed. The results were sorted by either phylum or closest clade as determined by the Uniport taxonomy. Each phylum or clade was then grouped into different kingdoms. While protists are no longer a wildly used kingdom and proto-animals is not a technical kingdom, these terms are used to describe each of the phylum or clades in relation to one another. Colors in the table are used to separate the kingdoms for better visualization. Red indicating animals, red-orange indicating proto-animals, yellow indicating protests, blue indicating amoeboa, green indicating plants, and purple indicating fungi. A gap is place between plants and fungi for clarity. An asterisk (*) indicates groups chosen that are not a phylum, which include either a class, order, or clade. Another note is that the clade description "early animals" is used to differentiate phylum that evolutionally appear before Deuterostomia and Protostomia.

LIN54 Percent Identity By Region

Kingdom	Clade	Phylum	LIN54 LxCxE
Animals	Deurterostomia	Chordata	-
		Echinodermata	-
	Protostomia	Annelida	-
		Arthropoda	-
		Brachiopoda	-
		Bryozoa	38.46
		Mollusca	-
		Nematoda	-
		Orthonectida	-
		Platyhelminthes	-
		Rotifera	-
		Tardigrada	-
	Early Animals	Cnidaria	-
		Placozoa	-
		Porifera	-
Proto-Animals		Choanoflagellata*	
		Filasterea*	46.15
		Rotosphaerida*	-
		Apusozoa*	-
Protists	SAR	Oomycota	38.46
		Alevolata*	-
		Cercozoa	-
		Endomyxa	69.23
		Ciliophora	-
		Bacillariophyta	42.86
		Stramenopiles*	53.85
	Excavata	Heterolobosea*	57.14
		Parabasalia	-
	Other	Cryptophyceae*	60
		Haptista*	66.67
Amoeboa		Evosea	57.14
		Tubulinea	-
		Discosea	-
Plants		Streptophyta	76.92
		Rhodophyta	57.14
		Chlorophyta	71.43
		Prasinodermophyta	53.85

Figure 7 LIN54 Percent Identity Table based of off *Arabidopsis thaliana* TCX5

Percent identity table of the LIN54 region that we examined. Each phylum's consensus for each LIN54 region was compared to the *Arabidopsis thaliana* TCX5 region being analyzed. The results were sorted by either phylum or closest clade as determined by the Uniport taxonomy. Each phylum or clade was then grouped into different kingdoms. While protists are no longer a wildly used kingdom and proto-animals is not a technical kingdom, these terms are used to describe each of the phylum or clades in relation to one another. Colors in the table are used to separate the kingdoms for better visualization. Red indicating animals, red-orange indicating proto-animals, yellow indicating protests, blue indicating amoeboa, green indicating plants, and purple indicating fungi. A gap is place between plants and fungi for clarity. An asterisk (*) indicates groups chosen that are not a phylum, which include either a class, order, or clade. Another note is that the clade description "early animals" is used to differentiate phylum that evolutionally appear before Deuterostomia and Protostomia.

Pocket Protein Percent Identity By Region

Kingdom	Clade	Phylum	PP LxCxE
Animals	Deurterostomia	Chordata	77.14
		Echinodermata	-
	Protostomia	Annelida	68.57
		Arthropoda	71.43
		Brachiopoda	71.43
		Bryozoa	62.86
		Mollusca	68.57
		Nematoda	57.14
		Orthonectida	60
		Platyhelminthes	30
		Rotifera	24.24
		Tardigrada	51.43
	Early Animals	Cnidaria	71.43
		Placozoa	60
		Porifera	-
Proto-Animals		Choanoflagellata*	-
		Filasterea*	31.43
		Rotosphaerida*	-
		Apusozoa*	50
Protists	SAR	Oomycota	20
		Alevolata*	-
		Cercozoa	-
		Endomyxa	-
		Ciliophora	33.33
		Bacillariophyta	-
		Stramenopiles*	28.57
	Excavata	Heterolobosea*	20.83
		Parabasalia	-
	Other	Cryptophyceae*	-
		Haptista*	34.48
Amoeboa		Evosea	28.57
		Tubulinea	37.93
		Discosea	29.03
Plants		Streptophyta	44.83
		Rhodophyta	32.14
		Chlorophyta	40
		Prasinodermophyta	42.86
Fungi		Chytridiomycota	26.92
		Cryptomycota	-
		Mucoromycota	30.77
		Ichthyosporea*	33.33

Figure 8 Pocket Protein Percent Identity Table based of off Human p107

Percent identity table of the pocket protein region of the that we examined. Each phylum's consensus for the pocket protein region was compared to the human p107 region being analyzed. The results were sorted by either phylum or closest clade as determined by the Uniport taxonomy. Each phylum or clade was then grouped into different kingdoms. While protists are no longer a wildly used kingdom and proto-animals is not a technical kingdom, these terms are used to describe each of the phylum or clades in relation to one another. Colors in the table are used to separate the kingdoms for better visualization. Red indicating animals, red-orange indicating proto-animals, yellow indicating protests, blue indicating amoeboa, green indicating plants, and purple indicating fungi. A gap is place between plants and fungi for clarity. An asterisk (*) indicates groups chosen that are not a phylum, which include either a class, order, or clade. Another note is that the clade description "early animals" is used to differentiate phylum that evolutionally appear before Deuterostomia and Protostomia.

3.3 Known MuvB interactions

We continued addressing the important question as to how the core interaction surfaces known in MuvB remain conserved. We used multiple sequence alignment of consensus sequences observed in species within each phylum analyzed in Figure 3. We split our alignments to view only the conservation of amino acids within the known protein interaction sequences described above in section 3.2. Altogether, our analysis addressed how known interactions between MuvB subunits and MuvB and the pocket protein diverge or remain conserved across species represented all major kingdoms of eukaryotes.

3.3.1 LIN9-LIN37 Interaction

We first chose to evaluate the LIN9-LIN37 interaction sites LIN9 93-129, 163-214, and LIN37 95-126. Using the recent structural data that discovered the LIN9-LIN37 interaction sites, we isolated the region of interest using the human LIN9 and LIN37 amino acid residues. We generated a consensus sequence from protein sequences from each species representing each phylum or clade outlined in Figure 3. Using the consensus sequence, we generated a multi-sequence alignment for each region of interest.

As noted in section 3.2, the percent identity of each region in LIN9 drops from animal into proto-animals (Figure 4). However, Porifera and proto-animals were around the same percent identity for LIN9 regions. Interestingly, the Evosea phylum in amoeboa kingdom was slightly more conserved, as compared to other non-animals at 63.64% for LIN9 94-130 and 44.19% for LIN9 163-214. Each of the non-animal kingdoms are around the same percent identity for each LIN9 region.

The percent identity profile of LIN37 96-126 followed a similar pattern (Figure 5). Unfortunately, only 2 phyla outside the animal kingdom contain an observed putative LIN37 homolog (Figure 5). We observed that the percent identity remains at 34.38% in Stramenopiles and 42.86% in Streptophyta, indicating this region is conserved. Altogether, these data suggest that that conservation is maintained into the plant and protist kingdoms. Not only is LIN37 found outside of the animal kingdom but the interaction with LIN9 has been maintained from plants to animals.

We noted that 3 animal phyla are outliers within the animal kingdom, specifically Nematoda and Rotifera for both LIN9 and LIN37 and Orthonectida for LIN9. Their percent identities for each region in LIN9 and the region in LIN37 is lower than the rest of the animal kingdom. The percent identities for LIN9 94-130 are 50% for Nematoda, 38.89% for Rotifera, and 36.67% for Orthonectida. For LIN9 163-214, 48% for Nematoda, 39.58% for Rotifera, and 36.73% for Orthonectida. Finally, the percent identities for LIN37 95-126 are 47.83% for Nematoda and 31.25% for Rotifera. This suggests that these three phyla might have undergone more rapid mutations as compared to other animal phyla resulting in more divergences in these regions.

To better understand the amino acid conservation or divergence within these regions, we examined the consensus sequence within each region known to mediate LIN9-LIN37

association. First, we examined the LIN9 93-129 region of interest and observed that region remained relatively conserved across all phyla (Figure 9). Mutations in human LIN9 E125, W126, and F127 resulted in the LIN9-LIN37 interaction being lost [Asthana, 2022]. Most of sequences in each phylum retained all 3 amino acids, with the exception of Apusozoa and Cryptophyceae. 10 phyla have a phenylalanine (F) instead of a tryptophan (W); however, this is observed across different kingdoms and not confined to one kingdom. Additional interaction studies will be necessary to determine if W126F is sufficient to retain LIN9-LIN37 association.

Next, we examined the LIN9 163-214 region of interest (Figure 10). Within this sequence, the important residues that may mediate the LIN9-LIN37 association include R174, R175, F180, and F181 [Asthana, 2022]. All the important interacting residues appear to be highly conserved across all phyla. F181 is the least conserved, as we observe its divergence outside the animal kingdom, with some exemptions in Nematoda, Orthonectida, and Rotifera. Rotifera is unique in that both F180 and F181 appear to be absent. Additional interaction studies will be necessary to determine if the divergences observed in these 3 phyla are sufficient to disrupt LIN9-LIN37 association.

Finally, we examined the reciprocal interaction surface in LIN37 95-126 that mediates LIN9-LIN37 association (Figure 11). Here the important residues that mediate the LIN9-LIN37 interaction are I97, L99, K100, V104, L106, F109, L115, Y116, I118, and W122. In contrast to our LIN9 region of interest conservation analysis, about half of the important residues were present across all the phyla. The amino acid sequence diverged around porifera. Whether the non-conserved regions are necessary for the interaction of LIN37 to LIN9 is unclear. However, given that the remaining half were observed in Porifera, Stramenopiles, and Streptophyta, we conclude that the interaction surface on LIN37 remains highly conserved. Altogether, our results indicate that the residues known to mediate the interaction between LIN9 and LIN37 are well conserved across all phyla in our analysis.

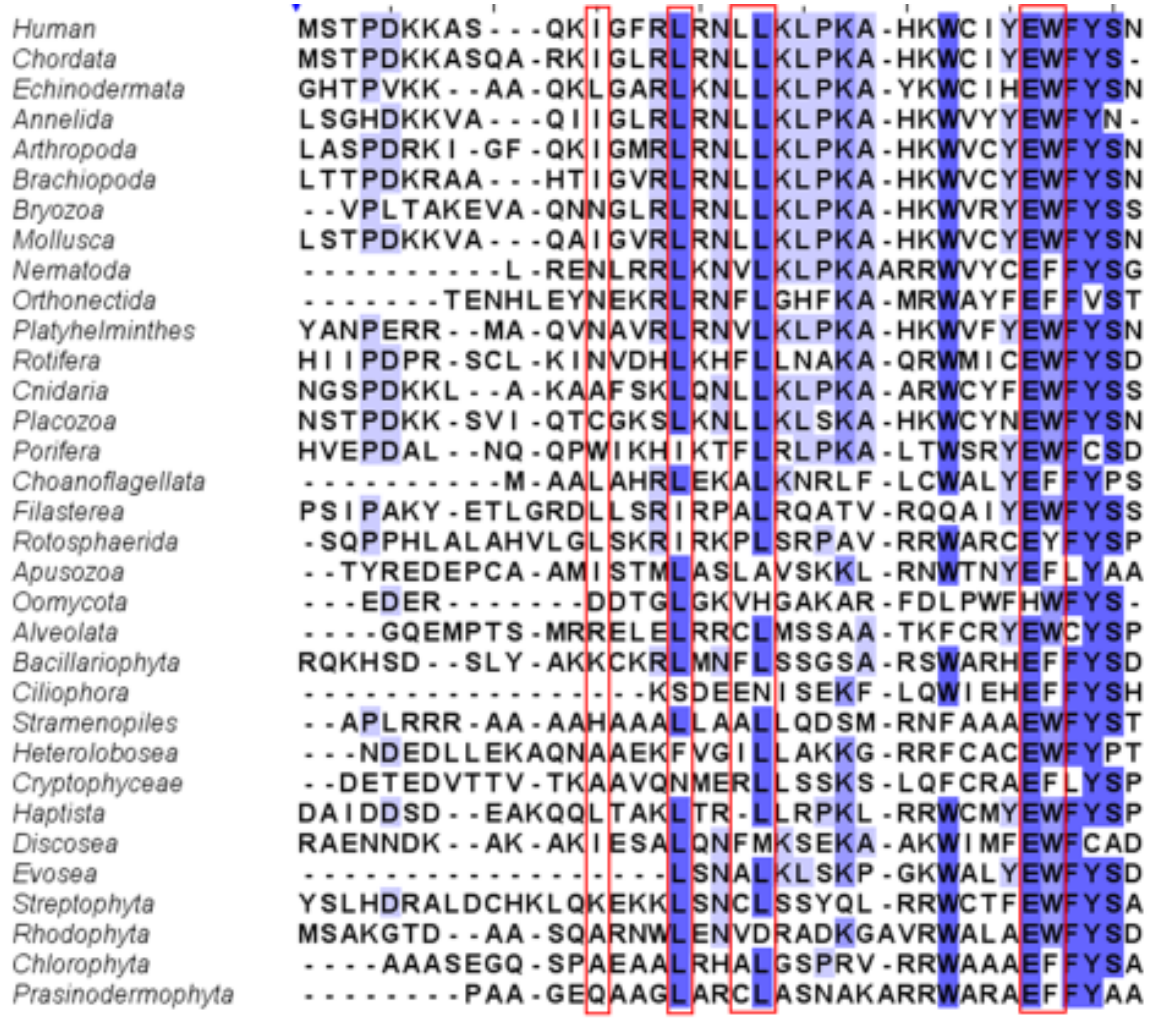

Figure 9 LIN9 Conservation of Human 93-129

Consensus of the human 93-129 LIN9 region where LIN37 interacts. The human region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. Important residues necessary for the LIN9-LIN37 interaction are shown in red boxes

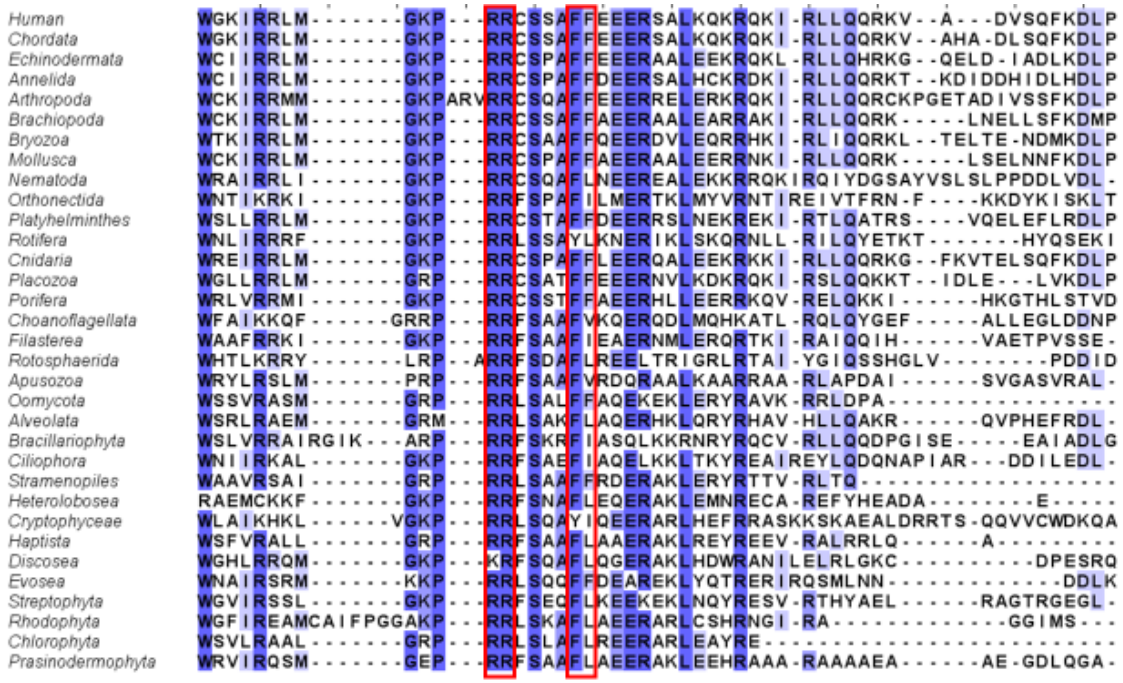

Figure 10 LIN9 Conservation of Human 163-214

Consensus of the human 163-214 LIN9 region where disruption of the region was known to inpact the binding of LIN37 to LIN9. The human region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. Important residues necessary for the LIN9-LIN37 interaction are shown in red boxes

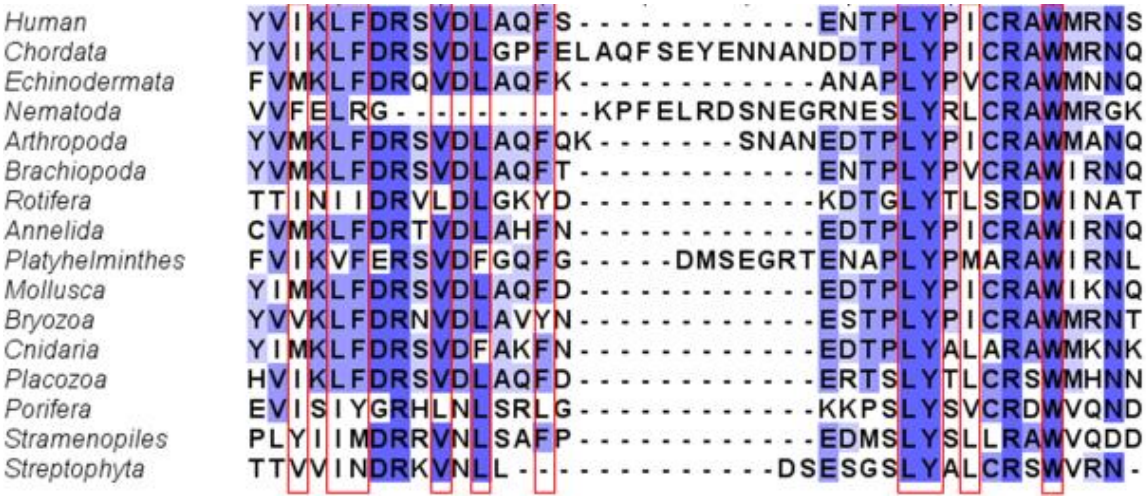

Figure 11 LIN37 Conservation of Human 95-126

Consensus of the human 95-126 region where LIN37 interacts with LIN9. The human region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. Important residues necessary for the LIN37-LIN9 interaction are shown in red boxes.

3.3.2 LIN9-LIN52 Interaction

Next, we analyzed the LIN9-LIN52 interaction sites, including the regions of interest LIN9 338-412 and LIN52 68-113. These regions also include a critical interaction site for B-Myb to bind to MuvB [Guiley, 2018].

Surprisingly, LIN9 338-412 was observed to be far less conserved than the LIN9-LIN37 regions of interest, as based on total percent identity (Figure 4). Outside of animals, the highest percent identity observed was in the amoeboa phylum Discosea with 31.08%. Many of the phyla in our complete analysis lacked any consensus in the region. However, in plants, the percent identity was observed to be between 25.33% to 22.64%. While plants are still not as well conserved compared to our previous analysis, 25.33% and 22.64% still indicates that conservation still does exist. For LIN52 68-113, few putative homologs were observed outside the animal kingdom. In contrast to the LIN9 interaction region, LIN52 68-113 did not drop below 37.5%, indicating that the region maintains a higher level of conservation (Figure 6). Surprisingly, Nematoda was less conserved compared to Porifera (43.75% vs 50%).

We first examined the LIN9 338-412 region of interest (Figure 12). Many residues remained conserved across all phyla in our study, including plants. The most important residue which have direct contact with LIN52 in human are K356, N367, E371, Q385, Y388, and A389 [Guiley, 2018]. Other than Q385, all important residues remained highly conserved and are present in every kingdom with only one or two phyla not conserving the region per residue. This suggests that the LIN9 interaction with LIN52 is maintained from plants to animals. Even though the percent identity of the region was low, on closer examination of the key residues required for the interaction, we observed a high level of conservation. A reason for the lower percent identity could be that the size of the region selected being quite large is a driving factor in the overall calculation.

We next examined the LIN52 68-113 region of interest (Figure 13). Overall, the region was relatively well conserved with the exception being from the putative LIN52 homolog observed from amoeboa. The 2 primary LIN52 residues that mediate the LIN9-LIN52 associate include G95 and E98 [Guiley, 2018]. Both important residues were observed in every phylum in our analysis. There are also many other residues that were also well conserved and may be important mediators of other interactions within the complex. These data suggests that the conservation of the LIN52 68-113 where LIN9 interacts is maintained in animals and possibly beyond. The reason why we observed such a high conservation of the LIN52 68-113 region based on percent identity is likely because our analysis did not detect homologs outside of the animal kingdom other than one in the amoeboa phylum Discosea.

While this region is also an interface for B-Myb to bind to as well, however the MuvB binding domain in B-Myb is not conserved past animals [Guiley, 2018] [Vorster, 2020]. Given the importance of the conserved region, it’s a surprise to find LIN9 maintain conservation into plants as well. While our analysis did not find LIN52 in plants, it was recently discovered that a homolog for LIN52 in plants does exist and the conserved

region detected is the LIN9-B-Myb interface [Lang, 2021]. We speculate that both LIN9 and LIN52 maintain each of their interface's conservation to allow for the two proteins to bind to each other. What that means is that B-Myb did not co-evolve at the same time to interface with LIN9 and LIN52 but instead evolved separately with the emergence of animals to be able to form the complex MMB after DREAM is disassembled. There is also the possibility of a completely different protein that interfaced to the region but was later replaced by B-Myb.

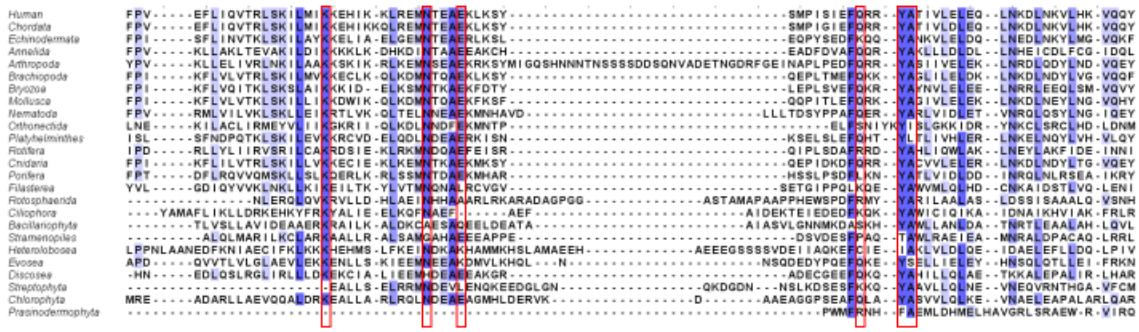

Figure 12 LIN9 Conservation of Human 338-412

Consensus of human LIN9 338-412 region where LIN9 interacts with LIN52. The human region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. Important residues necessary for the LIN52-LIN9 interaction are shown in red boxes

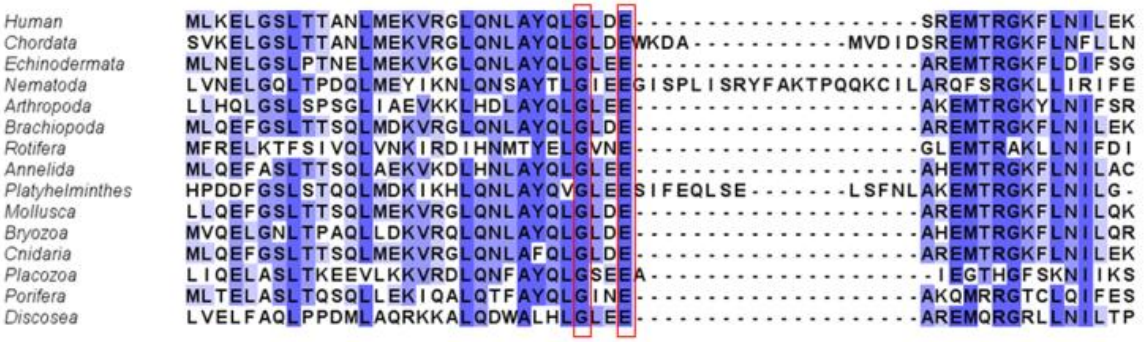

Figure 13 LIN52 Conservation of Human 68-113

Consensus of human LIN52 68-113 region where LIN52 interacts with LIN9. The human region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. Important residues necessary for the LIN52-LIN9 interaction are shown in red boxes

3.3.3 LIN52-Pocket Protein Interaction

For our last analysis of known interaction regions, we evaluated the LIN52-Pocket Protein interaction, including the LIN52 17-37 region of interest and the pocket protein's LxCxE pocket domain. We observed that the LIN52 17-45 region of interest was well conserved ranging from 93.1% in Chordata to 38.1% in Nematoda (Figure 6). Notably, the lowest percent identity was observed in Nematoda with 38.1%. In contrast, the more distantly related amoeboa phylum Discosea displayed a higher percent identity at 47.62%. This might be due to the LxCxE motif observed in Nematoda is diverged from the more common LxSxExL motif observed in the other phyla (Figure 14). The similarity between Chordata and Cnidaria was strikingly similar to each other (93.1% vs 80.95%), which indicates that this region is indeed highly conserved. Though Porifera does suffer a significant drop in identity at 40% which is surprising given how closely related Porifera is to Cnidaria. For the pocket protein's pocket domain, the conservation based on percent identity within animals remains highly conserved as well (Figure 8). While on average lower than LIN52's interaction site, there are only a few phyla in animals that are notably low, including Platyhelminthes at 30% and Rotifera at 24.24%. As compared to LIN52, more putative pocket protein homologs were observed outside of animals. But in our examination of the percent identity, we observed that the sequences observed outside animals were not well conserved. The lowest phylum was at 20% in the protist phylum Oomycota. Even in fungi and plants, the conservation was more highly maintained.

We first examined the LIN52 17-37 region of interest (Figure 14). We observed the highly conserved LxSxExL motif in all phlyas except for Nematoda, which instead encode for LxCxE motif. For phyla that contain the LxSxExL consensus motif, the corresponding phosphorylation site motif RxSP [Litovchick, 2011] is also relatively well conserved, except for the amoeba phylum Discosea (Figure 14). We then examined the pocket protein's interaction site and discovered that the region was relatively well conserved for almost all phyla. The exception to this is the area characterized as the phosphate binding pocket that LIN52 phosphorylated S28 uses to bind to the pocket protein [Guiley, 2015]. Here, we observed that the protists phyla or clades Oomycota, Ciliophora, Stramenopiles, and Heterolobosea were missing the phosphate binding pocket. However, all plant phyla except of Rhodophyta do maintain conservation of the phosphate binding pocket. These data suggest that the mechanism that allows for the MuvB subcomplex to bind to the pocket protein is conserved beyond animals and extends into the plants, protist, amoeboa, and most surprising fungi.

Figure 14 LIN52 Conservation of Human 17-37

Consensus of human LIN52 17-37 region where LIN52 interacts with the pocket protein. The human region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. Important residues necessary for the LIN52-Pocket Protein interaction are shown in red boxes

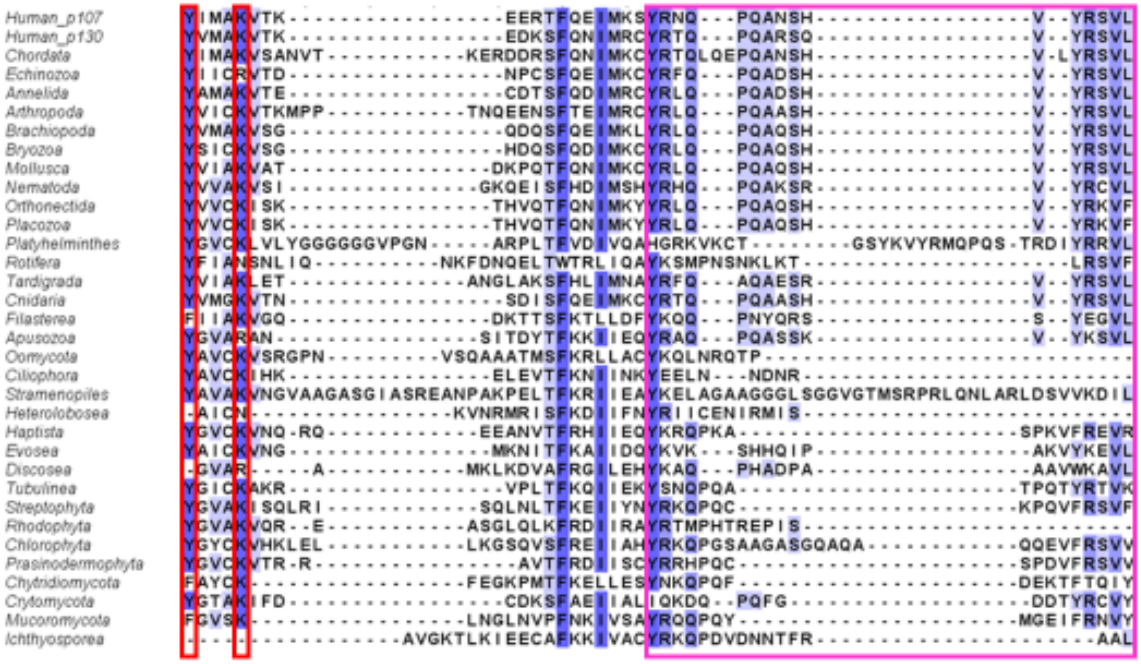

Figure 15 Pocket Protein Conservation of LxCxE Pocket Domain

Consensus of human p107 and p130 LxCxE binding pocket where the pocket protein interacts with the LxCxE/LxSxExL motif of either LIN52 or LIN54. The human regions were aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. Important residues necessary for the pocket protein and LIN52/54 interaction are shown in red boxes. The pink box highlights the phosphate pocket.

3.4 Discovering New Conserved Sites

We next investigated conservation of LIN37, as the protein is relatively understudied compared to the other MuvB subunits. We also aimed to test our protein conservation analysis to identify conserved regions that may indicate unknown interaction regions in the MuvB subcomplex. For this analysis, we evaluated one region near the N-terminus and one region near the C-terminus.

3.4.1 LIN37 ARxxL motif

In our percent identity table for the LIN37 N-terminus motif, we noticed that past Chordata there is a significant drop in conservation (Figure 5). For example, the closest relative Echinodermata only had LIN37 N-terminal regions with a percent identity of 28.21%, compared to the human LIN37 N-terminal region. This indicates that the LIN37 N-terminal region in not highly conserved. We noted that Nematoda has the second lowest observed percent identity in animals at 15.38% and Rotifera has no observed conservation. In plants, specifically Streptophyta, the percent identity was also low at 13.04%. Altogether, our results suggest that the LIN37 N-terminal region is not well conserved.

Upon analyzing the N-terminus conservation on a per residue level, we identified a small ARxxL motif that is conserved across most of the animal phylum except for Rotifera (Figure 16). Stramenopiles and Streptophyta notably lacked the conserved motif, suggesting that the region is exclusive to the animal kingdom (Figure 16). The ARxxL motif is located within human LIN37 residues 19-22. Not much is known about this region, as no structural analysis have been done before incorporating this region of LIN37. The fact that the ARxxL motif is close to the N-terminus suggests that it may act in similar function to LIN52's LxCxE motif to bind to another protein. Further analysis of the ARxxL motif will be required to discover its role in MuvB assembly and function

3.4.2 LIN37 RWK motif

Unlike the N-terminus motif, the C-terminus appeared more highly conserved on a percent identity basis (Figure 5), although we observed a significant drop in conservation from Chordata (100%) to Echinodermata (44.19%). We observed that Nematoda displayed the least conservation of the LIN37 C-terminus across the phyla analyzed, with 20.93% identity. Additionally, in contrast to the N-terminal results outlined above, putative Rotifera LIN37 homologs contained a minimally conserved C-terminal region at 23.26%.

Upon analyzing the C-terminus of LIN37, we identified a small RWK motif within the human LIN37 residues 213-215 (Figure 17). The RWK motif (LIN37 213-215) was observed across all phyla in the animal kingdom (Figure 17). Much like the identified N-terminal ARxxL motif, not much is known about this region. The RWK motif is located near potential phosphorylation sites that are in the LIN37 region 126-208 [Hornbeck, 2015], such that we speculate that the region may act as a switch to facilitate the RWK interaction motif, much like the LIN52 LxSxExL motif acts as a switch in mediating

MuvB association with the pocket protein. Further analysis of the RWK motif and the putative phosphorylation sites will be required to discover their role in MuvB assembly and function.

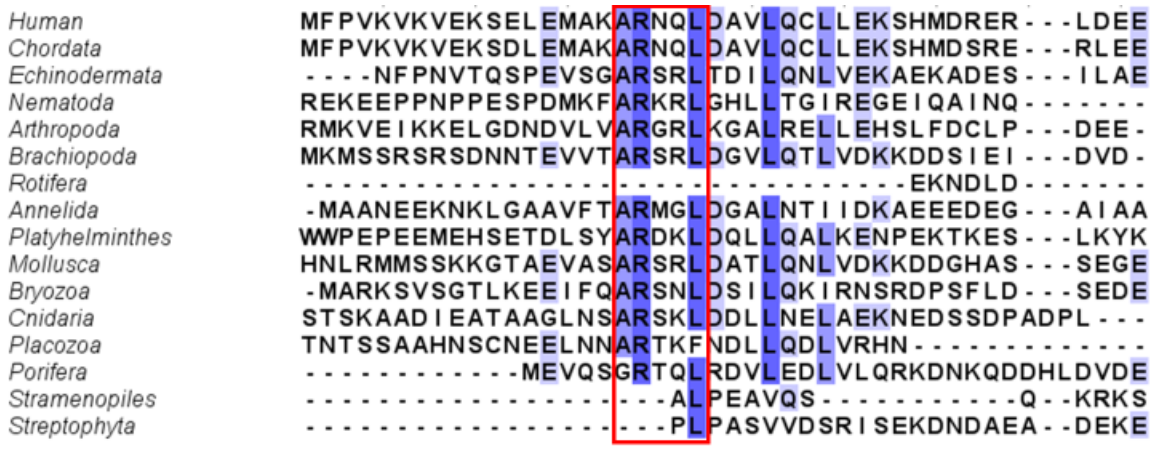

Figure 16 LIN37 Conservation of Human 1-43

Consensus of the human 1-43 LIN37 region where we hypothesis that interactions might exist. The human region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. The ARxxL motif is shown in a red box.

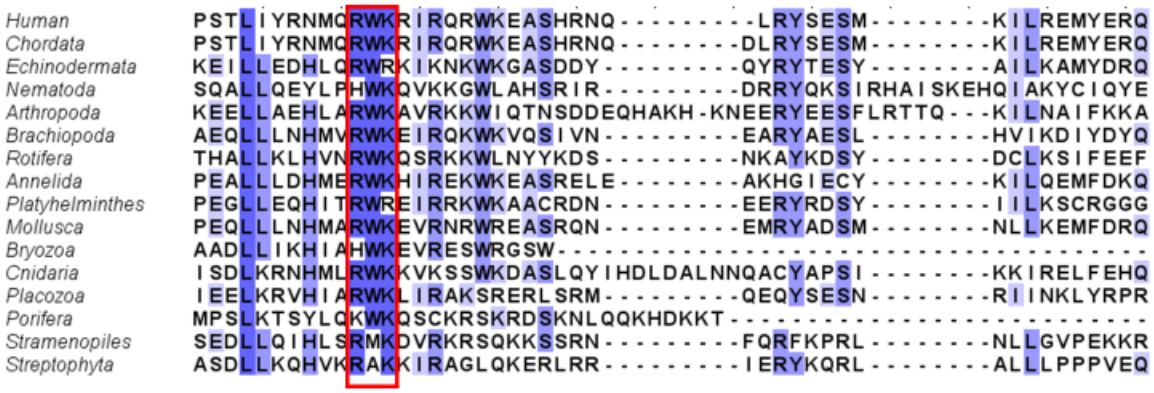

Figure 17 LIN37 Conservation of Human 203-246

Consensus of the human 203-246 LIN37 region where we hypothesis that interactions might exist. The human region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. The RWK motif is shown in a red box.

3.5 Nematoda LIN52

We noted that the consensus for Nematoda showed conservation of the phosphate pocket. This is interesting because it had been shown that *Caenorhabditis elegans* lack both the LIN52 phosphorylation site and the pocket protein's phosphate pocket (Figure 15) [Guiley, 2015]. In the following analysis, we investigated the conservation of the LIN52 LxCxE and RxSP motifs within the Nematoda genus to evaluate if *C. elegans* is a unique exception to the model of DREAM assembly whereby LIN52 phosphorylation induces MuvB association with the pocket protein, as observed in mammalian system. We noted previously that Nematoda LIN52 sequences contained both the unique LxCxE motif and the phosphate pocket in the pocket protein. Here, we explored the Nematoda phylum using our established pipeline to evaluate the conservation of this critical DREAM interaction interface.

3.5.1 LIN52 LxCxE motif is Unique to Nematoda

The LIN52 consensus observed in our phyla analysis showed that while most putative LIN52 homologs contained an LxSxExL motif, the phylum Nematoda was unique in that LIN52 contained a consensus LxCxE motif (Figure 14). Notability, much of the area just flanking the LxCxE motif highly diverged compared to the sequences observed in other phyla (Figure 14). We hypothesis that when LIN52 first appeared in the animal kingdom, the LxSxExL motif was the first to develop and that Nematoda diverged. Given that LxCxE motif is an overall stronger pocket protein association motif compared to the LxSxExL motif, its unknown why other phyla have not adopted the LxCxE motif. It might be because of the need for the ability to dissociate more easily is a desirable outcome.

3.5.2 Degradation of RxSP phosphorylation motif

We next investigated the LIN52 LxCxE motif within Nematoda phylum. We observed a degradation of the RxSP phosphorylation motif (Figure 18). The RxSP motif is phosphorylated at the S28 by DYRK1A in humans [Guiley, 2015], however, a DYRK1A homolog has not been detected in *C. elegans* [Litovchik, 2011]. Based on our analysis, the genus *Caenorhabditis* notably lost the phosphorylated serine residue (Figure 18). We suspect that in nematodes no selective pressure is required to maintain the RxSP phosphorylation motif because the LxCxE motif is sufficient for MuvB binding to the pocket protein.

Figure 18 LIN52 Conservation of Nematoda LxCxE motif along with the RxSP phosphorylation motif

Here, we highlight the Nematoda LxCxE and phosphorylation site. RxSP is degraded except for Spirurina which has maintained the motif. Tylenchina lacks a critical proline while Strongylida and Trichinellida lack the critical arginine. In *Caenorhibditis,* the complex had lost the serine which is where the phosphate would bind to.

3.6 LIN54 LxCxE Motif

We next asked the question of how widespread is the conservation of the LxCxE motif that has been observed recently in the *Arabidopsis thaliana* LIN54 homolog TCX5 [Lang, 2021]. We performed a similar analysis as with the known MuvB interaction regions, as described above, using the *A. thaliana* TCX5 LxCxE region as the seed for identifying consensus regions. We identified several phyla that contained a LIN54 LxCxE motif and generated consensus regions (Figure 19). Other than bryozoa, no other animal phylum contained the LxCxE motif in their putative LIN54 homologs. The motif was observed in amoeba and other plants as well as protests and proto-animals.

We compared each LxCxE motif region in the putative *A. thaliana* TXC5 homolog sequences (Figure 7). Within plants, the percent identity ranged from 53.85% to 76.92%. Within Protists, we observed a percent identity range between 38.46% to 69.23%. Within proto-animals, the only identified putative TXC5 homolog was identified in the clade Filasterea, with a percent identity of 46.15%.

While performing our consensus analysis, we noticed that sometimes many protein sequences in certain phylum or clades lacked the motif (Figure 20). For example, we observed that within protists only a few species in any given phylum or clade contained the motif in their putative LIN54 homolog. This observation may be due to how in the initial setup of our analysis we eliminate duplicates sequences if found within a species. In plants, multiple paralogs of LIN54 exists and likely appear in our initial analysis, but we selected the top sequence based on closest match. It's possible that the LxCxE motif is present within alternative paralogs within the same species. A future analysis that alters the initial construction of the HMM profile and include sequences that contain the motif might be required to explore this observed inconsistency.

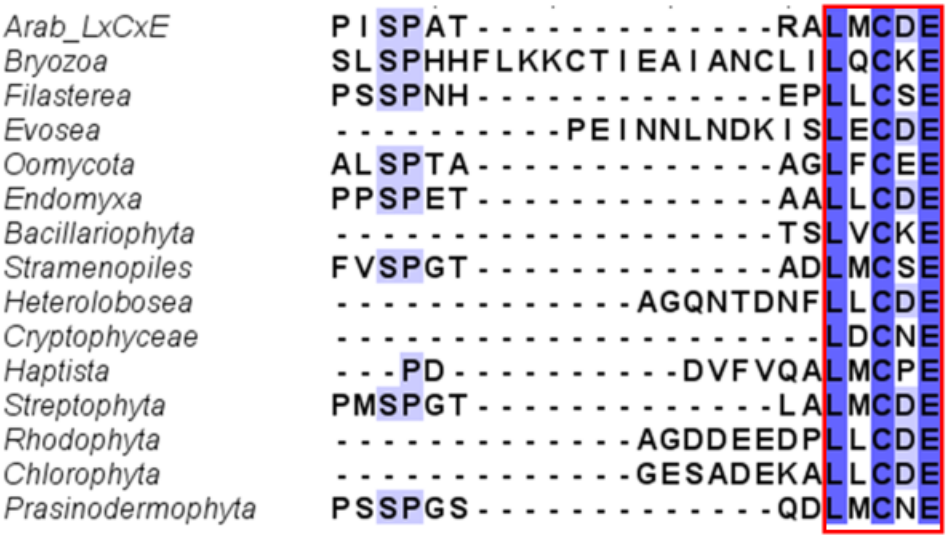

Figure 19 *Arabidopsis thaliana* TCX5 (homolog to LIN54) LxCxE motif conservation MSA

Consensus of the *Arabidopsis thaliana* TCX5 LxCxE motif region where the pocket protein interacts. The *Arabidopsis thaliana* TCX5 LxCxE region was aligned to each phylum or clade and the consensus of each phylum or clade was generated in Jalview. The blue highlight indicates percent identity. Phylum in which no alignment was observed is not shown. The red box shows where the LxCxE domain is located.

LIN54 Region Conservation Anaylsis

Kingdom	Clade	Phylum	LxCxE
Animals	Deurterostomia	Chordata	None
		Echinodermata	None
	Protostomia	Annelida	None
		Arthropoda	None
		Brachiopoda	None
		Bryozoa	Yes (1/1)
		Mollusca	None
		Nematoda	None
		Orthonectida	None
		Platyhelminthes	None
		Rotifera	None
		Tardigrada	None
	Early Animals	Cnidaria	None
		Placozoa	None
		Porifera	None
Proto-Animals		Choanoflagellata*	None
		Filasterea*	Yes (1/1)
		Rotosphaerida*	None
		Apusozoa*	None
Protists	SAR	Oomycota	Yes (22/34)
		Alevolata*	None
		Cercozoa	None
		Endomyxa	Yes (1/1)
		Ciliophora	None
		Bacillariophyta	Yes (1/17)
		Stramenopiles*	Yes (2/19)
	Excavata	Heterolobosea*	Yes (2/2)
		Parabasalia	None
	Other	Cryptophyceae*	Yes (1/7)
		Haptista*	Yes (1/7)
Amoeboa		Evosea	Yes (5/6)
		Tubulinea	None
		Discosea	None
Plants		Streptophyta	Yes (188/212)
		Rhodophyta	Yes (4/10)
		Chlorophyta	Yes (5/36)
		Prasinodermophyta	Yes (1/1)

Figure 20 *Arabidopsis thaliana* TCX5 (homolog to LIN54) LxCxE motif conservation table

Here we analysis the LIN54 MuvB subunit to determine the appearance of an LxCxE motif. The results were sorted by either phylum or closest clade as determined by the Uniport taxonomy results. Each phylum or clade was then grouped into different kingdoms. While protists are no longer a wildly used kingdom and proto-animals is not a technical kingdom, these terms are used to describe each of the phylum or clades with one another. Colors in the table are used to separate the kingdoms for better visualization. Red indicating animals, red-orange indicating proto-animals, yellow indicating protests, blue indicating amoeboa, green indicating plants, and purple indicating fungi. A gap is place between plants and fungi for clarity. A * indicates groups chosen that are not phylum. They can include class, order, or clade. Another note is that the clade early animals is used to differentiate phylum the evolutionally appear before Deuterostomia and Protostomia. We used a yes vs no system for the table. If the motif was present in the phylum, we counted the number of species in the phylum that contained the motif (shown in parentheses with the number of species found out of the total in the phylum).

4 Discussion

Our unique protein conservation analysis revealed key insights into MuvB subunit conservation across animals and beyond. Most notably, due to the essential role MuvB plays in the cell cycle, we had hypothesized that MuvB may represent the ancestral DREAM complex. However, our analysis refuted our hypothesis, as the pocket protein remains as conserved beyond the animal kingdom as well-defined subunits of the MuvB complex like LIN9 and LIN54. We also investigated key interaction sites known to mediate interactions between MuvB subunits and observed little change in LIN9-LIN37, LIN9-LIN52, and LIN52-Pocket Protein interactions even across different kingdoms. With confirmation that our analysis identified known interaction motifs between DREAM subunits, we investigated whether we could identify unique interaction regions. Our analysis revealed new regions in the LIN37 that may be important for MuvB function that will be investigated further. Altogether, our analysis reveals a new model for DREAM complex evolutionarily conservation across eukaryotic species.

In our analysis of each MuvB subunit, LIN52 was almost exclusively observed only in the animal kingdom, with the only exception being one hit in the amoeboa phylum Discosea. The species that was identified, *Acanthamoeba castellanii*, is known to have several genes incorporated through lateral gene transfer events [Clark, 2013]. This could be an explanation as to why we detected LIN52 and the percent identity of the LxSxExL region is higher than some animal phyla (Figure 6). Strangely, it was discovered that LIN52 homologs do exist outside of the animal kingdom as the homolog was identified in *Arabidopsis thaliana*, but the LIN52 homolog lacks the LxCxE motif that mediates binding to the pocket protein [Lang, 2021]. The lack of LxCxE motifs might explain why we did not find plant LIN52 homologs outside of animals and in only one amoeboa. Due to missing a critical component of what we know as LIN52, the question remains if the LIN52 homolog identified should be considered a homolog. We discovered that in LIN9's LIN52 interaction site, key amino acids remained conserved in most phyla including those in plants (Figure 12). We have observed a similar phenomenon regarding the LIN9 region that also binds to B-Myb in animals. Even though in the absence of B-Myb in *C. elegans*, the region retains the ability to associate with *Drosophila* B-Myb [Vorster, 2020]. Given that the LIN9-LIN52 interaction remains conserved outside of animals, we speculate that a possible B-Myb like protein interacts with this region in early animals and non-animals in later stages of the cell cycle to allow for MuvB to activate DREAM target genes. There is also the possibility that B-Myb itself evolved independently to interact with the LIN9-LIN52 region.

Where plants lack the pocket protein binding on their LIN52 homologs, they instead have the LxCxE motif located within the LIN54 homologs [Lang, 2021]. Upon analysis of the region in plants, we found similar motifs across other phyla including groups of protists, amoebozoa, and the proto-animal clade Filasterea. This raises the question of how the structural assembly of MuvB underwent a drastic change when the first animals began to appear. Not only did the pocket protein binding domain appear to shift from LIN54 to LIN52, but the configuration of the motif also changed from LxCxE to LxSxExL. LxSxExL is known to be a weaker binding configuration than LxCxE [Guiley, 2015]. In

humans, we know that the RxSP motif is also required to be phosphorylated by DYRK1A for DREAM assembly which raises the question of how a more complex mechanism arose when the need was not required before. Not only that but our analysis shows that within animals, Nematoda are the only phylum to possess an LxCxE motif in LIN-52. The question also remains of how Nematoda reverted to an LxCxE motif for binding to the pocket protein.

We are interested in whether DREAM complex evolution occurred through a core set of components first emerging, with additional units slowly are added over time. In the case of the MuvB subcomplex, our results support a model where LIN9, and LIN54 emerged first, with LIN37 and LIN52 being introduced over time. However, with recent research discoveries showing that all subunits are indeed conserved in plants, it raises the question of how the subcomplex emerged with all five proteins [Lang, 2021]. Perhaps a more ancestral form of MuvB subcomplex existed at one point in time but is now lost with billions of years of evolution resulting in a divergence of MuvB assembly across diverse eukaryotic species. It has been shown before how within protein complexes many are either all essential or none are essential in specific species [Ryan, 2013]. While the core function of DREAM appears to be similar across all forms of life, DREAM's structural assembly appears to be different between animals and plants. Here, we propose a model of plant and non-animal MuvB assembly with the pocket protein (Figure 21). The difference showing which subunit binds to the pocket protein compared to the standard animal model and how both LIN9 and LIN37 interact with all components in plants as opposed to animal MuvB where LIN37 does not have an apparent interaction with all subunits (Figure 21).

We also propose a potential evolutionary model highlighting the appearance of subunits of MuvB based on our analysis (Figure 22), with the ancestral core comprising of LIN9, LIN54, and RBAP48. From the core, additional units were added over time with most likely LIN37 being the first added and then LIN52 completing the 5-subunit subcomplex MuvB that we observe in the animal kingdom. We have based this on finding LIN9 and LIN54 in more unique species than any of the other MuvB subunits (Figure 3). We also propose a model of MuvB assembly divergence based on how the subunits arrange and highlight the possibility of a MuvB-like complex in fungi (Figure 23). Here, from the core ancestral complex, plants and non-animal DREAM assembled in a different configuration compared to animals. Mainly, the role of the pocket protein binding site in plants and non-animals is located in LIN54 homologs and in animals the role shifts to LIN52. Finally, the model points to a potential existence of a fungi MuvB subcomplex based on the findings of pocket protein homologs in fungi (Figure 3). More in-depth analysis will be performed to differentiate between the animal and non-animal MuvB subcomplexes and further investigate if a potential MuvB subcomplex exists in fungi.

With the newly acquired insight of the MuvB subcomplex conservation, we aim to apply our findings into further studies. With the newly found motifs in LIN37 (ARxxL and RWK), we intend to apply CRISPR/Cas9 mutagenesis into the model organism *C. elegans* and observe if the MuvB complex formation is disrupted or if the function of LIN37 itself has been disrupted. Similar work has been done in *C. elegans* with

disruption of the LxCxE motif in LIN52 [Goetsch, 2019]. Altogether, we determined the evolutionary conservation of the MuvB subcomplex of DREAM using our newly developed pipeline. Using our analysis, we uncovered additional conserved sites that had previously been unknown. In the future, testing the role of these new conserved sites in MuvB assembly will help us gain insight into how MuvB's function is evolutionarily protected.

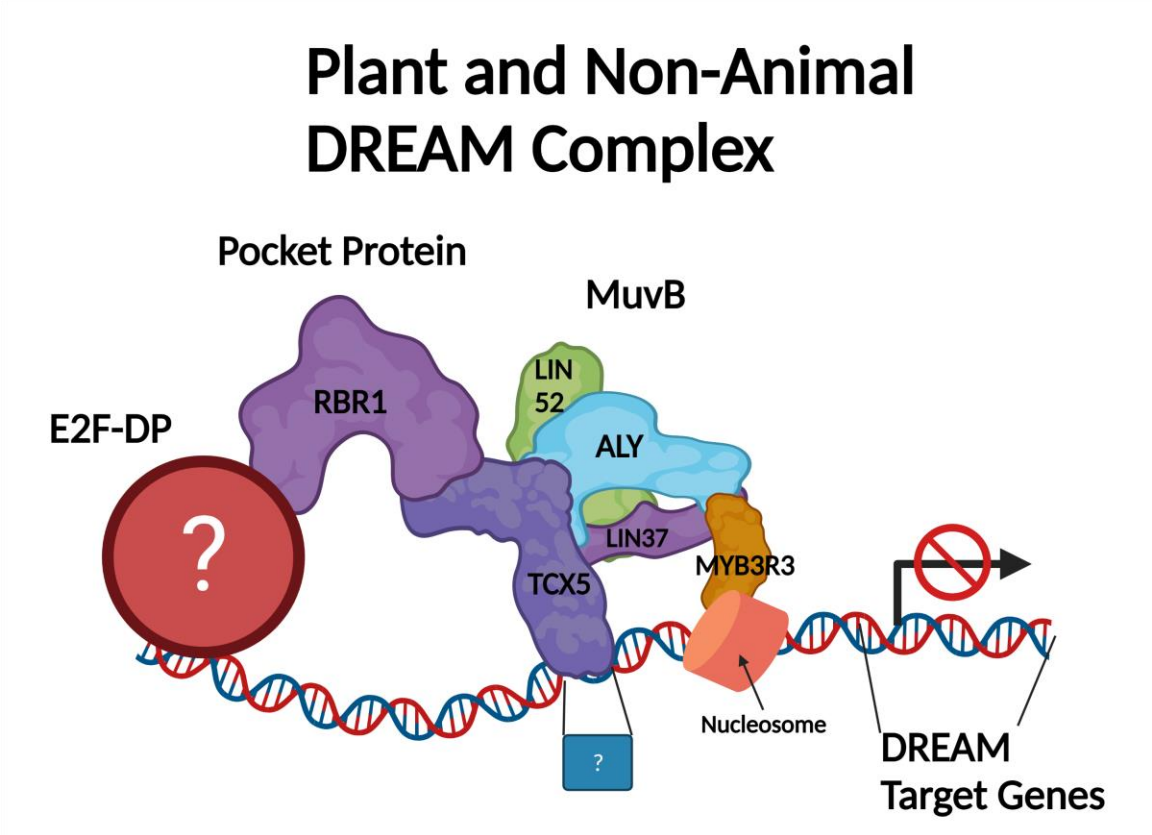

Figure 21 MuvB complex with the pocket protein in plants and other non-animals when assembled into DREAM

Here the model structural assembly of MuvB in plants is shown. With LIN54 acting as the adaptor protein to allow for the pocket protein to assemble into DREAM. LIN37 is also shown to be interacting with each of the subunits along with LIN9. LIN53 is shown to bind to a nucleosome and LIN54 is shown binding a site that would be like a CHR site, however this has yet to be identified as a CHR site. E2F-DP subcomplex is shown as a question mark as the exact subunits involved are unknown.

Created with BioRender.com

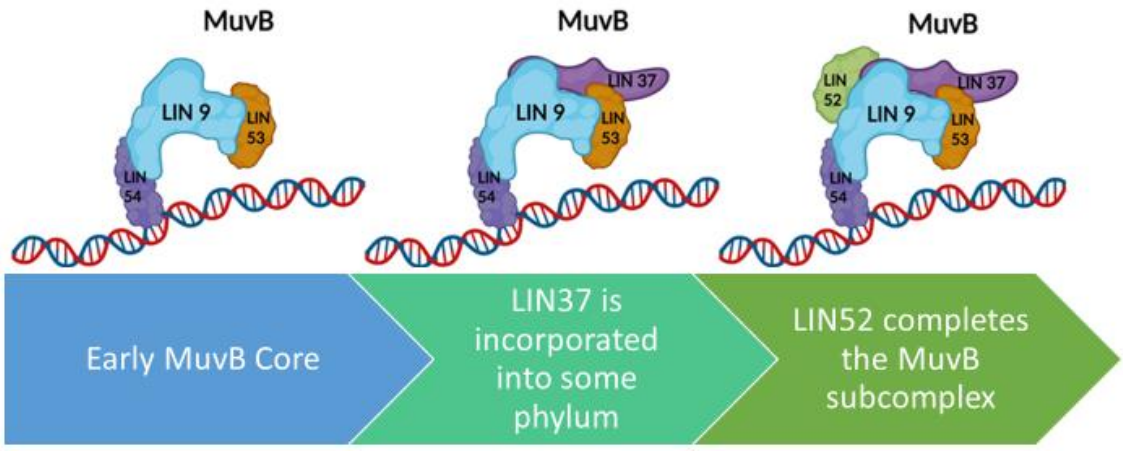

Figure 22 Model of MuvB subunit evolution

Here we propose that the earliest MuvB core consisted of the subunits LIN9, LIN54, and RBPA48 (shown as LIN53 until I fix it). From there, the LIN37 subunit was added to the complex to mediate MuvB into a role of both repression and activation of genes. The addition of LIN52 comes last in helping stabilize LIN9 and possibly B-Myb like proteins to the complex as well as assume the role of LIN54's pocket protein binding. The model of LIN54 binding to pocket protein in plants is described in Figure 15.

Created with BioRender.com

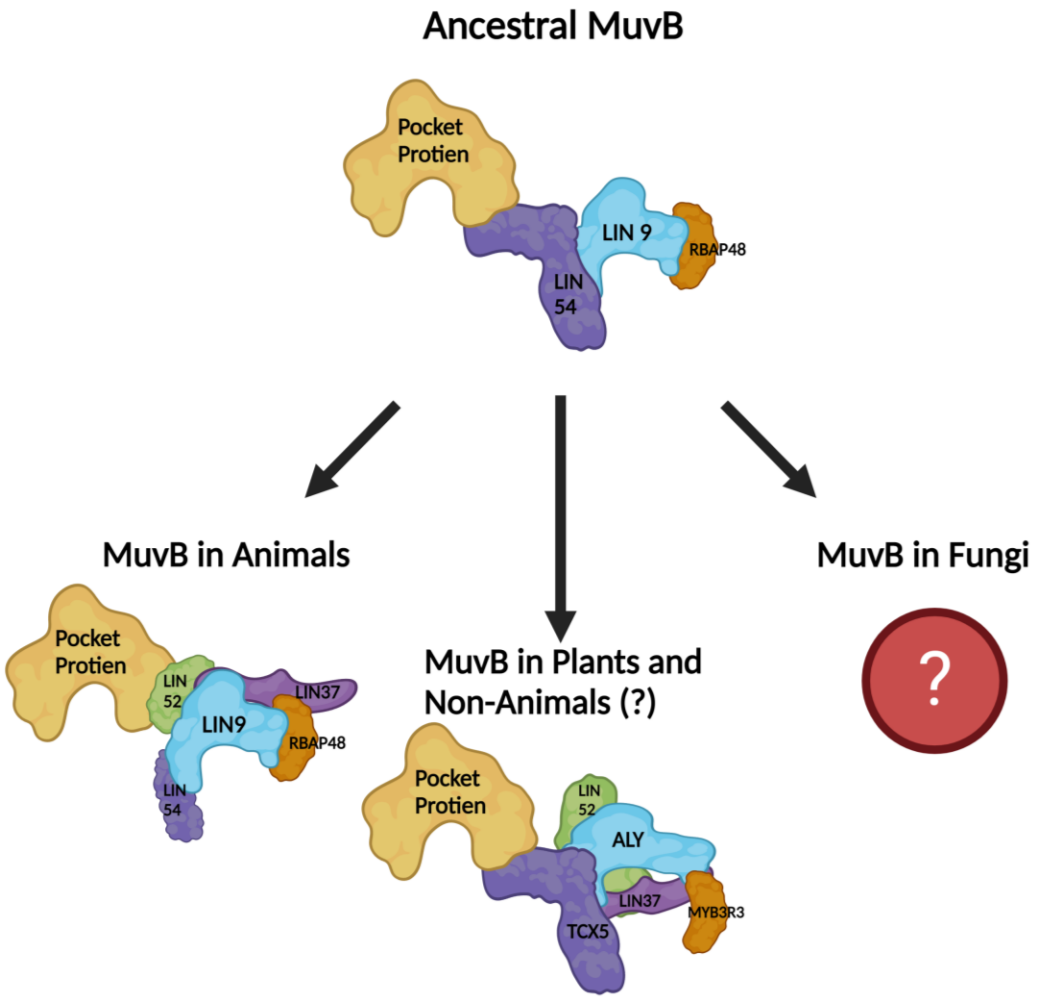

Figure 23 MuvB Structural Configuration Evolution over time

Ancestral MuvB core is shown with 3 subunits LIN9, LIN54, and RBAP48 (human homolog names are used here). From the ancestral MuvB, a divergence occurs with three different paths. MuvB in animals which is assembled like in Figure 1. MuvB in plants and non-animals resembles the MuvB complex in plants shown in Figure 21. Key difference between animal and non-animal is which MuvB subunit binds to the pocket protein with animals being LIN52 and non-animals being TCX5 (homlog of LIN54). The final branch points to a possible MuvB complex in fungi, however, no MuvB subunit has been detected and is only speculated due to the appearance of the pocket protein in fungi.

Created with BioRender.com

5 Reference List

Asthana, A. *et al.* The MuvB complex binds and stabilizes nucleosomes downstream of the transcription start site of cell-cycle dependent genes. *Nature Communications* **13**, doi:10.1038/s41467-022-28094-1 (2022).

Chen, B.-R. *et al.* LIN37-DREAM prevents DNA end resection and homologous recombination at DNA double-strand breaks in quiescent cells. *eLife* **10**, doi:10.7554/elife.68466 (2021).

Clarke, M. *et al.* Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biology* **14**, R11, doi:10.1186/gb-2013-14-2-r11 (2013).

Cobrinik, D. Pocket proteins and cell cycle control. *Oncogene* **24**, 2796-2809, doi:10.1038/sj.onc.1208619 (2005).

Esterlechner, J. *et al.* LIN9, a Subunit of the DREAM Complex, Regulates Mitotic Gene Expression and Proliferation of Embryonic Stem Cells. *PLoS ONE* **8**, e62882, doi:10.1371/journal.pone.0062882 (2013).

Fischer, M. & Müller, G. A. Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. *Critical Reviews in Biochemistry and Molecular Biology* **52**, 638-662, doi:10.1080/10409238.2017.1360836 (2017).

Forristal, C. *et al.* Loss of the Mammalian DREAM Complex Deregulates Chondrocyte Proliferation. *Molecular and Cellular Biology* **34**, 2221-2234, doi:10.1128/mcb.01523-13 (2014).

Fraser, H. B. & Plotkin, J. B. Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biology* **8**, R252, doi:10.1186/gb-2007-8-11-r252 (2007).

Goetsch, P. D., Garrigues, J. M. & Strome, S. Loss of the Caenorhabditis elegans pocket protein LIN-35 reveals MuvB's innate function as the repressor of DREAM target genes. *PLoS Genet* **13**, e1007088, doi:10.1371/journal.pgen.1007088 (2017).

Goetsch, P. D. & Strome, S. *DREAM Interrupted: Severing MuvB from DREAM's pocket protein in Caenorhabditis elegans impairs gene repression but not DREAM chromatin assembly* (Cold Spring Harbor Laboratory, 2019).

Guiley, K. Z. *et al.* Structural mechanism of Myb-MuvB assembly. *Proc Natl Acad Sci U S A* **115**, 10016-10021, doi:10.1073/pnas.1808136115 (2018).

Guiley, K. Z. *et al.* Structural mechanisms of DREAM complex assembly and regulation. *Genes Dev* **29**, 961-974, doi:10.1101/gad.257568.114 (2015).

Harrison, M. M., Ceol, C. J., Lu, X. & Horvitz, H. R. Some C. elegans class B synthetic multivulva proteins encode a conserved LIN-35 Rb-containing complex distinct from a NuRD-like complex. *Proc Natl Acad Sci U S A* **103**, 16782-16787, doi:10.1073/pnas.0608461103 (2006).

Hart, G. T., Lee, I. & Marcotte, E. M. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* **8**, 236, doi:10.1186/1471-2105-8-236 (2007).

Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research* **43**, D512-D520, doi:10.1093/nar/gku1267 (2015).

Jiang, J., Benson, E., Bausek, N., Doggett, K. & White-Cooper, H. Tombola, a tesmin/TSO1-family protein, regulates transcriptional activation in theDrosophilamale germline and physically interacts with Always early. *Development* **134**, 1549-1559, doi:10.1242/dev.000521 (2007).

Korenjak, M. *et al.* Native E2F/RBF complexes contain Myb-interacting proteins and repress transcription of developmentally controlled E2F target genes. *Cell* **119**, 181-193, doi:10.1016/j.cell.2004.09.034 (2004).

Lang, L. *et al.* The DREAM complex represses growth in response to DNA damage in Arabidopsis. *Life Science Alliance* **4**, e202101141, doi:10.26508/lsa.202101141 (2021).

Liban, T. J. *et al.* Conservation and divergence of C-terminal domain structure in the retinoblastoma protein family. *Proceedings of the National Academy of Sciences* **114**, 4942-4947, doi:10.1073/pnas.1619170114 (2017).

Litovchick, L., Florens, L. A., Swanson, S. K., Washburn, M. P. & Decaprio, J. A. DYRK1A protein kinase promotes quiescence and senescence through DREAM complex assembly. *Genes & Development* **25**, 801-813, doi:10.1101/gad.2034211 (2011).

Litovchick, L. *et al.* Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. *Mol Cell* **26**, 539-551, doi:10.1016/j.molcel.2007.04.015 (2007).

Mages, C. F., Wintsche, A., Bernhart, S. H. & Muller, G. A. The DREAM complex through its subunit Lin37 cooperates with Rb to initiate quiescence. *Elife* **6**, doi:10.7554/eLife.26876 (2017).

Marceau, A. H. *et al.* Structural basis for LIN54 recognition of CHR elements in cell cycle-regulated promoters. *Nat Commun* **7**, 12301, doi:10.1038/ncomms12301 (2016).

Matsuo, T., Kuramoto, H., Kumazaki, T., Mitsui, Y. & Takahashi, T. LIN54 harboring a mutation in CHC domain is localized to the cytoplasm and inhibits cell cycle progression. *Cell Cycle* **11**, 3227-3236, doi:10.4161/cc.21569 (2012).

Miao, X., Sun, T., Barletta, H., Mager, J. & Cui, W. Loss of RBBP4 results in defective inner cell mass, severe apoptosis, hyperacetylated histones and preimplantation lethality in mice†. *Biology of Reproduction* **103**, 13-23, doi:10.1093/biolre/ioaa046 (2020).

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research* **41**, e121-e121, doi:10.1093/nar/gkt263 (2013).

Müller, G. A. *et al.* The CHR promoter element controls cell cycle-dependent gene transcription and binds the DREAM and MMB complexes. *Nucleic Acids Research* **40**, 1561-1578, doi:10.1093/nar/gkr793 (2012).

Murzina, N. V. *et al.* Structural Basis for the Recognition of Histone H4 by the Histone-Chaperone RbAp46. *Structure* **16**, 1077-1085, doi:10.1016/j.str.2008.05.006 (2008).

Nabeel-Shah, S. *et al.* Functional characterization of RebL1 highlights the evolutionary conservation of oncogenic activities of the RBBP4/7 orthologue in Tetrahymena thermophila. *Nucleic Acids Research* **49**, 6196-6212, doi:10.1093/nar/gkab413 (2021).

Neer, E. J., Schmidt, C. J., Nambudripad, R. & Smith, T. F. The ancient regulatory-protein family of WD-repeat proteins. *Nature* **371**, 297-300, doi:10.1038/371297a0 (1994).

Osterloh, L. *et al.* The human synMuv-like protein LIN-9 is required for transcription of G2/M genes and for entry into mitosis. *The EMBO Journal* **26**, 144-157, doi:10.1038/sj.emboj.7601478 (2007).

Oti, M. & Brunner, H. The modular nature of genetic diseases. *Clinical Genetics* **71**, 1-11, doi:10.1111/j.1399-0004.2006.00708.x (2006).

Petrella, L. N. *et al.* synMuv B proteins antagonize germline fate in the intestine and ensure C. elegans survival. *Development* **138**, 1069-1079, doi:10.1242/dev.059501 (2011).

Pilkinton, M., Sandoval, R. & Colamonici, O. R. Mammalian Mip/LIN-9 interacts with either the p107, p130/E2F4 repressor complex or B-Myb in a cell cycle-phase-dependent context distinct from the Drosophila dREAM complex. *Oncogene* **26**, 7535-7543, doi:10.1038/sj.onc.1210562 (2007).

Reichert, N. *et al.* Lin9, a Subunit of the Mammalian DREAM Complex, Is Essential for Embryonic Development, for Survival of Adult Mice, and for Tumor Suppression. *Molecular and Cellular Biology* **30**, 2896-2908, doi:10.1128/mcb.00028-10 (2010).

Ryan, C. J., Krogan, N. J., Cunningham, P. & Cagney, G. All or Nothing: Protein Complexes Flip Essentiality between Distantly Related Eukaryotes. *Genome Biology and Evolution* **5**, 1049-1059, doi:10.1093/gbe/evt074 (2013).

Sadasivam, S., Duan, S. & Decaprio, J. A. The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. *Genes & Development* **26**, 474-489, doi:10.1101/gad.181933.111 (2012).

Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539, doi:10.1038/msb.2011.75 (2011).

Sim, C. K., Perry, S., Tharadra, S. K., Lipsick, J. S. & Ray, A. Epigenetic regulation of olfactory receptor gene expression by the Myb-MuvB/dREAM complex. *Genes & Development* **26**, 2483-2498, doi:10.1101/gad.201665.112 (2012).

Uxa, S. *et al.* DREAM and RB cooperate to induce gene repression and cell-cycle arrest in response to p53 activation. *Nucleic Acids Res* **47**, 9087-9103, doi:10.1093/nar/gkz635 (2019).

Vorster, P. J. *et al.* A long lost key opens an ancient lock: Drosophila Myb causes a synthetic multivulval phenotype in nematodes. *Biol Open* **9**, doi:10.1242/bio.051508 (2020).

Wang, H. *et al.* A Complex-based Reconstruction of the Saccharomyces cerevisiae Interactome. *Molecular & Cellular Proteomics* **8**, 1361-1381, doi:10.1074/mcp.m800490-mcp200 (2009).

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).

Zhang, Y. *et al.* Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes & Development* **13**, 1924-1935, doi:10.1101/gad.13.15.1924 (1999).

Zhang, Y. *et al.* SAP30, a Novel Protein Conserved between Human and Yeast, Is a Component of a Histone Deacetylase Complex. *Molecular Cell* **1**, 1021-1031, doi:10.1016/s1097-2765(00)80102-1 (1998).