REGULATION OF GENE EXPRESSION IN DROSOPHILA BY THE RB AND CTBP TRANSCRIPTIONAL COREPRESSORS

By

Ana-Maria Raicu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Cell and Molecular Biology – Doctor of Philosophy

2023

ABSTRACT

Elucidation of core biological processes requires an understanding of how transcription is regulated. Transcriptional regulation is critical for fine tuning gene expression to achieve precise temporal and spatial expression patterns. Key components of transcriptional regulation are repressors and corepressors—proteins that are recruited to DNA to reduce or turn off gene expression. Corepressors such as Rb and CtBP are essential metazoan transcription factors that regulate the expression of a diversity of genes. Retinoblastoma (Rb) was the first identified human tumor suppressor protein. It represses gene expression by binding to E2F transcriptional activators and recruiting histone modifiers, and has been shown to regulate genes through preferential interactions with promoter-proximal sequences. The C-terminal binding protein (CtBP) can function from both promoters and enhancers, and also regulates gene expression through recruitment of chromatin-modifying factors. CtBP is homologous to alpha-hydroxy acid dehydrogenases, and requires NADH binding and oligomerization to function. It has an unstructured C-terminal domain whose function in gene regulation is unknown. Both Rb and CtBP have diversified over evolutionary time through gene duplications in specific lineages, and CtBP has evolved unique protein isoforms through alternative splicing. The gene encoding Rb has undergone duplication in vertebrates and in the Drosophila lineage, and the gene encoding CtBP has duplicated multiple times in vertebrates, with alternative splice isoforms found across Metazoa, including in Drosophila. Such evolutionary variation is undoubtedly significant for Rb and CtBP gene regulatory functions, as many of the paralogs and certain alternative splice forms are highly conserved. The specific activities and positive selection for multiple Rb paralogs and CtBP isoforms is not well understood, however. To uncover differences between Rb paralogs and CtBP isoforms in the Drosophila system, I made use of a modified CRISPR/Cas9 technology,

called CRISPRi. In CRISPRi, a nuclease-dead Cas9 is fused to transcription factors and epigenetic modifiers to recruit them to genomic sequences via gene-specific guide RNAs for gene regulation. I adapted this technology to recruit Rb and CtBP proteins to gene promoters to characterize their repression properties and identify possible intrinsic differences between paralogs and isoforms. This direct, comparative approach to characterizing corepressor action *in vivo* takes advantage of powerful molecular genetic tools to shed light on how evolutionary forces have sculpted the activity of these conserved regulatory proteins. Additionally, through studies of Rb, we identified a form of gene regulation that we term "soft repression", in which a repressor functions to subtly, but precisely, modulate gene expression, rather than acting as an on/off switch. In a separate but related study that stemmed from structure-function questions about the CtBP intrinsically disordered C-terminal domain, I, along with a team of junior researchers, took a comparative phylogenetic approach to uncover the history of this protein's evolution. Through analysis of extant 'omic data, phylogenetic analyses, and precise molecular biology tools outlined here, we furthered our understanding of how these two key corepressors function in the Drosophila system, which we expect will inform future studies that have biomedical relevance.

This dissertation is dedicated to my parents,
who have been my biggest supporters.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude and appreciation to everyone that has supported me throughout my PhD journey. There are many people to thank, and everyone has contributed in some way to an unforgettable and truly wonderful PhD experience.

First and foremost, I would like to thank my advisor, Dr. David Arnosti, for his guidance, advice, and support throughout my training at Michigan State. Dr. Arnosti provided me with countless opportunities for success, trained me to think like a scientist, and extended me the freedom to explore different fields. His expertise and advice at every step of my journey was instrumental in my success as a developing scientist. His scientific curiosity is truly unmatched, and has driven me to be more curious and fascinated with the world around me as well.

My experience in the Arnosti lab would not have been so wonderful without the friendship of my labmates, especially Dr. Sandhya Payankaulam, who taught me much of what I know about Drosophila. I am thankful to the twenty junior students I had the pleasure of training especially Maddy Niblock, Aanchal Jain, Megha Suresh, and Patricia Castanheira, in whom I saw my younger self, eager to learn how to be a scientist, and who helped me learn how to be a better mentor. I must also acknowledge Pamela Himadewi, who was my first friend in the lab; although we overlapped for only one year, her work ethic and inquisitiveness drove me to be a better scientist.

I would also like to thank the Cell and Molecular Biology Program, of which I am a part, and the Department of Biochemistry and Molecular Biology. Thank you to Dr. Peggy Petroff, Alaina Burghardt, Jessica Lawrence, Pappan (Dr. Kaillathe Padmanab), and Carolyn Adams who have supported me in various ways throughout my academic career, and who made the difficult administrative work easier to deal with.

My graduate student years have been the best six years of my life, partly due to the wonderful friendships I made during my time in East Lansing. I especially want to highlight my friendship with Alice Chu, who was my first and closest friend at Michigan State, Basma Masraf Klump, Aiko Turmo, and Mylena Ortiz. Thank you for the talks, the laughs, the writing sessions, the workouts for stress relief, and many fun times we shared together.

My roommate, Andriana Manousidaki, has been a constant support and cheerleader for me all these years. I am so thankful for her kindness, her understanding, and compassion, especially during stressful periods. I have loved learning about Greece from her, as well as from the many Greek friends I made at MSU who made my time here culturally enriching. Christos, Ioannis, Manos, Mihalis, Dimitris and everyone else—thank you for giving me a wonderful few years in East Lansing.

The challenges of graduate school were ameliorated by my spiritual father, Fr. John Konkle, and the community at Holy Dormition Monastery and Holy Trinity Greek Orthodox Church. Thank you to all for your love, encouragement, and prayers.

I have sincerely appreciated the guidance of my Dissertation Committee, made up of Dr. Bill Henry, Dr. Amy Ralston, Dr. Michaela Smolle, Dr. Arjun Krishnan, and Dr. Eran Andrechek, who each helped me with various aspects of my dissertation, grant applications, and my overall graduate student experience.

Finally, I would like to acknowledge the love and support of my family. Thank you to my parents for their patience and understanding, their love, and guidance throughout this time. My father, Dr. Valerică Raicu, was a constant source of support as I made my journey into biological sciences, selected labs to rotate in, applied for grants, prepared for conferences, and applied for postdoc positions. Thank you for teaching me how to be a hardworking scientist. The prayers and hugs of my mother, Georgeta Raicu, have kept me going throughout the challenges of graduate school. When I left Wisconsin to pursue my graduate studies at Michigan State, it was incredibly hard to leave behind my five sisters, Laura, Adela, Naomi, Margaret, and Elaina. Having them in a different state was a challenge at times, but I appreciate the bonds we maintained all these years. I love you all and thank you for your friendship.

As I move on to my next step in my academic career, I will cherish the friendships and relationships I have made in East Lansing. Michigan State University is where I became a biologist, and I will never forget my experience here.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AA	Amino Acid
Acf	ATP-dependent chromatin assembly factor large subunit
bp	Base pair
CAT	chloramphenicol acetyl transferase
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation sequencing
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CryoEM	CryoElectron Microscopy
CtBP	C-terminal Binding Protein
CTD	C-terminal Domain
dCas9	nuclease dead Cas9
DNA	Deoxyribonucleic acid
DP	E2F dimerization partner
E2F	E2 promoter binding factor
E2F1	E2F transcription factor 1
E2F2	E2F transcription factor 2
GAL4	Galactose 4
HDAC	Histone deacetylase complex
IE	Instability element
kb	kilo base pair
kDa	kilo Dalton
NTD	N-terminal Domain

mcm6	Minichromosome maintenance 6
Mpp6	M-phase phosphoprotein 6
mRNA	messenger RNA
PCNA	Proliferating Cell Nuclear Antigen
Pex2	Peroxin 2
Rb	Retinoblastoma
Rbf1	Retinoblastoma family protein 1
Rbf2	Retinoblastoma family protein 2
RNA	Ribonucleic acid
RNAPII	RNA Polymerase II
RNA-seq	RNA sequencing
RT-qPCR	Real-time quantitative PCR
SDM	Site Directed Mutagenesis
TF	Transcription factor
TSS	Transcriptional start site
UAS	Upstream activation sequence
UTR	Untranslated Region
wg	wingless
WT	Wild Type
yw	yellow whit

CHAPTER 1: INTRODUCTION

BASIC COMPONENTS OF EUKARYOTIC GENE REGULATION

The importance of non-coding DNA sequence

Eukaryotic gene regulation is a highly complex and sophisticated process, involving numerous factors that either work broadly or in gene-specific ways (**Figure 1.1**; Reviewed in Lelli *et al.* 2012). Non-coding elements within the DNA sequence itself act in cis to direct the assembly and activity of factors (generally proteins and RNA) that impact gene regulation in trans. Essential DNA elements that act proximally to the start of gene transcription are promoters—regions of about 100 base pairs (bp) flanking a gene's transcriptional start site (TSS). Promoters mediate the binding of the transcriptional machinery, including the RNA polymerase that performs transcription, and a variety of associated transcription factors (TFs) that together form the pre-initiation complex (PIC). The promoter can be sufficient to drive gene expression in some organisms, such as prokaryotes, but is not sufficient in eukaryotes, which require additional regulatory DNA elements such as distally-acting enhancers, particularly for transcription of protein-coding genes by RNA polymerase II (Reviewed in Roeder, 2019).

Typically less than 1 kilobase (kb) in size, enhancers are DNA sequences that can be found even up to 1 megabase away from a gene's TSS in higher eukaryotes, and are located intergenically or even within the body of a gene (Reviewed in Kim and Wysocka, 2023). Enhancers can often function in an orientation-independent manner, and can be active or silent, depending on what types of TFs are binding to them. Through DNA looping and interactions with the Mediator complex (a complex of proteins that acts as a bridge between distal enhancers and promoters), enhancers are brought in proximity to promoters to assist with activating or repressing transcription (Reviewed in Richter *et al.* 2022). In higher eukaryotes, clusters of enhancers in domains of over 10 kb have been termed "stretch" or "super" enhancers, and have been proposed to possess unique properties related to the ability to concentrate regulatory factors that are key to gene regulation (Reviewed in Kim and Wysocka, 2023).

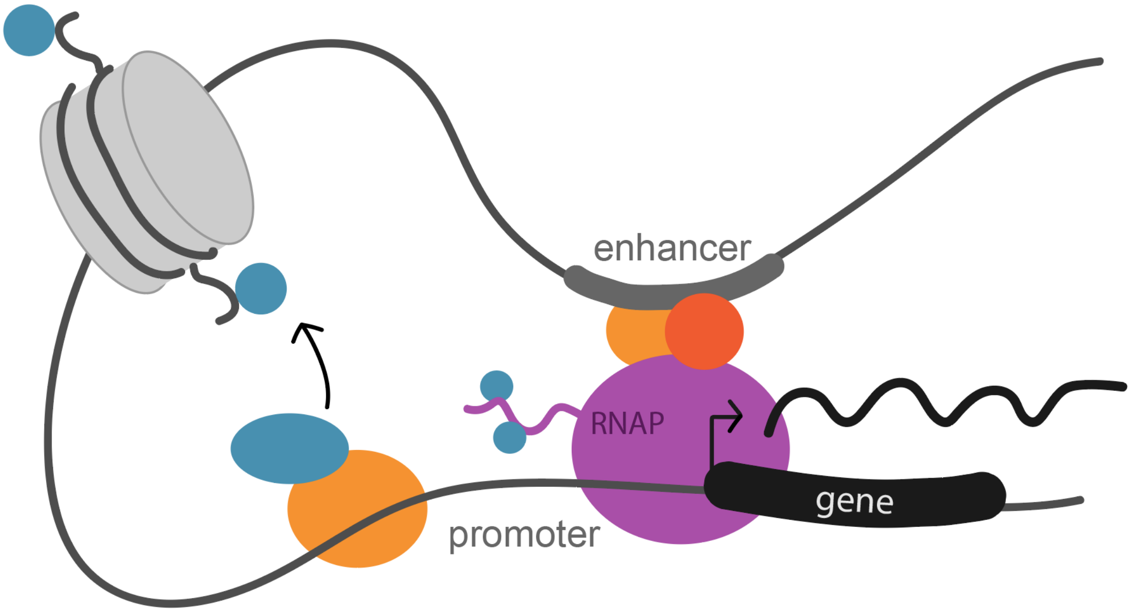


Figure 1.1. Schematic of protein-coding gene regulation in eukaryotes. Gene expression is regulated by promoter sequence, RNA polymerase (purple), distal enhancers (gray), associated transcription factors (orange) and cofactors (blue), and histone marks (blue circles). Additionally, chromatin compaction, nucleosome occupancy, and the 3D genome are also involved in regulating the turning on and off of genes.

Chromatin as another layer of complexity

Chromatin is a eukaryotic innovation that allows for compaction and protection of large genomes within the nucleus, and serves as a default regulator of most genes. The core components of chromatin structure are nucleosomes, which also play important roles in context-specific gene expression (Luger *et al*. 2012). Nucleosomes are made up of a histone octamer with unstructured histone tails that jut out of the octamer. Each nucleosome is wrapped by ~150 bp of DNA, and multiple nucleosomes adjacent to one another can function to occlude the underlying DNA sequence, making genes inaccessible to transcription. Remodeling of chromatin by chromatin remodeling enzymes involves energy-dependent movement and disassembly of nucleosomes to reveal underlying sequences to be accessed by TFs, allowing transcription to occur. Histone tails are modified by histone writers such as acetylases, kinases, and methyltransferases, which also play important roles; depending on the type and combination of modifications, these modifications can induce a repressive or an activated chromatin state. For instance, addition of acetylation marks on histone lysines is generally associated with gene activation, while removal of acetylation marks by histone deacetylases is repressive.

Added regulation by transcription factors

A sizable fraction of the genome is dedicated to encoding transcription factors, DNA-binding proteins that bind to promoters and enhancers to control transcript levels and the specificity of transcription. In humans, over 1,600 TFs have been identified (Lambert *et al.* 2018). These TFs may bind to hundreds or thousands of loci across the genome and are critical for gene expression, although the number of genes regulated by such TFs is typically much less than the number of bound loci, an observation that we return to in Chapter 3. TFs can be categorized as activators, which enhance gene expression, or repressors, which turn off gene expression; yet, some TFs serve both roles, depending on context (Lambert *et al.* 2018). TF action is mediated by both direct contact with factors of the core transcriptional machinery as well as modifications of the chromatin environment. Genome-wide studies and focused molecular biological investigations have revealed that activators and their associated factors can open chromatin, establishing an epigenetic state that allows for other TFs and the basal transcriptional machinery to be recruited to gene promoters, as well as promoting release of RNA polymerase from the PIC (Weake and Workman, 2010). Repressors, on the other hand, have been shown to induce a condensed chromatin state by compacting nucleosomes, altering histone tails, blocking the PIC formation, or blocking activation domains in the activators (Payankaulam, Li, and Arnosti, 2010). Most TFs show sequence-specific activity, binding to particular motifs to impact transcription across the genome. A complex cis-regulatory grammar underlies the action of ensembles of TFs, which may involve cooperative binding between adjacent TFs, interactions with nucleosomes, range-dependent activities of transcriptional effects, and more (Reviewed in Kim and Wysocka, 2023).

Corepressors as key players of gene regulation

Key players of gene regulation, which are the focus of this dissertation, are corepressors. Corepressors do not bind to DNA on their own as repressors do, but are recruited to DNA by other DNA-binding factors. Corepressors are diverse in their functions; some appear to lack intrinsic enzymatic activity, and instead function as scaffolds to assemble complexes of other factors such as histone deacetylases or methyltransferases (Payankaulam, Li, and Arnosti, 2010). Others possess enzymatic functions, which are not always well understood, as we discuss in Chapter 4. Yet others block the transactivation domain of activators or inhibit the recruitment of components of the PIC. Some corepressors use multiple means to exert their repressive effects across the genome, in a context-dependent manner. The large diversity in corepressor function and their pleiotropic effects make them challenging to study; corepressors may have tissue-specific effects, diverse interaction partners, promoter specificity, or function at particular developmental timepoints.

Long- and short-range repression mechanisms

Repression mechanisms have been described to be either long-range or short-range, which are functionally distinct pathways (Gray and Levine, 1996). Long-range repression is associated with repressors working over long distances from affected activators (>1 kb), loss of acetylation marks, and inhibition of RNA polymerase binding (Reviewed in Courey and Jia, 2001). In contrast, short-range repression is local, working within ~100 bp of an activator. It is associated with local deacetylation, nucleosome compaction, but no impact on PIC formation. Well-studied long-range repressors include Hairy, which interacts with the Groucho corepressor that spreads across a large locus and mediates long-distance silencing (Kok *et al*. 2015). Short-range repressors include gap gene encoded proteins such as Giant and Knirps, which can interact with corepressors such as CtBP for localized repression (Arnosti *et al*. 1996; Strunk *et al*. 2001). Interestingly, CtBP can also interact with Hairy, and Groucho with Knirps, suggesting that long- and short-range repression mechanisms do not absolutely need to use distinct cofactors, but may use cofactors in diverse ways, depending on the context (Payankaulam and Arnosti, 2009; Bianchi-Frias *et al*. 2004).

The impact of genome organization and phase separated condensates

Although not explored in my research, the three-dimensional architecture of the genome is critical for gene regulation. Architectural proteins such as CTCF and cohesin function to create Topologically Associated Domains (TADs), in which regions of the genome that are physically distinct associate with one another in specific 3D conformations (Reviewed in Kim and Wysocka, 2023). TADs include regions of the DNA that may be far away on the linear genome but are brought into close proximity, such as promoters and distal enhancers. This long-range promoter-enhancer interaction is important for transcriptional control, as regulatory elements and transcription factors on distal enhancers can communicate with the basal transcriptional machinery and assist with the formation of the PIC and RNA polymerase release. The interaction of enhancers and promoters within TADs may be mediated by phase-separation, in which TFs and other factors necessary for transcription are found in hubs called molecular condensates. These molecular, phase-separated transcriptional condensates are composed of a large cluster of TFs, the basal transcriptional machinery, and the Mediator complex (Richter *et al*. 2022). They are maintained through intrinsically disordered regions of the various TFs found within them. We hypothesize that the CtBP corepressor, with its intrinsically disordered C-terminal domain, may also participate in or play a role in the formation of transcriptional condensates, although this has not been formally studied (Soto *et al.* 2022).

Evolution's impact on gene regulation

The complexity of gene regulatory processes described above is further magnified by evolutionary forces acting on DNA both on cis-regulatory elements (CRE) and on trans-acting factors (Reviewed in Stern and Orgogozo, 2008). Over evolutionary time, random mutations can accumulate in CREs, altering transcription factor binding motifs, which can change the types of TFs that bind, or the affinity of a TF for its binding site (Pan *et al.* 2010). CRE mutations can lead to altered gene expression. Mutations in coding sequences can also impact the function of the encoded protein product, whether it be TFs or other transcriptional machinery. These mutations may lead to changes in TF interaction partners, the inhibition or enhancement of their typical activities, or to the gain or loss of functions. These evolutionary changes, especially those in CREs, have been demonstrated to drive morphological variations and transitions in diverse lineages (Stern and Orgogozo, 2008).

Throughout evolutionary time, the emergence of duplicated genes and of alternative splicing has also led to diversification of transcription factors. Gene duplication, which introduces additional paralogs to a TF family, impacts the activity of transcription factors, with diverse consequences (Conrad and Antonarakis, 2007). Duplication is most commonly followed by gene loss, as the second copy of the gene and encoded protein product are redundant. However, if the gene is not immediately lost, one of the paralogs may subfunctionalize, and take on a subset of the roles of the original copy, or it may neofunctionalize, and take on a novel function in the cell

(Conrad and Antonarakis, 2007). These two possibilities lead to the retention of the paralogous gene, as the duplicated copy has a new, important role. This has been observed for a variety of transcription factors, including genes encoding the well-characterized HOX proteins, which have duplicated up to eight times in fish, with duplicates taking on new roles in body plan organization (Ferrier, 2016). This is also the case with the retinoblastoma (Rb) corepressor protein, which duplicated multiple times in vertebrates and experienced an independent gene duplication in the genus Drosophila as well.

Additional diversification exists through generation of distinct splice forms producing variants of the same protein, created at the level of alternative mRNA splicing. One example of this is the Capicua (Cic) gene in Drosophila, whose transcript is alternatively spliced to produce a short (S) and a long (L) isoform. The S isoform has a shorter N-terminal domain, with an additional motif termed N2 (Forés *et al*. 2015). The presence of the N2 motif in Cic-S allows it to interact with Groucho for transcriptional repression, while Cic-L does not. Here, alternative splicing is clearly important in generating splice forms with differential impacts on gene regulation. This is also the case with CtBP, which produces two major splice variants in Drosophila: one that contains an unstructured C-terminal domain of about 130 amino acids (AA), and one that lacks this domain, which we explore further in Chapter 4 and 5. The selective advantage for retaining duplicated genes and for expressing alternatively spliced isoforms is still unknown. We will focus the rest of this introduction and dissertation on two key metazoan corepressors, Rb and CtBP, which have experienced diversification at many levels, and whose gene regulatory roles have undoubtedly been impacted by evolutionary change.

THE HUMAN RETINOBLASTOMA TUMOR SUPPRESSOR PROTEIN

Overview of Rb

The retinoblastoma tumor suppressor protein (Rb) is an ancient and highly conserved transcriptional cofactor, and expressed in most eukaryotes from unicellular green algae to humans. It is speculated that the ancestral Rb protein may have played a role in the transition from unicellular to multicellular life (Cao *et al*. 2010). Rb proteins have been identified across Eukaryota; aside from Metazoa, where the Rb family was first identified, Rb has also been found in plants and in unicellular eukaryotes such as Chlamydomonas (encoded by mat3). The yeast Rb-like protein, Whi5, does not bear sequence or structural homology to Rb but functions in a similar manner to regulate cell cycle progression, suggesting parallel evolution (Reviewed in Wirt and Sage, 2010). All vertebrates, including humans, encode at least three Rb proteins, with some even encoding up to six (Liban *et al*. 2017). In contrast, most eukaryotes encode a single Rb homolog that mediates all the conserved functions of this protein. Interestingly, in several other lineages, independent gene duplication events led to the diversification of the family. This includes the Drosophila lineage, which is the only known arthropod to have duplicated and retained both Rb gene (**Figure 1.2**). Independent duplications have also been identified in plants, sea urchins, and placozoans (Cao *et al*. 2010).

Rb was the first tumor suppressor to be identified. Studies of pediatric retinoblastoma revealed that mutations in both copies of a gene later found to be the RB1 gene triggers early childhood retinal cancer. Alfred Knudson's inference that a loss of heterozygosity of RB1 was the cause of retinoblastoma led to his "two-hit hypothesis", which proposed that cancer can be caused by as few as two mutations (Knudson, 1971). Since then, mutations in RB1 or in genes that regulate

Rb function have been found in a majority of human cancers, including osteosarcoma, small cell lung cancer, ovarian cancer, and breast cancer (Reviewed in Flores and Goodrich, 2022).

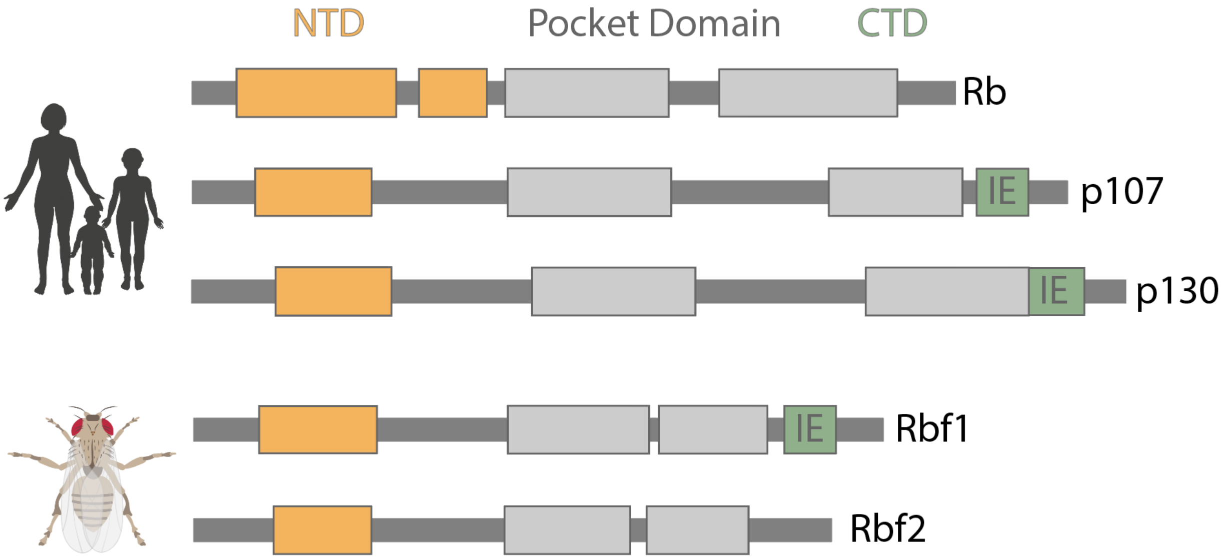


Figure 1.2. Schematic of Rb proteins in humans and in flies. The *Homo sapiens* genome encodes three Rb family members (Rb, p107 and p130). Drosophila duplicated Rb to produce the Rbf1 and Rbf2 paralogs. All Rb proteins contain a central pocket domain, a structured N-terminal domain (NTD), and a more diverged and unstructured C-terminal domain (CTD) that contains an Instability Element (in green), in some orthologs. Adapted from Sengupta *et al*. 2015.

Rb proteins are transcriptional corepressor proteins that regulate gene expression through binding to members of the E2F transcriptional activator family, which themselves bind to many gene promoters. The binding of Rb to the transactivation domain of E2F and its inhibition contributes to repression of target genes. Additionally, through direct contacts with the core transcriptional machinery, Rb can directly inhibit the formation of the PIC at transcriptional start sites (Ross *et al*. 1999; Reviewed in Payankaulam *et al*. 2010). Rb also recruits additional cofactors such as chromatin regulators to gene promoters, which can remodel the chromatin environment to induce heterochromatin formation, thereby silencing target genes.

The Rb association with E2F orchestrates the entry of cells into S phase. In G1, the hypophosphorylated form of Rb binds to E2F and represses cell cycle genes. When the cell has received cues to move into S phase, Rb is phosphorylated by cyclin-dependent kinases, and this releases it from E2F (reviewed in Weinberg, 1995; **Figure 1.3**). Regulation of Rb via phosphorylation is not as simple as it was first proposed; the 16 phosphorylation sites across the Rb protein can be mono-phosphorylated, with distinct downstream effects in G1 (Narasihma *et al*. 2014; Sanidas *et al*. 2019). Hyperphosphorylation in late G1 leads to Rb dissociation from E2F, activation of the E2F-regulated genes, and movement into S phase. In addition to this canonical cell cycle regulatory role, Rb proteins play roles in cell differentiation, DNA replication, apoptosis, and various metabolic processes (Reviewed in Chau and Wang, 2003, Dick and Rubin, 2013, Nicolay and Dyson, 2013).

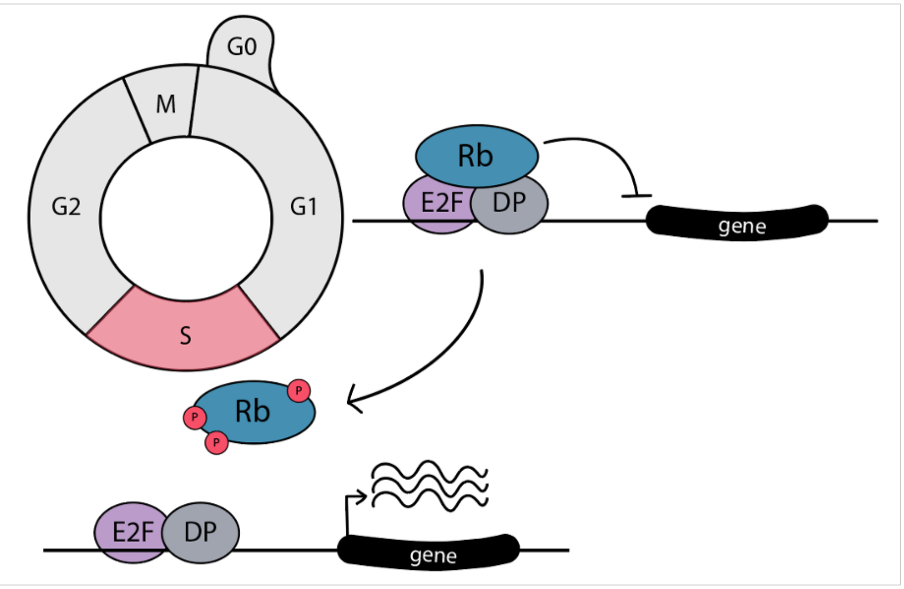


Figure 1.3. Canonical regulation of cell cycle genes by Rb. During the G1 phase of the cell cycle, Rb is found on E2F-bound promoters. Rb binding to E2F, which is colocalized with its dimerization partner (DP), inactivates E2F and represses expression of the nearby gene. At the end of G1 phase, when the cell receives the cues to release the cell cycle block, Rb is phosphorylated by cyclin-dependent kinases, and this releases Rb from E2F to allow for expression of the downstream gene. The expression of these genes allows for entry into S phase.

Aside from Rb, humans encode two other paralogs, called p107 and p130; together, they form the Rb family of pocket proteins (**Figure 1.2**; Weinberg, 1995). Their name originates from the fact that they contain a large pocket-like domain with which they bind to E2F and to chromatin modifiers using different residues. Rb family proteins are transcriptional corepressors, regulating cell cycle progression; Rb and p107 specifically regulate the G1 to S phase transition, while p130 is predominantly active in G0 phase (Jiang *et al*. 2000; Takahashi *et al*. 2000; Reviewed in Wirt and Sage, 2010). Rb is considered the predominant tumor suppressor; it is found in most cell types, and it is mutated in cancer more often than p107 and p130 (Reviewed in Flores and Goodrich, 2022). These three paralogs regulate some of the same sets of genes and processes, but also play non-redundant roles in the cell, as described below.

Structure of Rb family proteins

The 928 AA human Rb protein has a mass of ~105 kDa and is encoded by the RB1 gene located on Chromosome 13. p107 and p130 are similar in size and structure (**Figure 1.2**; Reviewed in Classon and Dyson, 2001). Rb proteins are made up of three modular domains: the N-terminal domain (NTD), the central pocket domain, and the C-terminal domain (CTD). The central pocket domain is made up of two subdomains called the cyclin A and the cyclin B folds. The pocket is necessary but not sufficient on its own for suppressing cell growth, as it still needs the CTD for full activity (Qin *et al*. 1992). The pocket is considered the minimal region that is necessary for viral oncoprotein binding, including binding by E1A, T antigen, and E7 (Kaelin *et al*. 1990). Importantly, it contains two binding sites that are crucial for its activity: the interface between the cyclin A and B folds (the “linker”, or “spacer”), which interacts with the E2F transactivation domain, and the LxCxE binding motif in the cyclin B fold, which binds a plethora of interacting partners such as chromatin remodelers and histone modifiers that contain an LxCxE motif (reviewed in Zhang and Dean, 2001). The spacer between the cyclin A and the cyclin B folds has diverged between the three Rbs, and the differences in sequence and length may lead to differential regulation of the proteins and differential interactions with cofactors (Classon and Dyson, 2001). For instance, cyclin A/cdk2 and cyclin E/cdk2 complexes specifically bind to the p107/p130 spacer regions, but not the Rb spacer (Mulligan and Jacks, 1998).

The NTD, on the other hand, is less well-characterized, but has a cyclin fold structure similar to that of the pocket domain (Hassler *et al.* 2007). The NTD has been deemed dispensable for Rb's activity as assayed on reporter genes (Goodrich, 2003). Still, several proteins can bind the NTD directly, and both missense and in-frame deletion mutations in the N-terminus are found in hereditary retinoblastoma, indicating that this domain plays a role in Rb activity (Hassler *et al.* 2007). In contrast, the C-terminal domain is highly unstructured and contains several residues that are involved in making contacts with the E2F "marked box" domain (Rubin *et al.* 2005). Other residues in the CTD are phosphorylation targets that regulate Rb conformation and activity (Reviewed in Rubin, 2013; Sanidas *et al.* 2019). For example, an alpha helix in the Rb CTD binds to cyclin D/cdk4,6, while the RxL docking motifs directly N-terminal to this structured region are used to interact with cyclin E and A complexes (Topacio *et al.* 2019). The CTD also contains the Instability Element (IE), found only in p107 and p130, which we will discuss later, as it was discovered in Drosophila. Removal of the IE in p107 and p130 leads to decreased repression activity in cell culture (Sengupta *et al.* 2015).

The Rb-E2F interaction

The human E2F family comprises eight unique proteins, most of which interact with the Rb paralogs. E2F1-3 are transcriptional activators, E2F4-6 are repressors, and E2F7 and E2F8 are atypical repressors (Reviewed in Kent and Leone, 2019 and Broeker and Andrechek, 2021). Only

E2F1-5 interact with the Rb proteins, while E2F6-8 have diverged and do not have known Rb interactions. E2F1-6 dimerize with DP, a dimerization partner that is required for DNA binding (Trimarchi and Lees, 2002). Rb interacts with E2F1-5 while p107 and p130 have preferential binding for E2F4 and E2F5 (**Figure 1.4**; Reviewed in Wirt and Sage, 2010). The interactions are predominantly mediated by the Rb pocket domain and the E2F transactivation domain (TAD); the E2F TAD inserts into the cleft between the cyclin A and cyclin B folds of the Rb pocket (Lee *et al*. 2002).

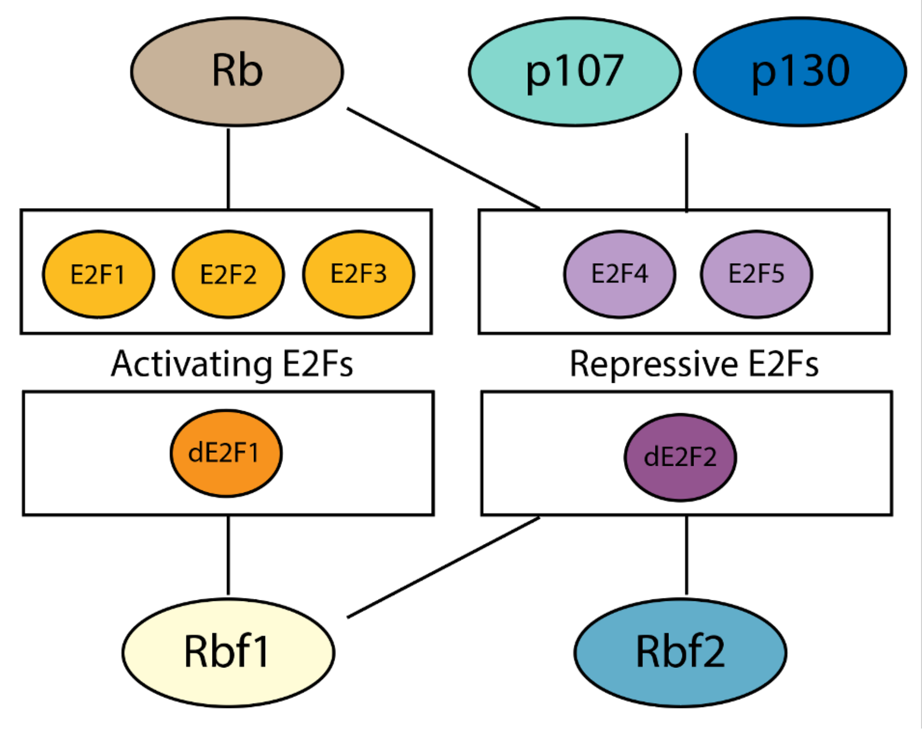


Figure 1.4. Schematic of Rb-E2F interactions in humans (above), and in Drosophila (below). Rb interacts with E2F1-3 activators and E2F4, 5 repressors. E2F6-8 do not bind Rb proteins and are not shown here. p107 and p130 preferentially interact with the repressor E2Fs. The Drosophila Rbf1 interacts with both dE2F1 and dE2F2, while Rbf2 preferentially interacts with dE2F2. Figure adapted from Du and Pogoriler (2006).

Additional interactions are found between the Rb CTD and the E2F marked box domain, which may provide further specificity (Xiao *et al*. 2003). All Rb family members share this property of binding to E2Fs, but with differences in binding affinities (Reviewed in Classon and Dyson, 2001). For instance, the p107 CTD binds to the E2F4 marked box domain with much greater affinity than for E2F1 (Liban *et al.* 2017). Conversely, Rb has similar affinity for E2F1 and E2F4. The typical E2F proteins regulate expression of genes involved in DNA repair, DNA replication, autophagy, and the cell cycle, among others; Rb inactivation of E2F proteins by binding its TAD leads to repression of these E2F target genes (Reviewed in Broeker and Andrechek, 2021).

Promoter-proximal mechanism of gene regulation by Rb proteins

Rb was initially identified as a factor that binds gene promoters, where E2F proteins are also located. ChIP-qPCR identified Rb paralog binding to endogenous cell cycle gene promoters, within 100 bp of the TSS (Takahashi *et al.* 2000). More recently, ChIP-seq performed in quiescent and senescent human cells showed that Rb family proteins are bound promoter-proximally, with most binding sites mapping within 1 kb of a gene's TSS (Chicas *et al.* 2010). Many Rb binding sites are known E2F target genes, or are enriched for E2F motifs (Chicas *et al.* 2010; Ertel *et al.* 2010; Nicolay *et al.* 2015). Yet, Rb binding to non-E2F regulated genes and sites without E2F motifs also exist, indicating both E2F-dependent and possibly E2F-independent modes of transcriptional regulation by Rb proteins.

Non-promoter binding has also been observed on enhancers and repeat elements such as LINEs and SINEs (Kareta *et al.* 2015, Ishak *et al.* 2016). A recent study by the Dyson lab showed that Rb binds E2F-enriched promoters, c-Jun-enriched enhancers, and CTCF-enriched insulator elements (Sanidas *et al.* 2022). Almost all the Rb-bound loci in their study performed in human retinal epithelial cells were co-localized with one of these three proteins, in a mutually exclusive manner. Interestingly, only 30% of the Rb bound sites in this tissue were near promoters. In a different cell type, promoters comprised 50% of bound sites, suggesting that Rb genome-wide localization is a cell-type dependent feature (Sanidas *et al.* 2022).

Studies using Rb fusions to the DNA-binding domain of the yeast GAL4 transcription factor have also been useful in identifying distances from which Rb proteins can function on transfected reporters (Weintraub *et al.* 1995; Bremner *et al.* 1995; Adnane *et al.* 1995). For example, GAL4-Rb represses reporter genes with UAS sites near the transcriptional start site. It is also able to repress the HSV tk promoter from distal locations, hundreds of base pairs upstream or downstream of the CAT reporter's TSS (Adnane *et al.* 1995). GAL4-Rb can also repress an SV40 enhancer from 2kb upstream (Weintraub *et al.* 1995). These studies have also identified that in many contexts, GAL4-p107 and GAL4-p130 function similarly to GAL4-Rb fusions (Bremner *et al.* 1995; Ferreira *et al.* 1998; Meloni *et al.* 1999). In general, a preponderance of evidence suggests that Rb proteins typically function promoter-proximally on endogenous loci to exert their gene regulatory effects, possibly due to their reliance on E2F proteins, which themselves are promoter proximal.

Rb gene targets

Rb was initially characterized as a repressor of cell-cycle genes required for entry into S phase (Weinberg, 1995). Depletion of Rb using RNAi in senescent human cells leads to upregulation of cell cycle and DNA replication genes as expected (Chicas *et al.* 2010). However, Rb regulation is not limited to cell cycle genes; it has been shown to regulate genes involved in diverse processes such as apoptosis, DNA repair, and cell differentiation as well (Gordon and Du, 2011). Additionally, there are a few reports of Rb functioning as an activator of gene expression, such as on the TGF-β promoter, and genes that are downregulated after Rb loss, indicating that Rb may either directly or indirectly also activate gene expression in some contexts (Kim *et al*. 1992; Chicas *et al*. 2010; Reviewed in Chinnam and Goodrich, 2011).

Rb interaction with chromatin modifiers and remodelers

To effect transcriptional repression, Rb proteins recruit chromatin and histone modifiers, which facilitate the remodeling of the chromatin environment at the level of histone tail modifications and nucleosome positioning (Reviewed in Fiorentino *et al*. 2013). The association between Rb and chromatin enzymes requires, in many cases, a particular region of the cyclin B fold in the Rb pocket, and the LxCxE motif of the Rb-interacting cofactor. The presence of an exogenous LxCxE peptide or specific viral oncoprotein can inhibit these interactions, disrupting Rb function (Brehm *et al*. 1998, Vandel *et al*. 2001). Yet, some cofactors can bind independent of this LxCxE motif (Zhang *et al*. 2000; Isaac *et al*. 2006).

Extensive *in vitro* biochemical assays uncovered the binding of cofactors, which are necessary for Rb's repressive activity on certain promoters and in certain cell types. For example, Rb associates directly with histone deacetylases, HDAC1 and HDAC2, which deacetylate H3 and H4 tails to create transcriptionally inactive chromatin (Brehm *et al*. 1998, Luo *et al*. 1998, Magnaghi-Jaulin *et al*. 1998). Trichostatin A inhibition of HDAC activity in cell culture inhibits Rb's ability to repress a reporter gene, confirming the reliance on deacetylase activity in certain contexts (Magnaghi-Jaulin *et al*. 1998). The nucleosome remodeling and deacetylase complex NuRD can also be recruited by Rb to deacetylate histones for heterochromatin formation (Montoya-Durango *et al*. 2016).

Rb proteins also interact with chromatin remodelers such as BRG1 and hBRM, components of the SWI/SNF remodeling complex (Dunaief *et al*. 1994, Trouche *et al*. 1997, Zhang *et al*. 2000). BRG1 activity is not required in all contexts, but some promoters are sensitive to BRG1-Rb repression complexes, and sometimes also require simultaneous HDAC interactions (*Zhang et al.* 2000). Interactions with histone methyltransferases (HMT) like Suv(39)H1, which places H3K9me marks, have also been observed (Vandel *et al.* 2001, Nielsen *et al.* 2001). The lack of identifiable LxCxE motif in BRG1 and in HMT suggests that these interactions may be indirect, or happen through other domains in Rb. Interestingly, Rb can also interact with the HP1 heterochromatin binding protein, which recognizes and binds the H3K9me2/3 histone marks (Nielsen *et al.* 2001). The interactions with Suv(39)H1 and HP1 may create a large repression complex that is necessary for heterochromatin formation (Isaac *et al.* 2006). Aside from chromatin modifiers and remodelers, Rb can interact with DNA methyltransferases like DNMT1, which further represses a CAT reporter (Robertson *et al.* 2000). The repression complex with DNMT1 may also be simultaneously bound by HDACs. The associations described above are not limited to Rb; p107 and p130 interact with many of the same factors (Ferreira *et al.* 1998, Reviewed in Fiorentino *et al.* 2013). Yet, there is evidence of paralog-specific protein interactions as well; for instance, p107 specifically interacts with Smad3 and Sp1, suggesting some paralog-specific mechanisms of gene regulation (Wirt and Sage, 2010).

Differences between Rb family paralogs

Although comparison of Rb paralogs in cell culture through GAL4 fusions has illustrated the many ways in which Rb family paralogs overlap in function and repressive effect, differences in activity have also been identified (Chow *et al.* 1996; Luo *et al.* 1998; Ferreira *et al.* 1998; Meloni *et al.* 1999). Like Rb, the p107 and p130 paralogs were discovered through associations with viral oncoproteins (Reviewed in Mulligan and Jacks, 1998). p107 and p130 have ~50% sequence conservation at the protein level, and are structurally more similar to each other than they are to Rb (Li *et al.* 1993). Based on similarities to Rb proteins in other lineages, p107 and p130 are likely to more closely resemble the last common ancestral sequence. Rb is more derived, with only ~30% sequence identity to p107 and p130 (Mulligan and Jacks, 1998).

In humans, only Rb loss causes retinoblastoma, and it is primarily Rb that is found to be mutated in most types of cancer, while p107 and p130 are not mutated as often (Chinnam and Goodrich, 2011). In mouse models, Rb null mice die around E13, while p107 or p130 null mice do not exhibit overt phenotypes (Reviewed in Lipinski and Jacks, 1999). However, mice lacking both p107 and p130 die shortly after birth, indicating that p107 and p130 have overlapping roles that are compensated for when one of them is missing. Perhaps the null phenotypes reflect the differential expression of mammalian Rb proteins; Rb is expressed at moderate levels in most cell types, while p130 is highly expressed in quiescent cells in G0, and p107 in cells stimulated to progress through the cell cycle (Reviewed in Classon and Dyson, 2001).

An additional important difference between the paralogs is that they have selective impact on gene promoters. Although there are reported instances of Rb family proteins binding and regulating the same gene promoters, many gene promoters are responsive to select Rb proteins at select times in the cell cycle. For example, the B-myb promoter is only responsive to p107 and p130 in 3T3 cells, but is not sensitive to Rb, while the p107 promoter is only responsive to Rb (Classon *et al.* 2000). At the level of the entire transcriptome, Rb family paralogs regulate a distinct set of genes in senescent human cells (Chicas *et al.* 2010). Rb depletion using RNAi upregulates canonical cell cycle and DNA replication genes, p107 depletion downregulates many metabolism-related genes including OXPHOS and mitochondrial genes, and p130 depletion upregulates various organelle-related genes (Chicas *et al.* 2010). The overlap of differentially expressed genes by all three Rb paralogs is only ~5%. Rb is the predominant regulator in these senescent cells; meanwhile, this pattern does not hold in quiescent cells, in which loss of one paralog is compensated for by the other paralogs, indicating that the paralogs may regulate a similar set of genes in this cell state. Whether the observed differences among the Rb paralogs are due to different intrinsic repression ability or simply due to the selective E2F interactions they are forming and the subsequent DNA recruitment has not been clarified.

RB IN DROSOPHILA

Drosophila melanogaster is a system that is well-suited for studying Rb function and the consequences of Rb gene duplication, as an independent gene duplication event in Drosophila led to the expression of two Rb paralogs. The Drosophila Rb family involves fewer components than the mammalian system, yet with highly conserved functions. Additionally, the *Drosophila melanogaster* model system is genetically tractable, and has molecular tools to facilitate the study of protein functions *in vivo*.

The ancestral Drosophila lineage experienced an independent gene duplication event that led to the expression of two paralogs, called Rbf1 and Rbf2, that function in similar ways to the three mammalian paralogs (Du *et al*. 1996; Stevaux *et al*. 2002). Rbf1 is structurally more similar to mammalian p107 and p130, but genetically functions more similarly to Rb, as the predominant regulator of cell cycle genes. The protein sequence of Drosophila Rbf2 is more derived, like that of Rb, but Rbf2's genetic activity may be more similar to p107 and p130, in that it is not required for viability, and may play a larger role in regulation of cell growth-related genes. Rbf2 may play a redundant role with Rbf1 on some promoters, although this idea remains to be validated in many cases. Both Drosophila Rb paralogs bind E2F proteins, with Rbf1 interacting with dE2F1 and dE2F2, while Rbf2 only interacts with dE2F2 (**Figure 1.4**; Stevaux *et al*. 2002).

Rbf1 is expressed throughout most developmental stages and in the majority of fly tissues, while Rbf2 shows more dynamic embryonic expression, with a peak in levels at 8-10 hours (Keller *et al.* 2005). Rbf1 and Rbf2 have similar patterns of expression in larval tissues, but in the adult, Rbf2 is mostly restricted to the female ovary (Stevaux *et al.* 2002). Loss of rbf1 results in larval lethality, while rbf2 null flies are viable but experience changes in fertility and lifespan (Du and Dyson, 1999; Steveaux *et al.* 2002; Steveaux *et al.* 2005; Mouawad *et al.* 2019). Although rbf2 is not an "essential" gene, it has been retained in the Drosophila lineage for ~50 million years, suggesting important roles.

Comparison across different Drosophila species

All Drosophila genomes sequenced to date encode Rbf1 and Rbf2, a duplication that is specific to the Drosophila genus; no other arthropods characterized thus far show evidence of an Rb gene duplication. Rbf2 is more highly derived than Rbf1; specifically, the spacer in the pocket domain, the N-terminus, and the C-terminus show much evolutionary variation among 12 Drosophila species, and the CTD does not closely resemble that of other Rb proteins (Mouawad *et al.* 2019). Unsurprisingly, the most highly conserved region of Rb proteins are the cyclin A and cyclin B folds of the pocket domain, which are also conserved across metazoans. The diversification of Rbf2 protein sequence suggests possible subfunctionalization or neofunctionalization of the protein, where structural changes in Rbf2 allowed it to take on different roles from Rbf1, perhaps through unique protein interactions.

Regulation of gene expression by Rb proteins

Like mammalian Rb, the Drosophila Rbf1 protein regulates cell cycle genes, particularly at the G1 to S phase transition of the cell cycle (Du *et al.* 1996; Du and Dyson, 1999; Dimova *et al.* 2003). Rbf2 has some overlapping roles with Rbf1 in gene regulation (Stevaux *et al.* 2002). For instance, co-expression of Rbf2 and E2F2 also leads to cells stalling in G1 (Stevaux *et al.* 2002). Yet, Rbf2 is unable to repress cell cycle genes such as *PCNA* very well, and RNAi depletion of rbf2 has few effects in cell culture, compared to the variety of misregulated genes after rbf1 depletion (Stevaux *et al.* 2002; Dimova *et al.* 2003). Thus, Rbf2 was initially labeled as a weaker repressor than Rbf1, possibly playing a predominantly redundant role.

Comparative exo-ChIP analysis of Rbf1 and Rbf2 binding sites in fly embryos revealed surprising binding profiles for these two proteins. In addition to cell cycle and DNA replication genes, Rbf1 binds to genes involved in signaling pathways such as insulin signaling, Hippo, and JAK/STAT signaling—a previously uncharacterized feature of Rb proteins (Acharya *et al.* 2012; Wei *et al.* 2015). Most of the Rbf1-bound genes are also bound by Rbf2, but Rbf2 has an additional ~2,000 unique binding sites, including ribosomal protein and mitochondrial protein genes (Wei *et al.* 2015). Some of these ribosomal protein genes are modestly repressed by Rbf2, such as tko, but not by Rbf1. Others are clearly regulated only by Rbf1 as indicated by RNAi and luciferase assays (Wei *et al.* 2015).

Overexpression of Rbf1 and Rbf2 in Drosophila embryos followed by RNA-seq also uncovered strikingly different effects of each paralog (Mouawad *et al.* 2019). Many genes were selectively repressed by Rbf2, including mitochondrial and ribosomal protein genes. In contrast, Rbf1 repressed canonical cell cycle genes which Rbf2 failed to repress. A subset of cell cycle genes are found to be upregulated upon Rbf2 overexpression, possibly due to competitive interactions between Rbf1 and Rbf2 for these promoters (Mouawad *et al.* 2019). We hypothesize that the Rbf2-specific binding and regulation of ribosomal protein genes may point to subfunctionalization, where Rbf2 took on cell growth control roles, and perhaps neofunctionalized for reproduction roles. It is still unclear what drives differential targeting of Rbf1 and Rbf2 to specific promoters, and whether the inherent biochemical activities mediated by Rbf1 and Rbf2 are identical. For instance, until I conducted studies with the CRISPRi system described in this dissertation, we had not been able to assess whether Rbf1 and Rbf2 can regulate any promoter, provided it is recruited to that site in the genome.

Rb interactions with histone modifiers

Rb interactions with histone and chromatin remodelers, as observed in the mammalian system, are also conserved in the fly. Mass spectrometry analysis performed on endogenous Rbf2 purified from embryonic Drosophila nuclei identified Rbf2 interactions with the Moira and BAF53 chromatin remodelers, the TRRAP histone acetyltransferase, and dREAM complex components such as Mip130 (Ullah *et al.* 2007). Surprisingly, it also purified with subunits of the COP9 signalosome, which is a conserved regulatory complex originally identified as a repressor of light-induced development in Arabidopsis. This COP9 interaction was found to be important for the stability of Rbf1 and Rbf2, protecting them from degradation by the proteasome (Ullah *et al.* 2007). We expect that these interactions are predominantly mediated by the pocket domain, which has high conservation to the mammalian Rb pockets.

Studies of mutant Rbf1 to understand structural features important for repression

Structure-function studies in our laboratory have identified aspects of Drosophila Rb proteins critical for their functions. Through the creation of mutant forms of Rbf1, followed by *in vivo* overexpression or transfection in cell culture assays, we uncovered details about the significance of the divergent CTD and the conserved pocket domain. A striking feature of Rbf1, which Rbf2 does not possess, is a C-terminal Instability Element (IE), which was discovered in Drosophila (Acharya *et al.* 2010). Removal of the ~60 residue IE from Rbf1 leads to Rbf1 accumulation. An Rbf1$^{\Delta IE}$ mutant is functionally attenuated: Rbf1$^{\Delta IE}$ overexpression in developing Drosophila eyes using *ey*-GAL4 leads to normal eye development while overexpression of WT Rbf1 leads to some eyes having severe defects, indicating that this mutation may be hypomorphic

(Zhang *et al.* 2014). Rbf1$^{\Delta IE}$ is also unable to repress cell cycle promoters, such as a *PCNA*-luciferase reporter (Raj *et al.* 2012a). Yet in some contexts, this mutant retains activity, as Rbf1$^{\Delta IE}$ overexpression in the developing wing leads to wings which are significantly larger than control, and Rbf1$^{\Delta IE}$ can still repress certain genes like *InR*, *Wts*, and *Pi3K* in cell culture (Raj *et al.* 2012a; Elenbaas *et al.* 2015). The reduced repressive potential in some contexts may be caused by its inability to bind to endogenous E2F proteins that it relies on for DNA recruitment. However, according to Co-IP, Rbf1$^{\Delta IE}$ is still able to bind to E2F, so perhaps the IE functions in repression, independent of recruitment, and may be a "repression domain" of some sort (Acharya *et al.* 2010). This is an idea we test in Chapter 2.

Another much more critical domain of Rb is the central pocket domain, for which Rb proteins are named. Studies of the human Rb pocket domain have deemed the pocket to be necessary for Rb's repressive activity (Weintraub *et al.* 1995). Just the NTD or CTD alone, when expressed in cultured cells, are not sufficient for Rb's transcriptional repressor activity on reporter genes. As described above, the pocket is the site of E2F interactions and interactions with cofactors like histone modifiers, which give Rb its repressive potential. Likewise, a Rbf1$^{\Delta pocket}$ mutant in the fly diminishes its ability to repress reporter genes such as *PCNA* and *DNApolα* (Acharya *et al.* 2010). This mutant is unable to interact with E2F1 in a Co-IP experiment, suggesting that lack of E2F binding and presumed lack of cofactor interactions make this mutant unable to function as a repressor. Interestingly, the Rbf1 pocket on its own is a good repressor in cell culture, but it does not possess the entire activity of WT Rbf1 (Acharya *et al.* 2010). Thus, the NTD and CTD may play some role in repression mechanisms. These experiments described here have underlined the commonalities between the Drosophila and mammalian Rb proteins, and have furthered our knowledge of functional properties of the Rb corepressor family.

OVERVIEW OF THE CTBP TRANSCRIPTIONAL COREPRESSOR

Like Rb, CtBP was identified through its association with the E1A viral oncoprotein in the 1990s. Its name is derived from the fact that it binds to the CTD of E1A (Boyd *et al.* 1993; Schaeper *et al.* 1995). CtBP functions as a transcriptional corepressor, and does not bind DNA on its own; rather, it relies on DNA-binding proteins to recruit it across the genome. It regulates genes involved in apoptosis, the epithelial to mesenchymal transition, and cell adhesion mostly through repression, but sometimes through activation as well (Grooteclaes *et al.* 2003; Fang *et al.* 2006; Jin *et al.* 2007; Paliwal *et al.* 2012). CtBP is implicated in a variety of human cancers, as many of its gene targets, when misregulated, can lead to tumor formation. CtBP mutations have been identified in breast, ovarian, and prostate cancers, among many others (Dcona *et al.* 2017). CtBP is an essential gene; its loss is lethal in both mouse and Drosophila (Hildebrand and Soriano, 2002; Poortinga *et al.* 1998). Unique among transcriptional corepressors, CtBP resembles alpha hydroxy acid dehydrogenases and binds to the NAD(H) cofactor, which ensures proper folding of the protein, and perhaps is necessary for sensing the metabolic state of the cell (Chinnadurai 2002; Kumar *et al.* 2002; Balasubramanian *et al.* 2003). It functions as a tetramer to exert its gene regulatory effects and relies on NAD(H) binding for oligomerization (Madison *et al.* 2013; Bellesis *et al.* 2018; Jecrois *et al.* 2021).

The CtBP family in mammals is made up of the CtBP1 and CtBP2 paralogs (Katsanis and Fisher, 1998). These two paralogs are highly conserved in the central dehydrogenase core region and have a largely overlapping expression profile (Hildebrand and Soriano, 2002). In contrast, invertebrates such as Drosophila encode a single CtBP protein which is alternatively spliced to produce two major isoforms (Nibu *et al.* 1998; Poortinga *et al.* 1998). These two isoforms differ in the length of their C-terminal domain; the “long” isoform, or CtBP(L), contains a ~130 amino acid unstructured CTD, while the "short" isoform, or CtBP(S), lacks most of this extension (**Figure 1.5**; Nibu *et al*. 1998; Poortinga *et al*. 1998; Sutrias-Grau and Arnosti, 2004). A number of alternative splice events lead to mRNAs encoding slightly different long and short forms of the proteins. The biological significance of these alternative isoforms in Drosophila and duplication of CtBP in mammals is not fully understood, especially with regards to the impact on regulation of gene expression.

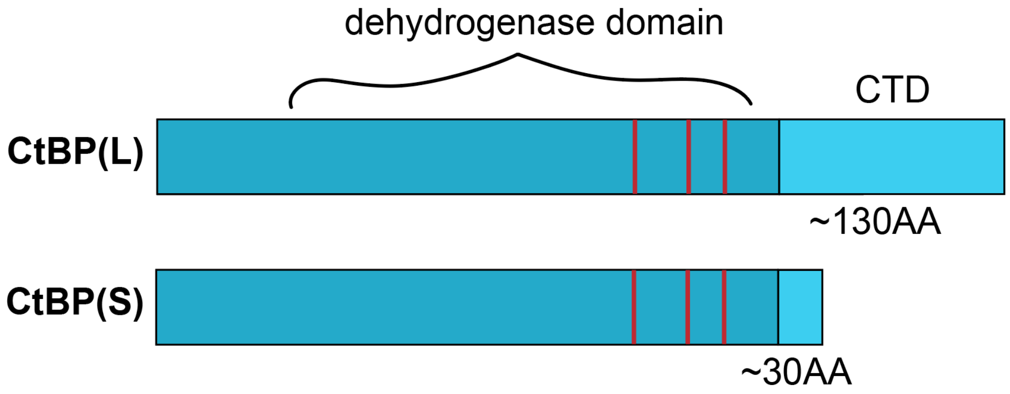


Figure 1.5. Schematic of CtBP protein isoforms encoded by the CtBP locus in *Drosophila melanogaster*. CtBP(L) is the long isoform, with a C-terminal domain (CTD) that is ~100 amino acids (AA) longer than CtBP(S), the short isoform. Orange vertical lines depict the conserved catalytic triad (R-E-H residues).

Structure of CtBP

CtBP is a protein just under 50 kDa in size. It is composed of three major domains: an N-terminal PLDLS-binding domain, a central NAD-binding domain, which is also its oligomerization domain, and a C-terminal unstructured domain (Kuppuswamy *et al*. 2008). The NTD is an important site for interactions with cofactors that harbor PLDLS motifs, such as the E1A oncoprotein (Boyd *et al*. 1993; Schaeper *et al*. 1995; Nibu *et al*. 1998; Turner and Crossley 2001). Various DNA-binding cofactors also use their PLDLS motif to interact with CtBP; for instance, Hairy and Knirps use their PLDLS motif to recruit CtBP to the Drosophila genome, and mediate long- and short-range repression, respectively (Nibu *et al.* 1998; Poortinga *et al.* 1998). Interactions with HDACs, the histone demethylase LSD1, histone methyltransferase G9a, and CoREST corepressor have also been shown to require the N-terminal substrate binding domain (Reviewed in Turner and Crossley, 2001; Shi *et al.* 2003; Kuppuswamy *et al.* 2008). Mammalian protein isoforms exist in which the NTD has a large extension, produced from an alternative, upstream transcriptional start site that lengthens the first protein-coding exon (Schmitz *et al.* 2000; Verger *et al.* 2006). This alternative NTD form is called RIBEYE, and this form of CtBP is found in synaptic ribbons, which are a specialized structure in vertebrate neurons.

The central NAD-binding domain is another critical component of this corepressor. NAD binds to CtBP and helps it to dimerize and tetramerize (Kumar *et al.* 2002; Balasubramanian *et al.* 2003; Madison *et al.* 2013; Bellesis *et al.* 2018; Jecrois *et al.* 2021). Mutating the NAD binding domain *in vitro* severely compromises dimerization ability, but not completely (Kumar *et al.* 2002; Thio *et al.* 2004; Nardini *et al.* 2009). In the fly, an NAD-binding mutant loses function, and does not rescue a CtBP null fly, indicating that NAD binding is necessary for the full function of CtBP (Sutrias-Grau and Arnosti, 2004; Zhang and Arnosti, 2011).

The NAD-binding portion of CtBP bears striking resemblance to NADH-dependent D-2-hydroxyacid dehydrogenases, and it contains a highly conserved catalytic triad, made up of the Arg-Glu-His residues (Chinnadurai *et al.* 2002; Kumar *et al.* 2002; Kuppsuwammy *et al.* 2008). Mutating a critical residues in the dehydrogenase domain does not significantly affect CtBP's repression ability in the mammalian and fly models, suggesting that its dehydrogenase activity is not required for gene regulation (Grooteclaes *et al.* 2003; Sutrias-Grau and Arnosti, 2004; Madison *et al.* 2013). Yet, some studies implicate the dehydrogenase activity in repression, as a dehydrogenase mutant protein is unable to repress a luciferase reporter to the same degree as the WT protein, and a dehydrogenase mutant in the fly does not fully rescue a CtBP null fly (Kumar *et al*. 2002; Zhang and Arnosti, 2011). Thus, the function of the dehydrogenase activity in repression is still unresolved. Directly adjacent to the NAD-binding domain is another, smaller PLDLS-binding domain, which works with the NTD to bind to cofactors.

The unstructured C-terminus

The C-terminal domain of CtBP is a highly unstructured domain, rich in glycine and proline residues, with unknown function (Nardini *et al*. 2006). This domain was originally suggested to be dispensable for CtBP's activity (Kumar *et al*. 2002). The CTD has not been structurally resolved using either X-ray crystallography or CryoEM; in fact, CtBP proteins lacking this domain are functional, stable, can still tetramerize to repress transcription, and are capable of being structurally resolved. One biochemical property that may be facilitated by the CTD is a slight enhancement of oligomerization *in vitro* although it is not necessary for dimerization (Madison *et al*. 2013; Bellesis *et al*. 2018; Jecrois *et al*. 2021).

The CtBP CTD is present in mammalian, worm, and fly homologs, but varies significantly from one species to another (Poortinga *et al*. 1998; Nicholas *et al*. 2008). Initial peptide alignments revealed that while human and fly CTDs exhibit some similarities, significant variation in this domain exists in the C. elegans homolog, which is several hundred amino acids longer than the others (Poortinga *et al*. 1998; Nicholas *et al*. 2008). Additional variation in the CTD exists in the form of CtBP isoforms that either retain or are missing this domain. This is well-characterized in *Drosophila melanogaster*, with its CtBP(S) and CtBP(L) isoforms that differ in the loss or retention of this domain. CtBP(S) terminates soon after the conserved NAD-binding domain, and is produced through alternative splicing (Mani-Telang and Arnosti, 2007). CtBP(S) is the more abundant isoform of the two throughout fly development, suggesting the CtBP(L) may not be essential for CtBP function (Mani-Telang and Arnosti, 2007). Yet, when tethered to GAL4, both CtBP(S) and CtBP(L) are able to repress reporter genes to the same extent, indicating little difference in transcriptional regulation as a function of the presence or absence of this domain (Sutrias-Grau and Arnosti, 2004). Thus, the functional significance of the presence of developmentally-regulated, alternatively spliced “long” and “short” isoforms of CtBP is unknown.

Comparison of mammalian CtBP paralogs

As mentioned above, CtBP loss is lethal in the mouse; CtBP2 null mice die by E10, while CtBP1 null mice are viable and fertile, yet 25% die by postnatal day 20 (Hildebrand and Soriano, 2002). These results suggest that like the Rb paralogs, CtBP paralogs in mammals are not identical in function. Indeed, the expression profile of the two paralogs is somewhat different in the developing mouse embryo (Furusawa *et al.* 1999). Yet, GAL4 fusions to the human CtBP1 and CtBP2 revealed that on a luciferase reporter, both CtBP paralogs mediate the same level of repression, suggesting that in some cases, they may have similar function (Madison *et al.* 2013).

CtBP gene targets

Early studies of CtBP showed that it functions as a corepressor, relying on DNA-binding proteins to recruit it to genomic sites. CtBP proteins have been found to regulate expression of genes involved in the epithelial to mesenchymal transition, as well as in apoptosis (Grooteclaes *et al.* 2003). Loss of both CtBP1 and CtBP2 in mouse embryonic fibroblasts led to global misregulation of gene expression (Grooteclaes *et al.* 2003). In particular, epithelial genes including E-cadherin and keratin-8 were upregulated, and proapoptotic genes including Bax and Noxa were also upregulated. RNAi depletion of both CtBP1 and CtBP2 in human MCF7 breast cancer cells leads to misregulation of thousands of genes involved in diverse processes, such as RNA processing, cellular metabolic processes, DNA damage response, cell cycle, cell proliferation, cell death, cell adhesion (Di *et al*. 2013). Only a small subset of the misregulated genes are directly bound by CtBP proteins, as identified by ChIP-seq in the same cell line. Interestingly, there were almost as many upregulated genes as downregulated; thus, while CtBP often functions as a corepressor, it also has the ability to activate genes, either in a direct or indirect manner.

Genetic manipulation of the fly CtBP, followed by transcriptomic analysis has not yet been performed. Thus, we can only infer whether the transcriptional profiles of CtBP overexpression or knockdown in the mammalian system are similar to those in the fly. Overall, very little is known about how CtBP proteins regulate gene expression on endogenous loci. Genetic manipulations *in vivo* and tests on transiently transfected reporters have uncovered some of their functions, as well as domains and features necessary for repression. Yet, much is left to be learned about CtBP paralogs and isoforms, in particular with regards to the relative contribution of the conserved, unstructured C-terminal domain.

CRISPR AS A NEW METHOD TO STUDY TRANSCRIPTION FACTOR ACTIVITY

Gene manipulations

Modern molecular genetic studies seek to understand gene regulatory networks at a global level, and to identify biochemical properties of regulatory factors, in particular, DNA-binding transcription factors. Basic approaches to uncover the function of transcription factors include genetic knockout or overexpression. These complementary approaches are typically followed up by ‘omics techniques that can shed light on where the TFs are binding across the genome (ChIP-seq and CUT&RUN), what kind of genes are misregulated in response (bulk or single cell RNA-seq), and how the chromatin environment is impacted (ATAC-seq, MNase, ChIP-seq and others). Sometimes the main goal is to understand how a gene regulatory program can be perturbed to modify cell states, such as in cellular reprogramming. Other times, the goal is to uncover how a known perturbation impacts cell physiology and function, as in the case of disease models, where changes in transcription factor activity promote cancer, for instance. Alternatively, understanding the molecular function of transcription factors can be a focal activity in itself, providing insights on the regulation of the factor and its target genes. For this latter goal, a variety of mostly *in vivo* approaches have been employed.

Methods to recruit TFs to target sites

To study transcription factor activity in particular contexts, fusion of the TF of interest to a DNA-binding protein like GAL4 or LexA has proved to be useful in directing the TF to certain genomic locations (Sadowski *et al*. 1992). This is particularly useful for factors such as corepressors or coactivators that cannot bind DNA on their own. GAL4 fusions are a commonly used tool; GAL4 is a yeast transcription factor that binds DNA at sequences called Upstream Activation Sequences (UAS). The GAL4-UAS system has been adapted in model organisms such as Drosophila and in mammalian cell culture as a way to test *in vivo* functional properties of GAL4-TF chimeras, as well as providing a tool for highly precise temporal and spatial regulation of gene expression, which I discuss in Chapter 2 (Brand and Perrimon, 1993). Briefly, UAS sites are engineered into a locus of interest, such as an endogenous promoter or a transiently transfected reporter, to recruit the GAL4-TF to the UAS and study the TF's impact on the downstream target gene, such as an endogenous gene or a reporter gene (CAT, GFP, luciferase etc.).

A recent study shows the power of GAL4-TF fusion assays: the Stark laboratory performed a STARR-seq screen with chimeric fusions of GAL4 to Rb paralogs, CtBP, Sin3, and CoREST (Jacobs *et al*. 2022 *bioRxiv*). This screen of thousands of putative enhancers was able to identify those that are specifically repressed by one or more of these corepressors, including identifying specific DNA sequence motifs or other features of the enhancers governing corepressor sensitivity. Overall, in this assay, CtBP had the fewest number of sensitive enhancers, while CoRest had the most, and conferred the strongest effects in many cases. Additionally, while Rbf1- and Rbf2-sensitive enhancers overlapped, there was paralog-specific enhancer sensitivity noted. This type of high-throughput analysis uncovers global rules of repression, and is useful in helping us understand how TFs function. Questions about promoter or sequence specificity, and the range from which TFs can function on a region of interest can be answered using this method. Clearly, a limitation of this approach is that genetic elements are assayed in an artificial context, on transiently transfected targets. A further disadvantage to this general method is the reliance on a UAS motif, which must be engineered into an endogenous locus, one motif at a time, or inserted into a plasmid for cell culture expression. Engineering such sites into the native chromosome of a complex, living organism such as Drosophila is challenging and time-consuming.

The basics of CRISPR/Cas9

Another more recent and highly versatile method for recruiting TFs to DNA sites of interest is the modified CRISPR/Cas system. CRISPR, or Clustered Regularly Interspersed Short Palindromic Repeats, was originally discovered as an adaptive immune response in bacteria to fight invading bacteriophages and mobile genetic elements (Barrangou *et al.* 2007). During a pathogenic invasion, the bacterial cell fragments the pathogenic DNA and inserts it into an array of repeats, interspersed with previously introduced pathogenic DNA, called spacers. This array of repeat and spacer sequences makes up the CRISPR locus; an adjacent region comprises several genes called CRISPR-associated, or Cas. During a subsequent infection, Cas endonuclease proteins bind to the two RNAs produced from the CRISPR locus, the tracrRNA and the crRNA, which together incorporate one repeat and one spacer sequence. The Cas protein/RNA complex is recruited to the invading pathogen's genome, and cleaves the target DNA, cutting the pathogen's life short (Lander, 2016).

Several types of CRISPR systems exist, with slight variations among them; the Type II CRISPR/Cas system makes use of three components described above. It can be simplified with a chimeric RNA that combines the tracrRNA and crRNA into a single guide RNA (gRNA), which guides the Cas9 with high specificity, searching for complementarity to a DNA sequence as well as an adjacent PAM sequence (Protospacer Adjacent Motif). For the Streptococcus pyogenes Cas9 nuclease, the PAM is an NGG sequence (Jinek *et al.* 2012). Once the gRNA/Cas9 heteroduplex has identified the correct target site, Cas9 uses its two endonuclease domains (HNH and RuvC) to cleave both strands of DNA (Nishimasu *et al.* 2014). This double stranded break is repaired by the cell through two main pathways: Homology Directed Repair (HDR), which can occur during S phase and uses a sister chromatid as a template for accurate repair of the break, or Non-Homologous End Joining (NHEJ), which is error-prone and can introduce indels.

The CRISPR/Cas system has recently been adapted as a versatile eukaryotic gene editing tool. By exploiting the error-prone NHEJ pathway, this system can knock out genes of interest. In separate, groundbreaking studies, the Doudna, Church, and Zhang labs independently demonstrated the ability of CRISPR/Cas9 to edit genes in human cells, using Streptococcus pyogenes Cas9 and a gene-specific gRNA (Jinek *et al.* 2013; Mali *et al.* 2013; Cong *et al.* 2013). Since this initial discovery, CRISPR/Cas has become a powerful tool for genome engineering, used widely for both basic research and gene therapy. It has largely replaced previous genome editing approaches and has been modified in a variety of ways for diverse applications as described below (reviewed in Wang and Doudna, 2023).

CRISPR for studying TFs

A major modification to the CRISPR/Cas technology is the catalytic inactivation of Cas9, effectively rendering it dead. This dead Cas9, or dCas9 can still bind DNA via gRNA recruitment, but does not cleave the target sequence (Qi *et al.* 2013). Two inactivating mutations, one in the HNH domain, and one in the RuvC domain, are sufficient for loss of endonuclease activity. When dCas9 is recruited to coding sequences, it inhibits transcription by functioning as a roadblock to the RNA polymerase and transcriptional machinery (Qi *et al.* 2013). Chimeric fusions of TFs to the dCas9 protein have also been generated to recruit repressors and activators to gene promoters and enhancers. Fusion of a repressor to dCas9 is called CRISPR interference, or CRISPRi, and fusion of an activator is called CRISPR activation, or CRISPRa. Since the first descriptions of CRISPRi and CRISPRa, hundreds of cofactors have been fused to dCas9 for various purposes.

For gene repression in mammalian cells, the Kruppel-associated box (KRAB) domain derived from zinc finger transcriptional repressors is commonly used, and was the initial repressor domain fused to dCas9 (Gilbert *et al.* 2013; Konerman *et al.* 2015). Since then, the goal in the CRISPRi field has been to identify the most potent and broadly-acting repressor domain for CRISPRi studies. KRAB fusion to MeCP2 and the ZIM3 KRAB domain have been identified in separate screens as an improvement of the original KRAB domain (Yeo *et al.* 2018; Alerasool *et al.* 2020; Replogle *et al.* 2022). Likewise, dCas9 fusions with activators started with VP64 (Gilbert *et al.* 2013; Perez-Pinera *et al.* 2013; Maeder *et al.* 2013), improved to dCas9-VPR (Chavez *et al.* 2015), and different variations also emerged, including dCas9-SAM (Konermann *et al.* 2015). In addition to identifying the optimal repressor or activator domain, researchers have improved individual gRNAs and gRNA libraries that lead to the greatest repression or activation of the target gene, with the lowest chance of off-target effects (Kampmann *et al.* 2018; Replogle *et al.* 2022).

The consensus is that a gRNA designed to bind within a few hundred base pairs of a gene's TSS is ideal, and depending on the effector domain used, one to a few gRNAs per promoter are sufficient to induce strong modulation of the target gene.

Additional applications of the dCas9 technology has been for recruitment of histone modifiers and chromatin remodelers, again with the goal of altering the chromatin environment to alter gene expression. dCas9-LSD1 has been used to recruit a histone demethylase, dCas9-p300 to recruit a histone acetyltransferase, dCas9-HDAC for histone deacetylase recruitment, dCas9-PRDM1 for histone methyltransferase recruitment, and a plethora of other cofactors to modify histones and alter the chromatin environment (Kearns *et al.* 2015; Hilton *et al.* 2015; Cano-Rodriguez *et al.* 2015; Kwon *et al.* 2017; O'Geen *et al.* 2017). Each effector is highly specific, with some cell-line specific activity, or greater potency on some genes than others. dCas9 fusions to DNA methyltransferases have also been reported to alter gene expression through CpG methylation (Stepper *et al.* 2016; Vojta *et al.* 2016).

Aside from human cell culture, dCas9-effector chimeras have also been tested in model systems such as *Drosophila melanogaster*, and CRISPR/Cas technologies have been adapted to the fly system for genome engineering, genetic screens, and many other applications (Reviewed in Zirin *et al.* 2022). One Drosophila-specific advantage is that the CRISPR components, Cas9 and the gRNA, can be spatially and temporally regulated using the GAL4-UAS system. This allows for the use of tissue-specific, developmental, or inducible enhancers to regulate expression of Cas9 or the gRNA, and provides a powerful resource for very regulated and specific perturbations and assays in a living organism (Lin *et al.* 2015; Ewen-Campen *et al.* 2017; Reviewed in Zirin *et al.* 2022).

Few studies have been aimed at using dCas9 for understanding mechanisms of long-range versus short-range repression, or for uncovering the ways in which particular TFs have acquired specificity for certain classes of genes or promoters. As described in Chapter 2 and 5, I used this method to study how Rb paralogs and CtBP isoforms function, in what promoter contexts they work, how they differ from each other, and what kind of range of activity they possess.

DISSERTATION PREVIEW

A critical gap in our knowledge of how transcriptional corepressors such as Rb and CtBP function is understanding how evolutionary variation in corepressors leads to differential gene regulatory functions. In addition, whether gene regulatory differences between Rb paralogs and CtBP isoforms is due to recruitment patterns or intrinsic repression activity is still unknown. Here, I have used the *Drosophila melanogaster* system to study the highly conserved Rb and CtBP transcriptional corepressors. In Chapter 2, I describe the creation of dCas9-Rb effectors and their implementation in both cell culture and in whole flies. I show that depending on the targeted promoter, Rbf1 and Rbf2 have similar or unique effects on gene expression, possibly consistent with "soft repression". The results presented in Chapter 2 also provide a context in which Rbf2 is a more potent repressor than Rbf1, and not just a weaker repressor as previously suggested. In Chapter 3 I discuss the hypothesis that corepressors such as Rb and SIN3 function to modulate gene expression in a "soft" manner, only modestly repressing certain gene targets, such as those involved in metabolism. Chapter 4 focuses on the use of a comparative phylogenetic approach to uncover interesting facets of CtBP biology. I propose that CtBP is a bilaterian innovation, and find that the long C-terminal domain is conserved across millions of years of evolution, suggesting a significant role in CtBP biology. In Chapter 5, I test some of the hypotheses generated in Chapter 4, by creating dCas9-CtBP fusion proteins and testing them in whole flies and in cell culture. I show that CtBP(S) is a more potent repressor than CtBP(L) on several promoters, but that these differences are not observed in transient transfection assays in cell culture. This disparity suggests that the long unstructured C-terminal domain may modulate CtBP activity in some contexts, particularly for a gene in a native chromatin state. In conclusion, in this dissertation, I used complementary molecular genetic and comparative phylogenetic approaches to reveal previously unknown properties of Rb paralogs and CtBP isoforms in Drosophila.

REFERENCES

Acharya, P., Negre, N., Johnston, J., Wei, Y., White, K. P., Henry, R. W., & Arnosti, D. N. (2012). **Evidence for Autoregulation and Cell Signaling Pathway Regulation From Genome-Wide Binding of the Drosophila Retinoblastoma Protein.** G3: Genes|Genomes|Genetics, 2(11), 1459–1472. https://doi.org/10.1534/g3.112.004424

Acharya, P., Raj, N., Buckley, M. S., Zhang, L., Duperon, S., Williams, G., Henry, R. W., & Arnosti, D. N. (2010). **Paradoxical Instability–Activity Relationship Defines a Novel Regulatory Pathway for Retinoblastoma Proteins.** Molecular Biology of the Cell, 21(22), 3890–3901. https://doi.org/10.1091/mbc.e10-06-0520

Adnane, J., Shao, Z., & Robbins, P. D. (1995). **The Retinoblastoma Susceptibility Gene Product Represses Transcription When Directly Bound to the Promoter.** Journal of Biological Chemistry, 270(15), 8837–8843. https://doi.org/10.1074/jbc.270.15.8837

Alerasool, N., Segal, D., Lee, H., & Taipale, M. (2020). **An efficient KRAB domain for CRISPRi applications in human cells.** Nature Methods, 17(11), 1093–1096. https://doi.org/10.1038/s41592-020-0966-x

Balasubramanian, P., Zhao, L.-J., & Chinnadurai, G. (2003). **Nicotinamide adenine dinucleotide stimulates oligomerization, interaction with adenovirus E1A and an intrinsic dehydrogenase activity of CtBP.** FEBS Letters, 537(1–3), 157–160. https://doi.org/10.1016/S0014-5793(03)00119-4

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., & Horvath, P. (2007). **CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes**. Science, 315(5819), 1709–1712. https://doi.org/10.1126/science.1138140

Bellesis, A. G., Jecrois, A. M., Hayes, J. A., Schiffer, C. A., & Royer, W. E. (2018). **Assembly of human C-terminal binding protein (CtBP) into tetramers**. Journal of Biological Chemistry, 293(23), 9101–9112. https://doi.org/10.1074/jbc.RA118.002514

Bianchi-Frias, D., Orian, A., Delrow, J. J., Vazquez, J., Rosales-Nieves, A. E., & Parkhurst, S. M. (2004). **Hairy Transcriptional Repression Targets and Cofactor Recruitment in Drosophila.** PLoS Biology, 2(7), e178. https://doi.org/10.1371/journal.pbio.0020178

Boyd, J. M., Subramanian, T., Schaeper, U., La Regina, M., Bayley, S., & Chinnadurai, G. (1993). **A region in the C-terminus of adenovirus 2/5 E1a protein is required for association with a cellular phosphoprotein and important for the negative modulation of T24-ras mediated transformation, tumorigenesis and metastasis.** The EMBO Journal, 12(2), 469–478. https://doi.org/10.1002/j.1460-2075.1993.tb05679.x

Brand, A. H., & Perrimon, N. (1993). **Targeted gene expression as a means of altering cell fates and generating dominant phenotypes.** Development, 118(2), 401–415. https://doi.org/10.1242/dev.118.2.401

Brehm, A., Miska, E. A., McCance, D. J., Reid, J. L., Bannister, A. J., & Kouzarides, T. (1998). **Retinoblastoma protein recruits histone deacetylase to repress transcription**. Nature, 391, 6.

Bremner, R., Cohen, B. L., Sopta, M., Hamel, P. A., Ingles, C. J., Gallie, B. L., & Phillips, R. A. (1995). **Direct Transcriptional Repression by pRB and Its Reversal by Specific Cyclins.** Molecular and Cellular Biology, 15(6), 3256–3265. https://doi.org/10.1128/MCB.15.6.3256

Broeker, C. D., & Andrechek, E. R. (2022). **E2F Transcription Factors in Cancer, More than the Cell Cycle.** In Comprehensive Pharmacology (pp. 277–311). Elsevier. https://doi.org/10.1016/B978-0-12-820472-6.00102-X

Cano-Rodriguez, D., Gjaltema, R. A. F., Jilderda, L. J., Jellema, P., Dokter-Fokkens, J., Ruiters, M. H. J., & Rots, M. G. (2016). **Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner.** Nature Communications, 7(1), 12284. https://doi.org/10.1038/ncomms12284

Cao, L., Peng, B., Yao, L., Zhang, X., Sun, K., Yang, X., & Yu, L. (2010). **The ancient function of RB-E2F Pathway: Insights from its evolutionary history.** Biology Direct, 5(1), 55. https://doi.org/10.1186/1745-6150-5-55

Chau, B. N., & Wang, J. Y. J. (2003). **Coordinated regulation of life and death by RB.** Nature Reviews Cancer, 3(2), 130–138. https://doi.org/10.1038/nrc993

Chavez, A., Scheiman, J., Vora, S., Pruitt, B. W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C. D., Wiegand, D. J., Ter-Ovanesyan, D., Braff, J. L., Davidsohn, N., Housden, B. E., Perrimon, N., Weiss, R., Aach, J., Collins, J. J., & Church, G. M. (2015). **Highly efficient Cas9-mediated transcriptional programming.** Nature Methods, 12(4), 326–328. https://doi.org/10.1038/nmeth.3312

Chicas, A., Wang, X., Zhang, C., McCurrach, M., Zhao, Z., Mert, O., Dickins, R. A., Narita, M., Zhang, M., & Lowe, S. W. (2010). **Dissecting the Unique Role of the Retinoblastoma Tumor Suppressor during Cellular Senescence.** Cancer Cell, 17(4), 376–387. https://doi.org/10.1016/j.ccr.2010.01.023

Chinnadurai, G. (2002). **CtBP, an Unconventional Transcriptional Corepressor in Development and Oncogenesis.** Molecular Cell, 9(2), 213–224. https://doi.org/10.1016/S1097-2765(02)00443-4

Chinnadurai, G. (2007). **CtBP Family Proteins. In GtBP Family Proteins** (pp. 1–17). Springer New York. https://doi.org/10.1007/978-0-387-39973-7_1

Chinnam, M., & Goodrich, D. W. (2011). **RB1, Development, and Cancer**. In Current Topics in Developmental Biology (Vol. 94, pp. 129–169). Elsevier. https://doi.org/10.1016/B978-0-12-380916-2.00005-X

Chow, K. N. B., Starostik, P., & Dean, D. C. (1996). **The Rb Family Contains a Conserved**

Cyclin-Dependent-Kinase-Regulated Transcriptional Repressor Motif. Molecular and Cellular Biology, 16(12), 7173–7181. https://doi.org/10.1128/MCB.16.12.7173

Classon, M., & Dyson, N. (2001). p107 and p130: **Versatile Proteins with Interesting Pockets.** Experimental Cell Research, 264(1), 135–147. https://doi.org/10.1006/excr.2000.5135

Classon, M., Salama, S., Gorka, C., Mulloy, R., Braun, P., & Harlow, E. (2000). **Combinatorial roles for pRB, p107, and p130 in E2F-mediated cell cycle control.** Proceedings of the National Academy of Sciences, 97(20), 10820–10825. https://doi.org/10.1073/pnas.190343497

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). **Multiplex Genome Engineering Using CRISPR/Cas Systems**. Science, 339(6121), 819–823. https://doi.org/10.1126/science.1231143

Conrad, B., & Antonarakis, S. E. (2007). **Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease.** Annual Review of Genomics and Human Genetics, 8(1), 17–35. https://doi.org/10.1146/annurev.genom.8.021307.110233

Courey, A. J., & Jia, S. (2001). **Transcriptional repression: The long and the short of it.** Genes & Development, 15(21), 2786–2796. https://doi.org/10.1101/gad.939601

Dcona, M. M., Morris, B. L., Ellis, K. C., & Grossman, S. R. (2017). **CtBP- an emerging oncogene and novel small molecule drug target: Advances in the understanding of its oncogenic action and identification of therapeutic inhibitors.** Cancer Biology & Therapy, 18(6), 379–391. https://doi.org/10.1080/15384047.2017.1323586

Di, L.-J., Byun, J. S., Wong, M. M., Wakano, C., Taylor, T., Bilke, S., Baek, S., Hunter, K., Yang, H., Lee, M., Zvosec, C., Khramtsova, G., Cheng, F., Perou, C. M., Ryan Miller, C., Raab, R., Olopade, O. I., & Gardner, K. (2013). **Genome-wide profiles of CtBP link metabolism with genome stability and epithelial reprogramming in breast cancer.** Nature Communications, 4(1), 1449. https://doi.org/10.1038/ncomms2438

Dick, F. A., & Rubin, S. M. (2013). **Molecular mechanisms underlying RB protein function.** Nature Reviews Molecular Cell Biology, 14(5), 297–306. https://doi.org/10.1038/nrm3567

Dimova, D. K. (2003). **Cell cycle-dependent and cell cycle-independent control of transcription by the Drosophila E2F/RB pathway.** Genes & Development, 17(18), 2308–2320. https://doi.org/10.1101/gad.1116703

Du, W., & Dyson, N. (1999). **The role of RBF in the introduction of G1 regulation during Drosophila embryogenesis.** The EMBO Journal, 18(4), 916–925. https://doi.org/10.1093/emboj/18.4.916

Du, W., & Pogoriler, J. (2006). **Retinoblastoma family genes.** Oncogene, 25(38), 5190–5200. https://doi.org/10.1038/sj.onc.1209651

Du, W., Vidal, M., Xie, J. E., & Dyson, N. (1996). **RBF, a novel RB-related gene that regulates E2F activity and interacts with cyclin E in Drosophila.** Genes & Development, 10(10), 1206–1218. https://doi.org/10.1101/gad.10.10.1206

Dunaief, L., Strober, E., Khavari, P. A., Begemann, M., Crabtree, R., & Medical, I. H. (1994). **The Retinoblastoma Protein and BRGI Form a Complex and Cooperate to Induce Cell Cycle Arrest.** Cell, 79, 119-130.

Elenbaas, J. S., Mouawad, R., Henry, R. W., Arnosti, D. N., & Payankaulam, S. (2015). **Role of Drosophila retinoblastoma protein instability element in cell growth and proliferation.** Cell Cycle, 14(4), 589–597. https://doi.org/10.4161/15384101.2014.991182

Ertel, A., Dean, J. L., Rui, H., Liu, C., Witkiewicz, A., Knudsen, K. E., & Knudsen, E. S. (2010). **RB-pathway disruption in breast cancer: Differential association with disease subtypes, disease-specific prognosis and therapeutic response.** Cell Cycle, 9(20), 4153–4163. https://doi.org/10.4161/cc.9.20.13454

Ewen-Campen, B., Yang-Zhou, D., Fernandes, V. R., González, D. P., Liu, L.-P., Tao, R., Ren, X., Sun, J., Hu, Y., Zirin, J., Mohr, S. E., Ni, J.-Q., & Perrimon, N. (2017). **Optimized strategy for in vivo Cas9-activation in Drosophila.** Proceedings of the National Academy of Sciences, 114(35), 9409–9414. https://doi.org/10.1073/pnas.1707635114

Fang, M., Li, J., Blauwkamp, T., Bhambhani, C., Campbell, N., & Cadigan, K. M. (2006). **C-terminal-binding protein directly activates and represses Wnt transcriptional targets in Drosophila.** The EMBO Journal, 25(12), 2735–2745. https://doi.org/10.1038/sj.emboj.7601153

Ferreira, R., Magnaghi-Jaulin, L., Robin, P., Harel-Bellan, A., & Trouche, D. (1998). **The three members of the pocket proteins family share the ability to repress E2F activity through recruitment of a histone deacetylase.** Proceedings of the National Academy of Sciences, 95(18), 10493–10498. https://doi.org/10.1073/pnas.95.18.10493

Ferrier, D. E. K. (2016). **Evolution of Homeobox Gene Clusters in Animals: The Giga-Cluster and Primary vs. Secondary Clustering.** Frontiers in Ecology and Evolution, 4. https://doi.org/10.3389/fevo.2016.00036

Fiorentino, F. P., Marchesi, I., & Giordano, A. (2013). **On the role of retinoblastoma family proteins in the establishment and maintenance of the epigenetic landscape.** Journal of Cellular Physiology, 228(2), 276–284. https://doi.org/10.1002/jcp.24141

Flores, M., & Goodrich, D. W. (2022). **Retinoblastoma Protein Paralogs and Tumor Suppression**. Frontiers in Genetics, 13, 818719. https://doi.org/10.3389/fgene.2022.818719

Forés, M., Ajuria, L., Samper, N., Astigarraga, S., Nieva, C., Grossman, R., González-Crespo, S., Paroush, Z., & Jiménez, G. (2015). **Origins of Context-Dependent Gene Repression by Capicua.** PLoS Genetics, 11(1), e1004902. https://doi.org/10.1371/journal.pgen.1004902

Furusawa, T., Moribe, H., Kondoh, H., & Higashi, Y. (1999). **Identification of CtBP1 and CtBP2 as Corepressors of Zinc Finger-Homeodomain Factor δEF1.** Molecular and Cellular Biology, 19(12), 8581–8590. https://doi.org/10.1128/MCB.19.12.8581

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weissman, J. S., & Qi, L. S. (2013). **CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes.** Cell, 154(2), 442–451. https://doi.org/10.1016/j.cell.2013.06.044

Goodrich, D. W. (2003). **How the other half lives, the amino-terminal domain of the retinoblastoma tumor suppressor protein.** Journal of Cellular Physiology, 197(2), 169–180. https://doi.org/10.1002/jcp.10358

Gordon, G. M., & Du, W. (2011). **Conserved RB functions in development and tumor suppression.** Protein & Cell, 2(11), 864–878. https://doi.org/10.1007/s13238-011-1117-z

Gray, S., & Levine, M. (1996). **Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila.** Genes & Development, 10(6), 700–710. https://doi.org/10.1101/gad.10.6.700

Grooteclaes, M., Deveraux, Q., Hildebrand, J., Zhang, Q., Goodman, R. H., & Frisch, S. M. (2003). **C-terminal-binding protein corepresses epithelial and proapoptotic gene expression programs.** Proceedings of the National Academy of Sciences, 100(8), 4568–4573. https://doi.org/10.1073/pnas.0830998100

Harbour, J. W., Lai, S.-L., Whang-Peng, J., Gazdar, A. F., Minna, J. D., & Kaye, F. J. (1988). **Abnormalities in Structure and Expression of the Human Retinoblastoma Gene in SCLC.** Science, 241(4863), 353–357. https://doi.org/10.1126/science.2838909

Hassler, M., Singh, S., Yue, W. W., Luczynski, M., Lakbir, R., Sanchez-Sanchez, F., Bader, T., Pearl, L. H., & Mittnacht, S. (2007). **Crystal Structure of the Retinoblastoma Protein N Domain Provides Insight into Tumor Suppression, Ligand Interaction, and Holoprotein Architecture.** Molecular Cell, 28(3), 371–385. https://doi.org/10.1016/j.molcel.2007.08.023

Hildebrand, J. D., & Soriano, P. (2002). **Overlapping and Unique Roles for C-Terminal Binding Protein 1 (CtBP1) and CtBP2 during Mouse Development.** Molecular and Cellular Biology, 22(15), 5296–5307. https://doi.org/10.1128/MCB.22.15.5296-5307.2002

Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2015). **Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers.** Nature Biotechnology, 33(5), 510–517. https://doi.org/10.1038/nbt.3199

Isaac, C. E., Francis, S. M., Martens, A. L., Julian, L. M., Seifried, L. A., Erdmann, N., Binné, U. K., Harrington, L., Sicinski, P., Bérubé, N. G., Dyson, N. J., & Dick, F. A. (2006). **The Retinoblastoma Protein Regulates Pericentric Heterochromatin.** Molecular and Cellular

Biology, 26(9), 3659–3671. https://doi.org/10.1128/MCB.26.9.3659-3671.2006

Ishak, C. A., Marshall, A. E., Passos, D. T., White, C. R., Kim, S. J., Cecchini, M. J., Ferwati, S., MacDonald, W. A., Howlett, C. J., Welch, I. D., Rubin, S. M., Mann, M. R. W., & Dick, F. A. (2016). **An RB-EZH2 Complex Mediates Silencing of Repetitive DNA Sequences.** Molecular Cell, 64(6), 1074–1087. https://doi.org/10.1016/j.molcel.2016.10.021

Jacobs, J., Pagani, M., Wenzl, C., & Stark, A. (2022). **Widespread regulatory specificities between transcriptional corepressors and enhancers in Drosophila** [Preprint]. https://doi.org/10.1101/2022.11.07.515017

Jecrois, A. M., Dcona, M. M., Deng, X., Bandyopadhyay, D., Grossman, S. R., Schiffer, C. A., & Royer, W. E. (2021). **Cryo-EM structure of CtBP2 confirms tetrameric architecture.** Structure, 29(4), 310-319.e5. https://doi.org/10.1016/j.str.2020.11.008

Jiang, H., Karnezis, A. N., Tao, M., Guida, P. M., & Zhu, L. (2000). **pRB and p107 have distinct effects when expressed in pRB-deficient tumor cells at physiologically relevant levels.** Oncogene, 19, 3878-3887.

Jin, W., Scotto, K. W., Hait, W. N., & Yang, J.-M. (2007). **Involvement of CtBP1 in the transcriptional activation of the MDR1 gene in human multidrug resistant cancer cells.** Biochemical Pharmacology, 74(6), 851–859. https://doi.org/10.1016/j.bcp.2007.06.017

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). **A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity.** Science, 337(6096), 816–821. https://doi.org/10.1126/science.1225829

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., & Doudna, J. (2013). **RNA-programmed genome editing in human cells.** eLife, 2, e00471. https://doi.org/10.7554/eLife.00471

Kaelin, W. G., Ewen, M. E., & Livingston, D. M. (1990). **Definition of the minimal simian virus 40 large T antigen- and adenovirus E1A-binding domain in the retinoblastoma gene product.** Molecular and Cellular Biology, 10(7), 3761–3769. https://doi.org/10.1128/MCB.10.7.3761

Kampmann, M. (2018). **CRISPRi and CRISPRa Screens in Mammalian Cells for Precision Biology and Medicine.** ACS Chemical Biology, 13(2), 406–416. https://doi.org/10.1021/acschembio.7b00657

Kareta, M. S., Gorges, L. L., Hafeez, S., Benayoun, B. A., Marro, S., Zmoos, A.-F., Cecchini, M. J., Spacek, D., Batista, L. F. Z., O'Brien, M., Ng, Y.-H., Ang, C. E., Vaka, D., Artandi, S. E., Dick, F. A., Brunet, A., Sage, J., & Wernig, M. (2015). **Inhibition of Pluripotency Networks by the Rb Tumor Suppressor Restricts Reprogramming and Tumorigenesis.** Cell Stem Cell, 16(1), 39–50. https://doi.org/10.1016/j.stem.2014.10.019

Kearns, N. A., Pham, H., Tabak, B., Genga, R. M., Silverstein, N. J., Garber, M., & Maehr, R. (2015). **Functional annotation of native enhancers with a Cas9–histone demethylase fusion.**

Nature Methods, 12(5), 401–403. https://doi.org/10.1038/nmeth.3325

Kent, L. N., & Leone, G. (2019). **The broken cycle: E2F dysfunction in cancer.** Nature Reviews Cancer, 19(6), 326–338. https://doi.org/10.1038/s41568-019-0143-7

Kim, S., & Wysocka, J. (2023). **Deciphering the multi-scale, quantitative cis-regulatory code.** Molecular Cell, 83(3), 373–392. https://doi.org/10.1016/j.molcel.2022.12.032

Kim, S.-J., Wagnert, S., Liut, F., O'Reilly, M. A., Robbins, P. D., & Greent, M. R. (1992). **Retinoblastoma gene product activates expression of the human TGF-/32 gene through transcription factor ATF-2**. Nature, 358, 331-334.

Knudson, A. G. (1971). **Mutation and Cancer: Statistical Study of Retinoblastoma.** Proceedings of the National Academy of Sciences, 68(4), 820–823. https://doi.org/10.1073/pnas.68.4.820

Kok, K., Ay, A., Li, L. M., & Arnosti, D. N. (2015). **Genome-wide errant targeting by Hairy**. eLife, 4, e06394. https://doi.org/10.7554/eLife.06394

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O., & Zhang, F. (2015). **Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex.** Nature, 517(7536), 583–588. https://doi.org/10.1038/nature14136

Kumar, V., Carlson, J. E., Ohgi, K. A., Edwards, T. A., Rose, D. W., Escalante, C. R., Rosenfeld, M. G., & Aggarwal, A. K. (2002). **Transcription Corepressor CtBP Is an NAD+-Regulated Dehydrogenase.** Molecular Cell, 10(4), 857–869. https://doi.org/10.1016/S1097-2765(02)00650-0

Kuppuswamy, M., Vijayalingam, S., Zhao, L.-J., Zhou, Y., Subramanian, T., Ryerse, J., & Chinnadurai, G. (2008). **Role of the PLDLS-Binding Cleft Region of CtBP1 in Recruitment of Core and Auxiliary Components of the Corepressor Complex.** Molecular and Cellular Biology, 28(1), 269–281. https://doi.org/10.1128/MCB.01077-07

Kwon, D. Y., Zhao, Y.-T., Lamonica, J. M., & Zhou, Z. (2017). **Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC.** Nature Communications, 8(1), 15315. https://doi.org/10.1038/ncomms15315

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). **The Human Transcription Factors.** Cell, 172(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Lander, E. S. (2016). **The Heroes of CRISPR.** Cell, 164(1–2), 18–28. https://doi.org/10.1016/j.cell.2015.12.041

Lee, C., Chang, J.H., Lee, H.S., & Cho, Y. (2002). **Structural basis for the recognition of the**

E2F transactivation domain by the retinoblastoma tumor suppressor. Genes & Development, 16(24), 3199–3212. https://doi.org/10.1101/gad.1046102

Lee, W.-H., Shew, J.-Y., Hong, F. D., Sery, T. W., Donoso, L. A., Young, L.-J., Bookstein, R., & Lee, E. Y.-H. P. (1987). **The retinoblastoma susceptibility gene encodes a nuclear phosphoprotein associated with DNA binding activity.** Nature, 329(6140), 642–645. https://doi.org/10.1038/329642a0

Lelli, K. M., Slattery, M., & Mann, R. S. (2012). **Disentangling the Many Layers of Eukaryotic Transcriptional Regulation.** Annual Review of Genetics, 46(1), 43–68. https://doi.org/10.1146/annurev-genet-110711-155437

Li, Y., Graham, C., Lacy, S., Duncan, A. M., & Whyte, P. (1993). **The adenovirus E1A-associated 130-kD protein is encoded by a member of the retinoblastoma gene family and physically interacts with cyclins A and E.** Genes & Development, 7(12a), 2366–2377. https://doi.org/10.1101/gad.7.12a.2366

Liban, T. J., Medina, E. M., Tripathi, S., Sengupta, S., Henry, R. W., Buchler, N. E., & Rubin, S. M. (2017). **Conservation and divergence of C-terminal domain structure in the retinoblastoma protein family.** Proceedings of the National Academy of Sciences, 114(19), 4942–4947. https://doi.org/10.1073/pnas.1619170114

Lin, S., Ewen-Campen, B., Ni, X., Housden, B. E., & Perrimon, N. (2015). **In Vivo Transcriptional Activation Using CRISPR/Cas9 in Drosophila.** Genetics, 201(2), 433–442. https://doi.org/10.1534/genetics.115.181065

Lipinski, M. M., & Jacks, T. (1999). **The retinoblastoma gene family in differentiation and development**. Oncogene, 18(55), 7873–7882. https://doi.org/10.1038/sj.onc.1203244

Luger, K., Dechassa, M. L., & Tremethick, D. J. (2012). **New insights into nucleosome and chromatin structure: An ordered state or a disordered affair?** Nature Reviews Molecular Cell Biology, 13(7), 436–447. https://doi.org/10.1038/nrm3382

Luo, R. X., Postigo, A. A., & Dean, D. C. (1998). **Rb Interacts with Histone Deacetylase to Repress Transcription.** Cell, 92(4), 463–473. https://doi.org/10.1016/S0092-8674(00)80940-X

Madison, D. L., Wirz, J. A., Siess, D., & Lundblad, J. R. (2013). **Nicotinamide Adenine Dinucleotide-induced Multimerization of the Co-repressor CtBP1 Relies on a Switching Tryptophan.** Journal of Biological Chemistry, 288(39), 27836–27848. https://doi.org/10.1074/jbc.M113.493569

Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H., & Joung, J. K. (2013). **CRISPR RNA–guided activation of endogenous human genes.** Nature Methods, 10(10), 977–979. https://doi.org/10.1038/nmeth.2598

Magnaghi-Jaulin, L., Groisman, R., Naguibneva, I., Robin, P., Lorain, S., Le Villain, J. P., Troalen, F., Trouche, D., & Harel-Bellan, A. (1998). **Retinoblastoma protein represses transcription by recruiting a histone deacetylase.** Nature, 391(6667), 601–605. https://doi.org/10.1038/35410

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., & Church, G. M. (2013). **RNA-Guided Human Genome Engineering via Cas9.** Science, 339(6121), 823–826. https://doi.org/10.1126/science.1232033

Mani-Telang, P., & Arnosti, D. N. (2007). **Developmental expression and phylogenetic conservation of alternatively spliced forms of the C-terminal binding protein corepressor.** Development Genes and Evolution, 217(2), 127–135. https://doi.org/10.1007/s00427-006-0121-4

Meloni, A. R., Smith, E. J., & Nevins, J. R. (1999). **A mechanism for Rb/p130-mediated transcription repression involving recruitment of the CtBP corepressor.** Proceedings of the National Academy of Sciences, 96(17), 9574–9579. https://doi.org/10.1073/pnas.96.17.9574

Montoya-Durango, D. E., Ramos, K. A., Bojang, P., Ruiz, L., Ramos, I. N., & Ramos, K. S. (2016). **LINE-1 silencing by retinoblastoma proteins is effected through the nucleosomal and remodeling deacetylase multiprotein complex.** BMC Cancer, 16(1), 38. https://doi.org/10.1186/s12885-016-2068-9

Mouawad, R., Prasad, J., Thorley, D., Himadewi, P., Kadiyala, D., Wilson, N., Kapranov, P., & Arnosti, D. N. (2019). **Diversification of Retinoblastoma Protein Function Associated with Cis and Trans Adaptations.** Molecular Biology and Evolution, 36(12), 2790–2804. https://doi.org/10.1093/molbev/msz187

Mouawad, R., Himadewi, P., Kadiyala, D., & Arnosti, D. N. (2020). **Selective repression of the Drosophila cyclin B promoter by retinoblastoma and E2F proteins.** Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1863(7), 194549. https://doi.org/10.1016/j.bbagrm.2020.194549

Mulligan, G., & Jacks, T. (1998). **The retinoblastoma gene family: Cousins with overlapping interests.** Trends in Genetics, 14(6), 223–229. https://doi.org/10.1016/S0168-9525(98)01470-X

Narasimha, A. M., Kaulich, M., Shapiro, G. S., Choi, Y. J., Sicinski, P., & Dowdy, S. F. (2014). **Cyclin D activates the Rb tumor suppressor by mono-phosphorylation.** eLife, 3, e02872. https://doi.org/10.7554/eLife.02872

Nardini, M., Svergun, D., Konarev, P. V., Spanò, S., Fasano, M., Bracco, C., Pesce, A., Donadini, A., Cericola, C., Secundo, F., Luini, A., Corda, D., & Bolognesi, M. (2006). **The C-terminal domain of the transcriptional corepressor CtBP is intrinsically unstructured.** Protein Science, 15(5), 1042–1050. https://doi.org/10.1110/ps.062115406

Nardini, M., Valente, C., Ricagno, S., Luini, A., Corda, D., & Bolognesi, M. (2009). **CtBP1/BARS Gly172 → Glu mutant structure: Impairing NAD(H)-binding and dimerization.** Biochemical and Biophysical Research Communications, 381(1), 70–74. https://doi.org/10.1016/j.bbrc.2009.02.010

Nibu, Y., Zhang, H., & Levine, M. (1998). **Interaction of Short-Range Repressors with Drosophila CtBP in the Embryo.** Science, 280(5360), 101–104. https://doi.org/10.1126/science.280.5360.101

Nicholas, H. R., Lowry, J. A., Wu, T., & Crossley, M. (2008). **The Caenorhabditis elegans Protein CTBP-1 Defines a New Group of THAP Domain-Containing CtBP Corepressors.** Journal of Molecular Biology, 375(1), 1–11. https://doi.org/10.1016/j.jmb.2007.10.041

Nicolay, B. N., Danielian, P. S., Kottakis, F., Lapek, J. D., Sanidas, I., Miles, W. O., Dehnad, M., Tschöp, K., Gierut, J. J., Manning, A. L., Morris, R., Haigis, K., Bardeesy, N., Lees, J. A., Haas, W., & Dyson, N. J. (2015). **Proteomic analysis of pRb loss highlights a signature of decreased mitochondrial oxidative phosphorylation.** Genes & Development, 29(17), 1875–1889. https://doi.org/10.1101/gad.264127.115

Nicolay, B. N., & Dyson, N. J. (2013). **The multiple connections between pRB and cell metabolism.** Current Opinion in Cell Biology, 25(6), 735–740. https://doi.org/10.1016/j.ceb.2013.07.012

Nielsen, S. J., Schneider, R., Bauer, U.-M., Bannister, A. J., Morrison, A., O'Carroll, D., Firestein, R., Cleary, M., Jenuwein, T., Herrera, R. E., & Kouzarides, T. (2001). **Rb targets histone H3 methylation and HP1 to promoters.** Nature, 412(6846), 561–565. https://doi.org/10.1038/35087620

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F., & Nureki, O. (2014). **Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA.** Cell, 156(5), 935–949. https://doi.org/10.1016/j.cell.2014.02.001

O'Geen, H., Ren, C., Nicolet, C. M., Perez, A. A., Halmai, J., Le, V. M., Mackay, J. P., Farnham, P. J., & Segal, D. J. (2017). **dCas9-based epigenome editing suggests acquisition of histone methylation is not sufficient for target gene repression.** Nucleic Acids Research, 45(17), 9901–9916. https://doi.org/10.1093/nar/gkx578

Paliwal, S., Ho, N., Parker, D., & Grossman, S. R. (2012). **CtBP2 Promotes Human Cancer Cell Migration by Transcriptional Activation of Tiam1.** Genes & Cancer, 1947601912463695. https://doi.org/10.1177/1947601912463695

Pan, Y., Tsai, C.-J., Ma, B., & Nussinov, R. (2010). **Mechanisms of transcription factor selectivity.** Trends in Genetics, 26(2), 75–83. https://doi.org/10.1016/j.tig.2009.12.003

Payankaulam, S., & Arnosti, D. N. (2009). **Groucho corepressor functions as a cofactor for the Knirps short-range transcriptional repressor.** Proceedings of the National Academy of Sciences, 106(41), 17314–17319. https://doi.org/10.1073/pnas.0904507106

Payankaulam, S., Li, L. M., & Arnosti, D. N. (2010). **Transcriptional Repression: Conserved and Evolved Features.** Current Biology, 20(17), R764–R771.

https://doi.org/10.1016/j.cub.2010.06.037

Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2013). **RNA-guided gene activation by CRISPR-Cas9–based transcription factors.** Nature Methods, 10(10), 973–976. https://doi.org/10.1038/nmeth.2600

Poortinga, G., Watanabe, M., & Parkhurst, S. M. (1998). **Drosophila CtBP: A Hairy-interacting protein required for embryonic segmentation and Hairy-mediated transcriptional repression.** The EMBO Journal, 17(7), 2067–2078. https://doi.org/10.1093/emboj/17.7.2067

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). **Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression.** Cell, 152(5), 1173–1183. https://doi.org/10.1016/j.cell.2013.02.022

Qin, X. Q., Chittenden, T., Livingston, D. M., & Kaelin, W. G. (1992). **Identification of a growth suppression domain within the retinoblastoma gene product.** Genes & Development, 6(6), 953–964. https://doi.org/10.1101/gad.6.6.953

Raj, N., Zhang, L., Wei, Y., Arnosti, D. N., & Henry, R. W. (2012). **Ubiquitination of Retinoblastoma Family Protein 1 Potentiates Gene-specific Repression Function.** Journal of Biological Chemistry, 287(50), 41835–41843. https://doi.org/10.1074/jbc.M112.422428

Replogle, J. M., Bonnar, J. L., Pogson, A. N., Liem, C. R., Maier, N. K., Ding, Y., Russell, B. J., Wang, X., Leng, K., Guna, A., Norman, T. M., Pak, R. A., Ramos, D. M., Ward, M. E., Gilbert, L. A., Kampmann, M., Weissman, J. S., & Jost, M. (2022). **Maximizing CRISPRi efficacy and accessibility with dual-sgRNA libraries and optimal effectors.** eLife, 11, e81856. https://doi.org/10.7554/eLife.81856

Richter, W. F., Nayak, S., Iwasa, J., & Taatjes, D. J. (2022). **The Mediator complex as a master regulator of transcription by RNA polymerase II.** Nature Reviews Molecular Cell Biology, 23(11), 732–749. https://doi.org/10.1038/s41580-022-00498-3

Robertson, K. D., Ait-Si-Ali, S., Yokochi, T., Wade, P. A., Jones, P. L., & Wolffe, A. P. (2000). **DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters.** Nature Genetics, 25(3), 338–342. https://doi.org/10.1038/77124

Roeder, R. G. (2019). **50+ years of eukaryotic transcription: An expanding universe of factors and mechanisms.** Nature Structural & Molecular Biology, 26(9), 783–791. https://doi.org/10.1038/s41594-019-0287-x

Ross, J. F., Liu, X., & Dynlacht, B. D. (1999). **Mechanism of Transcriptional Repression of E2F by the Retinoblastoma Tumor Suppressor Protein.** Molecular Cell, 3(2), 195–205. https://doi.org/10.1016/S1097-2765(00)80310-X

Rubin, S. M. (2013). **Deciphering the retinoblastoma protein phosphorylation code.** Trends in

Biochemical Sciences, 38(1), 12–19. https://doi.org/10.1016/j.tibs.2012.10.007

Rubin, S. M., Gall, A.-L., Zheng, N., & Pavletich, N. P. (2005). **Structure of the Rb C-Terminal Domain Bound to E2F1-DP1: A Mechanism for Phosphorylation-Induced E2F Release.** Cell, 123(6), 1093–1106. https://doi.org/10.1016/j.cell.2005.09.044

Sadowski, I., Bell, B., Broad, P., & Hollis, M. (1992). **GAL4 fusion vectors for expression in yeast or mammalian cells.** Gene, 118(1), 137–141. https://doi.org/10.1016/0378-1119(92)90261-M

Sanidas, I., Lee, H., Rumde, P. H., Boulay, G., Morris, R., Golczer, G., Stanzione, M., Hajizadeh, S., Zhong, J., Ryan, M. B., Corcoran, R. B., Drapkin, B. J., Rivera, M. N., Dyson, N. J., & Lawrence, M. S. (2022). **Chromatin-bound RB targets promoters, enhancers, and CTCF-bound loci and is redistributed by cell-cycle progression.** Molecular Cell, 82(18), 3333-3349.e9. https://doi.org/10.1016/j.molcel.2022.07.014

Sanidas, I., Morris, R., Fella, K. A., Rumde, P. H., Boukhali, M., Tai, E. C., Ting, D. T., Lawrence, M. S., Haas, W., & Dyson, N. J. (2019). **A Code of Mono-phosphorylation Modulates the Function of RB.** Molecular Cell, 73(5), 985-1000.e6. https://doi.org/10.1016/j.molcel.2019.01.004

Schaeper, U., Boyd, J. M., Verma, S., Uhlmann, E., Subramanian, T., & Chinnadurai, G. (1995). **Molecular cloning and characterization of a cellular phosphoprotein that interacts with a conserved C-terminal domain of adenovirus ElA involved in negative modulation of oncogenic transformation.** Proc. Natl. Acad. Sci. USA. Nov 7;92(23):10467-71. doi: 10.1073/pnas.92.23.10467.

Schmitz, F., Konigstorfer, A., & Sudhof, T.C. (2000). **RIBEYE, a Component of Synaptic Ribbons: A Protein's Journey through Evolution Provides Insight into Synaptic Ribbon Function.** Neuron, 28, 857-872.

Sengupta, S., Lingnurkar, R., Carey, T. S., Pomaville, M., Kar, P., Feig, M., Wilson, C. A., Knott, J. G., Arnosti, D. N., & Henry, R. W. (2015). **The Evolutionarily Conserved C-terminal Domains in the Mammalian Retinoblastoma Tumor Suppressor Family Serve as Dual Regulators of Protein Stability and Transcriptional Potency.** Journal of Biological Chemistry, 290(23), 14462–14475. https://doi.org/10.1074/jbc.M114.599993

Shi, Y., Sawada, J., Sui, G., Affar, E. B., Whetstine, J. R., Lan, F., Ogawa, H., Po-Shan Luke, M., Nakatani, Y., & Shi, Y. (2003). **Coordinated histone modifications mediated by a CtBP co-repressor complex.** Nature, 422(6933), 735–738. https://doi.org/10.1038/nature01550

Soto, L. F., Li, Z., Santoso, C. S., Berenson, A., Ho, I., Shen, V. X., Yuan, S., & Fuxman Bass, J. I. (2022). **Compendium of human transcription factor effector domains.** Molecular Cell, 82(3), 514–526. https://doi.org/10.1016/j.molcel.2021.11.007

Stepper, P., Kungulovski, G., Jurkowska, R. Z., Chandra, T., Krueger, F., Reinhardt, R., Reik, W., Jeltsch, A., & Jurkowski, T. P. (2017). **Efficient targeted DNA methylation with chimeric dCas9–Dnmt3a–Dnmt3L methyltransferase.** Nucleic Acids Research, 45(4), 1703–1713. https://doi.org/10.1093/nar/gkw1112

Stern, D. L., & Orgogozo, V. (2008). **The Loci of Evolution: How Predictable is Genetic Evolution?** Evolution, 62(9), 2155–2177. https://doi.org/10.1111/j.1558-5646.2008.00450.x

Stevaux, O. (2002). **Distinct mechanisms of E2F regulation by Drosophila RBF1 and RBF2.** The EMBO Journal, 21(18), 4927–4937. https://doi.org/10.1093/emboj/cdf501

Stevaux, O., Dimova, D. K., Ji, J.-Y., Moon, N. S., Frolov, M. V., & Dyson, N. J. (2005). **Retinoblastoma Family 2 is Required In Vivo for the Tissue-Specific Repression of dE2F2 target Genes.** Cell Cycle, 4(9), 1272–1280. https://doi.org/10.4161/cc.4.9.1982

Sutrias-Grau, M., & Arnosti, D. N. (2004). **CtBP Contributes Quantitatively to Knirps Repression Activity in an NAD Binding-Dependent Manner.** Molecular and Cellular Biology, 24(13), 5953–5966. https://doi.org/10.1128/MCB.24.13.5953-5966.2004

Takahashi, Y., Rayman, J. B., & Dynlacht, B. D. (2000). **Analysis of promoter binding by the E2F and pRB families in vivo: Distinct E2F proteins mediate activation and repression**. Genes & Development, 14(7), 804–816. https://doi.org/10.1101/gad.14.7.804

Thio, S. S. C. (2004). **The CtBP2 co-repressor is regulated by NADH-dependent dimerization and possesses a novel N-terminal repression domain.** Nucleic Acids Research, 32(5), 1836–1847. https://doi.org/10.1093/nar/gkh344

Topacio, B. R., Zatulovskiy, E., Cristea, S., Xie, S., Tambo, C. S., Rubin, S. M., Sage, J., Kõivomägi, M., & Skotheim, J. M. (2019). **Cyclin D-Cdk4,6 Drives Cell-Cycle Progression via the Retinoblastoma Protein's C-Terminal Helix.** Molecular Cell, 74(4), 758-770.e4. https://doi.org/10.1016/j.molcel.2019.03.020

Trimarchi, J. M., & Lees, J. A. (2002). **Sibling rivalry in the E2F family.** Nature Reviews Molecular Cell Biology, 3(1), 11–20. https://doi.org/10.1038/nrm714

Trouche, D., Le Chalony, C., Muchardt, C., Yaniv, M., & Kouzarides, T. (1997). **RB and hbrm cooperate to repress the activation functions of E2F1.** Proceedings of the National Academy of Sciences, 94(21), 11268–11273. https://doi.org/10.1073/pnas.94.21.11268

Turner, J., & Crossley, M. (2001). **The CtBP family: Enigmatic and enzymatic transcriptional co-repressors.** BioEssays, 23(8), 683–690. https://doi.org/10.1002/bies.1097

Ullah, Z., Buckley, M. S., Arnosti, D. N., & Henry, R. W. (2007). **Retinoblastoma Protein Regulation by the COP9 Signalosome.** Molecular Biology of the Cell, 18(4), 1179–1186. https://doi.org/10.1091/mbc.e06-09-0790

Vandel, L., Nicolas, E., Vaute, O., Ferreira, R., Ait-Si-Ali, S., & Trouche, D. (2001).

Transcriptional Repression by the Retinoblastoma Protein through the Recruitment of a Histone Methyltransferase. Molecular and Cellular Biology, 21(19), 6484–6494. https://doi.org/10.1128/MCB.21.19.6484-6494.2001

Verger, A., Quinlan, K. G. R., Crofts, L. A., Spanò, S., Corda, D., Kable, E. P. W., Braet, F., & Crossley, M. (2006). **Mechanisms Directing the Nuclear Localization of the CtBP Family Proteins.** Molecular and Cellular Biology, 26(13), 4882–4894. https://doi.org/10.1128/MCB.02402-05

Vojta, A., Dobrinić, P., Tadić, V., Bočkor, L., Korać, P., Julg, B., Klasić, M., & Zoldoš, V. (2016). **Repurposing the CRISPR-Cas9 system for targeted DNA methylation.** Nucleic Acids Research, 44(12), 5615–5628. https://doi.org/10.1093/nar/gkw159

Wang, J. Y., & Doudna, J. A. (2023). **CRISPR technology: A decade of genome editing is only the beginning.** Science, 379(6629), eadd8643. https://doi.org/10.1126/science.add8643

Weake, V. M., & Workman, J. L. (2010). **Inducible gene expression: Diverse regulatory mechanisms.** Nature Reviews Genetics, 11(6), 426–437. https://doi.org/10.1038/nrg2781

Wei, Y., Mondal, S. S., Mouawad, R., Wilczyński, B., Henry, R. W., & Arnosti, D. N. (2015). **Genome-Wide Analysis of Drosophila RBf2 Protein Highlights the Diversity of RB Family Targets and Possible Role in Regulation of Ribosome Biosynthesis.** G3 Genes|Genomes|Genetics, 5(7), 1503–1515. https://doi.org/10.1534/g3.115.019166

Weinberg, R. A. (1995). **The retinoblastoma protein and cell cycle control.** Cell, 81(3), 323–330. https://doi.org/10.1016/0092-8674(95)90385-2

Weintraub, S. J., Chow, K. N. B., Luo, R. X., Zhang, S. H., He, S., & Dean, D. C. (1995). **Mechanism of active transcriptional repression by the retinoblastoma protein.** Nature, 375(6534), 812–816. https://doi.org/10.1038/375812a0

Wirt, S. E., & Sage, J. (2010). **p107 in the public eye: An Rb understudy and more.** Cell Division, 5(1), 9. https://doi.org/10.1186/1747-1028-5-9

Xiao, B., Spencer, J., Clements, A., Ali-Khan, N., Mittnacht, S., Broceño, C., Burghammer, M., Perrakis, A., Marmorstein, R., & Gamblin, S. J. (2003). **Crystal structure of the retinoblastoma tumor suppressor protein bound to E2F and the molecular basis of its regulation.** Proceedings of the National Academy of Sciences, 100(5), 2363–2368. https://doi.org/10.1073/pnas.0436813100

Yeo, N. C., Chavez, A., Lance-Byrne, A., Chan, Y., Menn, D., Milanova, D., Kuo, C.-C., Guo, X., Sharma, S., Tung, A., Cecchi, R. J., Tuttle, M., Pradhan, S., Lim, E. T., Davidsohn, N., Ebrahimkhani, M. R., Collins, J. J., Lewis, N. E., Kiani, S., & Church, G. M. (2018). **An enhanced CRISPR repressor for targeted mammalian gene regulation.** Nature Methods, 15(8), 611–616. https://doi.org/10.1038/s41592-018-0048-5

Zhang, H. S., & Dean, D. C. (2001). **Rb-mediated chromatin structure regulation and transcriptional repression.** Oncogene, 20(24), 3134–3138. https://doi.org/10.1038/sj.onc.1204338

Zhang, H. S., Gavin, M., Dahiya, A., Postigo, A. A., Ma, D., Luo, R. X., Harbour, J. W., & Dean, D. C. (2000). **Exit from G1 and S Phase of the Cell Cycle Is Regulated by Repressor Complexes Containing HDAC-Rb-hSWI/SNF and Rb-hSWI/SNF.** Cell, 101(1), 79–89. https://doi.org/10.1016/S0092-8674(00)80625-X

Zhang, Y. W., & Arnosti, D. N. (2011). **Conserved Catalytic and C-Terminal Regulatory Domains of the C-Terminal Binding Protein Corepressor Fine-Tune the Transcriptional Response in Development.** Molecular and Cellular Biology, 31(2), 375–384. https://doi.org/10.1128/MCB.00772-10

Zirin, J., Bosch, J., Viswanatha, R., Mohr, S. E., & Perrimon, N. (2022). **State-of-the-art CRISPR for in vivo and cell-based studies in Drosophila.** Trends in Genetics, 38(5), 437–453. https://doi.org/10.1016/j.tig.2021.11.006

CHAPTER 2: RETINOBLASTOMA PROTEIN ACTIVITY REVEALED BY CRISPRI STUDY OF DIVERGENT RBF1 AND RBF2 PARALOGS

This work was published in the following preprint and adapted here:

Raicu, A.M., Castanheira, P., & Arnosti, D. N. (2023). **Retinoblastoma protein activity revealed by CRISPRi study of divergent Rbf1 and Rbf2 paralogs.** *bioRxiv*. 2023.05.19.541454; doi: https://doi.org/10.1101/2023.05.19.541454

ABSTRACT

Retinoblastoma tumor suppressor proteins are highly conserved transcriptional corepressors, regulating the key transition from G1 to S phase of the cell cycle. The mammalian Rb family is composed of Rb, p107, and p130, which possess both overlapping and unique roles in gene regulation. Likewise, the Drosophila lineage experienced a gene duplication event, leading to the expression of the Rbf1 and Rbf2 paralogs. To uncover the significance of the multiplicity of the Rb family, and how gene regulatory roles have been apportioned between the family members, we made use of the CRISPRi system. We engineered dCas9 fusions to Rbf1 and Rbf2, and deployed them to gene promoters in developing Drosophila tissue to study their relative impact on target promoters. On some genes, both Rbf1 and Rbf2 are able to mediate potent repression, and this happens in a distance-dependent manner. In general, a greater repression effect is noted for promoter-proximal sites, suggesting that Rb proteins have short-range effects, and may directly impact function of the basal transcriptional machinery. Depending on context, Rbf1 or Rbf2 was more potent, suggesting that the paralogs possess non-identical functions, possibly due to primary sequence divergence. Notably, with this dCas9 tethering system, an Rbf1 mutant lacking the entire pocket domain, that has been generally assumed to be necessary for function, retained repression activity. Significantly, direct comparison of Rb activity on endogenous genes and transiently transfected reporters showed that only qualitative, but not key quantitative aspects of repression were conserved, indicating that the native chromatin environment is essential for understanding the context-specific effects of Rb activity. The data presented here points to the complexity of Rb-mediated transcriptional regulation in a living organism, undoubtedly impacted by the different promoter landscapes and the evolution of the Rb proteins themselves.

INTRODUCTION

The Retinoblastoma (Rb) tumor suppressor protein is an ancient and highly conserved eukaryotic transcriptional corepressor. Rb is a member of a family of pocket proteins that includes the p107 and p130 paralogs. These proteins share a common conserved central pocket domain, through which they bind to E2F transcription factors found on gene promoters (Zhang and Dean, 2001). Rb binding to the E2F transactivation domain allows for transient and reversible inhibition of E2F activity, which is one method through which genes are repressed. At a secondary site in the pocket domain, Rb proteins recruit cofactors such as histone modifiers and chromatin remodelers, leading to chromatin modifications and subsequent repression of target genes (Zhang and Dean, 2001).

Canonical Rb targets are cell cycle genes—in particular, genes involved in the progression from G1 to S phase of the cell cycle. In G1, Rb represses these cell cycle genes through E2F binding; phosphorylation of Rb relieves the repression, allowing for movement into S phase. Rb proteins also regulate the expression of genes involved in DNA repair, transcription, apoptosis, polarity, diverse signaling pathways, and metabolic processes, among many others (Reviewed in Chau and Wang, 2003, Dick and Rubin, 2013, Nicolay and Dyson, 2013; Payankaulam *et al.* 2016).

Rb was the first tumor suppressor protein to be identified in humans, as loss of heterozygosity of the RB1 gene leads to pediatric retinoblastoma (Knudson, 1971). Mutations in the Rb pathway (including E2F, cyclins, and Cdk proteins) have been identified in most human cancers, implicating this pathway in both the initiation and the progression of cancer. Although all Rb proteins function in similar ways as transcriptional corepressors, Rb is more often found to be mutated in human cancer than p107 and p130, and is considered the predominant tumor suppressor

(Weinberg, 1992). It plays a leading role in gene regulatory processes, while p107 and p130 have some tissue-specific or promoter-specific roles, and do not substitute for the full spectrum of Rb activity in Rb null tissues.

Like humans, all other vertebrates encode at least three Rb proteins, with some expressing up to six (Liban *et al.* 2017). Interestingly, the Drosophila genus also experienced an independent Rb gene duplication event that dates back ~50 million years. In contrast, most eukaryotes express a single Rb protein that mediates all of the conserved functions of this family. The selective advantage for multiple RB genes in both vertebrates and in the Drosophila lineage is not well-understood. *Drosophila melanogaster* is an excellent system for studies aimed at uncovering the evolutionary significance of Rb paralogy, and determining how gene regulatory tasks are apportioned between Rb family members.

Initial studies of the Drosophila Rbf1 and Rbf2 paralogs labeled Rbf1 as the predominant corepressor, with Rbf2 being a weaker version of Rbf1. *rbf1* null flies are larval lethal, but *rbf2* null flies are viable, exhibiting some fertility and lifespan defects (Du and Dyson, 1999; Steveaux *et al.* 2002; Steveaux *et al.* 2005; Mouawad *et al.* 2019). Additionally, in the adult, Rbf2 is localized mostly to the ovary, while Rbf1 is expressed more widely (Stevaux *et al.* 2002, Keller *et al.* 2005). We and others have shown that when assayed on specific cell cycle promoters, Rbf1 is a much more potent repressor than Rbf2 (Steveaux *et al.* 2002, Mouawad *et al.* 2019). Additionally, suppression of *rbf2* by RNAi in S2 cells does not misregulate many genes, but *rbf1* depletion leads to upregulation of many genes, including cell cycle genes (Dimova *et al.* 2003). Despite its weaker activity in these contexts, we found that Rbf2 binds to twice as many genomic targets as Rbf1 in embryos, with an enrichment of mitochondrial protein and ribosomal protein genes being specifically bound and regulated by the Rbf2 paralog (Acharya *et al.* 2012, Wei *et al.*

2015, Mouawad *et al.* 2019). Additionally, we recently showed that Rbf2 is a stronger repressor than Rbf1 on certain promoters, such as *tko* and *cycB* (Wei *et al.* 2015; Mouawad *et al.* 2020).

Although Rb proteins have been found to interact with sites distal to gene promoters, binding is often close to transcriptional start sites (TSS). Genome-wide ChIP studies performed on Drosophila embryos identified preferential binding within 500 bp of gene TSSs, with a peak at -200 bp (Acharya *et al.* 2012; Wei *et al.* 2015). ChIP-seq in Drosophila third instar larvae and in human senescent cells confirmed the promoter-proximal preference of Rb proteins (Korenjak *et al.* 2012; Chicas *et al.* 2010). The significance of this binding preference remains an outstanding question. Rb proteins may directly target the basal machinery as short-range repressors, or they may be constrained by preferential binding to the E2F transcription factors, which themselves are promoter-associated. The recent mapping of Rb proteins to enhancers and insulators suggests additional complexity, at least in mammals (Sanidas *et al.* 2022).

Whether Rb functions as a short-range or long-range repressor on endogenous loci has not been definitively answered. The impact of long- versus short-range repression can have profound consequences for gene expression: long-range repressors can impact multiple distal regulatory elements for long-distance silencing whereas short-range repressors act locally to provide precise, tunable regulatory effects (Gray and Levine 1996; Barolo and Levine, 1997). Initial studies tethering Rb to the GAL4 DNA binding domain allowed for testing the ability of Rb proteins to function from proximal or distal positions. GAL4-Rb can repress an SV40 enhancer from both proximal and distal positions up to 2 kb away, and from both upstream and downstream sites relative to the TSS, suggesting a possible long-range repression mechanism (Weintraub *et al.* 1995; Bremner *et al.* 1995; Adnane *et al.* 1995; Luo *et al.* 1998). More recently, the Stark lab used GAL4-Rb fusions along with STARR-seq in Drosophila S2 cells to compare the impact of Rbf1 and Rbf2 paralogs. They revealed that thousands of enhancers can be repressed by the Drosophila paralogs from a position adjacent to the tested regulatory element, but distal to the TSS, indicating that TSS proximity is not essential for repression and that Rb also has repressive effects on enhancers (Jacobs *et al.* 2022, *bioRxiv*). Yet, such transiently transfected reporters may not fully recapitulate Rb biology on endogenous loci.

Rather than relying on global perturbations or genome-wide binding assays, a method for recruiting Rb paralogs to the same genomic locus in a living organism would be ideal for answering questions about intrinsic repression activity associated with Rb gene duplications. Here, we used CRISPRi to recruit Rb paralogs across the Drosophila genome and compare effects by each paralog. Using gene-specific guide RNAs, we targeted dCas9-Rb chimeras to diverse gene promoters in a tissue-specific manner. We find that Rbf1 and Rbf2 have diverse effects on promoters; in some cases, the corepressors act similarly, while in others Rbf1 or Rbf2 is more potent. This system also permitted a structure-function analysis of Rbf1, revealing a role for a particular motif in the CTD in promoter targeting, but not repression activity. Overall, we find that Rb proteins are not promiscuous, dominant repressors, but rather, they are highly responsive to the specific promoter landscape in which they act. Our identification of context-specific roles of Rb paralogs *in vivo* points to differences in intrinsic repressive activity, shaped by evolutionary changes in Rb paralogs over time.

RESULTS

Deploying dCas9-Rb chimeras in Drosophila

To investigate comparative Rb paralog activity on endogenous genes in Drosophila, we created FLAG-epitope-tagged dCas9-Rbf1 and dCas9-Rbf2 chimeras, and FLAG-tagged dCas9 alone as a negative control (**Figure 2.1**). We also employed the dCas9-VPR chimera, which functions as a strong activator in Drosophila (Ewen-Camben *et al.* 2017). dCas9-Rb chimeric proteins have not been tested before in any system, necessitating a proof-of-principle approach in which we could quickly screen dCas9-Rb impacts *in vivo*. To test the impact of dCas9-Rb recruitment to gene promoters, we used the GAL4/UAS system for tissue-specific expression of the chimeras in the developing wing. Specifically, we used the *nubbin* driver, which is expressed predominantly in the L3 wing disc, to screen for transcriptional impacts in larvae and phenotypic impacts in the adult wing. To this end, we created homozygous transgenic fly lines expressing two copies of *nub*-GAL4 and two copies of UAS:dCas9-Rb effectors. The *nub*-GAL4>UAS:dCas9-Rb flies were crossed to homozygous flies expressing two tandem gRNAs for a gene's promoter, to create flies expressing a single copy of each of the three transgenes (**Figure 2.1B**; see Materials & Methods). We confirmed genotypes of flies by PCR and tested mRNA expression level of the transgenes in the wing disc using RT-qPCR (**Figure 2.1C; S2.1B, C**). The dCas9-Rb effectors are expressed at similar levels of mRNA and also at similar protein levels *in vivo* (**Figure S2.1**). Crossing *nub*-GAL4>UAS:dCas9-Rb flies to a non-targeting gRNA control fly line (QUAS; Ewen-Camben *et al.* 2017), we observed phenotypically wild type (WT) adult wings, indicating that expression of a single copy of each of the transgenes does not impact normal wing development (**Figure S2.2F**).

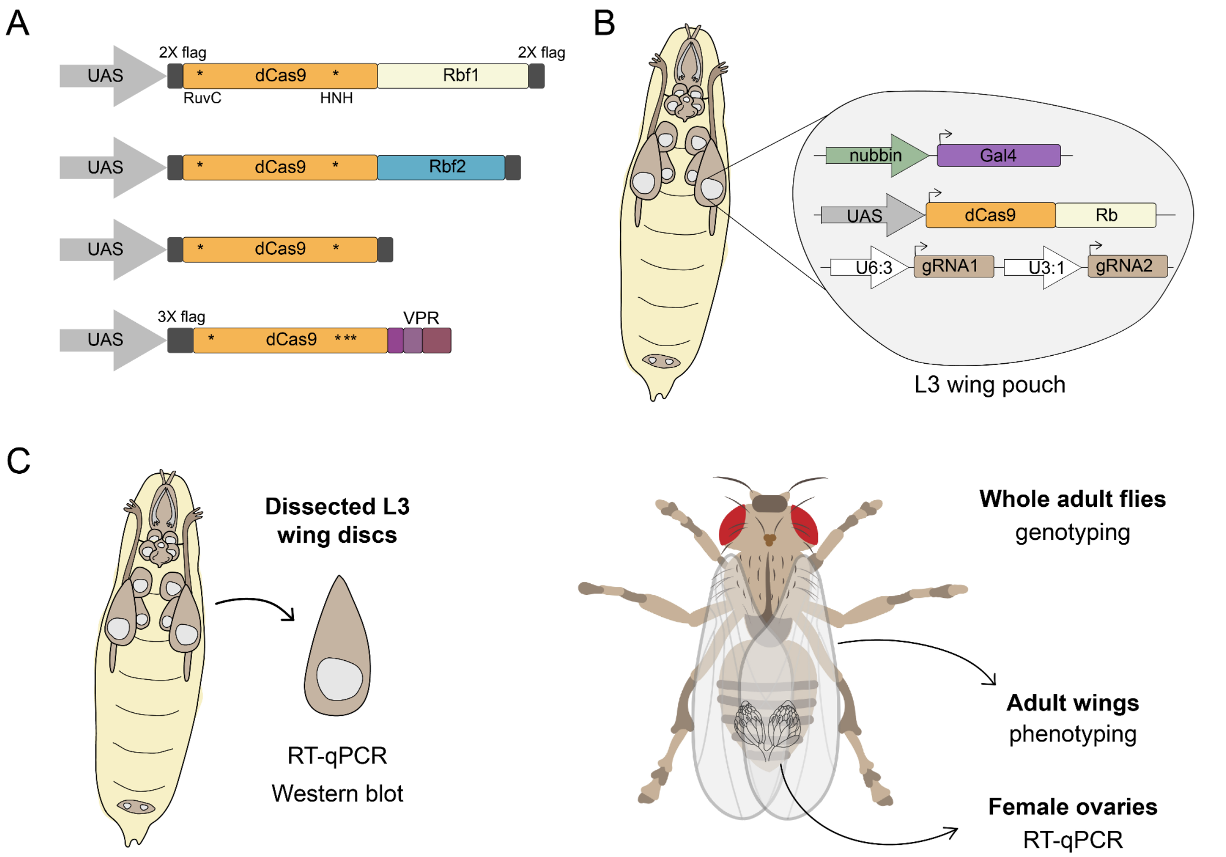


Figure 2.1. Creation of an *in vivo* system for targeting Rb paralogs to gene promoters using CRISPRi. A) The fly Rbf1 and Rbf2 FLAG-tagged coding sequences were fused to the C-terminus of the *S. pyogenes* nuclease dead Cas9 (dCas9; D10A mutation in RuvC catalytic domain and H840A mutation in HNH catalytic domain), and placed under UAS expression. FLAG-tagged dCas9 was created as a negative control. dCas9-VPR was obtained as a fly line; VPR is a tripartite activator that was previously characterized as a Drosophila transcriptional activator (Ewen-Campen *et al*. 2017). Asterisks indicate mutation sites in dCas9. **B)** *Drosophila melanogaster* expressing three transgenes were generated for tissue-specific expression of dCas9-Rb effectors using GAL4-UAS. Flies express dCas9-Rb chimeras in the *nubbin* expression pattern (begins in L2 and then localized to the wing pouch of L3 wing discs), with ubiquitous expression of two tandem gRNAs designed to target a single gene's promoter. Flies used in experiments express one copy of each of the three transgenes. **C)** Schematic of experimental procedures performed with different developmental stages. L3 wing discs were used for RT-qPCR analysis of gene expression changes and measuring mRNA and protein expression levels of the dCas9 effectors. Whole adult flies were used for genotyping, adult wings for phenotyping, and female ovaries for RT-qPCR analysis of gene expression changes.

Gene-specific effects of Rbf1 and Rbf2 recruitment by dCas9

Studies of Rb paralogs indicate that Rbf1 and Rbf2 are not fully redundant, and may have promoter-specific effects (Wei *et al.* 2015; Mouawad *et al.* 2019). For instance, Rbf1 is found on ~2000 gene promoters in the embryo, while Rbf2 is found on ~4000 promoters. Additionally, Rbf1 associates with both E2F1 and E2F2 while Rbf2 only interacts with E2F2 (Stevaux *et al.* 2002). The molecular basis for preferential recruiting of one paralog versus the other is unresolved. To test the impact of each Rb paralog on endogenous gene promoters, we recruited them to diverse loci as described above (**Figure 2.1B**). We tested 28 promoters, either known targets of one or both Rb paralogs, or genes with no known Rb regulation based on ChIP-seq and RNA-seq data (**Table S2.1**; Wei *et al.* 2015; Mouawad *et al.* 2019). The selected genes span a variety of functions, including cell cycle, DNA replication, signaling pathways, and ribosomal protein genes. Overall, phenotypes differed based on the targeted gene, and sometimes only a few of the effectors produced a phenotype on a particular gene. In some instances, the phenotype was fully penetrant, while in others, only a subset of collected wings displayed the observed phenotype (**Table S2.1**).

Here, we detail the effects on eight of these genes, where stronger morphological phenotypes were observed: *E2F2*, *Mpp6* (cell cycle), *InR*, *wg*, *dpp* (signaling pathways), *Acf* (histone remodeler), *mcm6* (DNA replication), and *Pex2* (peroxisome protein). Adult wing phenotypes ranged from mild (missing anterior or posterior crossvein, ACV and PCV) to moderate (supernumerary bristles, ectopic venation) to severe (severe morphological defect; **Figure 2.2, S2.2**). Targeting Rbf1 or Rbf2 to the *E2F2* promoter leads to a fully penetrant, severe morphological defect (**Figure 2.2B, C**). This is specific to Rb recruitment, as dCas9 alone does not produce an observable phenotype, and neither does the VPR activator. The two gRNAs designed to target *E2F2* actually bind closer to *Mpp6*, which is a divergently transcribed gene from

E2F2. Therefore, the observed phenotype of Rbf1 and Rbf2 targeting may involve changes in expression of either or both of these genes, which we explore later. In contrast to these severe Rb-specific effects, recruitment of the chimeric dCas9-Rb effectors to *wg* does not produce strong, visible phenotypes. Here, the VPR activator produces severe morphological defects, consistent with previous CRISPRa reports (**Figure 2.2B, D**; Ewen-Campen *et al.* 2017). As discussed below, we further see the contrast between the different chimeras on this locus when we measure transcriptional effects.

Targeting the *InR* promoter leads to a diversity of mild to moderate phenotypes, specifically by dCas9 and dCas9-Rbf2 (**Figure 2.2B, E**). This is an instance where the dCas9 protein by itself impacts the targeted gene, perhaps disrupting the promoter through steric effects. The lack of phenotype from dCas9-VPR may be because the non-specific inhibitory effects produced by dCas9 are ameliorated by the VPR activation domain. Interestingly, the different phenotypic impact by Rbf1 versus Rbf2 suggests that here, Rbf2 may be mediating more potent regulation of *InR* than Rbf1, which we explore later. On the *Acf*, *Pex2*, *mcm6*, and *dpp* promoters, an assortment of effects are observed (**Figure S2.2**). In some cases, Rbf1 and Rbf2 are equally penetrant, and in others, they differ in effects. Some of these promoters, such as *dpp*, are more significantly impacted by the VPR activator than by the repressors, while others, such as *mcm6* are equally affected by all the chimeras.

In conclusion, by targeting 28 gene promoters *in vivo*, we show that there are promoter-specific effects, cases where the Rb paralogs can generate the same effect, and instances where they differ in phenotypic impact. The diversity of phenotypes suggests that the Rb corepressors may work in a context-specific manner; therefore, it is important to dissect the mechanistic differences of these paralogs through assessing transcriptional impact.

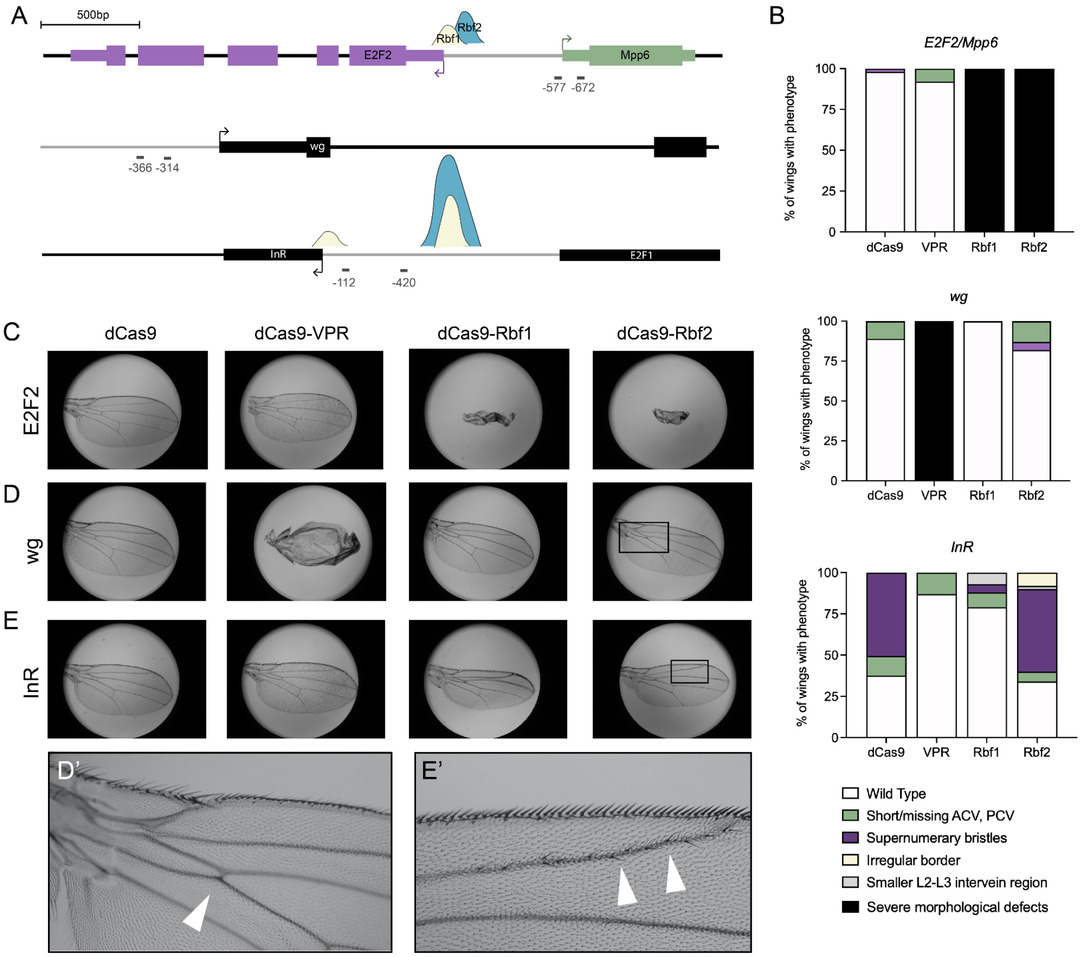


Figure 2.2. Rbf1 and Rbf2 targeting to gene promoters in the wing leads to unique phenotypic effects based on target gene. A) Diagram of gRNA binding sites on the promoters of *E2F2*, *wg*, and *InR*. Arrows indicate the TSS. gRNAs targeting *E2F2* overlap with the nearby gene, *Mpp6*. Thick black bars are exons, thinner bars are Untranslated Regions (UTR), and black horizontal lines are introns. Gray horizontal lines indicate intergenic regions, and short lines below the genes are the locations of the gRNAs, with approximate distance from the indicated TSS below. These gRNAs were obtained from Harvard TRiP (Zirin *et al*. 2020). Peaks indicate Rb binding sites from embryo exo-ChIP (Wei *et al*. 2015). **B)** Adult wings from ~50 flies (n=100) in which dCas9 effectors were targeted to the *E2F2*, *wg*, and *InR* genes were analyzed for phenotypic effect after targeting. Legend below indicates the observed phenotypes. Values are indicated as a proportion of the total number of wings. **C)** Representative images from targeting the *E2F2* promoter. Rbf1 and Rbf2 led to severe morphological defects with 100% penetrance. In contrast, dCas9 alone and dCas9-VPR targeting produced WT wings. **D)** Representative images from targeting the *wg* locus. The VPR activator led to severe morphological defects with 100% penetrance, while Rbf2 had mild effects and Rbf1 had no impact on wing development. **D'** Inset indicates the "short ACV" phenotype seen with Rbf2. **E)** Representative images from targeting the InR locus. All effectors caused mild effects, including dCas9 alone. Here, effects from dCas9 and dCas9-Rbf2 are indistinguishable, while Rbf1 targeting led to a smaller number of wings with a phenotype. **E'** Inset shows the "supernumerary bristles" phenotype seen with many of the effectors.

Transcriptional effects of chimeric dCas9-Rb corepressors

To test the transcriptional impact of Rb recruitment on the *E2F2/Mpp6* locus, we collected L3 wing discs and performed RT-qPCR. The tandem gRNAs used above are indicated as 4 and 5 (**Figure 2.3A**). Based on previously described RNA-seq experiments performed in the embryo and in wing discs, we know that when overexpressed as non-dCas9-fusion proteins, both Rb paralogs have effects on *E2F2* and *Mpp6*, depending on the context; thus, we expected to observe repression of both genes by the paralogs (**Table S2.1**). Indeed, both genes were repressed and both Rb effectors showed transcriptional repression activity. *E2F2* was more modestly repressed than *Mpp6* overall, and Rbf1 effects were greater than Rbf2 effects. Both paralogs repressed *Mpp6* >50% which was greater than the effect of dCas9 alone (**Figure 2.3B**). VPR had no transcriptional effect, possibly because it offset the dCas9-mediated repression (data not shown). We also measured the impact of targeting this promoter in a different tissue context, to compare the results to the wing disc measurements. Because we know that Rbf2 is localized to the gonads of adults, we used the *traffic jam*-GAL4 driver to express the dCas9 chimeras in the follicle cells of the adult ovary. Using the same gRNAs, we recruited Rb chimeras to the *E2F2/Mpp6* bidirectional promoter, and measured gene expression changes in whole ovaries (**Figure 2.3C**). Consistent with the wing disc results, the Rb-specific effects (compared to dCas9 alone) for *Mpp6* were greater than for *E2F2*. As in the wing disc, Rbf1 again was a better repressor than Rbf2 on this locus.

We measured the transcript levels of *wg* as well, where Rb paralogs did not show developmental phenotypes, while VPR did. As expected, *wg* levels were significantly upregulated by VPR, with 3-fold higher expression than control (**Figure S2.3A**). Unexpectedly, *wg* was significantly repressed by both Rbf1 and Rbf2, with Rbf1 effects apparently slightly greater than those of Rbf2, similar to the pattern noted on *E2F2/Mpp6*. *wg* is a gene for which Rb ChIP peaks were not identified in embryos; while Rb may not be recruited there by endogenous E2F proteins, our results show that ectopic recruitment of these corepressors via dCas9 does lead to Rb-mediated gene repression (Wei *et al*. 2015). This repression seems to not disrupt wing development, while *wg* upregulation by VPR does.

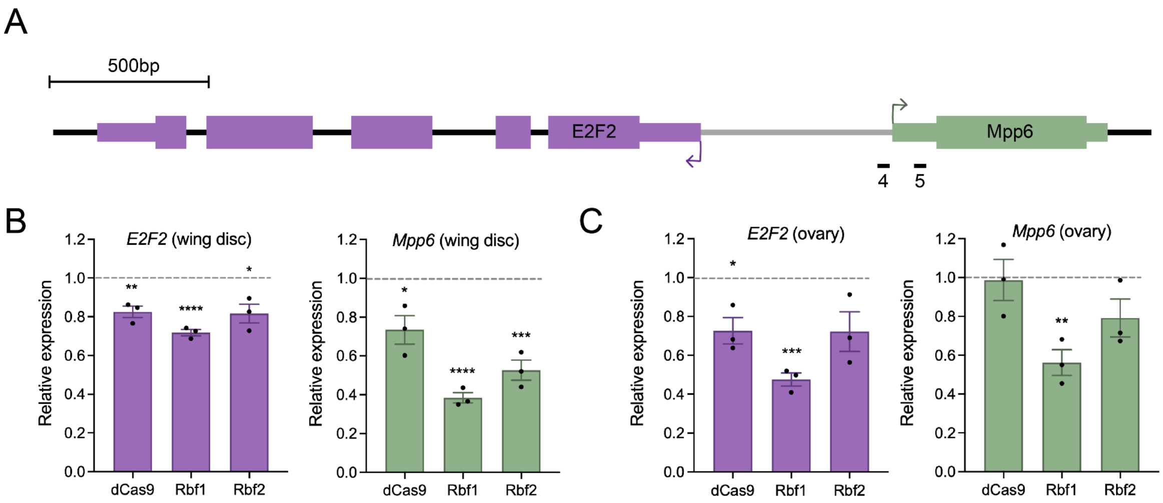


Figure 2.3. Rbf1 and Rbf2 differentially repress the *E2F2* and *Mpp6* genes in wing discs and ovaries. A) Schematic of the *E2F2/Mpp6* locus in the fly, with *E2F2* indicated in purple and *Mpp6* in green. The two tandem gRNAs from **Figure 2.2A** are indicated now as 4 and 5. **B)** L3 wing discs were dissected for RT-qPCR analysis after targeting this locus. *E2F2* and *Mpp6* expression were significantly repressed by Rb paralog recruitment. In both cases, Rbf1 represses more than Rbf2. dCas9 repressed as well, but the level of repression was not enough to create a phenotype. The control samples (set to 1) are wings in which dCas9 was targeted using the QUAS gRNA, which does not target the Drosophila genome (Ewen-Campen *et al*. 2017). **C)** Ovaries were collected from 3-4 day old females for RT-qPCR analysis after targeting using gRNAs 4 + 5 and the traffic jam-GAL4 driver. As with the wing discs, Rbf1 is a better repressor in the ovary, and both genes show significant repression. Error bars indicate SEM. * p<0.05, ** p<.01, *** p<.001, **** p<.0001.

On the *InR* locus, we did not measure a decrease in expression as we expected for dCas9 and dCas9-Rbf2, which produced adult wing phenotypes (**Figure S2.3B**). Instead, we measured slight activation by Rbf1 and Rbf2, while dCas9 and VPR had no effect on the transcript levels. It is possible that Rb protein recruitment on this locus interferes with an endogenous repressor. As was observed with *wg*, the transcriptional effects early in development do not correlate with the phenotypic impact in the adult stage. It is clear that in this context, there is no evidence of repression of *InR* by dCas9-Rb chimeras. Taken altogether, simply positioning an Rb close to a TSS does not guarantee repression, yet we observe that in many contexts, transcriptional repression is an Rb-dependent process. What is it about the promoter that allows for Rb-mediated repression activity? To answer this question, we tested whether the site of recruitment and distance from the TSS play an important role in repression potency.

Position-sensitive Rb paralog repression on the *E2F2/Mpp6* bidirectional promoter

In vivo, Rb proteins tend to bind proximal to a gene's TSS; for Rbf1 and Rbf2, enrichment is observed at -200 bp (Acharya *et al.* 2012; Korenjak *et al.* 2012; Wei *et al.* 2015). Rb may bind in a promoter-proximal position simply because E2F proteins must work from a promoter-proximal position, or because Rb activity may require promoter proximity for optimal repression. The fact that in both wing discs and in ovarian follicle cells, Rb-mediated repression of *Mpp6* is greater than *E2F2* may be a consequence of the gRNAs being positioned closer to the *Mpp6* TSS. To test whether Rb distance to the TSS affects repression ability of these two genes and whether moving the gRNAs closer to the *E2F2* TSS would have a greater impact on *E2F2* while relieving *Mpp6* repression, we designed gRNAs spanning the intergenic region and UTRs (**Figure 2.4A**). We generated new transgenic fly lines expressing one of nine single gRNAs, and crossed them to the *nub*-GAL4>UAS:dCas9-Rb homozygous flies to generate triple transgenic flies for the experiments described below.

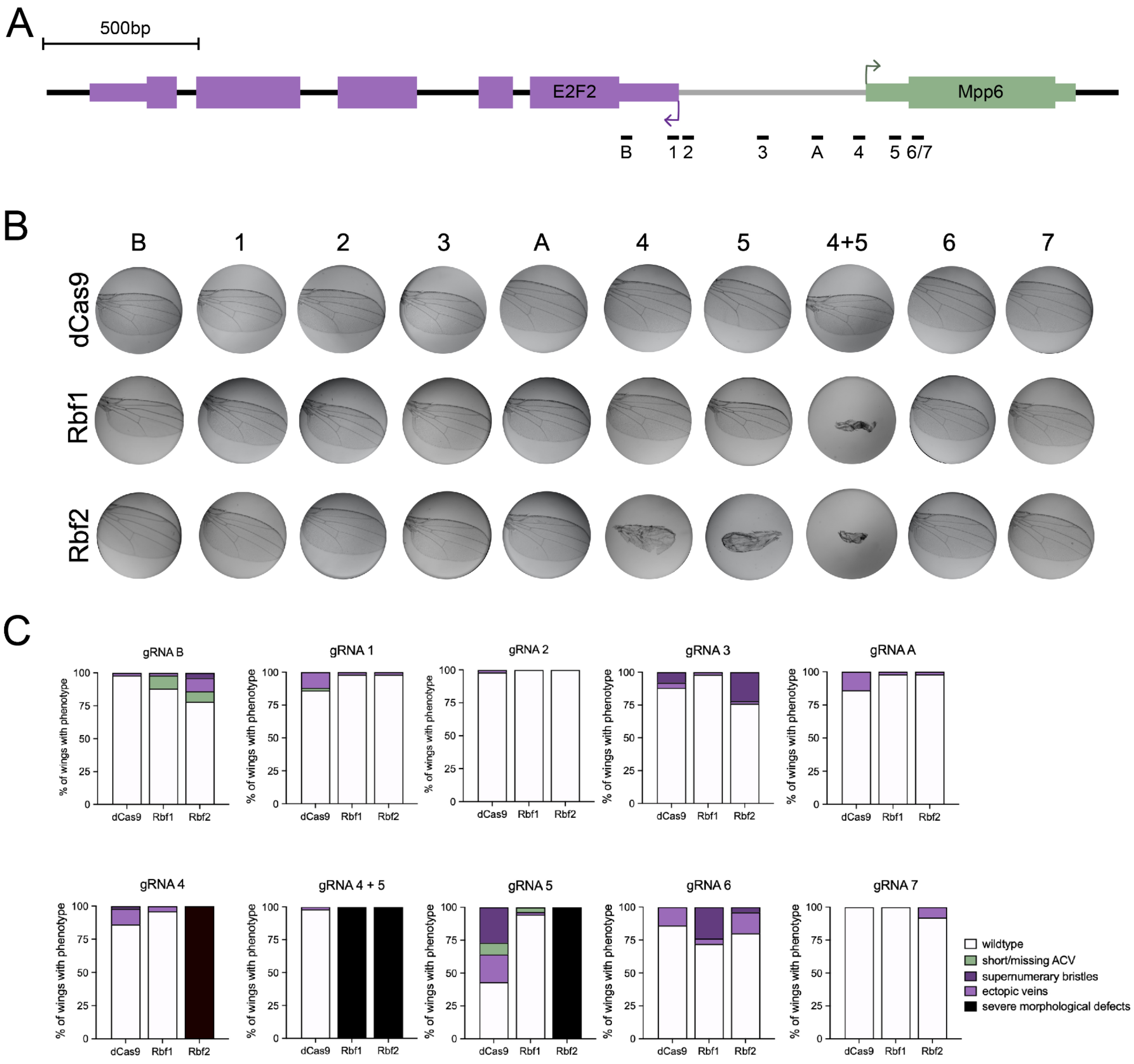


Figure 2.4. Repression of *E2F2/Mpp6* in a distance-dependent manner in the wing. Adult wings were dissected after targeting by dCas9, dCas9-Rbf1 or dCas9-Rbf2 using gRNAs designed to bind to different positions on the *E2F2/Mpp6* bidirectional promoter. **A)** Schematic of gRNA positions on the *E2F2/Mpp6* promoter used *in vivo* in the fly. dCas9 effectors were recruited to these sites using one gRNA at a time. **B)** Representative wing images from recruiting dCas9, dCas9-Rbf1 or dCas9-Rbf2 to different positions on the *E2F2/Mpp6* promoter. **C)** Graphs indicating the % of wings with particular phenotypes.

We first screened for a phenotype in the adult wings. Strong developmental effects were observed only with gRNAs in positions 4 or 5 (**Figure 2.4B, C**). Unexpectedly, the severe morphological defect seen with 4 + 5 was only observed with the single gRNA 4 or 5 when targeting with Rbf2, but not with Rbf1. In contrast, the other gRNAs generated mild or no phenotypic effects (**Figure 2.4B, C**). The lack of biological activity with the single gRNAs is not simply because they are incapable of targeting this locus, as we demonstrate below with reporter assays. Additionally, recruiting the effectors closer to the *E2F2* TSS using gRNAs 1 or 2 did not lead to a severe phenotype, suggesting that the severe phenotype observed with 4 + 5 may be caused by repression of *Mpp6*.

Next, we used RT-qPCR to measure changes in *E2F2* and *Mpp6* expression after targeting the effectors to each of these gRNA positions (**Figure 2.5**). We anticipated the greatest repression of *Mpp6* from near its TSS, with potential de-repression from other sites. Indeed, for *Mpp6*, positioning the chimeric proteins at position 4 or 5, or 4 + 5 had the greatest effect (**Figure 2.5C, E**). With single gRNA 4 or 5 alone, Rbf2 was more potent in repressing *Mpp6*; however, Rbf1 did cause modest repression as well. Moving the Rb paralogs further 5' from the *Mpp6* TSS, the gene is de-repressed in a distance-dependent manner, although a slight but significant Rbf1 effect was seen from position B, within the transcription unit of *E2F2* (**Figure 2.5C, E**). Additionally, weak, non-specific effects were observed at position 6, and no effects from position 7, just ~50 bp downstream of gRNA 5. Interestingly, from position 1, we measure statistically non-significant but apparent upregulation of *Mpp6* by both paralogs. Taken together, these results on *Mpp6* indicate that optimal repression by Rb paralogs is observed from promoter proximal sites, consistent with a short-range, and possibly promoter-specific, mechanism of repression.

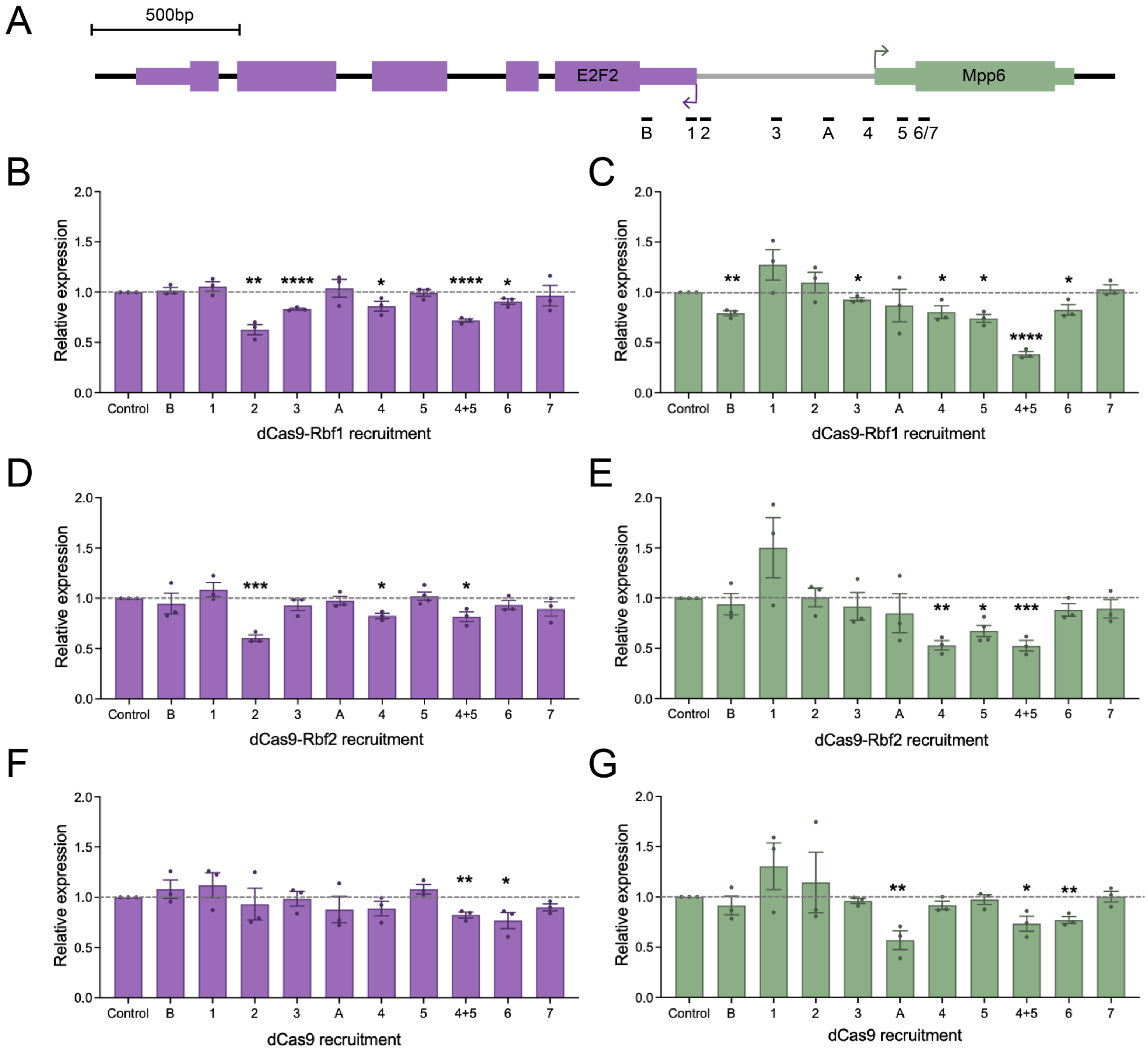


Figure 2.5. Rb regulation of *E2F2* and *Mpp6* from different positions on the promoter. **A)** Schematic of gRNA positions on the *E2F2*/*Mpp6* promoter used. dCas9 effectors were recruited to these sites using one gRNA at a time, and L3 wing discs were dissected for RT-qPCR analysis. **B)** Expression of *E2F2* after recruitment of dCas9-Rbf1 to each gRNA site. **C)** Expression of *Mpp6* after recruitment of dCas9-Rbf1 to each gRNA site. **D)** Expression of *E2F2* after recruitment of dCas9-Rbf2 to each gRNA site. **E)** Expression of *Mpp6* after recruitment of dCas9-Rbf2 to each gRNA site. **F)** Expression of *E2F2* after recruitment of dCas9 to each gRNA site. **G)** Expression of *Mpp6* after recruitment of dCas9 to each gRNA site. Error bars indicate SEM, and * indicates p<0.05, ** p<.01, *** p<.001, **** p<.0001.

The *E2F2* gene, while it was repressed modestly from 4 + 5, was most significantly repressed from position 2, which is also adjacent to its TSS (**Figure 2.5B, D**). Interestingly, the nearby position 1 did not have any effect, and neither did most other sites. Like *Mpp6*, it appears that *E2F2* is repressed by Rb paralogs through a short-range mechanism, possibly through direct contacts with transcription factors or the RNA polymerase itself, which has previously been observed *in vitro* (Ross *et al.* 1999). Few of the other sites led to *E2F2* repression, aside from the Rbf1 and dCas9-specific repression from position 6, which is puzzling. Perhaps there are elements in region 6 that regulate this gene's expression. The fact that position 2 leads to significant repression of *E2F2* but no change in *Mpp6* expression, coupled with the lack of phenotypic effects at this site, is further evidence that the severe wing phenotype is linked to *Mpp6* repression. The results on this chromosomally embedded locus must be considered in light of the chromatin environment at this site, which is rich in TF binding and histone marks (**Figure S2.4**). The chromatin environment may play an important role in how this shared regulatory region is read out by transcriptional machinery and Rb proteins.

Position-sensitive Rb paralog repression in cell culture

Initial studies of mammalian Rb relied on observations *in vitro* or in cell culture, and on measuring changes in expression of transiently transfected reporters after Rb perturbations. Even more recent high-throughput analysis of Drosophila Rb paralog activity has relied on uncovering differential impact on enhancers by using non-integrated reporters in S2 cell culture (Jacobs *et al.* 2022, *bioRxiv*). To compare our *in vivo* results, which are the first of their kind in the Drosophila system, to traditional transient transfection assays in cell culture, we turned to Drosophila S2 cells. We generated two luciferase reporters using the *E2F2* and *Mpp6* promoter sequences, and recruited dCas9-Rb effectors to the various sites on the promoters using the gRNAs (**Figure 2.6A, E**).

Given the wing disc results, we anticipated repression from 4 and 5 for the *Mpp6* reporter. Instead, we measured the strongest repression from positions 2 and 3, with Rbf1 and Rbf2 having almost an indistinguishable effect (**Figure 2.6B, C**). These gRNAs bind at around -600 bp and -400 bp, respectively, from the *Mpp6* TSS, so Rb paralogs are able to repress from sites distal to the TSS in this assay. We also measured significant repression from position 5, an effect that was observed with dCas9 alone as well, suggesting steric hindrance from dCas9 complexes located 3' of the TSS (**Figure 2.6D**). These S2 cell data suggest that the differences in the chromatin environment between the endogenous target genes in the natural chromosomal environment and reporter genes in transiently transfected plasmids have an impact on Rb protein function, although the differences in cell types must also be considered. Additionally, the quantitative differences between Rbf1 and Rbf2 observed in the wing system are not fully recapitulated in the transfection assay.

Interestingly, the effects observed from targeting the dCas9 proteins to the *Mpp6* promoter are distinct from those observed for the *E2F2* promoter, which is driven by the same cis-regulatory region, but in the opposite orientation (**Figure 2.6E-H**). The nonspecific CRISPRi effect at gRNA 5 is absent for *E2F2*, while gRNA 2, which lies close to the *Mpp6* TSS, now shows non-specific interference. Not surprisingly, gRNA 1 and B, which are immediately 3' of the *Mpp6* TSS, mediate nonspecific repression. Regulation from gRNA 3 is still Rb-specific, but weaker than for *E2F2*. The reasons are evident for differential responses to promoter-obstructing dCas9 molecules, considering the different TSS for *Mpp6* and *E2F2*; however, the different impact of targeting site 3 suggests that the joint cis-regulatory region, and its transformation after Rb recruitment, are differently read out by each gene. Taken altogether, the range of action on the *Mpp6* reporter suggests that Rb proteins can function from longer distances (~600 bp) than was optimal for the

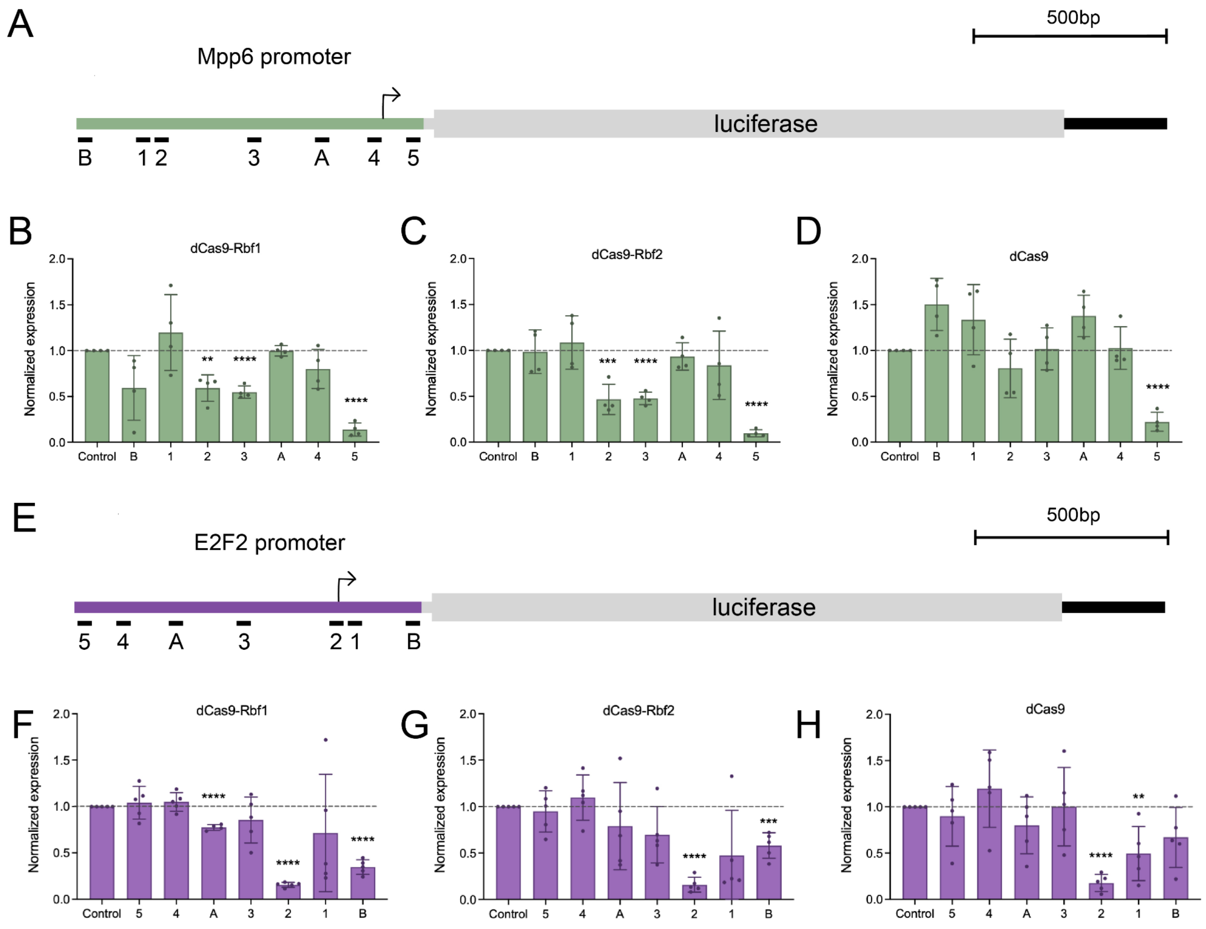


Figure 2.6. Testing positional effects of recruiting Rb paralogs to the *Mpp6* promoter in S2 cells. For all experiments described here, S2 cells were transfected with actin-GAL4, a luciferase reporter, one of the dCas9 effectors, and a single gRNA. **A)** Schematic of luciferase reporter that was designed to be regulated by the *Mpp6* promoter. **B)** dCas9-Rbf1 has position-specific effects. Position 5 caused the same level of repression as dCas9 alone (D), suggesting steric hindrance. **C)** dCas9-Rbf2 has a similar pattern of repression as dCas9-Rbf1. **D)** dCas9 has little effect on this promoter aside from position 5, which suggests steric hindrance. **E)** Schematic of luciferase reporter that was designed to be regulated by the *E2F2* promoter. Here, the gRNAs are in opposite orientation from what is shown in panel A. **F)** dCas9-Rbf1 has position-specific effects. Position 2 caused the same level of repression as dCas9 alone (H), suggesting steric hindrance. **G)** dCas9-Rbf2, has a similar pattern of repression as dCas9-Rbf1. **H)** dCas9 has little effect on this promoter, aside from position 2, 1, and possibly B, which suggest that this promoter is more sensitive to recruitment of a large protein than the promoter in the opposite orientation. Error bars indicate SEM, and * indicates p<0.05, ** is p<.01, *** p<.001, **** p<.0001.

native chromosomally integrated genes, and some of the distinctions between Rbf1 and Rbf2 in the wing disc were not evident for these reporter assays. While transient reporters may not provide a complete picture of repression potential, they remain a useful tool—these experiments provided a useful test for some gRNAs inactive in the disc, such as gRNA 1, showing that the gRNA was at least capable of mediating steric blocking in S2 cells.

Effect of non-chimeric Rbf1 and Rbf2 proteins on the *E2F2/Mpp6* promoter

Given the diverse activities of dCas9-Rb chimeras on these regulatory regions, it was important to test the effect of Rb proteins interacting with native E2F binding sites as well. Thus, we expressed Rb paralogs (not fused to dCas9) in cell culture, to measure their abilities to regulate the *Mpp6* and *E2F2*-luciferase reporters. Both Rbf1 and Rbf2 significantly repressed the *Mpp6* promoter, with stronger effects by Rbf1 (**Figure 2.7B**). On the *E2F2* promoter, Rbf1 significantly repressed but Rbf2 had only very mild effects (**Figure 2.7E**). Intriguingly, for this divergently transcribed region, the impacts of Rbf2 differ greatly, whereas Rbf1 seems to have a similar impact on both promoters.

We identified a highly conserved E2F motif in the 5' UTR of *E2F2*, adjacent to the Rbf1 and Rbf2 ChIP peaks in the embryo (**Figure S2.5B, 2.2A**). In the case of mammalian Rb, removal of E2F motifs in transfected reporters inhibits its ability to repress a reporter gene, underscoring the importance of E2F motifs for Rb activity (Weintraub *et al*. 1992; Lam and Watson, 1993; Dynlacht *et al*. 1994; Qin *et al*. 1995). To determine whether this particular E2F motif plays a role in the differential impacts observed by the Rb paralogs, we mutated the E2F motif in the two luciferase plasmids (**Figure 2.7A, D, S2.5A, E**). Overexpression of Rb paralogs on the *Mpp6*-luciferase ΔE2F reporter led to similar levels of repression as on the WT promoter (**Figure 2.7C**). Both Rb paralogs still similarly repressed, but the magnitude was somewhat dampened, indicating that for *Mpp6* expression, this E2F motif is meaningful but dispensable. Another mutant, E2F^{4X}

also had the same mild effect (**Figure S2.5A, D**). Other reports from mammals have also indicated partial redundancy of E2F motifs on promoters (Burkhart *et al*. 2010a).

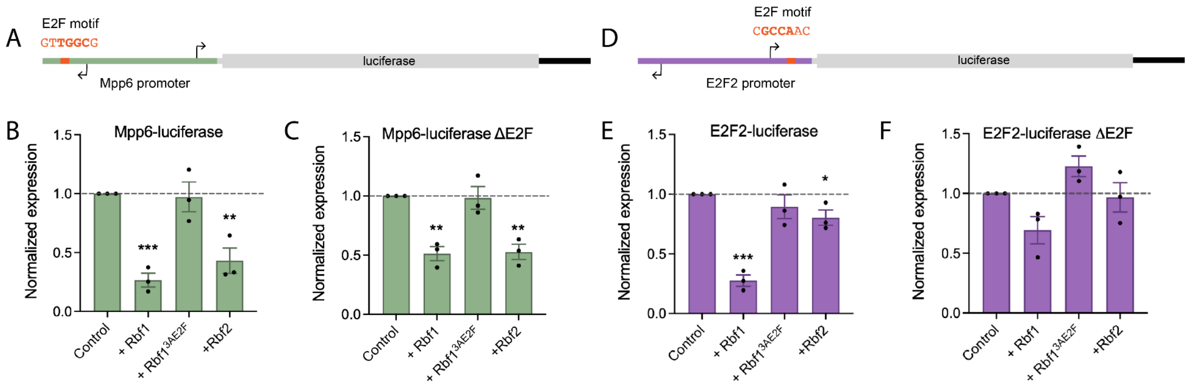


Figure 2.7. Differential impact of non-chimeric Rbf1 and Rbf2 on the *E2F2/Mpp6* promoter. A) Schematic of the *Mpp6*-luciferase reporter used in transfections with Rb plasmids. The orange box indicates a highly conserved E2F motif (**Figure S2.5**), with mutated nucleotides bolded in orange**.** Arrows indicate TSS. The arrow in the opposite direction indicates the *E2F2* TSS. **B)** Co-transfection of the *Mpp6*-luciferase reporter with Rbf1 and Rbf2 (untethered to dCas9) leads to significant repression by both Rbf1 and Rbf2, but not by the Rbf1^{3AE2F} mutant discussed in **Figure 2.8**. **C)** Removal of the E2F motif from the *Mpp6* promoter (ΔE2F) does not significantly impact repression ability by Rbf1 and Rbf2, although the magnitude of repression by Rbf1 is slightly dampened. **D)** Schematic of the *E2F2*-luciferase reporter used in transfections with Rb plasmids. The orange box indicates the same highly conserved E2F motif indicated in panel A, with mutated nucleotides bolded in orange. Arrows indicate TSS. The arrow in the opposite direction indicates the *Mpp6* TSS. **E)** Co-transfection of the *E2F2*-luciferase reporter with Rbf1 and Rbf2 (untethered to dCas9) leads to significant repression by Rbf1, but not much by Rbf2, in contrast to what was seen with the promoter in the opposite orientation in panel B. **F)** Removal of the E2F motif from the *E2F2* promoter (ΔE2F) impact repression ability by both Rbf1 and Rbf2, as they are unable to significantly repress this promoter without the E2F motif intact. Error bars indicate SEM, and * indicates $p<0.05$, ** is $p<.01$, *** $p<.001$.

On the *E2F2*-luciferase ΔE2F reporter, repression by the Rb proteins is attenuated, with no response to Rbf2 overexpression (**Figure 2.7F**). In this context, the E2F motif is critical for regulation by both Rb paralogs. Thus, the relative contribution of this E2F motif for the *Mpp6* and the *E2F2* promoters differs; for the two genes, the common cis-regulatory sequence is read out in disparate ways. Unexpectedly, the *E2F2*-luciferase E2F^{4X} mutant was not only not repressed by

Rb overexpression, but showed upregulation in some instances, possibly because the mutation may have introduced other regulatory information into the gene (**Figure S2.5G**).

Functional analysis of Rbf1 features necessary for repression

Previous data indicate that Rbf1 and Rbf2 can show similar or dissimilar activities depending on the promoter context. One idea that has been adduced is that differential regulation by Rbf2 is due to divergence between the paralogs' C-terminal domains. The Rbf1 CTD differs from that of Rbf2 in that Rbf1 contains an Instability Element (IE)—an ancestral feature of this protein family that is also found in mammalian p107 and p130 (Acharya *et al.* 2010; Sengupta *et al.* 2015). Residues within the p107 IE are responsible for E2F4-specific contacts, allowing for differential binding; therefore, this IE may be chiefly important for dictating binding specificity between Rb and E2F proteins (Liban *et al.* 2016; Liban *et al.* 2017).

In the fly, the overexpression of Rb proteins in different contexts has demonstrated that the IE can impact transcriptional activity but has not differentiated between effects on promoter binding versus inherent transcription activity. An Rbf1$^{\Delta IE}$ mutant can still interact with E2F1 and E2F2, but is unable to repress cell cycle promoters such as *PCNA* (Acharya *et al.* 2010; Raj *et al.* 2012a). Overexpression phenotypes in the fly eye and wing suggests that it may be a hypomorph (Acharya *et al.* 2010; Elenbaas *et al.* 2015). Indeed, according to RNA-seq, there is limited repression of various genes after overexpression of Rbf1$^{\Delta IE}$ in wing discs, in comparison to the WT Rbf1 (**Figure S2.6A, B**). Still, there are some contexts in which Rbf1$^{\Delta IE}$ is a good repressor: reporter assays in S2 cells showed that Rbf1$^{\Delta IE}$ is still capable of repressing *InR*, *Wts*, and *Pi3K* (Raj *et al.* 2012a). Thus, on some gene promoters, the IE is necessary for repression, but on others it is dispensable for activity.

To test the ability of Rbf1$^{\Delta IE}$ to repress when tethered to dCas9, an experiment which clearly differentiates between targeting and differential transcriptional activity, we created a fly line expressing UAS:dCas9-Rbf1$^{\Delta IE}$ and recruited it to the *E2F2*/*Mpp6* promoter using gRNAs 4 + 5 (**Figure 2.8A, C**). Intriguingly, Rbf1$^{\Delta IE}$ caused the same severe wing phenotype as dCas9-Rbf1, and produced similar levels of repression of *E2F2* and *Mpp6* in the wing disc (**Figure 2.8B, D, E**). Thus, on this promoter, the IE is not required for repression; as a dCas9 chimera, the Rbf1$^{\Delta IE}$ mutant protein is as potent as WT Rbf1. These results are corroborated by transient transfections in S2 cells, where dCas9-Rbf1$^{\Delta IE}$ repressed the *Mpp6*-luciferase reporter to the same extent as dCas9-Rbf1 (**Figure S2.7B**). Taken together, these results suggest that the IE is not required for repression, but instead is involved in recruitment to the genome via E2F. We therefore conclude that the inability of Rbf1$^{\Delta IE}$ to repress genes in our previous Rbf1 overexpression assays is due to a defect in recruiting, rather than inherent repression activity mediated by the IE.

The C-terminal domain can influence E2F-mediated interactions; however, the pocket domain is absolutely necessary for E2F binding. The cleft between the cyclin A and cyclin B domains binds to the transactivation domain of E2F, and as discussed previously, the pocket also interacts with a variety of chromatin modifying complexes. In the fly, an Rbf1$^{\Delta pocket}$ mutant only weakly represses a *PCNA*-luciferase reporter, and its interaction with E2F1 is abolished (Acharya *et al*. 2010). A plethora of studies using mammalian pocket domain mutants corroborate these findings that without the pocket, Rb cannot function as a corepressor (Bremner *et al*. 1995). Thus, it has been hypothesized that the pocket domain is key to Rb's transcriptional repressive activities. Yet, contradicting data shows that pocket mutants that have abolished LxCxE-binding or E2F binding still function *in vivo*, putting into question the requirement of a pocket and the significance of the remaining N- and C-terminal domains (Sellers *et al*. 1998; Sun *et al*. 2006; Cecchini *et al*. 2014).

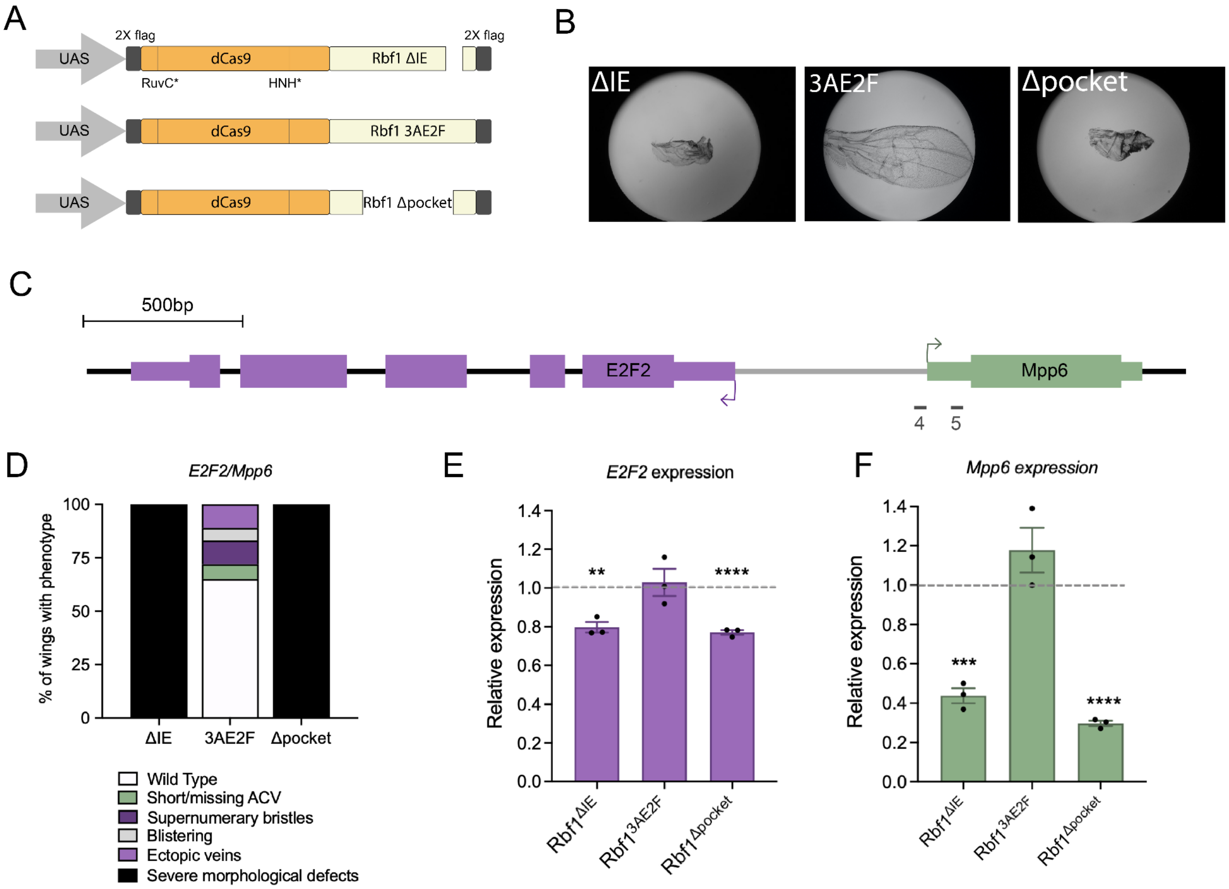


Figure 2.8. Targeting Rbf1 mutants to gene promoters indicates that the IE and pocket are not necessary for repression. A) Schematic of three Rbf1 mutants which were fused to dCas9 for targeting. ΔIE refers to removal of residues 728 to 786 from the C-terminus, 3AE2F is a novel E2F binding mutant, and Δpocket removes residues 376 to 728. **B)** Recruitment to the *E2F2/Mpp6* locus using the tandem gRNAs 4 + 5 led to severe morphological defects with ΔIE and Δpocket, but milder effects with 3AE2F. **C)** Representative images. **D)** RT-qPCR data from targeting Rbf1 mutants to gRNA positions 4 + 5. *Mpp6* expression is repressed equally well by ΔIE and Δpocket, as compared to WT Rbf1 (**Figure 2.2B**), suggesting that the Instability Element and the pocket domain are not required for repression. However, the 3AE2F mutant is unable to mediate repression, suggesting that without E2F binding, Rb cannot function as a repressor. Similar trends are seen for *E2F2*, where 3AE2F is unable to repress. Error bars indicate SEM.

To test the significance of the pocket domain for intrinsic repression activity, we generated a fly line expressing UAS:dCas9-Rbf1$^{\Delta pocket}$. Strikingly, recruitment of Rbf1$^{\Delta pocket}$ to the *E2F2/Mpp6* promoter using gRNAs 4 + 5 resulted in the same severe phenotype as dCas9-Rbf1, and repressed both *E2F2* and *Mpp6* to the same magnitude as the WT protein (**Figure 2.8**). These results were also corroborated in cell culture, where the pocket mutant functioned similar to WT Rbf1 (**Figure S2.7D**). No study has measured such levels of repression by an Rb protein lacking the pocket domain before. A key takeaway from these data is that the more variable portions of Rb proteins, the N- and C-terminal domains, appear to be intact and functional in this context, perhaps allowing for interactions with cofactors. Indeed, the NTD resembles the cyclin folds of the pocket, and the CTD plays a role in interactions with E2F proteins; thus, they may retain the ability to create a repression complex. Interactions with cofactors have been reported outside of the pocket domain, making this a real possibility (Hassler *et al.* 2007).

Another question that has not been clearly addressed yet is whether the Rb-E2F interactions are wholly separable from Rb interactions with cofactors, that is, whether possible conformational effects upon E2F binding facilitate cofactor interactions with Rb proteins. Thus, we created a mutant version of Rbf1 that lacks E2F binding but retains an intact pocket. Based on *in vitro* work testing the mammalian Rb pocket binding to E2F1, we identified three conserved residues in Rbf1 that may be essential for high affinity binding to E2F proteins (Lee *et al.* 2002). Based on this work, we made three mutations, F476A, E541A, and K588A, to create an E2F-binding mutant form of Rbf1 (Rbf1^{3AE2F}). We generated flies expressing UAS:dCas9-Rbf1^{3AE2F} and tested this mutant on the *E2F2/Mpp6* promoter using gRNAs 4 + 5 (**Figure 2.8A, C**). Targeting led to significant morphological defects in the adult wing, although not as numerous or as severe as those induced by WT Rbf1 or the other Rbf1 mutants (**Figure 2.8B, D**). Interestingly, this mutant protein did not show transcriptional repression of *E2F2* or *Mpp6* in the wing disc. It is possible that these particular mutations affect stability or activity of Rbf1. Yet, when tested on the *Mpp6*-luciferase reporter in S2 cells, it exhibited specific repression activity, similar to that of WT Rbf1. Taken together, these data suggest that Rb binding to E2F does not appear to be required for repression activity, as this mutant still retains the ability to impact wing development and gene expression in cell culture (**Figure S2.7C**).

DISCUSSION

Rb paralogy

Gene duplication is associated with functional diversification among paralogs. For transcription factors and cofactors, such diversification may reflect expression changes, whereby paralogs are differentiated by promoter evolution. At the same time, there are numerous examples of changes in activity due to protein evolution, which may impact genomic targeting, or inherent transcriptional activity. Rb paralogs in vertebrates and Drosophila share common structural and functional properties, but are clearly differentiated by unique expression patterns (Jiang *et al.* 1997; Keller *et al.* 2005). Furthermore, they have unique genomic binding patterns, which may explain divergent gene regulatory effects in knock-out experiments (Chicas *et al.* 2010; Wei *et al.* 2015). The net result of the combined changes is a distinct set of regulatory outputs: p130 has a predominant regulation during G0, while p107 and Rb regulate genes in G1 phase. Depletion of each family member in human senescent cells leads to only a 5% overlap of misregulated genes, with unique sets of genes misregulated by each paralog (Chicas *et al.* 2010). Similarly, Rbf1 and Rbf2 bind to some genes in common, but many genes are occupied uniquely by one paralog, in particular Rbf2, which has about twice the number of targeted promoters (Wei *et al.* 2015). Misexpression of the paralogs accordingly produces very different gene regulatory effects. Even for the promoters bound by both Rbf1 and Rbf2, however, misexpression or knockdown of each Rb paralog usually has divergent effects, indicating that the proteins' functions are not identical.

To what extent has protein evolution driven changes in Rb paralog activity? At the primary sequence-level, p107 and p130 share ~50% sequence identity, but only ~30% with Rb, with greatest divergence in the N- and C-terminal domains (Mulligan and Jacks, 1998). Divergence in the CTD is functionally significant, as it leads to differential interactions and affinity for E2F family proteins; the Rb CTD binds preferentially to E2F1, while p107 features CTD contacts that drive a preference for E2F4 (Liban *et al.* 2017). Rbf1 and Rbf2 also have differential interactions with E2F activators; Rbf1 interacts with both E2F1 and E2F2, but Rbf2 only interacts with the E2F2 repressor (Stevaux *et al.* 2002). Thus, it is possible that like the mammalian Rb proteins, the divergence in their C-terminal domains led to preferential interactions with E2Fs. By directly testing Rbf1 isoforms in a side-by-side manner with our CRISPRi approach, we demonstrated that the IE domain of the CTD is not essential for repression activity, but likely plays a role in differential targeting - notably, the loss of the IE is a derived feature of the Rbf2 CTD.

Until now, there has been relatively little information on whether structural divergences also impact inherent transcriptional activity. For instance, Rbf1 and Rbf2 have different abilities to repress specific promoters, but this may be strictly a function of differential promoter binding (Wei *et al.* 2015; Mouawad *et al.* 2019). Some of the earliest molecular biology studies pointed to invariant mechanisms of repression. For instance, GAL4-Rb paralog chimeras act in a similar manner on reporter genes (Bremner *et al.* 1995; Ferreira *et al.* 1998; Meloni *et al.* 1999). Chimeras interchanging the Rb and p107 pockets retain function as GAL4 fusions, indicating overlapping gene-regulatory functions (Chow *et al.* 1996). Thus, most functional studies have concentrated on the similarities between the proteins. However, structural innovations among the Rb paralogs may impact actual repression potential, once bound to a locus. The recent study by Jacobs *et al.* reported an extensive comparison of corepressors, including Rbf1 and Rbf2, in their activity to modulate Drosophila enhancers in a high-throughput assay. They found that Rbf1 and Rbf2 chimeras were much more similar to each other than to other diverse corepressors, including CtBP, SIN3 and CoREST (Jacobs *et al.* 2022, *bioRxiv*). It is notable that our assays of dCas9-Rb repressors uncovered similarities but also important functional distinctions, particularly when assayed on endogenous targets.

The evolutionary selection that has preserved three Rb paralogs in vertebrates, and two paralogs in the Drosophila lineage, is likely to reflect effects on diverse biological processes. Does the relatively new "experiment" in Drosophila reflect a subfunctionalization, i.e. division of labor, recapitulating an ancient molecular event that occurred in our own piscine progenitors? Some lines of evidence do support a model for parallel evolution, for instance, the conserved feature of preferential binding of ribosomal protein genes by a derived Rb paralog in both flies and humans, or the loss of the IE in derived CTD of Rbf2 and (partially) in Rb itself, which may impact targeting. These specializations, in both vertebrates and Drosophila, may have allowed for certain paralogs to focus regulation on cell growth related processes, such as protein synthesis and establishment of polarity, while other paralogs predominate for an ancestral function of cell cycle control. On the other hand, Drosophila-specific phenomena may have played an important role in shaping Rb paralog structure and function; for instance, the variety of developmental cues used in diverse Drosophila to impact reproductive capacity (Sarikaya *et al.* 2019). It is notable that of the two Rb paralogs in Drosophila, Rbf2, which is highly expressed in the female ovary, and whose loss impacts reproduction, is the more evolutionary divergent of the two. More detailed studies that measure quantitative Rb paralog contributions to diverse sets of putative gene targets, including those regulated by "soft repression", will build a deeper understanding of how Rb paralogy pursues novel and conserved patterns of sub- and neo-functionalization. (Mitra, Raicu, *et al.* 2021).

CRISPRi contributions to studies of transcriptional repression

Apart from the work of the Stark laboratory, most functional analyses of Rb repression activity have involved either single gene assays or perturbation of native Rb proteins, with global assessment of the transcriptome. These latter experiments are a joint measure of protein targeting and transcriptional activity, often combining direct and indirect effects. Thus, there has been a major need for precise assessment of protein function on a greater spectrum of promoter contexts, in order to identify and characterize the context-specific effects that we are beginning to appreciate. An important advantage of our CRISPRi system is its ability to directly query the response of dozens of endogenous genes to Rbf1 and Rbf2 targeting, with both transcriptional as well as sensitive developmental readouts, which proved to be related, but sometimes in a complex manner. From our survey of effects at the transcriptional level, we find that most frequently, Rbf1 is a more potent repressor, although Rbf2 is more effective in specific contexts. For instance, solely Rbf2 was able to cause the severe wing phenotype when targeting *E2F2/Mpp6* using single gRNA 4 or 5, possibly because of a greater ability to work as a single nucleation site for binding of chromatin modifying factors; Rbf1 may have weaker interactions that are stimulated by positioning multiple Rbf1 molecules on the promoter. The CRISPRi approach also proved to be highly informative about the utility of transient reporter assays used in so many functional studies. By deploying the identical gRNAs and CRISPRi effectors on the *E2F2* and *Mpp6* regulatory region, we found that the reporters faithfully reflected only certain properties of regulation *in vivo*. Specific distance-dependent effects were not reproduced in this system, and a non-specific promoter-blocking action of the CRISPRi molecules was only found in transiently transfected reporters, likely because of the less defined chromatin state of these plasmid-borne genes. The transient transfection approach was a useful complement for other reasons as well; we were able to directly compare CRISPRi effects to regulation by free, non-dCas9-fused Rb proteins that presumably are binding E2F transcription factors. The superior repression by native Rbf1 versus Rbf2 of *E2F2* (but less so *Mpp6*) indicates that E2F targeting may influence the different response, rather than inherent transcriptional potential.

Soft repression by Rb

One of the striking results from this study, though perhaps disappointing from a biotechnology perspective, is that for many instances, positioning active Rb repressors near endogenous promoters had little or no effect. To a certain degree, this may reflect the robustness of the wing developmental program; transcript levels were reduced for *wg* CRISPRi, but only the Cas9-VPR activation of the gene, not its repression, impacted wing development. Out of the 28 promoters that we targeted with Rbf1 and Rbf2, only one-third (10/28) led to a wing phenotype. Even targeting rbf1 and rbf2 promoters showed little or no effect, although we know that the wing is impacted by perturbation in Rbf1 levels (Payankaulam *et al.* 2016). Where repression was measured, this was generally 50% or less, although that may be a reflection of the limited pouch-specific pattern of expression of the *nubbin*-GAL4 transgene in the entire wing disc. For genome engineering purposes, it may be desirable to have a repression effector, such as the KRAB domain in mammals, that uniformly and decisively silences any target. However, from the perspective of biologically-relevant regulation, such specificity and context-dependence may be exactly what is required for optimal gene regulation by these Rb corepressors.

We have suggested that Rb factors may at times act as "soft repressors" (Mitra, Raicu *et al.* 2021), which do not function solely as an on/off switches for gene expression, but rather modulate gene expression by dialing down expression modestly, within physiologically relevant levels. This is in line with the Stark laboratory's experimental results that find GAL4-Rb proteins repress only half of the enhancers they are tested against, with mostly mild repression (Jacobs *et al.* 2022, *bioRxiv*). Soft repression would come into play when a target gene sits at a junction of metabolic pathways, including signaling elements such as InR and protein synthesis machinery, which are widely yet dynamically expressed in tune with demands of cellular physiology. The positional dependence and even promoter-specific effects noted for the divergent *E2F2/Mpp6* regulatory region align with a mechanism of gene regulation that is sensitive to subtleties of chromatin, neighboring factors, and basal promoter constitution.

MATERIALS AND METHODS

Cloning of dCas9 and gRNA constructs

The FLAG-tagged (DYKDDDDK) coding sequence for Rbf1 was obtained from the pRbf1 plasmid described previously (Acharya *et al.* 2010). Full length dCas9 (with D10A, H840A substitutions) was obtained from pcDNA-dCas9 (gift from Charles Gersbach; Addgene plasmid #47106; Perez-Pinera *et al.* 2013). The dCas9 coding sequence (with two 5' FLAG epitope tags) was subcloned into the pRbf1 parent vector, 5' of the Rbf1 sequence. During cloning, a linker between the 3' end of dCas9 and 5' end of Rbf1 was inserted, introducing PacI for simpler downstream cloning of other Rb effectors (linker sequence is 5'-GGCTTAATTAATAGTACC-3'). The dCas9-Rbf1 ORF was removed using 5' NotI and 3' XbaI sites and cloned into the pUASTattB vector. The final UAS:dCas9-Rbf1 vector was used to subclone other Rb FLAG-tagged sequences into the PacI-XbaI site, including Rbf2, Rbf1$^{\Delta IE}$, Rbf1^{3AE2F}, and Rbf1$^{\Delta pocket}$

(Acharya *et al.* 2010; Mouawad *et al.* 2020). All coding sequences for these effectors were removed from their parent vector using PCR amplification with 5’ PacI and 3’ XbaI sites introduced on either end. The dCas9 control was created by removing Rbf1 from UAS:dCas9-Rbf1 using 5’ PacI and 3’ XbaI, and inserting a gBlock containing an identical 3’ FLAG found in Rbf1 into the 3’ end of dCas9. Rbf1^{3AE2F} was created here through Site Directed Mutagenesis (Expand Long Template PCR system) to sequentially introduce F476A, E541A, and K588A into the Rbf1 coding sequence. Single gRNAs for the *E2F2/Mpp6* promoter were designed using the DRSC Find CRISPRs tool (www.flyrnai.org/crispr/) and cloned into the pCFD3.1 vector (Addgene #123366), cutting with BbsI and inserting the annealed oligos.

Transgenic flies

Flies were fed on standard lab food (molasses, yeast, corn meal) and kept at RT in the lab, under normal dark-light conditions. The *nubbin*-GAL4 fly line was obtained from the Bloomington Drosophila Stock Center (BDSC; #25754), and *traffic jam*-GAL4 was a gift from Sally Horne-Badovinac. GAL4 lines were maintained as a homozygous line with a Chr 3 balancer obtained from BDSC #3704 (w[1118]/Dp(1; Y)y[+]; CyO/Bl[1]; TM2, e/TM6B, e, Tb[1]). Homozygous dCas9-effector flies were generated by using the ϕC31 integrase service at Rainbow Transgenic Flies Inc. #24749 embryos were injected with each dCas9-effector construct to integrate into Chr 3 landing site 86Fb. Successful transgenic flies were selected through the mini-white selectable marker expression in-house, and maintained as a homozygous line with Chr 2 balancer (from BDSC #3704). driver-GAL4 and dCas9-effector flies were crossed to generate double homozygotes (driver-GAL4>UAS:dCas9-Rb). These flies were maintained in the lab and used for crossing to homozygous gRNA flies described below. *nub*-GAL4>UAS:dCas9-Rbf1 homozygous flies had a notched wing phenotype, consistent with Rbf1 overexpression in the wing (Elenbaas *et* *al*. 2015). *nub*-GAL4>UAS:dCas9-Rbf2 did not have a wing phenotype, and neither did any of the other homozygous fly lines. Two Rbf1 mutant lines (Rbf1^{K774A} and Rbf1^{3SA}) were also generated but not used further in this study, as the homozygous flies had severe morphological wing defects, which did not permit us to cross them to gRNA lines (Coding sequences described in Acharya *et al*. 2010, Zhang *et al*. 2014). The dCas9-VPR construct was obtained as a fly line from the BDSC #67055 (w[*]; P{w[+mW.hs]=GawB}nubbin-AC-62; P{y[+t7.7] w[+mC]=UAS-3xFLAG.dCas9.VPR}attP2). This dCas9 differs from the dCas9 used in our study, as it has two additional point mutations aside from D10A and H840A. sgRNA fly lines were created by Harvard TRiP, and obtained from the BDSC (fly line numbers indicated in **Table S2.1**). For gRNAs designed in this study (**Table S2.2**), #25709 embryos were injected with each gRNA plasmid through the ϕC31 integrase service at Rainbow Transgenic Flies Inc., to integrate into Chr 2 landing site 25C6, which is the same landing site as the TRiP gRNA lines. Successful transgenic flies were selected through the mini-white selectable marker expression in-house, and maintained as a homozygous line. Homozygous *nub*-GAL4>UAS:dCas9-Rb flies were crossed to homozygous gRNA flies to generate triple heterozygotes (-/-; *nubbin*-GAL4/sgRNA; UAS:dCas9-Rb/+) that are used for all fly experiments described here.

Genotyping flies

All flies generated in this study were genotyped at the adult stage. Flies of each genotype were homogenized (1 fly/tube) in squish buffer (1M Tris pH 8.0, 0.5M EDTA, 5M NaCl with 1μl of 10mg/mL Proteinase K for each fly). Tubes were set at 37C for 30 minutes, 95C for 2 mins, centrifuged at 14,000RPM for 7 minutes, and stored at 4C. Following PCR amplification, amplicons were cleaned using Wizard SV-Gel and PCR Clean-Up System, and sent for Sanger sequencing.

Imaging adult wings

Adult wings were collected from ~50 male and female 1-3 day-old adults. They were stored in 200 proof ethanol in -20C until mounted. Wings were removed, mounted onto ASi non-charged microscope slides using Permount, and photographed with a Canon PowerShot A95 camera mounted onto a Leica DMLB microscope. Images were all taken at the same magnification (10X) and using the same software settings.

Wing disc dissections and RT-qPCR

50 third instar wing discs were dissected from L3 larvae and placed in 200μl Trizol (ambion TRIzol Reagent) and stored in -80C until use. RNA was extracted using chloroform and the QIAGEN maXtract High Density kit, and stored in -80C. cDNA synthesis was performed using applied biosystems High Capacity cDNA Reverse Transcription Kit. RT-qPCR was performed using SYBR green (PerfeCTa SYBR Green FastMix Low ROX by Quantabio) and measured using the QuantStudio 3 machine by applied biosystems. Three control genes were averaged (Rp49, RpS13, CG8636) for all samples with control obtained from crossing dCas9 or dCas9 effectors to a non-targeting gRNA (QUAS). Primers used are indicated in **Table S2.3**. RT-qPCR was performed on 3-4 biological replicates with two technical duplicates. Student's t-test (two tailed, $p<0.05$) was used to measure statistical significance. Error bars indicate SEM.

Ovary dissections and RT-qPCR

For determining changes in gene expression in the ovary, traffic jam-GAL4>dCas9-Rb flies were crossed to gRNA flies. Female progeny were mated with male progeny for 5-6 days, and were fed a normal diet supplemented with yeast. 20-25 ovaries/replicate were dissected in 0.3% Triton-PBS, and were placed in 1X PBS before transferring to Trizol. RNA was extracted as described above, was immediately treated with DNase (TURBO DNA-free kit) and with RNAse inhibitor during the cDNA synthesis stage. RT-qPCR was performed as described above.

Luciferase reporter assays

Luciferase reporter constructs were created by amplifying yw^{67} (Mani-Telang and Arnosti, 2007) genomic DNA with primers flanking the 1kb promoter that is shared between *E2F2* and *Mpp6*, including the 5' UTR of *E2F2* and *Mpp6* but no coding sequence. The promoter was inserted into 5' NotI and 3' HindIII sites in the luciferase reporter plasmid described previously (Wei *et al*. 2016) in both orientations, generating *Mpp6*-luciferase (the *Mpp6* ATG is proximal to luciferase TSS) and *E2F2*-luciferase (the E2F2 ATG is proximal to luciferase TSS) reporters (Wei *et al*. 2015). Drosophila S2 cells were grown in 25C in Schneider Drosophila medium supplemented with glutamine (gibco) containing 10% FBS and 1% penicillin-streptomycin, and equal amounts of plasmids (250ng of each: luciferase reporter, actin-GAL4 (Addgene #24344), one dCas9-Rb constructs, and one sgRNA) were co-transfected into 1.5 million cells with QIAGEN Effectene transfection reagent. Cells were harvested 72 hours later and luciferase assay was performed using the Biotium Steady-Luc HTS assay kit. Briefly, cell media was removed after centrifugation, and cells were resuspended in PBS. The lysate was used in triplicate (70μl of lysate added to 70μl of luciferase reagent) and quantified using the Veritas microplate luminometer (Turner Biosystems). Values were normalized to a non-targeting gRNA control (empty pCFD3 plasmid). Three technical replicates were averaged for each biological replicate. Three to five biological replicates were compared via Student's t-test to test for significance.

Western blot

~15 L3 wing discs were dissected in 1X PBS and placed in 35μl cell lysis buffer (50mM Tris, pH 8.0; 150 mM NaCl; 1% Triton X-100). Tubes were flash frozen in liquid nitrogen and stored in -

80C. Discs were homogenized and protein was extracted by freezing, thawing, and vortexing several times, followed by boiling in Laemmli buffer. 20μl of lysate was used in the western blot as described below. Western blot was also performed on lysates from Drosophila S2 cells. S2 cells were grown in 25C in Schneider Drosophila medium with glutamine (Gibco/BRL) containing 10% FBS and 1% penicillin-streptomycin. 1.5 million cells were co-transfected with Effectene Transfection Reagent (Qiagen), according to manufacturer's protocol. 250ng of actin-GAL4 (Addgene #24344) and 250ng of UAS:dCas9-Rb were co-transfected in 6-well plates. Cells were harvested 3 days post-transfection and lysed using S2 lysis buffer (50mM Tris, pH 8.0; 150 mM NaCl; 1% Triton X-100), followed by boiling with Laemmli buffer. 100ug of cell lysates were separated on a 4-20% resolving gel (Bio-Rad Mini-PROTEAN TGX Precast Gel #456-1094), transferred to a PVDF membrane for analysis using α-FLAG antibody (Sigma Aldrich #F3165, 1:10,000), and α-CtBP (DNA208; Keller *et al*. 2005). Blocking with both primary and secondary antibodies was performed in 5% milk-TBST (500mM Tris-HCl, pH 7.4, 1500 mM NaCl, 0.1% Tween 20). Blots were developed using HRP-conjugated GαM and GαR secondary antibody (Pierce), and imaged using SuperSignal West Pico PLUS chemiluminescent substrate.

<u>E2F motif search</u>

The E2F motif, described previously (Acharya *et al*. 2012), was identified in the *E2F2* gene promoter using MAST (MEME-suite v5.0.2) using p<0.0005 cut off.

<u>Identification of promoter landscape</u>

We used the UCSC Genome Browser, Flybase.org, and ChIP-Atlas to identify the chromatin environment on the *E2F2*/*Mpp6* bidirectional promoter in both S2 cells and L3 larval discs.

ACKNOWLEDGEMENTS

We would like to thank Tyler Chau and Matthew Chang for their assistance analyzing transcriptomic and bioinformatic data, Megha Suresh for generating the mutant luciferase plasmids, members of the Arnosti lab for their thoughtful suggestions, and the Michigan State University RTSF and plasmidsaurus for sequencing plasmids.

FUNDING

This work was supported by the National Institute for General Medical Sciences grant (number R01GM124137) to D.N.A., and the National Institute of Child Health and Development grant (number F31HD105410) to A.M.R.

REFERENCES

Acharya, P., Raj, N., Buckley, M. S., Zhang, L., Duperon, S., Williams, G., Henry, R. W., & Arnosti, D. N. (2010). **Paradoxical Instability–Activity Relationship Defines a Novel Regulatory Pathway for Retinoblastoma Proteins.** Molecular Biology of the Cell, 21(22), 3890–3901. https://doi.org/10.1091/mbc.e10-06-0520

Acharya, P., Negre, N., Johnston, J., Wei, Y., White, K. P., Henry, R. W., & Arnosti, D. N. (2012). **Evidence for Autoregulation and Cell Signaling Pathway Regulation From Genome-Wide Binding of the Drosophila Retinoblastoma Protein.** G3: Genes|Genomes|Genetics, 2(11), 1459–1472. https://doi.org/10.1534/g3.112.004424

Adnane, J., Shao, Z., & Robbins, P. D. (1995). **The Retinoblastoma Susceptibility Gene Product Represses Transcription When Directly Bound to the Promoter.** Journal of Biological Chemistry, 270(15), 8837–8843. https://doi.org/10.1074/jbc.270.15.8837

Brand, A. H., & Perrimon, N. (1993). **Targeted gene expression as a means of altering cell fates and generating dominant phenotypes.** Development, 118(2), 401–415. https://doi.org/10.1242/dev.118.2.401

Bremner, R., Cohen, B. L., Sopta, M., Hamel, P. A., Ingles, C. J., Gallie, B. L., & Phillips, R. A. (1995). **Direct Transcriptional Repression by pRB and Its Reversal by Specific Cyclins.** Molecular and Cellular Biology, 15(6), 3256–3265. https://doi.org/10.1128/MCB.15.6.3256

Chicas, A., Wang, X., Zhang, C., McCurrach, M., Zhao, Z., Mert, O., Dickins, R. A., Narita, M., Zhang, M., & Lowe, S. W. (2010). **Dissecting the Unique Role of the Retinoblastoma Tumor Suppressor during Cellular Senescence.** Cancer Cell, 17(4), 376–387. https://doi.org/10.1016/j.ccr.2010.01.023

Cecchini, M. J., Thwaites, M. J., Talluri, S., MacDonald, J. I., Passos, D. T., Chong, J.-L., Cantalupo, P., Stafford, P. M., Sáenz-Robles, M. T., Francis, S. M., Pipas, J. M., Leone, G., Welch, I., & Dick, F. A. (2014). **A Retinoblastoma Allele That Is Mutated at Its Common E2F Interaction Site Inhibits Cell Proliferation in Gene-Targeted Mice.** Molecular and Cellular Biology, 34(11), 2029–2045. https://doi.org/10.1128/MCB.01589-13

Chow, K. N. B., Starostik, P., & Dean, D. C. (1996). **The Rb Family Contains a Conserved Cyclin-Dependent-Kinase-Regulated Transcriptional Repressor Motif.** Molecular and Cellular Biology, 16(12), 7173–7181. https://doi.org/10.1128/MCB.16.12.7173

Chau, B. N., & Wang, J. Y. J. (2003). **Coordinated regulation of life and death by RB.** Nature Reviews Cancer, 3(2), 130–138. https://doi.org/10.1038/nrc993

Dick, F. A., & Rubin, S. M. (2013). **Molecular mechanisms underlying RB protein function.** Nature Reviews Molecular Cell Biology, 14(5), 297–306. https://doi.org/10.1038/nrm3567

Dimova, D. K. (2003). **Cell cycle-dependent and cell cycle-independent control of transcription by the Drosophila E2F/RB pathway.** Genes & Development, 17(18), 2308–2320. https://doi.org/10.1101/gad.1116703

Du, W., & Dyson, N. (1999). **The role of RBF in the introduction of G1 regulation during Drosophila embryogenesis.** The EMBO Journal, 18(4), 916–925. https://doi.org/10.1093/emboj/18.4.916

Dynlacht, B. D., Flores, O., Lees, J. A., & Harlow, E. (1994). **Differential regulation of E2F transactivation by cyclin/cdk2 complexes.** Genes & Development, 8(15), 1772–1786. https://doi.org/10.1101/gad.8.15.1772

Elenbaas, J. S., Mouawad, R., Henry, R. W., Arnosti, D. N., & Payankaulam, S. (2015). **Role of Drosophila retinoblastoma protein instability element in cell growth and proliferation.** Cell Cycle, 14(4), 589–597. https://doi.org/10.4161/15384101.2014.991182

Ewen-Campen, B., Yang-Zhou, D., Fernandes, V. R., González, D. P., Liu, L.-P., Tao, R., Ren, X., Sun, J., Hu, Y., Zirin, J., Mohr, S. E., Ni, J.-Q., & Perrimon, N. (2017). **Optimized strategy for in vivo Cas9-activation in Drosophila.** Proceedings of the National Academy of Sciences, 114(35), 9409–9414. https://doi.org/10.1073/pnas.1707635114

Ferreira, R., Magnaghi-Jaulin, L., Robin, P., Harel-Bellan, A., & Trouche, D. (1998). **The three members of the pocket proteins family share the ability to repress E2F activity through recruitment of a histone deacetylase.** Proceedings of the National Academy of Sciences, 95(18), 10493–10498. https://doi.org/10.1073/pnas.95.18.10493

Gray, S., & Levine, M. (1996). **Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila.** Genes & Development, 10(6), 700–710. https://doi.org/10.1101/gad.10.6.700

Jacobs, J., Pagani, M., Wenzl, C., & Stark, A. (2022). **Widespread regulatory specificities between transcriptional corepressors and enhancers in Drosophila** [Preprint]. Molecular Biology. https://doi.org/10.1101/2022.11.07.515017

Jiang, Z., Zacksenhaus, E., Gallie, B. L., & Phillips, R. A. (1997). **The retinoblastoma gene family is differentially expressed during embryogenesis.** Oncogene, 14(15), 1789–1797. https://doi.org/10.1038/sj.onc.1201014

Jiang, H., Karnezis, A. N., Tao, M., Guida, P. M., & Zhu, L. (2000). **pRB and p107 have distinct effects when expressed in pRB-deficient tumor cells at physiologically relevant levels.** Oncogene, 19, 3878-3887.

Keller, S. A., Ullah, Z., Buckley, M. S., William Henry, R., & Arnosti, D. N. (2005). **Distinct developmental expression of Drosophila retinoblastoma factors.** Gene Expression Patterns, 5(3), 411–421. https://doi.org/10.1016/j.modgep.2004.09.005

Korenjak, M., Anderssen, E., Ramaswamy, S., Whetstine, J. R., & Dyson, N. J. (2012). **RBF Binding to both Canonical E2F Targets and Noncanonical Targets Depends on Functional dE2F/dDP Complexes.** Molecular and Cellular Biology, 32(21), 4375–4387. https://doi.org/10.1128/MCB.00536-12

Knudson, A. G. (1971). **Mutation and Cancer: Statistical Study of Retinoblastoma.** Proceedings of the National Academy of Sciences, 68(4), 820–823. https://doi.org/10.1073/pnas.68.4.820

Lam, E. W., & Watson, R. J. (1993). **An E2F-binding site mediates cell-cycle regulated repression of mouse B-myb transcription.** The EMBO Journal, 12(7), 2705–2713. https://doi.org/10.1002/j.1460-2075.1993.tb05932.x

Lee, C., Chang, J.H., Lee, H.S., & Cho, Y. (2002). **Structural basis for the recognition of the E2F transactivation domain by the retinoblastoma tumor suppressor.** Genes & Development, 16(24), 3199–3212. https://doi.org/10.1101/gad.1046102

Liban, T. J., Thwaites, M. J., Dick, F. A., & Rubin, S. M. (2016). **Structural Conservation and E2F Binding Specificity within the Retinoblastoma Pocket Protein Family.** Journal of Molecular Biology, 428(20), 3960–3971. https://doi.org/10.1016/j.jmb.2016.08.017

Liban, T. J., Medina, E. M., Tripathi, S., Sengupta, S., Henry, R. W., Buchler, N. E., & Rubin, S. M. (2017). **Conservation and divergence of C-terminal domain structure in the retinoblastoma protein family.** Proceedings of the National Academy of Sciences, 114(19), 4942–4947. https://doi.org/10.1073/pnas.1619170114

Luo, R. X., Postigo, A. A., & Dean, D. C. (1998). **Rb Interacts with Histone Deacetylase to Repress Transcription.** Cell, 92(4), 463–473. https://doi.org/10.1016/S0092-8674(00)80940-X

Magnaghi-Jaulin, L., Groisman, R., Naguibneva, I., Robin, P., Lorain, S., Le Villain, J. P., Troalen, F., Trouche, D., & Harel-Bellan, A. (1998). **Retinoblastoma protein represses transcription by recruiting a histone deacetylase.** Nature, 391(6667), 601–605. https://doi.org/10.1038/35410

Mani-Telang, P., & Arnosti, D. N. (2007). **Developmental expression and phylogenetic conservation of alternatively spliced forms of the C-terminal binding protein corepressor.** Development Genes and Evolution, 217(2), 127–135. https://doi.org/10.1007/s00427-006-0121-4

Meloni, A. R., Smith, E. J., & Nevins, J. R. (1999). **A mechanism for Rb/p130-mediated transcription repression involving recruitment of the CtBP corepressor.** Proceedings of the National Academy of Sciences, 96(17), 9574–9579. https://doi.org/10.1073/pnas.96.17.9574

Mitra, A., Raicu, A., Hickey, S. L., Pile, L. A., & Arnosti, D. N. (2021). **Soft repression: Subtle transcriptional regulation with global impact.** BioEssays, 43(2), 2000231. https://doi.org/10.1002/bies.202000231

Mouawad, R., Prasad, J., Thorley, D., Himadewi, P., Kadiyala, D., Wilson, N., Kapranov, P., &

Arnosti, D. N. (2019). **Diversification of Retinoblastoma Protein Function Associated with Cis and Trans Adaptations.** Molecular Biology and Evolution, 36(12), 2790–2804. https://doi.org/10.1093/molbev/msz187

Mouawad, R., Himadewi, P., Kadiyala, D., & Arnosti, D. N. (2020). **Selective repression of the Drosophila cyclin B promoter by retinoblastoma and E2F proteins.** Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1863(7), 194549. https://doi.org/10.1016/j.bbagrm.2020.194549

Mulligan, G., & Jacks, T. (1998). **The retinoblastoma gene family: Cousins with overlapping interests.** Trends in Genetics, 14(6), 223–229. https://doi.org/10.1016/S0168-9525(98)01470-X

Nicolay, B. N., & Dyson, N. J. (2013). **The multiple connections between pRB and cell metabolism.** Current Opinion in Cell Biology, 25(6), 735–740. https://doi.org/10.1016/j.ceb.2013.07.012

Payankaulam, S., Yeung, K., McNeill, H., Henry, R. W., & Arnosti, D. N. (2016). **Regulation of cell polarity determinants by the Retinoblastoma tumor suppressor protein.** Scientific Reports, 6(1). https://doi.org/10.1038/srep22879

Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2013). **RNA-guided gene activation by CRISPR-Cas9–based transcription factors.** Nature Methods, 10(10), 973–976. https://doi.org/10.1038/nmeth.2600

Potter, C. J., Tasic, B., Russler, E. V., Liang, L., & Luo, L. (2010). **The Q System: A Repressible Binary System for Transgene Expression, Lineage Tracing, and Mosaic Analysis.** Cell, 141(3), 536–548. https://doi.org/10.1016/j.cell.2010.02.025

Qin, X.-Q., Livingston, D. M., Ewen, M., Sellers, W. R., Arany, Z., & Kaelin, W. G. (1995). The Transcription Factor E2F-1 Is a Downstream Target of RB Action. Molecular and Cellular Biology, 15(2), 742–755. https://doi.org/10.1128/MCB.15.2.742

Raj, N., Zhang, L., Wei, Y., Arnosti, D. N., & Henry, R. W. (2012). **Ubiquitination of Retinoblastoma Family Protein 1 Potentiates Gene-specific Repression Function.** Journal of Biological Chemistry, 287(50), 41835–41843. https://doi.org/10.1074/jbc.M112.422428

Ross, J. F., Liu, X., & Dynlacht, B. D. (1999). **Mechanism of Transcriptional Repression of E2F by the Retinoblastoma Tumor Suppressor Protein.** Molecular Cell, 3(2), 195–205. https://doi.org/10.1016/S1097-2765(00)80310-X

Sarikaya, D. P., Church, S. H., Lagomarsino, L. P., Magnacca, K. N., Montgomery, S. L., Price, D. K., Kaneshiro, K. Y., & Extavour, C. G. (2019). **Reproductive Capacity Evolves in Response to Ecology through Common Changes in Cell Number in Hawaiian Drosophila.** Current Biology, 29(11), 1877-1884.e6. https://doi.org/10.1016/j.cub.2019.04.063

Sellers, W. R., Novitch, B. G., Miyake, S., Heith, A., Otterson, G. A., Kaye, F. J., Lassar, A. B., & Kaelin, W. G. (1998). **Stable binding to E2F is not required for the retinoblastoma protein to activate transcription, promote differentiation, and suppress tumor cell growth.** Genes & Development, 12(1), 95–106. https://doi.org/10.1101/gad.12.1.95

Sun, H., Chang, Y., Schweers, B., Dyer, M. A., Zhang, X., Hayward, S. W., & Goodrich, D. W. (2006). **An E2F Binding-Deficient Rb1 Protein Partially Rescues Developmental Defects Associated with Rb1 Nullizygosity.** Molecular and Cellular Biology, 26(4), 1527–1537. https://doi.org/10.1128/MCB.26.4.1527-1537.2006

Sanidas, I., Lee, H., Rumde, P. H., Boulay, G., Morris, R., Golczer, G., Stanzione, M., Hajizadeh, S., Zhong, J., Ryan, M. B., Corcoran, R. B., Drapkin, B. J., Rivera, M. N., Dyson, N. J., & Lawrence, M. S. (2022). **Chromatin-bound RB targets promoters, enhancers, and CTCF-bound loci and is redistributed by cell-cycle progression.** Molecular Cell, 82(18), 3333-3349.e9. https://doi.org/10.1016/j.molcel.2022.07.014

Sengupta, S., Lingnurkar, R., Carey, T. S., Pomaville, M., Kar, P., Feig, M., Wilson, C. A., Knott, J. G., Arnosti, D. N., & Henry, R. W. (2015). **The Evolutionarily Conserved C-terminal Domains in the Mammalian Retinoblastoma Tumor Suppressor Family Serve as Dual Regulators of Protein Stability and Transcriptional Potency.** Journal of Biological Chemistry, 290(23), 14462–14475. https://doi.org/10.1074/jbc.M114.599993

Stevaux, O. (2002). **Distinct mechanisms of E2F regulation by Drosophila RBF1 and RBF2.** The EMBO Journal, 21(18), 4927–4937. https://doi.org/10.1093/emboj/cdf501

Stevaux, O., Dimova, D. K., Ji, J.-Y., Moon, N. S., Frolov, M. V., & Dyson, N. J. (2005). **Retinoblastoma Family 2 is Required In Vivo for the Tissue-Specific Repression of dE2F2 target Genes.** Cell Cycle, 4(9), 1272–1280. https://doi.org/10.4161/cc.4.9.1982

Wei, Y., Mondal, S. S., Mouawad, R., Wilczyński, B., Henry, R. W., & Arnosti, D. N. (2015). **Genome-Wide Analysis of Drosophila RBf2 Protein Highlights the Diversity of RB Family Targets and Possible Role in Regulation of Ribosome Biosynthesis.** G3 Genes|Genomes|Genetics, 5(7), 1503–1515. https://doi.org/10.1534/g3.115.019166

Weinberg, R. A., Hinds, P. W., Mittnacht, S., Dulic, V., Arnold, A., & Reed, S. I. (1992). **Regulation of retinoblastoma protein functions by ectopic expression of human cyclins.** Cell, 70(6), 993–1006. https://doi.org/10.1016/0092-8674(92)90249-C

Weinberg, R. A. (1995). **The retinoblastoma protein and cell cycle control.** Cell, 81(3), 323–330. https://doi.org/10.1016/0092-8674(95)90385-2

Weintraub, S. J., Chow, K. N. B., Luo, R. X., Zhang, S. H., He, S., & Dean, D. C. (1995). **Mechanism of active transcriptional repression by the retinoblastoma protein.** Nature, 375(6534), 812–816. https://doi.org/10.1038/375812a0

Zhang, H. S., & Dean, D. C. (2001). **Rb-mediated chromatin structure regulation and**

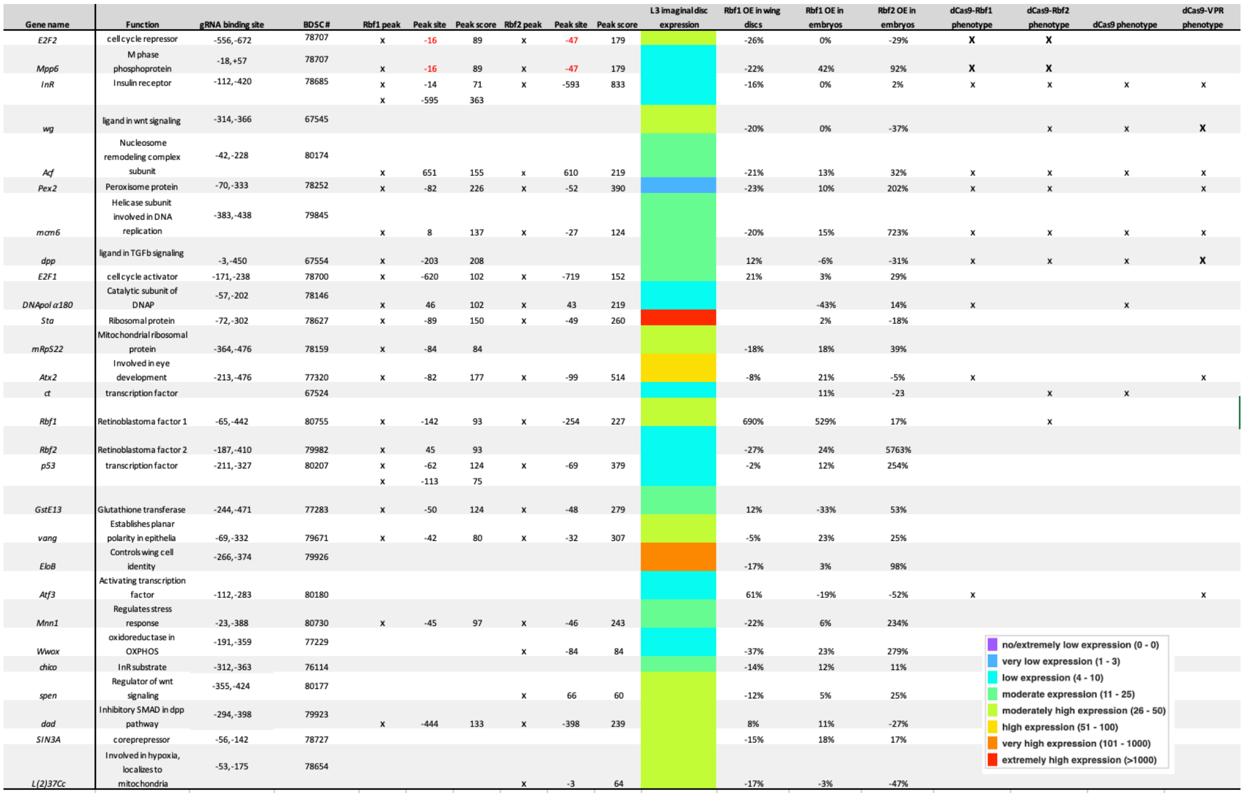
transcriptional repression. Oncogene, 20(24), 3134–3138. https://doi.org/10.1038/sj.onc.1204338

Zirin, J., Bosch, J., Viswanatha, R., Mohr, S. E., & Perrimon, N. (2022). **State-of-the-art CRISPR for in vivo and cell-based studies in Drosophila.** Trends in Genetics, 38(5), 437–453. https://doi.org/10.1016/j.tig.2021.11.006

APPENDIX A: SUPPLEMENTARY MATERIALS

This work was published as supplementary materials in the following preprint:

Raicu, A.M., Castanheira, P., & Arnosti, D. N. (2023). **Retinoblastoma protein activity revealed by CRISPRi study of divergent Rbf1 and Rbf2 paralogs.** *bioRxiv*. 2023.05.19.541454; doi: https://doi.org/10.1101/2023.05.19.541454



Gene name	Function	gRNA binding site	BDSC#	Rbf1 peak	Peak site	Peak score	Rbf2 peak	Peak site	Peak score	L3 imaginal disc expression	Rbf1 OE in wing discs	Rbf1 OE in embryos	Rbf2 OE in embryos	dCas9-Rbf1 phenotype	dCas9-Rbf2 phenotype	dCas9 phenotype	dCas9-VPR phenotype
E2F2	cell cycle repressor	-556, -672	78707	x	-16	89	x	-47	179	🟢	-26%	0%	-29%	X	X		
Mpp6	M phase phosphoprotein	-18, +57	78707	x	-16	89	x	-47	179	🟦	-22%	42%	92%	X	X		
InR	Insulin receptor	-112, -420	78685	x	-14	71	x	-593	833	🟦	-16%	0%	2%	x	x	x	x
				x	-595	363											
wg	ligand in wnt signaling	-314, -366	67545							🟢	-20%	0%	-37%		x	x	X
Acf	Nucleosome remodeling complex subunit	-42, -228	80174	x	651	155	x	610	219	🟩	-21%	13%	32%	x	x	x	x
Pex2	Peroxisome protein	-70, -333	78252	x	-82	226	x	-52	390	🔵	-23%	10%	202%	x	x		x
mcm6	Helicase subunit involved in DNA replication	-383, -438	79845	x	8	137	x	-27	124	🟩	-20%	15%	723%	x	x	x	x
dpp	ligand in TGFb signaling	-3, -450	67554	x	-203	208				🟩	12%	-6%	-31%	x	x	x	X
E2F1	cell cycle activator	-171, -238	78700	x	-620	102	x	-719	152	🟩	21%	3%	29%				
DNApol α180	Catalytic subunit of DNAP	-57, -202	78146	x	46	102	x	43	219	🟦		-43%	14%	x		x	
Sta	Ribosomal protein	-72, -302	78627	x	-89	150	x	-49	260	🟥		2%	-18%				
mRpS22	Mitochondrial ribosomal protein	-364, -476	78159	x	-84	84				🟢	-18%	18%	39%				
Atx2	Involved in eye development	-213, -476	77320	x	-82	177	x	-99	514	🟨	-8%	21%	-5%	x			x
ct	transcription factor		67524							🟦		11%	-23		x	x	
Rbf1	Retinoblastoma factor 1	-65, -442	80755	x	-142	93	x	-254	227	🟢	690%	529%	17%		x		
Rbf2	Retinoblastoma factor 2	-187, -410	79982	x	45	93				🟦	-27%	24%	5763%				
p53	transcription factor	-211, -327	80207	x	-62	124	x	-69	379	🟦	-2%	12%	254%				
				x	-113	75											
GstE13	Glutathione transferase	-244, -471	77283	x	-50	124	x	-48	279	🟩	12%	-33%	53%				
vang	Establishes planar polarity in epithelia	-69, -332	79671	x	-42	80	x	-32	307	🟢	-5%	23%	25%				
EloB	Controls wing cell identity	-266, -374	79926							🟧	-17%	3%	98%				
Atf3	Activating transcription factor	-112, -283	80180							🟦	61%	-19%	-52%	x			x
Mnn1	Regulates stress response	-23, -388	80730	x	-45	97	x	-46	243	🟩	-22%	6%	234%				
Wwox	oxidoreductase in OXPHOS	-191, -359	77229				x	-84	84	🟦	-37%	23%	279%				
chico	InR substrate	-312, -363	76114							🟩	-14%	12%	11%				
spen	Regulator of wnt signaling	-355, -424	80177				x	66	60	🟢	-12%	5%	25%				
dad	Inhibitory SMAD in dpp pathway	-294, -398	79923	x	-444	133	x	-398	239	🟢	8%	11%	-27%				
SIN3A	corepressor	-56, -142	78727							🟢	-15%	18%	17%				
L(2)37Cc	Involved in hypoxia, localizes to mitochondria	-53, -175	78654				x	-3	64	🟢	-17%	-3%	-47%				

🟪 no/extremely low expression (0 - 0)
🔵 very low expression (1 - 3)
🟦 low expression (4 - 10)
🟩 moderate expression (11 - 25)
🟢 moderately high expression (26 - 50)
🟨 high expression (51 - 100)
🟧 very high expression (101 - 1000)
🟥 extremely high expression (>1000)

Table S2.1. Genes targeted with dCas9-Rb effectors in this study. Target gene names are indicated, along with their function, where the gRNAs obtained from Harvard TRiP bind relative to the gene's annotated TSS, and whether Rbf1 and Rbf2 ChIP peaks are present on the promoter (Wei *et al*. 2015). Expression level of each gene in L3 imaginal discs is indicated by colors shown in the legend (obtained from Flybase), effect on gene's expression after Rbf1 overexpression (OE) in L3 wing discs (Mouawad *et al*. unpublished), effect of Rbf1 and Rbf2 OE in embryos (Mouawad *et al*. 2019), and whether (x) or not an adult wing phenotype was observed after recruitment of dCas9-effectors in L3 wing discs. Large X indicates a more severe phenotype than the smaller x. Small x encompasses a variety of phenotypes including missing ACV, ectopic veins, and supernumerary bristles. Wings from genes listed after dpp were not dissected and imaged as the top eight were, thus we cannot exclude possible presence of subtle phenotypes.

gRNA	gRNA sequence(s)
1	GAAAAAAATGACATAAATGG
2	GCAGTGCGCACGAAGAATAG
3	GGACAATAAACCTTAACGAA
4	TTCAGCCTAGCTAGAAAACG
5	AAAACTAGGGCGAAACCATC
6	AGTCTCGGCTTTGATTTGGA
7	TCAAAGCCGAGACTTTCGCG
A	AGTTGAGCTTGTTTGTCAGT
B	TATAGTCATCGAGTCGATTG

Table S2.2. Sequences of gRNAs designed for the *E2F2/Mpp6* bidirectional promoter. gRNA label corresponds to positions indicated in **Figure 2.4**.

Gene	Primers (Forward and Reverse)
CG8636	F: GATCCGCTGCTAGATCCCAC R: CCCTTGTACGGGCAGTTGA
Rp49	F: ATCGGTTACGGATCGAACAAGC R: GTAAACGCGGTTCTGCATGAGC
Rps13	F: GGTCGTATGCACGCTCCT R: CATCTGCGTTCAGTTTCAGC
E2F2	F: GACGAGGAAGTAGATATCAAGCG R: TCAAAGAACCCATCCACATCG
InR	F: ACGACAACAAAACCGTTGC R: TTCACGTGATCTCAATCATGC
Mpp6	F: GCTCGGTCATTCTGCTTTTG R: CTCGGCTTTGATTTGGATGG
wg	F:TCGGATTCGGGTTCAAGTTC R:CACTCCTGTCGCATCTCC

Table S2.3. RT-qPCR primers used in this study. F indicates Forward primer, R indicates Reverse primer.

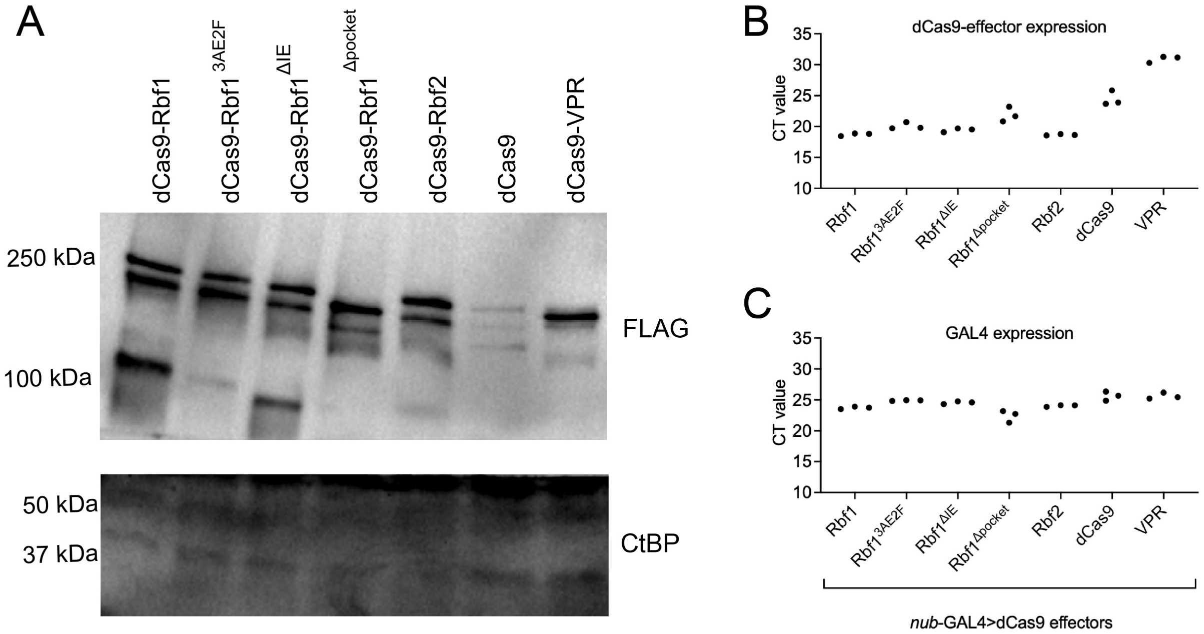


Figure S2.1. Evidence for expression of dCas9 isoforms *in vivo*. A) Western blot of dCas9 effectors expressed in L3 wing discs (genotypes shown in **Figure 2.1**) shows similar levels of expression of effectors at the protein level, aside from dCas9 alone which is expressed at lower levels. **B)** mRNA levels of dCas9-Rb effectors expressed in L3 wing discs were measured using RT-qPCR. Expression level (CT value) indicates that the effectors are expressed at relatively similar levels, aside from dCas9 and dCas9-VPR which are expressed at lower levels. **C)** mRNA levels of the *nubbin*-GAL4 transgene from L3 wing discs indicates that the GAL4 driver levels are similar from one genotype to the next, as expected.

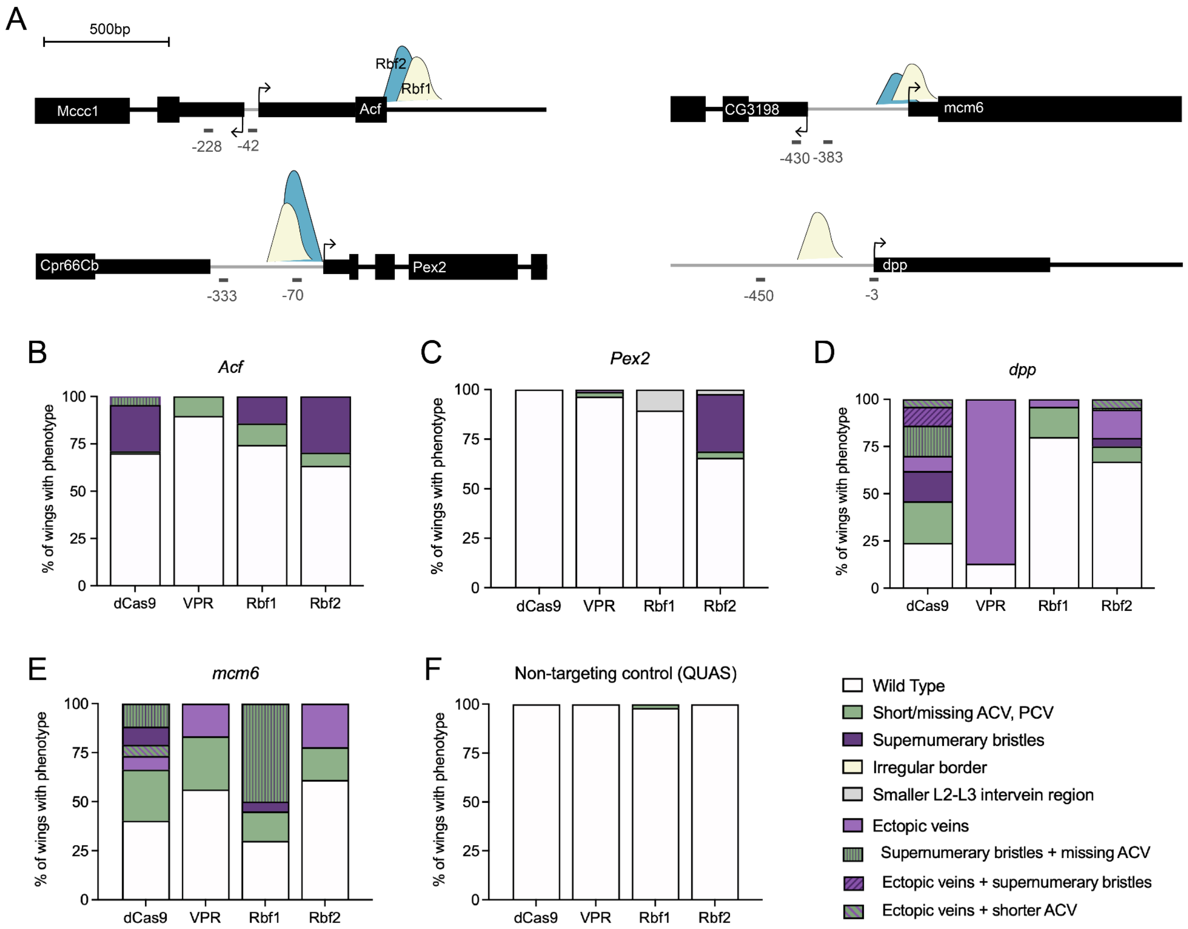


Figure S2.2. Rbf1 and Rbf2 targeted to diverse gene promoters produce gene-specific effects. A) Diagram of gRNA binding sites on the promoters of Acf, Pex2, mcm6, and dpp. Arrows indicate the TSS. Thick black bars are exons, thinner bars are UTR, and black lines are introns. Gray horizontal lines indicate intergenic regions, and gray lines below the genes are the locations of the gRNAs, with approximate distance from the TSS indicated below. These gRNAs were obtained from Harvard TRiP. Yellow peaks indicate Rbf1 binding sites and blue peaks indicate Rbf2 binding sites from embryo exo-ChIP (Wei *et al*. 2015). **B)** Targeting Acf caused mild phenotypes such as supernumerary bristle formation by dCas9, Rbf1, and Rbf2. **C)** Targeting Pex2 caused mild phenotypes that are different between Rbf1 and Rbf2, but not very penetrant. **D)** Targeting dpp caused some effector-specific phenotypes, such as ectopic venation in almost all wings by the VPR activator. Rb paralogs had mild effects in ~25% of wings. **E)** Targeting mcm6 caused a wide variety of phenotypes by all effectors, suggesting steric hindrance regardless of effector. **F)** Expression of effectors with a non-targeting gRNA control (QUAS) was used to determine the background of expressing dCas9-Rb in the L3 wing discs. The QUAS control demonstrates that effectors not targeted to any locus on the genome do not cause any adult wing phenotypes.

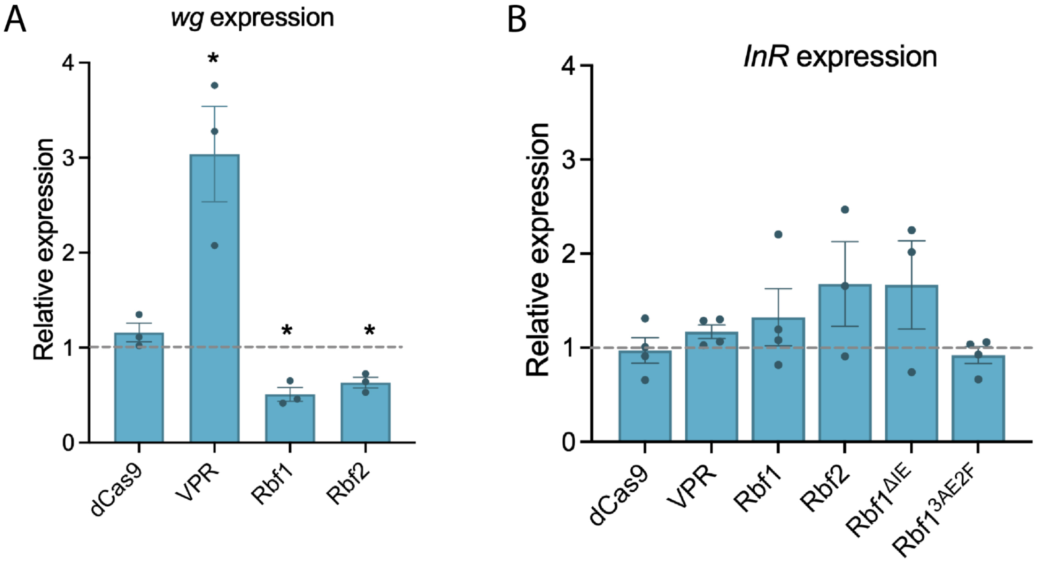


Figure S2.3. Rb impact on expression of other gene targets. A) The *wg* promoter was targeted by the Rb paralogs and the VPR activator. The VPR activator led to severe morphological defects (see **Figure 2.1C**), which is corroborated by a 3-4X increase in expression at the RNA level. Rbf1 and Rbf2 targeting did not lead to any overt wing phenotype, but transcript levels decreased by half. This suggests that Rb proteins can regulate at the transcript level, but depending on the gene that is being targeted, it may or may not lead to developmental defects. **B)** The *InR* promoter was targeted by Rb paralogs, Rb mutants, and the VPR activator. None of the effectors caused a decrease in *InR* expression; instead, Rbf1 and Rbf2 may have caused de-repression of the gene. Rbf1$^{\Delta IE}$ functions like the WT Rbf1, and Rbf1^{3AE2F} has no effect, similar to results described in **Figure 2.8**. Error bars indicate SEM, and * indicates p<0.05.

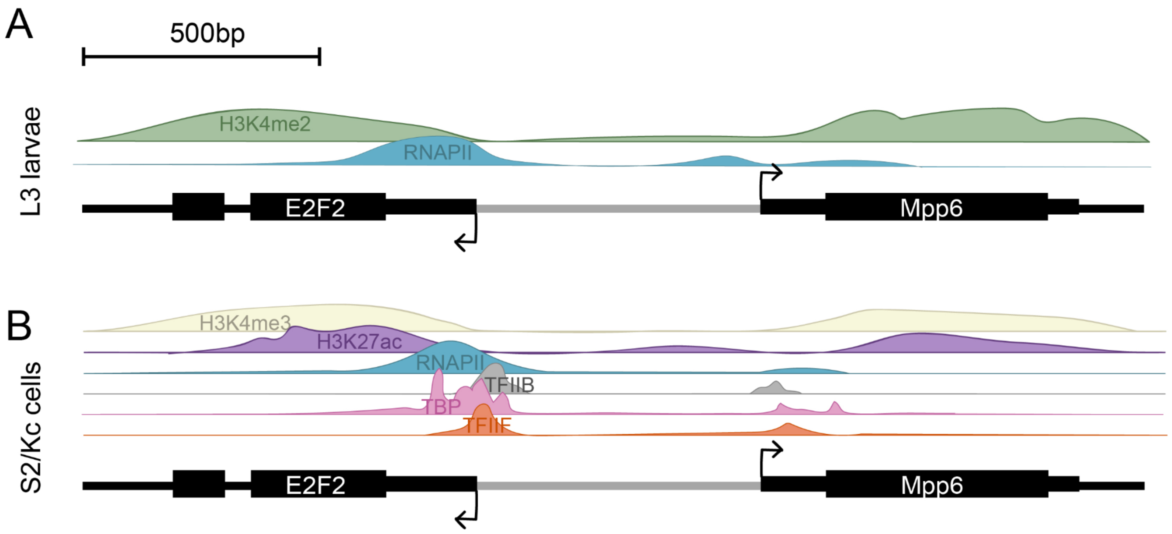


Figure S2.4. ***E2F2/Mpp6* promoter landscape.** Additional complexity on this locus stems from the fact that both *E2F2* and *Mpp6* actually have two TSSs, each of which is about 400bp upstream of the indicated TSS (arrow). However, RAMPAGE-seq and CAGE-seq data indicate that the downstream start sites are most often used, both in S2 cells and *in vivo* (UCSC Genome Browser). **A)** Chromatin landscape in L3 larvae. RNAPII binding is indicated in blue, over the two TSSs. H3K4me2 marks, which are indicative of transcriptionally active genes, are indicated in green. **B)** Chromatin landscape in S2 and Kc cells. H3K4me3 marks (yellow) and H3K27ac marks (purple) overlap the gene bodies, and are marks of active genes. RNAPII binding (blue) overlaps both TSSs, with a bigger peak over *E2F2*, where TFIIB (gray), TBP (pink), and TFIIF (orange) are also found.

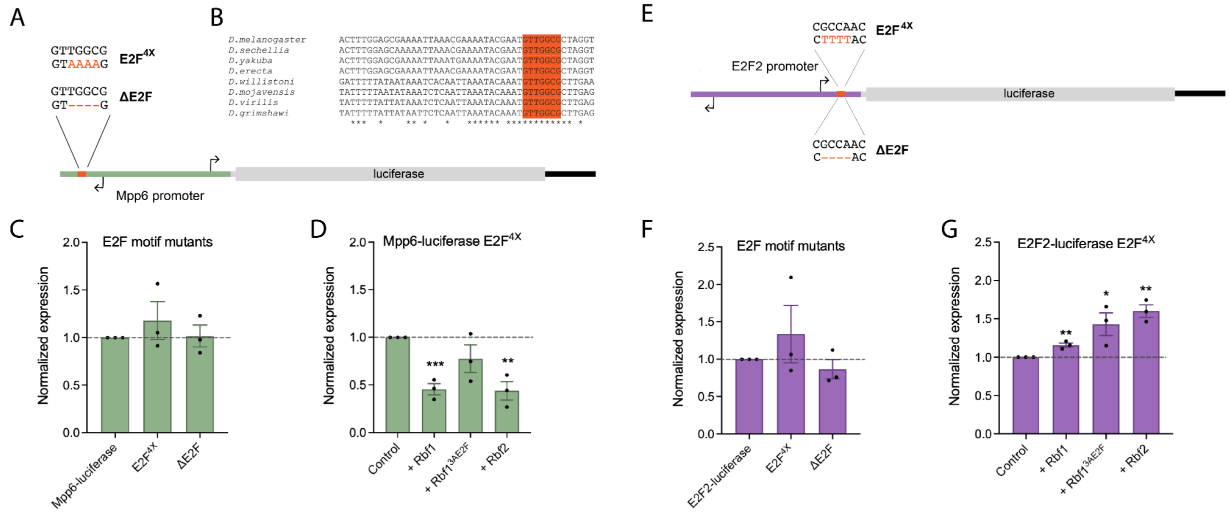


Figure S2.5. The conserved E2F motif is not necessary for *E2F2* and *Mpp6* expression. A) Schematic of the *Mpp6*-luciferase reporter used in transfections with Rb plasmids. The orange box indicates a highly conserved E2F motif that was mutated in two ways: E2F^{4X} changes 4 nucleotides to A, and ΔE2F removes 4 nucleotides from the motif. **B)** The E2F motif is perfectly conserved across multiple Drosophila species. **C)** Expression levels of the two mutant promoters relative to the WT *Mpp6* promoter. Both types of mutations to the E2F motif do not affect the normal expression level of the promoter. **D)** The E2F^{4X} mutant has minimal impact on how Rb proteins regulate the *Mpp6* promoter. **E)** Schematic of the *E2F2*-luciferase reporter used in transfections with Rb plasmids. The orange box indicates the same highly conserved E2F motif shown in panel B, but in the opposite orientation. The motif was mutated in two ways: E2F^{4X} changes 4 nucleotides to T, and ΔE2F removes 4 nucleotides from the motif. **F)** Expression levels of the two mutant promoters relative to the WT *E2F2* promoter. Both types of mutations to the E2F motif do not affect the normal expression level of the promoter. **G)** The E2F^{4X} mutant abolishes Rbf1 and Rbf2's ability to repress this promoter. Error bars indicate SEM, and * indicates $p<0.05$, ** is $p<.01$, *** $p<.001$.

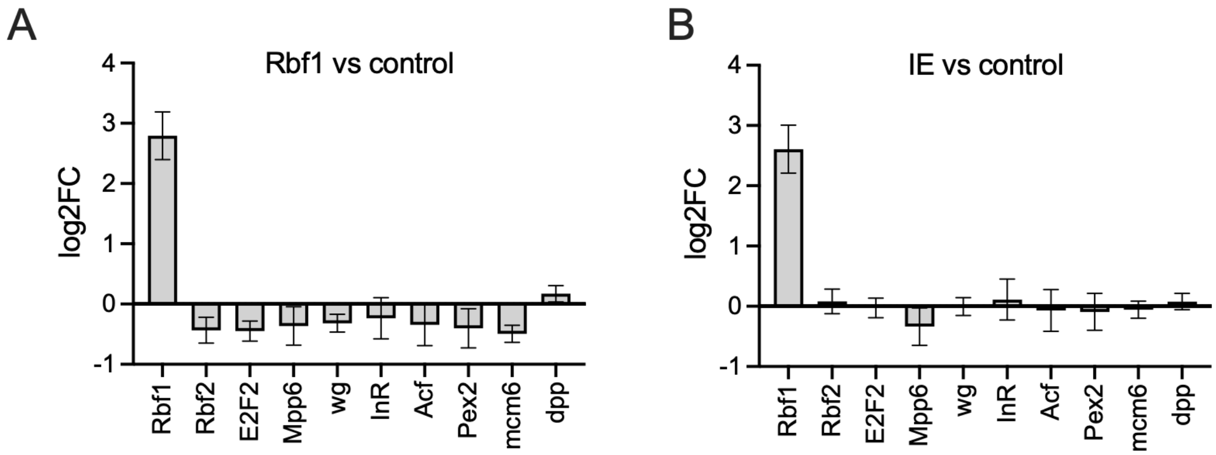


Figure S2.6. Effect of removing the IE on Rbf1's ability to repress gene expression in the wing disc. A) Overexpression of UAS:Rbf1 in the L3 wing disc using pen-Gal4 leads to repression of listed genes. All genes are mildly repressed, ~25% compared to the control (overexpression of UAS:GFP) **B)** Overexpression of UAS:Rbf1$^{\Delta IE}$ in the L3 wing disc using pen-Gal4 leads to little impact on the listed genes. These data are obtained from an RNA-seq experiment performed in L3 wing discs (See Materials & Methods in Elenbaas *et al*. 2015; Mouawad *et al*. unpublished). Rbf1 level is also shown, to illustrate that it is indeed overexpressed in this tissue.

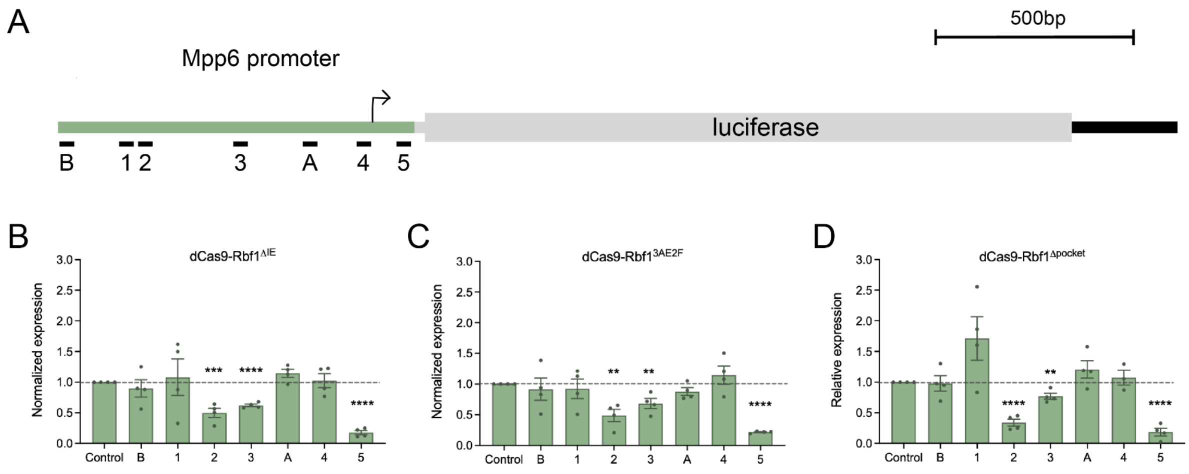


Figure S2.7. Testing Rbf1 mutants in a reporter assay in S2 cells. For all experiments described here, S2 cells were transfected with actin-GAL4, a luciferase reporter, one of the dCas9 effectors, and a single gRNA. **A)** Schematic of luciferase reporter that was designed to be regulated by the *Mpp6* promoter. **B)** dCas9-Rbf1$^{\Delta IE}$ has position-specific effects that are similar to dCas9-Rbf1. **C)** dCas9-Rbf1^{3AE2F} has position-specific effects that are similar to dCas9-Rbf1. **D)** dCas9-Rbf1$^{\Delta pocekt}$ has position-specific effects that are similar to dCas9-Rbf1. Error bars indicate SEM.

APPENDIX B: ADDITIONAL CRISPRI EXPERIMENTS

I. Generation of additional dCas9-Rbf1 chimeras

In addition to the constructs generated in Chapter 2, I also created dCas9 chimeras to three other Rbf1 mutants, which were not included in the manuscript (**Figure B.1**). Rbf1^{3SA} is a mutant in which three possible phosphorylatable serines are converted to alanine (Zhang *et al*. 2014). The Rbf1^{3SA} mutant has been shown to be a hypermorph, as it produces a more severe eye phenotype than WT Rbf1 when overexpressed there. I created a novel fly line expressing *nub*-GAL4>UAS:dCas9-Rbf1^{3SA} to test its impact on gene expression. However, this homozygous fly line had severe wing defects without a gRNA targeting any genomic locus. Thus, these flies were not used for further studies.

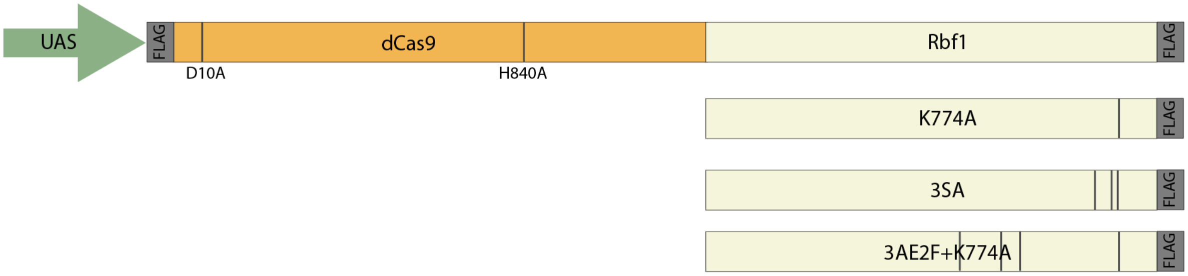


Figure B.1. Schematic of additional Rbf1 mutants that I designed and generated fly lines for. K774A is a single point mutation, 3SA is three point mutations, and 3AE2F + K774A is a combination of the previously described 3AE2F with a K774A point mutation.

The Rbf1^{K774A} mutant, in which a lysine in the CTD is converted to alanine, has also been shown to be a hypermorph (Acharya *et al*. 2010). Like Rbf1^{3SA}, the flies expressing *nub*-GAL4>UAS:dCas9-Rbf1^{K774A} had severe wing defects on their own, and they were not used for further studies. The Rbf1$^{3AE2F+K774A}$ mutant is a combination of the Rbf1^{3AE2F} mutant described in

Chapter 2 and Rbf1^{K774A} described here. The goal was to test the K774A hypermorph with the inability to bind to E2F, potentially not leading to the hypermorphic effects. I created a fly line expressing *nub*-GAL4>UAS:dCas9-Rbf1$^{3AE2F+K774A}$ which had normal wing development, unlike the other two hypermorphs described here. I crossed this fly line to the *E2F2* tandem gRNAs and performed phenotypic and transcriptional analysis of targeting this locus (**Figure B.2**). I found that this mutant has little to no effect on adult wing development and does not change the expression levels of *E2F2* and *Mpp6*. When assayed in cell culture, this mutant's protein levels were very low compared to those of the other Rbf1 constructs, suggesting possible protein instability, so this mutant was not used for further analysis (**Figure B.3A**).

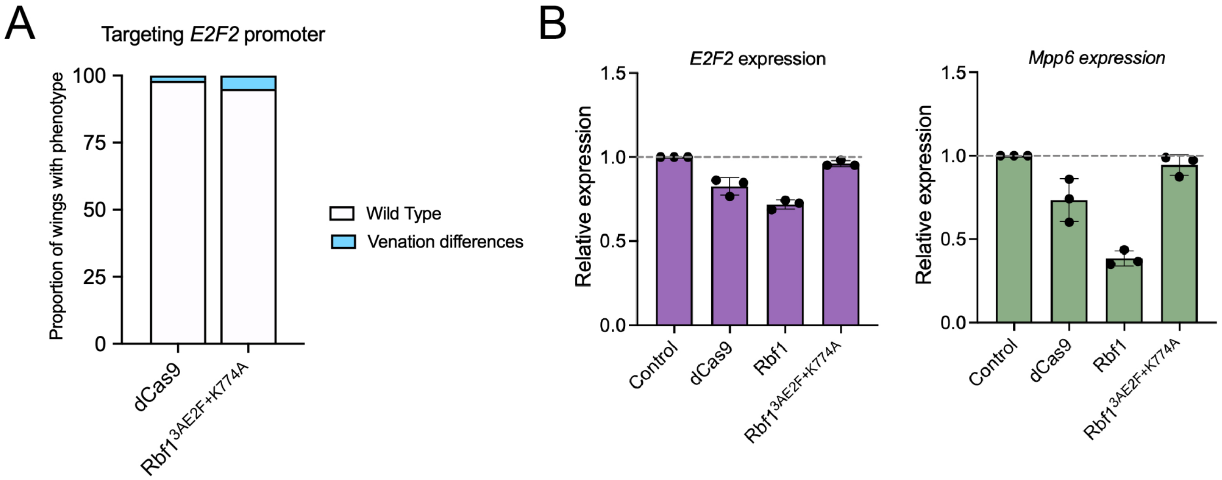


Figure B.2. Effect of the dCas9-Rbf1$^{3AE2F+K774A}$ mutant on wing phenotype and gene expression after targeting *E2F2/Mpp6*. A) Very minor wing effects are observed after targeting the *E2F2/Mpp6* shared promoter. **B)** Graphs indicate expression level of *E2F2* and *Mpp6* after targeting using gRNA 4 + 5. All data in these graphs has been presented in Chapter 2, aside from the data on the mutant, shown here to have no effect on expression level of either of these genes, in comparison to WT Rbf1 or dCas9 alone.

II. Design of gRNAs targeting other genes

In addition to generating gRNA constructs and flies for the *E2F2/Mpp6* locus as described in Chapter 2, I also designed novel gRNAs targeting the *PCNA*, *cycB*, and *InR* promoters (**Figure**

B.3). I created novel fly lines expressing each one of these gRNAs to target the dCas9 effectors to these genes' promoters. I chose *PCNA* because it is a promoter with Rbf1-specific repression, and I chose *cycB* because it is a promoter that's more sensitive to Rbf2. I expected to see paralog-specific effects on these promoters. Yet, when crossing *nub*-GAL4>UAS:dCas9-Rbf1, Rbf2, VPR flies to the gRNA expressing flies for *cycB* and *PCNA*, I observed no adult wing phenotypes.

I also designed new gRNAs targeting the *InR* locus, including internal enhancers which we have identified in Wei *et al*. 2016. When crossing *nub*-GAL4>UAS:dCas9-Rbf1, Rbf2, VPR flies to the gRNA expressing flies for *InR*, I observed no wing phenotypes here either. I did not follow up on these flies at the transcriptional level.

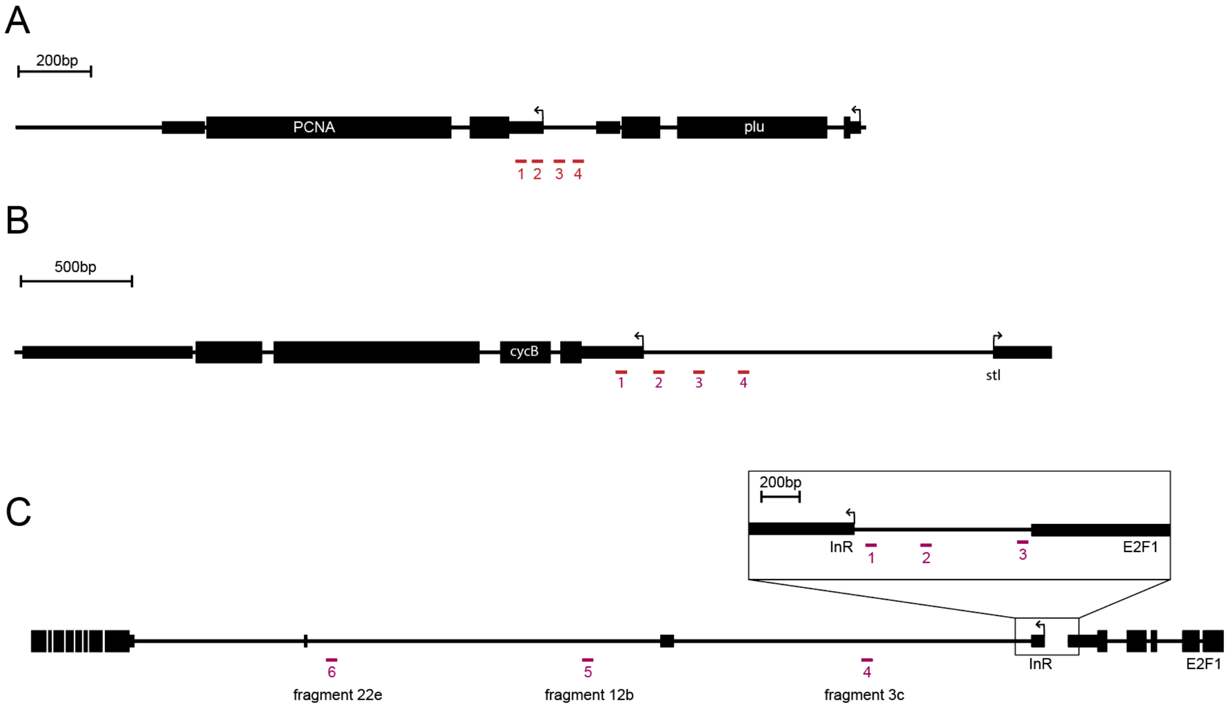


Figure B.3. Schematic of novel gRNAs targeting *PCNA* (A), *cycB* (B), and *InR* (C). Red bars indicate the position of each gRNA.

III. Targeting the *InR* promoter with Rbf1 mutants

I tested the effect of targeting Rbf1 IE, 3AE2F and pocket mutants on the *InR* promoter using the tandem gRNAs described in Chapter 2 (**Figure B.4)**. Here, I show that the IE mutant has few phenotypic effects, similar to what was seen for WT Rbf1. Interestingly, 3AE2F and pocket have almost identical effects. The presence of a phenotype due to their recruitment suggests that on this promoter, they have some impact on *InR* expression.

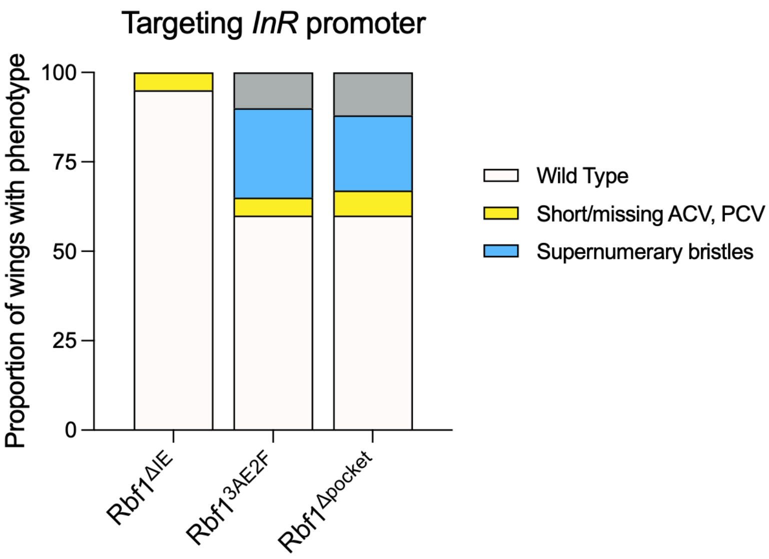


Figure B.4. Effect of targeting Rbf1 mutant proteins on the *InR* promoter using dCas9 fusions. Similar results were observed for the IE mutant as for WT Rbf1 (Chapter 2), while the other two mutants have similar results to Rbf2.

CHAPTER 3: SOFT REPRESSION: SUBTLE TRANSCRIPTIONAL REGULATION WITH GLOBAL IMPACT

This work was published in the following manuscript and adapted here:

Mitra A.*, Raicu, A.M.*, Hickey, S. L., Pile, L. A., & Arnosti, D. N. (2020). **Soft repression: Subtle transcriptional regulation with global impact.** BioEssays, e2000231. https://doi.org/10.1002/bies.202000231
*co-first authors

ABSTRACT

Pleiotropically acting eukaryotic corepressors such as retinoblastoma and SIN3 have been found to physically interact with many widely expressed “housekeeping” genes. Evidence suggests that their roles at these loci are not to provide binary on/off switches, as is observed at many highly cell-type specific genes, but rather to serve as governors, directly modulating expression within certain bounds, while not shutting down gene expression. This sort of regulation is challenging to study, as the differential expression levels can be small. We hypothesize that depending on context, corepressors mediate “soft repression,” attenuating expression in a less dramatic but physiologically appropriate manner. Emerging data indicate that such regulation is a pervasive characteristic of most eukaryotic systems, and may reflect the mechanistic differences between repressor action at promoter and enhancer locations. Soft repression may represent an essential component of the cybernetic systems underlying metabolic adaptations, enabling modest but critical adjustments on a continual basis.

INTRODUCTION

The first transcriptional regulation characterized in bacterial systems involves repressors described to function as on/off switches. Indeed, phage lambda repressor delivers tight repression to maintain lysogeny, while the LacI repressor can silence an otherwise highly transcribed operon, depending on nutritional status. Interestingly, subsequent studies have shown that *lacZ* expression can be delicately tuned over a wide range, depending on graded input from activators and repressors (Setty *et al*. 2003). In eukaryotes, transcriptional repression reflects input from a wider set of regulators: inherent chromatin barriers, histone modifications that facilitate heterochromatin formation, and combined action of DNA-binding transcription factors and corepressors that they recruit, including the evolutionarily conserved retinoblastoma (Rb) and SIN3 family proteins.

The action of transcriptional repressors and corepressors has played a central role in many studies of developmental biology, where such proteins are essential mediators of tissue-specific gene expression, as well as controllers of cell cycle, and circadian regulation (Barolo and Posakony, 2002; Bertoli *et al.* 2013; Cox *et al.* 2019). Inducible gene expression required for physiological response to environmental fluctuations also involves the deactivation of repression complexes, for instance in the upregulation of stress-responsive genes (Howe *et al.* 2018). In many systems, the effectiveness of repression is essentially complete, and depends on reaching a critical concentration of relevant transcription factors, or intensity of signaling pathways that permit the assembly (or disassembly) of repression complexes. To achieve a cell-type specific response, target genes are repressed below some threshold that ensures establishment and/or maintenance of a specific cell state. For instance, Blimp-1/PRDM1 is a tissue-specific repressor whose key role is in driving plasma cell differentiation and silencing genes involved in immune response (Shapiro-Shelef *et al.* 2003; Ulmert *et al.* 2020). Its loss in maternal uterine tissue has been shown to upregulate hundreds of genes that are normally silenced (Goolam *et al.* 2020). Likewise, the RE-1 silencing transcription factor (REST) is a regulator of cell differentiation. REST is ubiquitously expressed in nonneuronal cells for the silencing of neuronal genes, while mostly absent from differentiated neurons (Ballas *et al.* 2005). Its loss in quiescent neural progenitors leads to neural differentiation, suggesting its role is to prevent neural differentiation through gene silencing (Mukherjee *et al.* 2016). In contrast to this choice between silencing or activity, molecular genetic studies have identified numerous genomic targets of repression complexes that may be less dramatically impacted by the presence of such regulatory factors. Metazoan transcription factors and their corepressors are typically found to physically interact with thousands of genes, yet perturbation experiments frequently show only a small subset with significantly altered expression.

This disparity is usually ascribed to some degree of "off target" interactions, whereby these complexes do not have a significant function at some loci (Kok and Arnosti, 2015). An additional possibility is that there are context-dependent interactions, in which the binding to some genes may be essential for regulation only in certain cell types, or under specific conditions that may not have been assessed in a particular experiment.

A nonexclusive, alternative explanation to the presence of certain physical repressor complex interactions is that the type of repression that is biologically significant is of a form that is inherently "soft," that is, altering expression, but not in an absolute on/off fashion. Such regulation may be especially important for widely-expressed "housekeeping" genes, where expression rarely, if ever, is silenced. As we have argued with respect to Rb corepressors, binding and coordinate regulation of ribosomal protein genes may represent just such a case (Wei *et al*. 2015). A second example is that of regulation of genes in methionine catabolism by the SIN3 cofactor, where perturbation to SIN3 levels induces approximately two-fold changes in relevant pathway genes (Liu and Pile, 2017). Significant for the analysis of such datasets, the amount of repression may be subtle, and in some cases, less than the extent typically required to clearly differentiate signal from noise.

Here, we discuss studies of Rb and SIN3, two essential and conserved corepressor protein families, providing a picture of the diverse targets with which these transcriptional regulators are physically and functionally associated. We propose that soft repression is a major contributor to gene regulatory control and plays a key role in metabolic adaptation. The unique soft transcriptional responses of some genes to corepressor regulation may result from the complexity of signaling at the respective promoters, or from the different effects of repressor complexes acting at promoters versus enhancers. Importantly, we suggest that the action of these corepressors may represent a wider, unappreciated phenomenon impacting a great number of eukaryotic regulatory factors and pathways. Further investigation will uncover the significance of this second, less dramatic form of transcriptional regulation.

The hypothesis formalized: Canonical models for the action of transcriptional repressor proteins often emphasize the possibilities for tight control through on/off action, enabling exquisite tissue-specificity and physiological control. A number of genomic studies, however, have increasingly pointed to a pleiotropic "soft repression" mechanism of action on widely expressed genes, whose modulation may be subtle. Using two well-studied corepressor families, Retinoblastoma and SIN3, we hypothesize that some promoter proximal corepressors function to modestly attenuate gene expression in a biologically meaningful way–a mechanism that may be especially prominent on genes featuring multiple regulatory inputs.

Testing the hypothesis: To better understand soft repression, we propose the application of diverse technologies: (1) using high throughput RNA-sequencing methods with deeper sequencing and greater number of biological replicates to increase resolution and discern the difference between noise and soft repression; (2) single cell transcriptomic studies, which will circumvent discrepancies that might arise from heterogeneous cell populations; (3) nuclear run-on based technologies such as GRO-seq, to allow for the assessment of immediate transcriptional impacts of corepressor perturbation; (4) targeting the corepressor directly to single gene promoters to perturb a specific circuit and avoid pleiotropic effects from global manipulations of the repressor; (5) a thorough computational consideration of soft-repression in interpreting population- and species-level cis-regulatory variation. We urge gene expression researchers to consider soft repression as a significant and biologically relevant form of transcriptional regulation in future studies and test the hypothesis using these proposed methods.

The retinoblastoma tumor suppressor protein mediates both hard and soft repression

The retinoblastoma tumor suppressor protein (Rb) is a conserved transcriptional corepressor present in most eukaryotes including plants, animals, and microbes. The study of the RB1 gene stems from research by Knudson, who linked mutations in the gene to retinoblastoma, an eye cancer presenting in early childhood (Knudson, 1971). Since then, its role in cancer, development, and normal physiology has been extensively studied in a variety of systems. Most eukaryotes express a single Rb protein, but the gene has been duplicated in select lineages including in vertebrates and Drosophila. In humans, the Rb family comprises Rb, p107, and p130, which exhibit partially overlapping as well as non-redundant functions in development and cancer (Dick and Rubin, 2013). Similarly in Drosophila, paralogs Rbf1 and Rbf2 represent an ancient duplication event, where Rbf1 appears to have more roles in cell cycle regulation, while Rbf2 may interact with and regulate an extensive set of genes linked to growth control and metabolism, including ribosomal protein genes (Mouawad *et al*. 2019). Below, we summarize basic properties of these proteins with a focus on work from Drosophila; vertebrate paralogs of Rb have been similarly examined in countless studies in the context of development and disease.

Rb proteins regulate genes by binding to E2F family transcription factors found on promoters. E2F factors have a canonical role in the regulation of cell cycle genes that are transiently induced during the cell cycle. In Drosophila G1, Rb binds to the E2F-DP heterodimer and blocks E2F from activating expression of downstream genes such as *cycA*, *cdc2*, and *DNApolα*, which are required for S phase entry (Duronio *et al*. 1995). Rb mediated repression is relieved later in G1 as Rb is inhibited via phosphorylation, and the cell enters S phase. Similar regulation appears to apply to promoters active later in the cell cycle, such as *cycB*. This cell-cycle regulatory role by Rb proteins is highly conserved in eukaryotes (Cao *et al*. 2010). Initial characterization of Rb function derived from its cancer-associated phenotype, and led to cell cycle regulation as a central area of study. However, genome-level studies soon revealed a plethora of other regulatory roles.

Transcriptomics studies have uncovered diverse classes of genes that are differentially expressed after Rb loss or overexpression. Pioneering studies using cultured Drosophila S2 cells showed that *Rbf1* knockdown affected canonical cell cycle, DNA replication, and DNA repair genes, but also a host of non-cell cycle-related genes (Dimova *et al.* 2003). Interestingly, specific promoters tested in transfection assays were differentially sensitive to the Rb paralogs – most, but not all, cell cycle genes being more sensitive to repression by Rbf1 (Mouawad *et al.* 2019; Stevaux *et al.* 2005; Mouawad *et al.* 2020). In contrast, Rbf2 has preferential action on certain ribosomal protein promoters, with which it is prominently associated *in vivo*, although the extent of regulation is much more modest than that seen for cell cycle genes (Wei *et al.* 2015; Mouawad *et al.* 2019). Knockdown studies in human fibroblasts similarly illustrate that loss of each human Rb family member misregulates diverse classes of genes (Chicas *et al.* 2010). *RB1* knockdown led to upregulation of DNA replication, DNA metabolism, and cell cycle genes; *p107* knockdown led to downregulation of genes involved in oxidative phosphorylation, electron transport, and NADH dehydrogenase activity; genes involved with organelles were upregulated after *p130* knockdown. Although the regulation was not shown to be direct in all cases, these data indicate unique roles of Rb proteins related to cell metabolic processes. More recently, Rb loss was implicated in reprogramming of glucose tolerance, oxidative metabolism, glutathione synthesis, glutamine catabolism, and nucleotide metabolism in Drosophila (Nicolay and Dyson, 2013). We, and others, have found that the extent to which these target genes are repressed by Rb varies. From cell culture assays performed in our laboratory, we suggest that significant and potent decreases in expression of certain genes such *PCNA* represent hard repression, in which the gene is turned off for a period of time, while more moderate decreases observed, as is the case with *InR*, represent what we term soft repression (**Figure 3.1**).

Overall, comparison of genes directly bound by Rb family proteins in diverse organisms suggests that at least a portion of targeting is widely conserved, extending beyond the canonical cell cycle category. For instance, the human p130 protein is especially enriched on mitochondrial and cytoplasmic ribosomal protein promoters, similar to the pattern for Rbf1/2 in Drosophila, suggesting that this class of genes may represent a common target (Wei *et al*. 2015; Chicas *et al*. 2010). While Rb family members are typically found proximal to the transcriptional start site, human Rb proteins have also been found to localize to gene enhancers, and DNA repeat elements like LINEs and SINEs, where they recruit cofactors to change histone marks (Kareta *et al*. 2015; Ishak *et al*. 2016). Still, Rb proteins are preferentially localized to promoter proximal regions, perhaps due to reliance on recruitment by E2F transcription factors, which are also generally located near the transcriptional start site (Wei *et al*. 2015). Overall, emerging research indicates that retinoblastoma proteins exhibit highly conserved functional roles, which may include both hard and soft repression on distinct targets.

The sin3 corepressor regulates genes involved in essential cellular processes

SIN3 is a broadly acting transcriptional regulator conserved from yeast to humans. Pioneering genetic studies identified this factor as a key regulator of mating type switching in *Saccharomyces cerevisiae* (Nasmyth *et al*. 1987; Sternberg *et al*. 1987). The SIN3 protein itself does not possess enzymatic activity but rather acts as a scaffold that interacts with other factors including histone deacetylases and histone demethylases (Nakayama and Hayakawa, 2011). Some eukaryotes, such as budding yeast and *C. elegans*, possess a single SIN3 gene, as does Drosophila

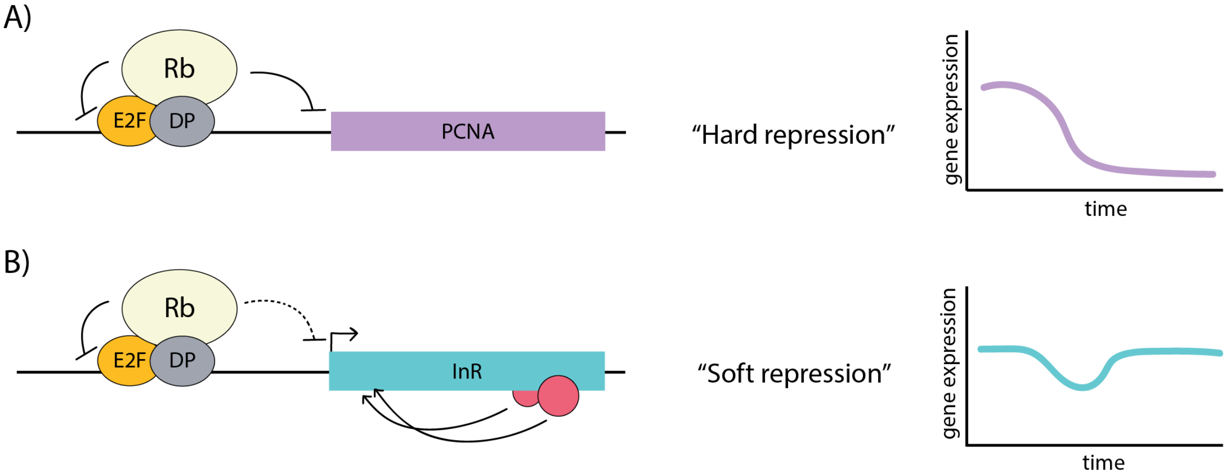


Figure 3.1. Comparison of hard versus soft repression. A) Rb can function as a potent repressor on certain genes such as the cell cycle-related *PCNA* by blocking the E2F transactivation domain and inducing a repressed chromatin state. **B)** In contrast, on other genes such as *InR*, Rb functions in concert with other factors that may have to balance each other's activities, leading to more moderate repression of the gene.

(*Sin3A*), where alternative splicing produces multiple protein isoforms. Diversity is achieved in *Schizosaccharomyces pombe* and vertebrate species through multiple SIN3 paralogs, including the mammalian *SIN3A* and *SIN3B* genes. Mutations and altered expression of these mammalian paralogs are found in diseases such as breast cancer, pancreatic cancer, and Witteveen-Kolk syndrome, a neurodegenerative disorder (Rielland *et al*. 2014; Leiws *et al*. 2016; Narumi-Kishimoto *et al*. 2019).

Early genome-wide transcriptomic analyses from multiple species revealed that SIN3 acts both as a corepressor and as a coactivator of transcription (Bernstein *et al*. 2000; Pile *et al*. 2003; Dannenberg *et al*. 2005). SIN3 was found to regulate genes involved in many different cellular processes, including cell cycle, metabolism, DNA replication, and stress response. Regulatory roles of the factor are evolutionarily conserved; for instance, SIN3 regulates cell cycle progression in organisms from yeast to mammals. In Drosophila S2 cells, reduction of *Sin3A* results in a change in expression of a number of cell cycle regulators, including a decrease in the level of *string*, required for the G2 to M transition, which halts the progression of cell cycle (Pile et al. 2003; Swaminathan and Pile, 2010). This same step in the cell cycle is affected by knockout of *mSin3a* in mouse embryonic fibroblasts (Dannenberg *et al.* 2005; Cowley *et al.* 2005). The SIN3B paralog is not required for cell proliferation but rather interacts with the DREAM complex to repress essential genes and enable maintenance of quiescence (David *et al.* 2008; Bainor *et al.* 2018).

Studies from Drosophila show that genes of housekeeping pathways are also bound and regulated by Sin3A. Sin3A influences mitochondrial function by regulating the expression of multiple nuclear encoded mitochondrial genes that encode electron transport chain subunits (Pile *et al.* 2003; Gajan *et al.* 2016). By regulating expression of mitochondrial genes, as well as genes and metabolites of the glutathione pathway, Sin3A influences overall fitness and response to oxidative stress (Pile *et al.* 2003; Barnes *et al.* 2014; Liu *et al.* 2020). Expression of enzymes involved in energy production are also regulated by the Sin3A complex (Pile *et al.* 2003; Gajan *et al.* 2016). *Sin3A* knockdown in Drosophila S2 cells affects the gene expression and metabolite levels of several glycolytic and TCA cycle intermediates (Pile *et al.* 2003; Liu *et al.* 2020). The reduction of *Sin3A* also affects expression of genes encoding enzymes of methionine metabolism and leads to a decrease in levels of the methyl donor S-adenosylmethionine (SAM) (Liu *et al.* 2017). These findings indicate that an important function of the Sin3A complex is to regulate expression of genes encoding enzymes of metabolic pathways to maintain cellular homeostasis.

SIN3 also regulates response to stress in both fly and mammalian models (Barnes *et al.* 2014; Kadamb *et al.* 2015). The knockdown of *Sin3A i*n Drosophila leads to reduction in expression of genes encoding proteins required for glutathione synthesis as well as increased susceptibility to oxidative stress, a sensitivity that is rescued by glutathione supplementation (Barnes *et al.* 2014). A study in mammalian cancer cell lines showed that SIN3B is important in the stress response to treatment with different DNA-damaging agents (Kadamb *et al.* 2015). Following treatment, there is an increase in expression of SIN3B at the transcript and protein level, which is p53-dependent. Additionally, when *Sin3B* is knocked down during damage, p53 target stress response genes are affected, linking SIN3B to the p53-mediated response to DNA damage.

As noted for Rb family proteins, ChIP studies from worms, flies, and mice show that SIN3 is generally localized to promoter proximal sequences of target genes, and not at distal enhancers (Williams *et al.* 2011; Saha *et al.* 2016; Beurton *et al.* 2019). Consistent with this pattern, SIN3 and other components of the complex immunoprecipitate with H3K4me3, a promoter-associated mark (Engelen *et al.* 2015). In addition to recruitment to the promoter, SIN3 has been reported, in mouse and yeast cells, to localize to gene bodies of some active genes through association with a complex of proteins distinct from those found at promoters. The Rpd3S complex in yeast, which contains Sin3, the Rpd3 HDAC, and two additional factors, is recruited through interactions of complex subunits with histone H3K36me3 (Keogh *et al.* 2005; Carrozza *et al.* 2005). The predominant role of Rpd3S is to facilitate deacetylation of nucleosomes after the passage of RNA polymerase II, suppressing cryptic initiation along the gene body (Keogh *et al.* 2005; Carrozza *et al.* 2005). In the mouse, the Sin3B isoform, along with HDAC1 and homologs of the other two yeast Rpd3S complex members, is found enriched at sites downstream of the promoter of select housekeeping genes (Jelinic *et al.* 2011). Knockdown of *Sin3B* leads to an increase in RNA polymerase II levels in the gene body and approximately two-fold activation of *GAPDH* and *RPL13α* expression. The mechanism of action of SIN3 in gene regulation along gene bodies is likely distinct from that of SIN3 localized to the promoter proximal region. As a scaffold, SIN3 interacts with a number of protein partners. Localization is likely to reflect recruitment by different sequence-specific DNA binding transcription factors and through recognition of specific combinations of histone modifications by chromatin binding domains within proteins of the SIN3 complex (Chaubal and Pile, 2018). SIN3 and associated proteins of the complex bind a diverse set of targets to modulate the expression of genes to impact multiple biological processes. In summary, targets of SIN3 regulation represent a wide variety of cellular functions. One report noted the silencing of cell cycle related genes by Sin3 acting with E2F4 and RBP2 in differentiated mouse myoblasts; however, the regulatory effects of SIN3 have generally not been reported to include outright silencing of most target genes, pointing to a modulating role rather than an outright on/off switch (van Oevelen *et al*. 2008).

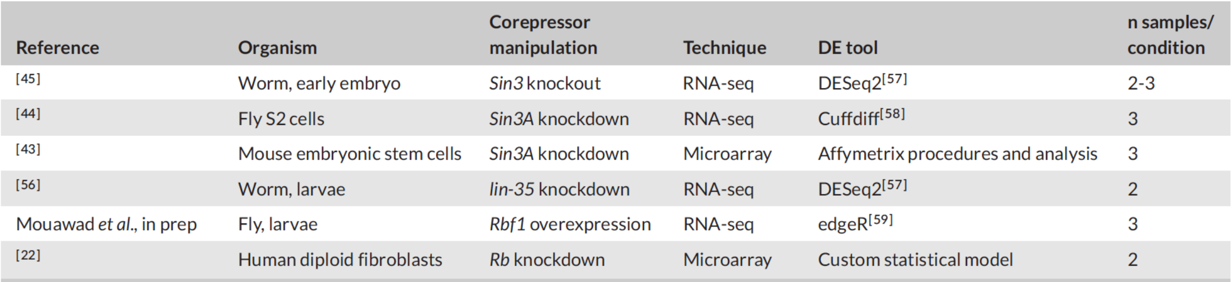
Global regulation through soft repression

From a global perspective, Rb and SIN3 share several characteristics, including their preferential association with promoter proximal sequences, their widespread expression, and the diversity of physically and functionally targeted genes. These corepressors differ in that Rb family proteins are regulated by conserved cyclin kinases that affect Rb-E2F association, thus dynamically modulating repression activity, whereas a similar control of SIN3 proteins has not been observed. However, a variety of post-translational modifications do affect SIN3 proteins, and may exert similar regulation (Hasan *et al*. 2015). In addition, expression of distinct SIN3 isoforms may adjust SIN3 activity over longer developmental times.

Another common characteristic of these two corepressors, which has been less appreciated, comes from examination of target gene responsiveness to perturbation of Rb or SIN3 proteins. From examination of transcriptomic data, it is apparent that a large portion of their regulons consist of widely active genes, which are subject to fine-tuned regulation. To draw a distinction with a typical silencing action commonly associated with repressors, we call this regulatory activity soft repression, and describe it as the action of repressors/corepressors to modulate or fine-tune gene expression without effectively silencing the promoter. This regulatory action differs from the usual understanding of transcriptional repressors that function as an on/off switch, which for Rb has been demonstrated on cell cycle genes (Chinnam and Goodrich, 2011). As discussed below, a closer analysis of these two highly conserved corepressors, Rb and SIN3, indicates that soft repression may be a common, yet underappreciated activity of transcriptional regulators in general.

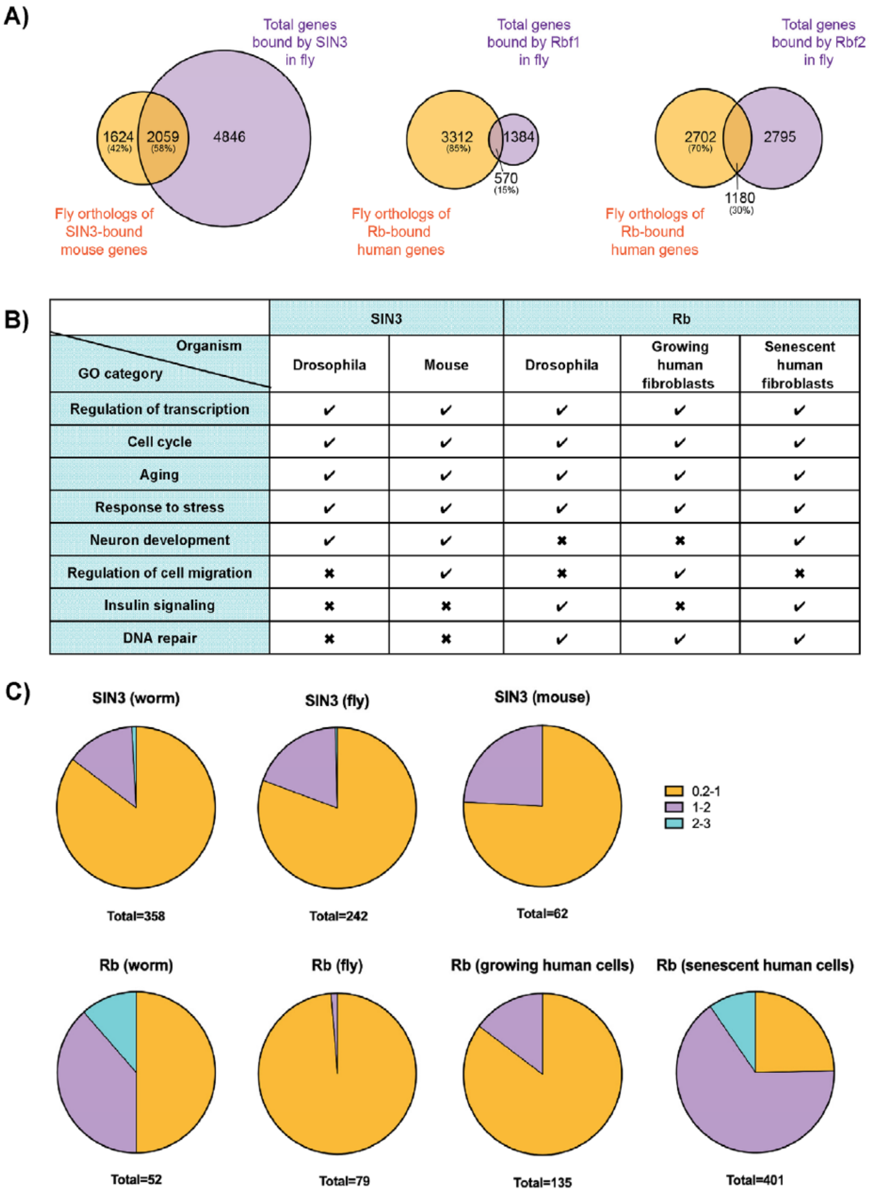
We performed a comparative examination of published ChIP-seq and transcriptomic data for Rb and SIN3, which supports this new classification of repression mechanisms. We first considered to what extent global regulatory roles of each of these corepressors may be evolutionarily conserved, and compared promoter occupancy of orthologous genes in Drosophila and mammals. Over 50% of all of the genes bound by SIN3A in the mouse that possess a fly ortholog were similarly bound by Sin3A in the fly (**Figure 3.2A**). A smaller, but still substantial, fraction of genes bound in human cells by Rb with an ortholog in the fly are bound by Rbf1 and/or Rbf2. The overlap in directly bound targets indicates that both Rb and SIN3 may have conserved roles, although binding does not always predict function. Therefore, to consider functional effects, we combined the ChIP-seq data with available microarray or RNA-seq data from worms, flies, and the mammalian systems (**Table 3.1**). We identified functional classes of genes significantly enriched in these lists of direct, repressed targets (**Figure 3.2B**). We found that knockdown of *Sin3A* in Drosophila and mouse embryonic stem cells leads to misregulation of genes involved in multiple overlapping processes such as transcriptional regulation, cell cycle, and aging (Williams *et al*. 2011; Saha *et al*. 2016). Similarly, in the fly and human cells, regulatory effects of retinoblastoma proteins involved common classes of genes, including transcriptional regulation, cell cycle, and aging, and also included classes not seen to be regulated by SIN3 such as insulin signaling and DNA repair (Chicas *et al*. 2010; Korenjak *et al*. 2012; Mouawad *et al*. In prep.).

Interestingly, most of the direct, repressed genes showed modest but significant changes in expression after perturbation of Rb or SIN3 (**Figure 3.2C**). For example, for both Rbf1 and Sin3A in Drosophila, over 80% of affected genes fell within the lowest category of less than or equal to a two-fold change in expression level (log2-fold change, 0.2-1). The knockdown of the Rb worm homolog, lin-35, caused a larger range in gene expression changes, 50% having greater than log2-fold change of 1. The prevalence of modestly impacted genes suggests that the effects have a biological basis, and that soft repression is observable in these transcriptomic measurements. We note that the bias towards small fold changes may also be the result of incomplete removal of Rb and SIN3 in perturbation experiments, or indirect genetic effects (although we only consider direct ChIP-seq target genes in this analysis). Overall, however, the observed levels of modulation are far from complete silencing of transcription, but do fall well within the levels that are associated with significant biological effects, such as the twofold changes associated with haploinsufficiency or changes in dosage compensation, both of which can have major consequences. Together, the frequency of soft repression-like effects, along with evidence of evolutionary conservation of physical and functional targeting, suggests that this type of regulation constitutes an important role for both SIN3 and Rb proteins.



Reference	Organism	Corepressor manipulation	Technique	DE tool	n samples/ condition
[45]	Worm, early embryo	*Sin3* knockout	RNA-seq	DESeq2[57]	2-3
[44]	Fly S2 cells	*Sin3A* knockdown	RNA-seq	Cuffdiff[58]	3
[43]	Mouse embryonic stem cells	*Sin3A* knockdown	Microarray	Affymetrix procedures and analysis	3
[56]	Worm, larvae	*lin-35* knockdown	RNA-seq	DESeq2[57]	2
Mouawad *et al.*, in prep	Fly, larvae	*Rbf1* overexpression	RNA-seq	edgeR[59]	3
[22]	Human diploid fibroblasts	*Rb* knockdown	Microarray	Custom statistical model	2

Table 3.1. Methods used to analyze extent of differential gene expression after genetic manipulations of Rb and SIN3. Reference 22: Chicas *et al*. 2010; 43: Williams *et al*. 2011; 44: Saha *et al*. 2016; 45: Beurton *et al*. 2019; 56: Latorre *et al*. 2015.



	SIN3		Rb		
Organism GO category	Drosophila	Mouse	Drosophila	Growing human fibroblasts	Senescent human fibroblasts
Regulation of transcription	✓	✓	✓	✓	✓
Cell cycle	✓	✓	✓	✓	✓
Aging	✓	✓	✓	✓	✓
Response to stress	✓	✓	✓	✓	✓
Neuron development	✓	✓	✖	✖	✓
Regulation of cell migration	✖	✓	✖	✓	✖
Insulin signaling	✖	✖	✓	✖	✓
DNA repair	✖	✖	✓	✓	✓

Figure 3.2. Conserved, direct targets of SIN3 and Rb exhibit soft, but significant repression. A) SIN3 and Rb bind to a substantial number of the same genes in both the fly and mammalian systems, which indicates conservation of genome-wide binding by these corepressors. To determine this, we used the BioMart datamining tool and analyzed the intersection of fly and mammalian ChIP-seq datasets (Smedley *et al*. 2015). Mouse genes for which orthologs can be identified in the fly were overlapped with fly genes bound by SIN3. Similarly, human genes for which orthologs can be identified in the fly were overlapped with fly genes bound by Rb. **B)** Chart indicates GO categories misregulated after overexpression or knockdown of SIN3 or Rb in fly and mammalian systems. **C)** Pie charts indicate the log2-fold change of direct, repressed genes after Rb or SIN3 manipulations in worm, fly, and mammalian models. Totals listed are the number of genes misregulated for each organism and corepressor. Data obtained from Chicas *et al*. 2010, Williams *et al*. 2011, Saha *et al*. 2016, Beurton *et al*. 2019, Korenjak *et al*. 2012, Latorre *et al*. 2015, Mouawad *et al*. In prep.

Considered on a genome-wide scale, the regulatory action of Rb and SIN3 appears to be largely dedicated to fine-tuning gene activity, although in the case of Rb family proteins, the cell cycle effects have a disproportionate impact on described phenotypes. What evidence would support the important, even predominant, role for soft repression for SIN3 and Rb proteins? First and foremost, SIN3, which apparently is largely restricted to soft repression, is essential in many organisms (Dannenberg *et al*. 2005; Cowley *et al*. 2005; David *et al*. 2008; Neufeld *et al*. 1998; Pennetta *et al*. 1998). For Rb, the case is more complex, because standard genetic approaches do not differentiate the impacts of misregulated gene expression involving hard and soft effects. The most compelling evidence comes from Drosophila, where gene duplication and subfunctionalization has apparently partly divided the cell cycle and soft regulatory tasks between Rbf1 and Rbf2. Rbf2—which has many targets within the soft targeting category—is dispensable for development but never lost over longer evolutionary time, likely due to pleiotropic effects (Mouawad *et al*. 2019; Stevaux *et al*. 2005). Secondly, regulation of a variety of pathways that have been examined in detail show the significant phenotypic consequences of these less than all-or-nothing effects. For Drosophila Sin3A, RNAi knockdown upregulates multiple methionine catabolism genes by approximately two-fold (Liu *et al*. 2017). This difference in expression is associated with a change in the level of the key methionine donor SAM, and an increase in H3K4me3, a histone modification linked to gene expression. In the mouse, conditional knockout of Sin3a in forebrain excitatory neurons results in a small yet reproducible (20–25%) increase in expression of *Homer1* and cyclin-dependent kinase *Cdk5*, two genes encoding factors involved in memory consolidation, associated with an increase in hippocampal activity (Bridi *et al*. 2020).

In the case of Rb regulation, one of the most attractive sets of genes for further investigation of soft repression are the over 100 ribosomal protein promoters that are extensively targeted by Rb family members. Ribosomal protein promoters are widely active, thus considered housekeeping in nature, yet respond sensitively to changes in nutrient availability, as well as signals in development and disease (Gorenstein and Warner, 1976; Powers and Walter, 1999; Guimaraes and Zavolan, 2016). The two-fold or less modulation of expression of these genes exerts pleiotropic effects on cellular growth, as evidenced by the minute phenotypes caused by haploinsufficiency (Marygold *et al.* 2007). While the contribution of Rb proteins to overall expression of these genes still needs to be established, a soft repression level of control is likely to impact cellular growth in a significant way. We propose that this collection of small but reproducible changes contributes to Rb and SIN3 acting as essential global transcriptional regulators that modulate gene expression, rather than fully repress their target genes, to produce measurable biological outcomes.

The where and how of soft repression

Based on gene ontology analysis, targets of soft repression may disproportionately represent housekeeping genes, although this term can be misleading, as it does not mean that the genes are constitutively active in an unregulated manner. To characterize the types of expression patterns observed for Rb and SIN3 target genes, we examined extant gene expression data using modENCODE data accessed through FlyBase. Focusing on conserved target genes from mammals and flies, we found the majority of these genes to be expressed at all developmental stages, but not in all tissue types. This suggests that these corepressors regulate genes that are widely expressed throughout development, and they potentially modulate their expression in particular contexts. A unifying characteristic of these genes is that despite the presence of the corepressor, they continue to be expressed. To determine whether these target genes are considered stably expressed, we used the tau metric as a computation of gene expression variability (Yanai *et al.* 2005; Kryuchkova-Mostacci and Robinson-Rechavi, 2017). A tau value of 0 is given to a gene that is expressed at the same level across the developmental time points and tissue types assayed. A tau value of 1 indicates the gene's expression is specific to one stage or one tissue type. Using a cutoff of 0.25 as the definition for a "stable gene," we found that over 50% of conserved Rb or SIN3 targets are considered stable genes throughout development, while only about 25% are stable throughout the tissue types assayed. A hypergeometric test indicates that the stable genes are significantly over-represented within the bound gene sets for both Rb and SIN3. Perhaps a typical promoter that is regulated by soft repression is one that is widely expressed and stable throughout developmental stages.

How is it that Rb and SIN3 complexes can be involved in soft repression, when the general types of associated factors – transcription factors and chromatin modifiers such as HDACs – are also involved in more dramatic on/off transcriptional regulation? Our model posits that genes that are fully repressed, perhaps through constitutive and facultative heterochromatin or subject to dominant regulation of distal enhancers, are not subject to control by soft repression (**Figure 3.3A, B**). In contrast, the large majority of Rb and SIN3 binding occurs near the transcription start site, rather than at distal enhancers. We speculate that soft repression activity may be exerted strictly from promoter proximal positions, and that specific properties of the promoter region can predispose the regulation to be partial, rather than all-or-nothing (**Figure 3.3C, D**). For instance, localization of a repression complex at a promoter may be better suited for partial interference with transcriptional initiation or release, if the biochemical mechanisms invoked (e.g., deacetylation of histone tails, an activity associated with both Rb and SIN3 associated factors) modestly impact nucleosome loading or density. A somewhat inhibited rate of binding or release of RNA polymerase II may then result, possibly changing kinetic constants without blocking essential steps in promoter firing. By contrast, a hard repression effect, such as seen on the *PCNA* promoter, may

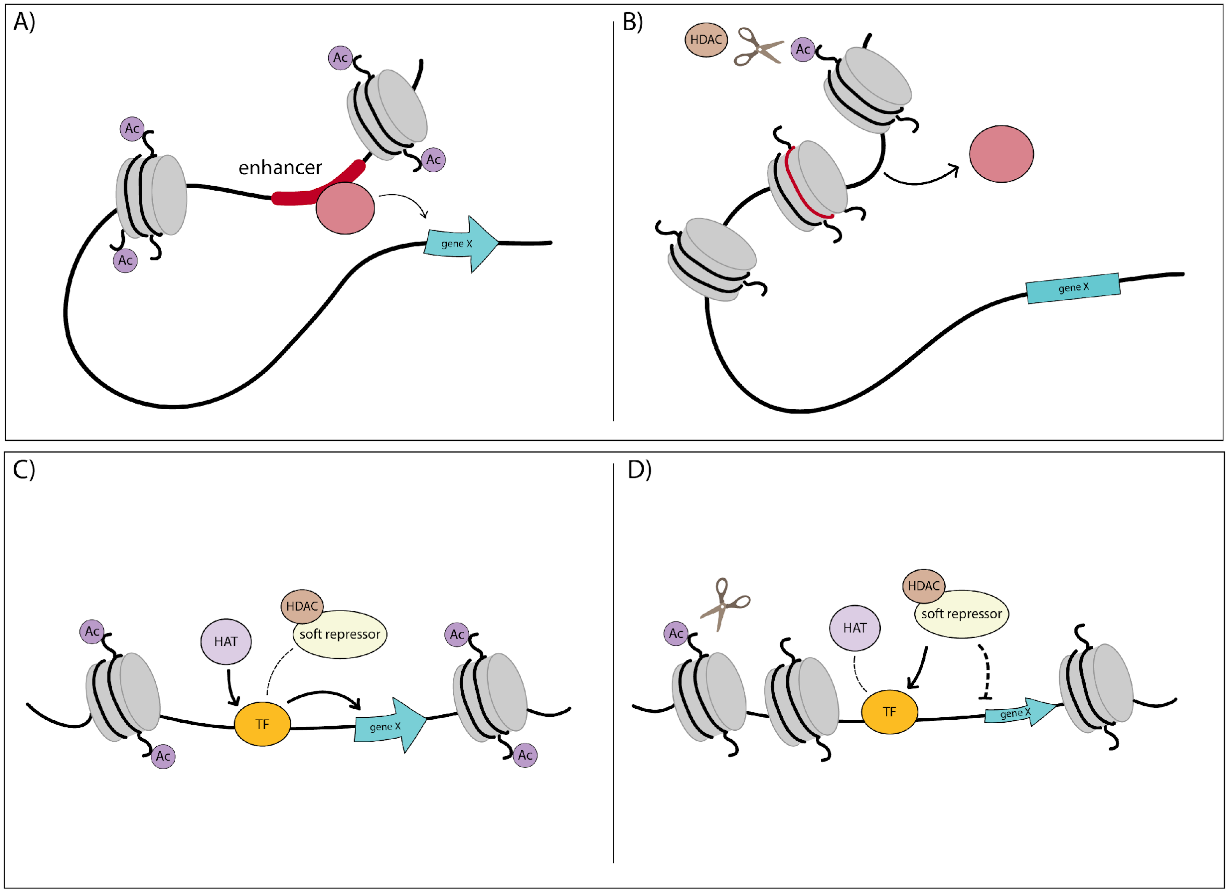


Figure 3.3. Contrasting models for transcriptional repression: enhancer based hard repression (A, B) and promoter based soft repression (C, D). A) At enhancers, activators and repressors work in a binary fashion, turning gene expression on and off in response to availability of binding sites and interaction with distal gene promoters. When an activator binds an available enhancer, it can turn on gene expression through chromatin looping. **B)** If an enhancer is occluded through nucleosome remodeling and chromatin compaction, activator access is inhibited and the gene is turned off. **C)** At promoters, soft repressors can fine-tune expression from the proximity to the transcriptional start site, sometimes through interaction with transcription factors (TF) bound to DNA. There is an interplay between activators and repressors, which compete for DNA recruitment to impact the chromatin environment and modulate gene expression. On the left, a promoter proximal TF interacts with cofactors such as histone acetyltransferases (HAT) to turn on expression of gene X, while on the right, **D)** the soft repressor complex, which many times includes a histone deacetylase (HDAC) in the case of Rb and SIN3, is more potent than the activator. The complex deacetylates nearby histone tails and dials down expression of gene X, but does not completely turn it off.

result from inhibition of the E2F transcriptional activation domain, blocking key interactions with the Mediator or TFIID. It is possible that deployment of similar chromatin modifiers to distal regions may interfere with transcription factor binding and enhancer-promoter looping, resulting in an all-or-nothing effect (**Figure 3.3A, B**). Notably, the same biochemical pathways may be involved in hard or soft regulation, and the architecture of the regulatory region would be decisive in dictating the outcome. An alternative and non-exclusive possibility is that soft repression reflects an inherent balance of competitive interactions with the basal machinery, and promoters that feature multiple inputs from other regulatory factors and elements would not be prone to complete silencing (**Figure 3.1B**). In summary, it is likely that the context of the regulatory regions in which Rb and SIN3 corepressors operate have a decisive impact on the ability of these factors to play modulating, rather than on/off roles. Further biochemical and molecular biological investigations will be required to elucidate the mechanisms in play. Such studies will likely yield important insights into the dynamic regulation of many constitutively active genes key to metabolism and disease.

Soft repression—a general property of transcriptional control?

Our studies of soft repression in the context of SIN3 and Rb proteins stemmed from consideration of the many physical targets from genes in shared pathways, which did not exhibit dramatic transcriptional responses, but appeared to explain pleiotropic phenotypes. How general might such regulation be? The physical occupancy of other components of transcriptional regulatory machinery provides clues that continuous repressive activities may be a common feature of many housekeeping promoters. Some of the first genome-wide studies of histone deacetylases revealed that these enzymes, central for repression, are commonly associated with active promoters, in a pattern similar to that of activation-associated histone acetyltransferases (Wang *et al.* 2009). This study did not address the functionality of HDAC at these active loci, although the occupancy positively correlated with gene activity, and the loss of HDAC expression allowed ectopic acetylation of silent transcriptional start sites. One role for HDAC1 at active loci was recently demonstrated to be in the release of paused RNA polymerase II to promote the elongation phase of transcription (Vaid *et al*. 2020). As both Rb and SIN3 physically interact with HDAC1, a major mechanism through which these factors exert soft repression could be through regulation of the transition to productive elongation.

An additional aspect of promoter proximal soft repression is that the genes involved may feature multiple regulatory inputs, so that loss of any single regulatory control feature results in a small change in expression. Indeed, many positive and negative regulators of transcription have been found to co-associate with areas of active gene transcription. Using DAM-ID to characterize chromatin association of factors, the van Steensel group performed a genome-wide binding analysis of 53 Drosophila chromatin associated proteins (Filion *et al*. 2010). Certain repressive chromatin marks and factors (e.g., factors of the Polycomb complex) did co-cluster with inactive regions; however, Sin3A and HDAC1 were found to generally colocalize to two subtypes of euchromatin that contain the majority of active genes, and are enriched in binding of components of the general transcription machinery. This study was conducted in cultured cells, where one can expect that most cells are in a similar state of developmental identity. For other types of ChIP-seq analyses of transcriptional regulators that used complex tissues or whole organisms, the co-occurrence of transcriptional repressors with areas of active gene expression may also be a reflection of the heterogeneity of cell types. Thus, to discern whether particular repressors and corepressors may be actively and continuously engaging in modulation of gene expression from specific promoters, it is important to know whether subpopulations of cells are present. Clearly, for widely-active housekeeping-type genes, this concern is of lesser importance. Overall, functional studies – from genetic to genomic – are needed to determine whether the promiscuous association of many factors with certain segments of the genome represents a complex regulatory playing field, or just the noise of binding as countless factors seek their functional targets.

CONCLUSIONS AND OUTLOOK

From a consideration of SIN3 and Rb, we suggest that many repressors may exert soft repression. There may not be a sharp division between hard and soft repression, but rather a continuum of regulatory action, dependent on promoter architecture, regulatory inputs, and modifications to the corepressors. A focus on soft repression is critical for three reasons: First, this subtle regulation reveals critical direct biological effects, overlooked if stringent quantitative cut-offs are applied. Second, SIN3 and other factors may predominantly act this way, so understanding these factors requires measuring such soft effects. Third, global, systems-level network studies depend on accurate descriptions of genes (nodes) and regulatory interactions (edges), and knowledge of regulation via soft repression will enhance the power of these models.

We see four challenges for progress: First, the magnitude of soft repression makes it hard to spot, since transcriptomic analyses typically focus on more robust effects. Beyond more and deeper sequencing, we need complementary approaches to test the functional importance of potential direct regulation, including high-throughput methods to subtly perturb promoters and repressors and corepressors. Second, pleiotropic effects from misregulation of soft repressors make it difficult to differentiate primary and secondary responses. Approaches that specifically decouple a soft repressor from a target locus in the context of a wild-type cell will be helpful. Third, cell-to-cell variation may complicate distinguishing noise from mild effects. Single cell transcriptomic studies could address this limitation. Finally, a thorough consideration of regulatory variation at the population and species level should be employed to discern possible soft repression effects. Perhaps specific promoter proximal SNPs associated with complex traits impact soft repression.

We propose that gene expression research formally consider the impacts of soft repression in myriad settings, to better uncover the basis of gene regulation in development and disease.

ACKNOWLEDGEMENTS

David N. Arnosti and Lori A. Pile conceived the idea of the manuscript. Anindita Mitra, Ana-Maria Raicu and Stephanie L. Hickey contributed equally to data analysis and interpretation. Anindita Mitra, Ana-Maria Raicu, Lori A. Pile, David N. Arnosti contributed to manuscript writing and editing. Ana-Maria Raicu generated figures with input from Anindita Mitra, Stephanie L. Hickey, Lori A. Pile, and David N. Arnosti. All authors read and approved the final manuscript. This work was supported by National Institutes of Health Grant R01GM088886 and from a Wayne State University research support grant to Lori A. Pile and National Institutes of Health Grant R01GM124137 to DNA.

REFERENCES

Bainor, A. J., Saini, S., Calderon, A., Casado-Polanco, R., Giner-Ramirez, B., Moncada, C., . . . David, G. (2018**). The HDAC-associated Sin3B protein represses DREAM complex targets and cooperates with APC/C to promote quiescence.** Cell reports, 25, 2797.

Ballas, N., Grunseich, C., Lu, D. D., Speh, J. C., & Mandel, G. (2005). **REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis.** Cell, 121, 645.

Barnes, V. L., Bhat, A., Unnikrishnan, A., Heydari, A. R., Arking, R., & Pile, L. A. (2014). **SIN3 is critical for stress resistance and modulates adult lifespan.** Aging (Albany. NY), 6, 645.

Barolo, S., & Posakony, J.W. (2002). **Three habits of highly effective signaling pathways: Principles of transcriptional control by developmental cell signaling.** Genes Dev, 16, 1167.

Bernstein, B. E., Tong, J.K., & Schreiber, S. L. (2000). **Genome wide studies of histone deacetylase function in yeast.** Proc. Natl. Acad. Sci. U. S. A., 97, 13708.

Bertoli, C., Skotheim, J. M., & De Bruin, R. A. M. (2013). **Control of cell cycle transcription during G1 and S phases.** Nat. Rev. Mol. Cell Biol., 14, 518.

Beurton, F., Stempor, P., Caron, M., Appert, A., Dong, Y., Chen, R. A. J., . . . Palladino, F. (2019). **Physical and functional interaction between SET1/COMPASS complex component CFP-1 and a Sin3S HDAC complex in C. elegans.** Nucleic Acids Res, 47, 11164.

Bridi, M., Schoch, H., Florian, C., Poplawski, S. G., Banerjee, A., Hawk, J. D., . . . Abel, T. (2020). **Transcriptional corepressor SIN3A regulates hippocampal synaptic plasticity via Homer1/mGluR5 signaling.** JCI insight, 5, e92385. https://doi.org/10.1172/jci.insight.92385.

Cao, L., Peng, B., Yao, L., Zhang, X., Sun, K., Yang, X., & Yu, L. (2010). **The ancient function of RB-E2F Pathway: Insights from its evolutionary history.** Biol. Direct, 5, 55. https://doi.org/10.1186/1745-6150-5-55.

Carrozza, M. J., Li, B., Florens, L., Suganuma, T., Swanson, S. K., Lee, K. K., . . . Workman, J. L. (2005). **Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription.** Cell, 123, 581.

Chaubal, A., & Pile, L. A. (2018). **Same agent, different messages: Insight into transcriptional regulation by SIN3 isoforms.** Epigenetics Chromatin, 11, 17. https://doi.org/10.1186/s13072-018-0188-y.

Chicas, A., Wang, X., Zhang, C., Mccurrach, M., Zhao, Z., Dickins, R. A., Narita, M., Zhang, M., Lowe, S.W. (2010). **Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence.** Cancer cell, 17, 376.

Chinnam, M., & Goodrich, D.W. (2011). **RB1, development, and cancer.** Curr. Top. Dev. Biol., 94, 129.

Cowley, S. M., Iritani, B. M., Mendrysa, S. M., Xu, T., Cheng, P. F., Yada, J., . . . Eisenman, R. N. (2005). **The mSin3A chromatin-modifying complex is essential for embryogenesis and T-Cell development.** Mol. Cell Biol., 25, 6990.

Cox, K. H., & Takahashi, J. S. (2019). **Circadian clock genes and the transcriptional architecture of the clock mechanism.** J. Mol. Endocrinol., 63, R93.

Dannenberg, J. H., David, G., Zhong, S., Van Der Torre, J., Wong, W. H., & Depinho, R. A. (2005). **mSin3A corepressor regulates diverse transcriptional networks governing normal and neoplastic growth and survival.** Genes Dev, 19, 1581.

David, G., Grandinetti, K. B., Finnerty, P. M., Simpson, N., Chu, G. C., & Depinho, R. A. (2008). **Specific requirement of the chromatin modifier mSin3B in cell cycle exit and cellular differentiation.** Proc. Natl. Acad. Sci. U. S. A., 105, 4168.

Dick, F. A., & Rubin, S. M. (2013). **Molecular mechanisms underlying RB protein function.** Nat. Rev. Mol. Cell Biol., 14, 297.

Dimova, D. K., Stevaux, O., Frolov, M. V., & Dyson, N. J. (2003). **Cell cycle-dependent and cell cycle-independent control of transcription by the Drosophila E2F/RB pathway.** Genes Dev, 17, 2308.

Duronio, R. J., O'Farrell, P. H., Xie, J. E., Brook, A., & Dyson, N. (1995). **The transcription factor E2F is required for S phase during drosophila embryogenesis.** Genes Dev, 9, 1445.

Engelen, E., Brandsma, J. H., Moen, M. J., Signorile, L., Dekkers, D. H., Demmers, J., et al. (2015). **Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry.** Nat. Commun., 6, 7155. https://doi.org/10.1038/ncomms8155.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., et al. (2010). **Systematic protein location mapping reveals five principal chromatin types in drosophila cells.** Cell, 143, 212.

Gajan, A., Barnes, V. L., Liu, M., Saha, N., & Pile, L. A. (2016). **The histone demethylase dKDM5/LID interacts with the SIN3 histone deacetylase complex and shares functional similarities with SIN3.** Epigenetics Chromatin, 9, 4. https://doi.org/10.1186/s13072-016-0053-9.

Goolam, M., Xypolita, M. E., Costello, I., Lydon, J. P., DeMayo, F. J., Bikoff, E. K., . . . Mould, A. W. (2020). **The transcriptional repressor Blimp1/PRDM1 regulates the maternal decidual response in mice.** Nat Commun, 11, 2782.

Gorenstein, C., & Warner, J. R. (1976). **Coordinate regulation of the synthesis of eukaryotic ribosomal proteins.** Proc. Natl. Acad. Sci. U. S. A., 73, 1547.

Guimaraes, J. C., & Zavolan, M. (2016). **Patterns of ribosomal protein expression specify normal and malignant human cells.** Genome Biol, 17, 236.

Hasan, T., & Saluja, D. (2015). in (In: L.R. Singh, T.A. Dar, P. Ahmad Eds.), **Proteostasis and chaperone surveillance.** (pp. 3–24). New Delhi, India: Springer.

Howe, G. A., Major, I. T., & Koo, A. J. (2018). **Modularity in jasmonate signaling for multistress resilience.** Annu. Rev. Plant Biol., 69, 387.

Ishak, C. A., Marshall, A. E., Passos, D. T.,White, C. R., Seung, J., Cecchini, M. J., et al. (2016). **An RB-EZH2 complex mediates silencing of repetitive DNA sequences.** Mol. Cell, 64, 1074.

Jelinic, P., Pellegrino, J., David, G. (2011). **A novel mammalian complex containing Sin3B mitigates histone acetylation and RNA polymerase II progression within transcribed loci.** Mol Cell Biol, 31, 54.

Kadamb, R., Mittal, S., Bansal, N., & Saluja, D. (2015). **Stress mediated Sin3B activation leads to negative regulation of subset of p53 target genes.** Biosci. Rep., 35, e00234. https://doi.org/10.1042/BSR20150122.

Kareta, M. S., Gorges, L. L., Hafeez, S., Benayoun, B. A., Marro, S., Zmoos,A. F., . . . Wernig, M. (2015). **Inhibition of pluripotency networks by the Rb tumor suppressor restricts reprogramming and tumorigenesis.** Cell Stem Cell, 16, 39.

Keogh, M. C., Kurdistani, S. K., Morris, S. A., Ahn, S. H., Podolny, V., Collins, S. R., . . . Krogan, N. J. (2005). **Cotranscriptional Set2 methylation of histone h3 lysine 36 recruits a repressive Rpd3 complex.** Cell, 123, 593.

Knudson, A. G. (1971). **Mutation and cancer: Statistical study of retinoblastoma.** Proc. Natl. Acad. Sci. U. S. A., 68, 820.

Kok, K., & Arnosti, D.N. (2015). **Dynamic reprogramming of chromatin: Paradigmatic palimpsests and HES factors.** Front. Genet., 6, 29. https://doi.org/10.3389/fgene.2015.00029.

Korenjak, M., Anderssen, E., Ramaswamy, S., Whetstine, J. R., & Dyson, N. J. (2012). **RBF binding to both canonical E2F targets and noncanonical targets depends on functional dE2F/dDP complexes.** Mol. Cell. Biol., 32, 4375.

Kryuchkova-Mostacci, N., & Robinson-Rechavi, M. (2017). **A benchmark of gene expression tissue-specificity metrics.** Brief. Bioinformatics, 18, 205.

Latorre, I., Chesney, M. A., Garrigues, J. M., Stempor, P., Appert, A., Francesconi, M., et al. (2015). **The DREAM complex promotes gene body H2A.Z for target repression.** Genes Dev, 29, 495.

Lewis, M. J., Liu, J., Libby, E. F., Lee, M., Crawford, N. P., & Hurst, D. R. (2016). **SIN3A and SIN3B differentially regulate breast cancer metastasis.** Oncotarget, 7, 78713.

Liu, M., & Pile, L. A. (2017). **The transcriptional corepressor SIN3 directly regulates genes involved in methionine catabolism and affects histone methylation, linking epigenetics and metabolism.** J. Biol. Chem., 292, 1970.

Liu, M., Saha, N., Gajan, A., Saadat, N., Gupta, S. V., & Pile, L. A. (2020). **A complex interplay between SAM synthetase and the epigenetic regulator SIN3 controls metabolism and transcription.** J. Biol. Chem., 295, 375.

Love, M. I., Huber, W., & Anders, S. (2014). **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** Genome Biol, 15, 550.

Marygold, S. J., Roote, J., Reuter, G., Lambertsson, A., Ashburner, M., Millburn, G. H., et al. (2007). **The ribosomal protein genes and minute loci of drosophila melanogaster.** Genome Biol., 8, R216.

Mouawad, R., Himadewi, P., Kadiyala, D., & Arnosti, D.N. (2020). **Selective repression of the Drosophila cyclin B promoter by retinoblastoma and E2F proteins.** Biochim. Biophys. Acta – Gene Regul. Mech., 1863, 194549. https://doi.org/10.1016/j.bbagrm.2020.194549.

Mouawad, R., Prasad, J., Thorley, D., Himadewi, P., Kadiyala, D., Wilson, N., . . . Arnosti, D. N. (2019). **Diversification of retinoblastoma protein function associated with cis and trans adaptations.** Mol. Biol. Evol., 36, 2790.

Mukherjee, S., Brulet, R., Zhang, L., & Hsieh, J. (2016). **REST regulation of gene networks in adult neural stem cells.** Nat Commun, 7. https://doi.org/10.1038/ncomms13360.

Nakayama, J. I., & Hayakawa, T. (2011). **Physiological roles of class I HDAC complex and histone demethylase.** J. Biomed. Biotechnol.,1-10, 2011. https://doi.org/10.1155/2011/129383.

Narumi-Kishimoto, Y., Araki, N., Migita, O., Kawai, T., Okamura, K., Nakabayashi, K., . . . Hata, K. (2019). **Novel SIN3Amutation identified in a Japanese patient with Witteveen-Kolk syndrome.** Eur. J. Med. Genet., 62, 103547. https://doi.org/10.1016/j.ejmg.2018.09.014.

Nasmyth, K., Stillman, D., & Kipling, D. (1987). **Both positive and negative regulators of HO transcription are required for mother-cell specific mating-type switching in yeast.** Cell, 48, 579.

Neufeld, T. P., Tang, A. H., & Rubin, G. M. (1998). **A genetic screen to identify components of the sina signaling pathway in drosophila eye development.** Genetics, 148, 277-286.

Nicolay, B. N., & Dyson, N. J. (2013). **The multiple connections between pRB and cell metabolism.** Curr. Opin. Cell. Biol., 25, 735.

Pennetta, G., & Pauli, D. (1998). **The Drosophila Sin3 gene encodes a widely distributed transcription factor essential for embryonic viability.** Dev. Genes Evol., 208, 531.

Pile, L. A., Spellman, P. T., Katzenberger, R. J., & Wassarman, D. A. (2003). **The SIN3 deacetylase complex represses genes encoding mitochondrial proteins.** J. Biol. Chem., 278, 37840.

Powers, T., & Walter, P. (1999). **Regulation of ribosome biogenesis by the rapamycin-sensitive TOR-signaling pathway in saccharomyces cerevisiae.** Mol. Biol. Cell, 10, 987.

Rielland, M., Cantor, D. J., Graveline, R., Hajdu, C., Mara, L., De Diego Diaz, B., . . . David, G. (2014). **Senescence-associated SIN3B promotes inflammation and pancreatic cancer progression.** J. Clin. Invest., 124, 2125.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: **A Bioconductor package for differential expression analysis of digital gene expression data.** Bioinformatics, 26, 139.

Saha, N., Liu, M., Gajan, A., & Pile, L. A. (2016). **Genome-wide studies reveal novel and distinct biological pathways regulated by SIN3 isoforms.** BMC genomics, 17. https://doi.org/10.1186/s12864-016-2428-5.

Setty, Y., Mayo, A. E., Surette, M. G., & Alon, U. (2003). **Detailed map of a cis-regulatory input function.** Proc. Natl. Acad. Sci. U. S. A., 100, 7702.

Shapiro-Shelef, M., Lin, K. I., McHeyzer-Williams, L. J., Liao, J., McHeyzer-Williams, M. G., & Calame, K. (2003). **Blimp-1 Is Required for the formation of immunoglobulin secreting plasma cells and preplasma memory B cells.** Immunity, 19, 607.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). **The BioMart community portal: An innovative alternative to large, centralized data repositories.** Nucleic Acids Res, 43, W589.

Sternberg, P.W., Stern, M. J., Clark, I., & Herskowitz, I. (1987). **Activation of the yeast HO gene by release from multiple negative controls.** Cell, 48, 567.

Stevaux, O., Dimova, D. K., Ji, J. Y., Moon, N. S., Frolov, M. V., & Dyson, N. J. (2005). **Retinoblastoma family 2 is required in vivo for the tissue specific repression of dE2F2 target genes.** Cell Cycle, 4, 1272.

Swaminathan, A., & Pile, L. A. (2010). **Regulation of cell proliferation and wing development by drosophila SIN3 and string.** Mech. Dev., 127, 96.

Trapnell, C., Williams, B. A., G. Pertea, Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L.,Wold, B. J., & Pachter, L. (2010). **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** Nat Biotechnol, 28, 511.

Ulmert, I., Henriques-Oliveira, L., Pereira, C. F., & Lahl, K. (2020). **Mononuclear phagocyte regulation by the transcription factor Blimp-1 in health and disease.** Immunology, https://doi.org/10.1111/imm.13249.

Vaid, R., Wen, J., & Mannervik, M. (2020). **Release of promoter proximal paused Pol II in response to histone deacetylase inhibition.** Nucleic Acids Res, 48, 4877.

van Oevelen, C.,Wang, J., Asp, P., Yan, Q., Kaelin, W. G. Jr, Kluger, Y., & Dynlacht, B. D. (2008). **A role for mammalian sin3 in permanent gene silencing.** Mol Cell, 32, 359.

Wang, Z., Zang, C., Cui, K., Schones, D. E., Barski, A., Peng,W., & Zhao, K. (2009). **Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes.** Cell, 138, 1019.

Wei, Y., Mondal, S. S., Mouawad, R., Wilczyński, B., Henry, R. W., & Arnosti, D. N. (2015). **Genome-wide analysis of drosophila RBf2 protein highlights the diversity of RB family targets and possible role in regulation of ribosome biosynthesis.** G3, 5, 1503.

Williams, K., Christensen, J., Pedersen, M. T., Johansen, J. V., Cloos, P. A., Rappsilber, J., & Helin, K. (2011). **TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity.** Nature, 473, 343.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R. et al. (2005). **Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.** Bioinformatics, 21, 650.

CHAPTER 4: THE CYNOSURE OF CTBP: EVOLUTION OF A BILATERIAN TRANSCRIPTIONAL COREPRESSOR

This work was published in the following manuscript and adapted here:

Raicu, A.M., Kadiyala, D., Niblock, M., Jain, A., Yang, Y., Bird, K.M., Bertholf, K., Seenivasan, A., Siddiq, M., Arnosti, D.N. **The Cynosure of CtBP: Evolution of a Bilaterian Transcriptional Corepressor.** Mol Biol Evol. 2023 Feb 3;40(2):msad003. doi: 10.1093/molbev/msad003. PMID: 36625090.

ABSTRACT

Evolution of sequence-specific transcription factors clearly drives lineage-specific innovations, but less is known about how changes in the central transcriptional machinery may contribute to evolutionary transformations. In particular, transcriptional regulators are rich in intrinsically disordered regions that appear to be magnets for evolutionary innovation. The C-terminal Binding Protein (CtBP) is a transcriptional corepressor derived from an ancestral lineage of alpha hydroxyacid dehydrogenases; it is found in mammals and invertebrates, and features a core NAD-binding domain as well as an unstructured C-terminus (CTD) of unknown function. CtBP can act on promoters and enhancers to repress transcription through chromatin-linked mechanisms. Our comparative phylogenetic study shows that CtBP is a bilaterian innovation whose CTD of about 100 residues is present in almost all orthologs. CtBP CTDs contain conserved blocks of residues and retain a predicted disordered property, despite having variations in the primary sequence. Interestingly, the structure of the C-terminus has undergone radical transformation independently in certain lineages including flatworms and nematodes. Also contributing to CTD diversity is the production of myriad alternative RNA splicing products, including the production of "short" tailless forms of CtBP in Drosophila. Additional diversity stems from multiple gene duplications in vertebrates, where up to five CtBP orthologs have been observed. Vertebrate lineages show fewer major modifications in the unstructured CTD, possibly because gene regulatory constraints of the vertebrate body plan place specific constraints on this domain. Our study highlights the rich regulatory potential of this previously unstudied domain of a central transcriptional regulator.

Key words: C-terminal binding protein, CtBP, transcription, corepressor, intrinsically disordered region, evolution.

INTRODUCTION

The C-terminal Binding Protein (CtBP) is a transcriptional corepressor that plays critical roles in development, tumorigenesis, and cell fate (Boyd *et al.* 1993; Schaeper *et al.* 1995; reviewed in Chinnadurai 2007; Stankiewicz 2014). CtBP has been implicated in human cancer, and is being investigated as a potential drug target (Nadauld *et al.* 2006; Barroilhet *et al.* 2013; Deng *et al.* 2013; Dcona *et al.* 2017). CtBP was first identified as a protein that binds the C-terminus of the adenoviral E1A oncoprotein and was later found to interact with diverse cellular transcription factors via their PLDLS motif, creating complexes that alter chromatin (Boyd *et al.* 1993; Schaeper *et al.* 1995; Nibu *et al.* 1998; Turner and Crossley 2001; Shi *et al.* 2003). This cofactor transcriptionally regulates genes involved in apoptosis, cell adhesion, and the epithelial-to-mesenchymal transition, functioning as a repressor in most cases, although it can directly activate promoters in some contexts (Grooteclaes *et al.* 2003; Fang *et al.* 2006; Jin *et al.* 2007; Paliwal *et al.* 2012). Unique among transcriptional coregulators, CtBP structurally resembles D-2-hydroxyacid dehydrogenases and binds the NAD(H) cofactor (Chinnadurai 2002; Kumar *et al.* 2002). *In vitro*, CtBP proteins can use a variety of alpha hydroxyacids as substrates, but the natural *in vivo* substrate, if any, remains unknown (Kumar *et al.* 2002; Balasubramanian *et al.* 2003; Achouri *et al.* 2007). Mammalian cell culture studies have shown that the residues required for *in vitro* catalytic activity are not required for transcriptional repression or the apoptotic activities of CtBP, but in the fly, residues of the dehydrogenase active site are required for normal activity of this repressor (Grooteclaes *et al.* 2003; Zhang and Arnosti 2011).

CtBP binding to NAD(H) is necessary for its normal functions, and substantial evidence indicates that NAD(H) binding supports CtBP dimerization and tetramerization (Kumar *et al.* 2002; Balasubramanian *et al.* 2003; Sutrias-Grau & Arnosti 2004; Mani-Telang and Arnosti 2007;

Madison *et al.* 2013; Bellesis *et al.* 2018; Jecrois *et al.* 2021). Tetramerization has been shown to be required for transcriptional repression, as tetramer- destabilizing mutants have compromised transcriptional regulatory activity (Bhambhani *et al.* 2011; Ray *et al.* 2017; Bi *et al.* 2018; Jecrois *et al.* 2021). Additionally, it is the oligomeric form of CtBP that associates with other factors, suggesting this is the relevant form for transcriptional regulation (Shi *et al.* 2003; Jecrois *et al.* 2021). Both NAD+ and the reduced NADH cofactor promote oligomerization, but their relative binding affinities to CtBP have been disputed: one study found NADH to have >100-fold stronger binding than NAD+, whereas other studies indicate no differences in binding affinity (Zhang *et al.* 2002; Fjeld *et al.* 2003; Bellesis *et al.* 2018). More recently, the Royer lab has shown through ultracentrifugation and ITC that there is only a 9-fold binding affinity difference, and that CtBP is saturated with NAD+, suggesting that it does not respond to cellular redox levels (Erlandsen *et al.* 2022). Currently, it is not clear whether NAD(H) binding to CtBP is merely a structural element, or whether the presence of the cofactor may bridge cellular metabolism and gene regulation.

CtBP proteins contain a variety of functional elements, including an N-terminal (NTD) substrate binding domain that overlaps with the conserved dehydrogenase domain, and the flexible and unstructured C-terminal domain (CTD). Differential promoter usage and alternative splicing produces distinct mammalian CtBP isoforms, with a CtBP2 RIBEYE variant having a sizable N-terminal extension that is unrelated to domains found in other CtBP isoforms (Schmitz *et al.* 2000). A hydrophobic cleft toward the N-terminus binds the E1A PLDLS motif, with additional interactions with cellular factors also mediated by the RRT-binding surface groove (Schaeper *et al.* 1995; Quinlan *et al.* 2006; Kuppuswamy *et al.* 2008). The central dehydrogenase domain includes a Arg-Glu-His (REH) catalytic triad and a Rossman fold involved in NAD(H) binding (Kumar *et al.* 2002).

The structural element that seems to distinguish CtBP proteins most clearly from more distantly related alpha hydroxyacid dehydrogenases is the unstructured CTD, which has not been structurally resolved (Nardini *et al.* 2006). This portion of the protein is the site of posttranslational modifications such as phosphorylation and sumoylation (Kumar *et al.* 2002; reviewed in Chinnadurai 2007; Jecrois *et al.* 2021). Dimeric and tetrameric forms of the protein lacking the entire C-terminal region can be obtained *in vitro*, indicating that this domain is not essential for this structural aspect of CtBP (Jecrois *et al.* 2021). One study suggested that an intact CTD was required for CtBP tetramerization, but Royer and colleagues demonstrated through SEC-MALS and cryoEM that the minimal dehydrogenase domain, without a CTD, can tetramerize in the presence of NAD(H) (Madison *et al.* 2013; Bellesis *et al.* 2018; Jecrois *et al.* 2021). Regarding function, the mammalian CtBP1 lacking the final 86 residues can still function as a repressor in cell culture (Kuppuswamy *et al.* 2008). Additionally, a "short" CtBP isoform is sufficient to rescue lethality of *dCtBP* loss in Drosophila, further indicating that core functions are possible in the absence of this domain (Zhang and Arnosti 2011). Thus, the roles of the CTD in oligomerization, transcriptional regulation, and other nuclear activities still remain to be defined.

Diverse CtBP proteins are found within Metazoa; invertebrates can express several isoforms from a single locus through alternative splicing and alternative promoter usage, whereas vertebrates have additional diversity through gene duplications that produced two or more paralogous CtBP genes. In mammals, multiple isoforms are expressed from each CtBP1 and CtBP2 paralog, which have both overlapping and unique genetic roles in the cell—both nuclear and extracellular (Katsanis and Fisher 1998; Schmitz *et al.* 2000; Hildebrand and Soriano 2002; reviewed in Chinnadurai 2007). CtBP2 is an essential gene in the mouse, with null mutants showing embryonic lethality. CtBP1 null mice are viable, but exhibit developmental phenotypes

(Hildebrand and Soriano 2002). In contrast, Drosophila possesses a single *CtBP* gene that expresses diverse CtBP isoforms through alternative splicing, affecting in particular the CTD (Nibu *et al*. 1998; Poortinga *et al*. 1998). Two major isoforms, the "long" and "short" forms, differ mainly in the C-terminus and are differentially expressed in development (Sutrias-Grau and Arnosti 2004; Mani-Telang and Arnosti 2007). The long version (CtBP(L)) contains a ~90 residue extension not found in the short protein (CtBP(S)). CtBP(S) is the most abundant isoform in Drosophila, and it represses just as well as CtBP(L) when tethered to Gal4 *in vivo* (Sutrias-Grau and Arnosti 2004; Mani-Telang and Arnosti 2007). Loss of *CtBP* is lethal in Drosophila; this phenotype can be rescued by expression of either a CtBP(S) or CtBP(L) transgene (Zhang and Arnosti 2011). However, there is an indication that expression of both isoforms is important; in this system, rescue by CtBP(L) leads to significant changes in several target genes, not seen with rescue by CtBP(S) (Zhang and Arnosti 2011).

The deeper biological significance of the CTD encoded in CtBP genes, and reason for its conservation, are still unknown. We hypothesize that the CTD may play a role in regulation and/or turnover, interactions with cofactors to regulate transcription, or protein localization. To lay the groundwork for experimental analysis of the CtBP CTD, we have undertaken a comprehensive comparative approach and assessed characteristics of the C-terminal portion of the protein across the animal kingdom. Investigating richly-resourced dipteran and other arthropod genomic resources and extending to invertebrates and vertebrates in general, we describe the conservation and variation found in divergent clades, pointing to likely functional aspects of this domain.

RESULTS

Origin of CtBP

Sequence conservation and functional similarities support the orthology of well-studied CtBP genes from mammals and Drosophila. The high degree of sequence divergence noted in the CTD of the *Caenorhabditis elegans* ortholog raises a question of what features most reliably support orthology in this gene family (Nicholas *et al.* 2008). A high level of sequence similarity is found across the ~330 amino acid dehydrogenase domain (arthropod to vertebrate CtBP1 >70% identity; **Figure S4.1**). However, whether genes with lower sequence identities in other organisms are orthologs has not been comprehensively assessed. The Arabidopsis ANGUSTIFOLIA (AN) gene encodes a divergent homolog of CtBP that is not likely to be orthologous to animal genes; AN has a lower (~30%; **Figure S4.1**) level of sequence similarity across the core dehydrogenase domain, lacks conserved catalytic residues, and has cytoplasmic functions related to microtubule regulation, membrane trafficking, and stress response (Folkers *et al.* 2002; Kim *et al.* 2002; Bhasin and Hulskamp 2017). This protein does not mediate repression in heterologous animal assays, although mutants show changes in gene expression (Kim *et al.* 2002; Stern *et al.* 2007; Xie *et al.* 2020). Similarly, fungal and choanoflagellate CtBP homologs have low (~30%; **Figure S4.1**) levels of sequence similarity across the dehydrogenase domain, and best hits using mammalian or dipteran searches identify genes encoding proteins that are annotated as dehydrogenases, lacking any unstructured CTD.

We asked at which point in the metazoan phylogeny we could identify CtBP-encoding genes with high levels of sequence identity, similar to those observed in mammalian-insect alignments. We did not identify homologs with such levels of similarity, or extended unstructured CTD, in representative genomes from Cnidaria or Porifera (**Figure S4.1**). In these genomes, the most similar homologs exhibited much lower levels of sequence identity (~30%), comparable to those of fungi and plants. Additionally, homologs from these species lack a C-terminal domain as is found in flies and humans. The first CtBP gene therefore likely arose in a common ancestor to bilaterians, a decisive point in animal evolutionary history, when new combinations of gene batteries appeared that regulated novel morphological traits. As we report here, certain unique features of CtBP appear to be conserved, though not entirely, across protostomes and deuterostomes. Diversity in CtBP has been achieved over time through gene duplication in Vertebrata, and generation of alternative isoforms through transcriptional and splicing variation. To better understand the molecular processes underlying CtBP diversity, we first considered these genes in Drosophila, where extensive genomics combined with experimental work inform our understanding.

Drosophila melanogaster Expresses Alternatively Spliced CtBP Isoforms

The *D. melanogaster* CtBP gene, which is essential for development, produces a number of alternatively spliced transcripts (Poortinga *et al*. 1998; Mani-Telang and Arnosti 2007). Ten mRNA isoforms have been reported, differing in their transcriptional start sites (TSS), length of their untranslated regions (UTR), and inclusion of 3′ exons. They encode seven total protein variants, ranging in size from 379 to 481 residues, and produce two general types of proteins: the CtBP "long" (CtBP(L)) and CtBP "short" (CtBP(S)), named based on the length of their CTD (**Figure 4.1A, B**). The CtBP(L) isoforms incorporate the 3′ most protein-coding exons, terminating in the same sequence motif; they differ in alternative splicing of three short motifs within the core and CTD (**Figure 4.1C**). In contrast, the ORFs of the CtBP(S) isoforms end shortly after the catalytic core and terminate in one of two ways: reading through a splice donor site after protein-coding exon 5 into the adjacent intron, or through usage of an alternative splice acceptor site 5′ of protein-coding exon 7, which encodes the last portion of the CtBP(L) isoforms in a different reading frame. In addition, these short isoforms also differ in the retention or deletion of a short VFQ tripeptide found near the start of the CTD (**Figure 4.1D**). The remaining NTD and core sequences are identical among CtBP isoforms in the fly. Of interest is the clear difference between the short and long isoforms, which we have shown to differ in their spatial and temporal expression in *D. melanogaster*, and which may play unique roles in development (Zhang and Arnosti 2011). In fact, the two major isoforms are developmentally regulated, and CtBP(S) is believed to be the predominant form expressed across development (Mani-Telang and Arnosti 2007; Zhang and Arnosti 2011).

Long and Short CtBP Isoforms are Conserved Throughout Drosophila

We next assessed the conservation of variant CtBP isoforms produced through alternative splicing in 11 additional Drosophila species (**Figure 4.2A**). For every species, multiple mRNA sequences exist for both long and short isoforms, and they differ in the retention or loss of the same short segments encoding VFQ, LNGGYYTG, and VSSQS observed in *D. melanogaster* (data not shown). The conservation of these variants suggests that expression of long and short isoforms of CtBP, as well as the exclusion/retention of short motifs, are functionally important. The NTD and catalytic core sequences are highly conserved, whereas CTD sequences themselves show more evolutionary variation, particularly in the center of the CTD, with the presence or the absence of alanine- and proline-rich sequences (**Figure 4.2B**). All species express mRNAs encoding CtBP-short proteins ending in AP/SECARP, using a conserved alternative splice acceptor site. Some also are found to produce mRNAs that create short isoforms terminating in SNQEK by reading through a splice donor site into the next intron. Additional unique short endings, created through alternative splicing, are seen in some species (**Figure 4.2C, S4.2A**). Surprisingly, the nucleotide

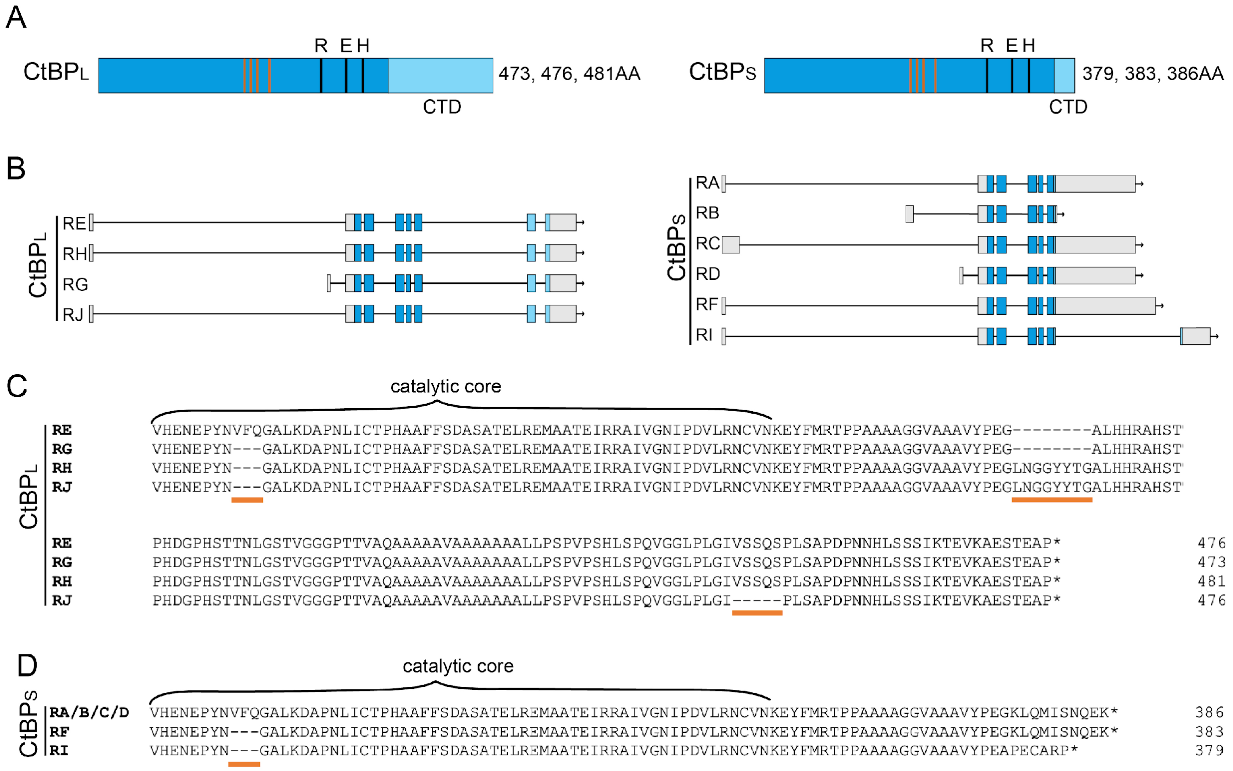
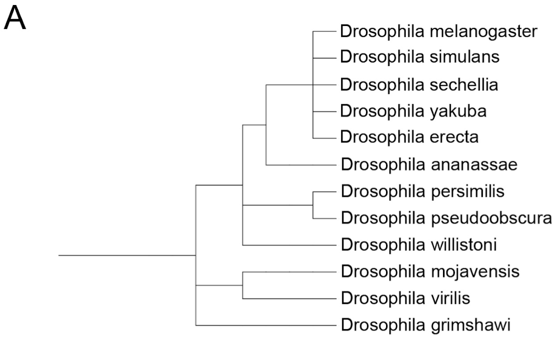

A
CtBPL
R E H
473, 476, 481AA
CTD
CtBPS
R E H
379, 383, 386AA
CTD
B
CtBPL
RE
RH
RG
RJ
CtBPS
RA
RB
RC
RD
RF
RI
C
catalytic core
CtBPL
RE VHENEPYNVFQGALKDAPNLICTPHAAFFSDASATELREMAATEIRRAIVGNIPDVLRNCVNKEYFMRTPPAAAAGGVAAAVYPEG--------ALHHRAHST
RG VHENEPYN---GALKDAPNLICTPHAAFFSDASATELREMAATEIRRAIVGNIPDVLRNCVNKEYFMRTPPAAAAGGVAAAVYPEG--------ALHHRAHST
RH VHENEPYN---GALKDAPNLICTPHAAFFSDASATELREMAATEIRRAIVGNIPDVLRNCVNKEYFMRTPPAAAAGGVAAAVYPEGLNGGYYTGALHHRAHST
RJ VHENEPYN---GALKDAPNLICTPHAAFFSDASATELREMAATEIRRAIVGNIPDVLRNCVNKEYFMRTPPAAAAGGVAAAVYPEGLNGGYYTGALHHRAHST
RE PHDGPHSTTNLGSTVGGGPTTVAQAAAAAVAAAAAAALLPSPVPSHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSSSIKTEVKAESTEAP* 476
RG PHDGPHSTTNLGSTVGGGPTTVAQAAAAAVAAAAAAALLPSPVPSHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSSSIKTEVKAESTEAP* 473
RH PHDGPHSTTNLGSTVGGGPTTVAQAAAAAVAAAAAAALLPSPVPSHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSSSIKTEVKAESTEAP* 481
RJ PHDGPHSTTNLGSTVGGGPTTVAQAAAAAVAAAAAAALLPSPVPSHLSPQVGGLPLGI-----PLSAPDPNNHLSSSIKTEVKAESTEAP* 476
D
catalytic core
CtBPS
RA/B/C/D VHENEPYNVFQGALKDAPNLICTPHAAFFSDASATELREMAATEIRRAIVGNIPDVLRNCVNKEYFMRTPPAAAAGGVAAAVYPEGKLQMISNQEK* 386
RF VHENEPYN---GALKDAPNLICTPHAAFFSDASATELREMAATEIRRAIVGNIPDVLRNCVNKEYFMRTPPAAAAGGVAAAVYPEGKLQMISNQEK* 383
RI VHENEPYN---GALKDAPNLICTPHAAFFSDASATELREMAATEIRRAIVGNIPDVLRNCVNKEYFMRTPPAAAAGGVAAAVYPEAPECARP* 379

Figure 4.1. The CtBP locus in *D. melanogaster* produces variant transcripts and proteins with different C-terminal lengths. A) Two general types of CtBP proteins are produced in *D. melanogaster*: long (CtBP(L)) and short (CtBP(S)). The proteins are almost identical in their N-terminus and central catalytic domain (blue), and differ in the sequences and lengths of the C-terminal domain (light blue). Proteins of three different sizes are predicted to be produced for each isoform. Orange vertical bars indicate the four residues involved in NAD binding, and black lines indicate the three residues making up the catalytic triad (REH). **B)** Schematic representation of the 10 transcripts produced from the CtBP locus. Gray boxes indicate 5′ and 3′ UTRs, blue boxes indicate protein-coding exons, and horizontal lines are introns. Isoforms E, H, G, J encode long versions of the protein and use two different TSSs. Isoforms A, B, C, D, F, and I encode short versions of the protein and use three different TSSs. **C)** Alignment of the C-terminal region of the conserved core and the CTD indicates that four different long proteins are encoded, which differ with the inclusion or deletion of three small motifs in the core and CTD: a VFQ tripeptide, an LNGGYYTG motif, and VSSQS motif (orange horizontal bars), all of which are spliced out of the mRNA in different combinations. **D)** Alignment of the CTD of the short isoforms indicates that three different proteins are predicted to be produced, which differ with the inclusion or deletion of the same VFQ tripeptide and terminate with SNQEK or APECARP.

sequence encoding the terminal SNQEK derived from the 3′ end of exon 5 and adjacent intron is 100% conserved in all species, which is a much higher level of conservation than noted for other coding regions, which harbor mostly synonymous changes (**Figure S4.2B**). Although isoforms ending in SNQEK are not reported in all of these species, the absolute conservation of this specific portion of the intron suggests that the capacity to generate these isoforms is conserved. The absolute conservation may reflect an RNA structure that would influence the use of this splice donor site to produce a short or long isoform. We predicted the structure of this conserved sequence of RNA using the RNAstructure software (Xu and Mathews 2016), and found that the splice donor site that is used to create CtBP(L), but is suppressed for CtBP(S), folds into a hairpin that may sequester the GU donor site in a stem loop structure (**Figure S4.2C**). Other Drosophila genes have been shown to exhibit a high level of conservation in certain intronic regions that can form RNA hairpin structures to influence alternative splicing events (Raker *et al.* 2009). In contrast, the nucleotide sequence encoding the AP/SECARP short-form variants is not as highly conserved, with both synonymous and nonsynonymous substitutions present in the protein-coding exon, and high divergence in the preceding intron (data not shown). This indicates that RNA secondary structure is not important for this canonically spliced isoform. In summary, although the CTD region of CtBP is more evolutionarily variable than the core dehydrogenase domain, it is likely that the diversity of CTD structure is an important aspect of CtBP proteins in Drosophila.

Conservation of Long and Short Forms of CtBP in Diptera

Drosophila are members of the suborder Brachycera, which also include agriculturally important tephritids and houseflies. Nematocera include gnats, midges, and mosquitoes, which also have extensive genomic resources. To assess CTD structure across Diptera, we selected 11 Brachycera and 11 Nematocera. All Diptera express CtBP(L) isoforms and many also express short variants (**Figure 4.3A**). Two regions exhibit a higher level of conservation among the CTD long forms; a “Central Block” containing a motif featuring hydrophobic and aromatic residues (YSEGINGGYY) with an adjacent H/S/T-rich sequence (AHSTTPHD), and a “Terminal Block”



C

	Short ending form			
Species name	SNQEK	**AP/S**ECARP	DNTAR	AKK
D. melanogaster	x	x		
D. simulans	x	x		
D. sechellia	x	x		
D. yakuba	x	x		
D. erecta	x	x		
D. ananassae		x		x
D. persimilis	x	x	x	
D. pseudoobscura		x	x	
D. willistoni		x		
D. mojavensis		x		
D. virilis		x		
D. grimshawi		x		

B

```
D.melanogaster/PH     NCVNKEYFMRTPPAAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSTVGGGPTT---------------------
D.simulans/X1         NCVNKEYFMRTPPAAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSTVGGGPTT---------------------
D.sechellia/X1        NCVNKEYFMRTPPAAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSTVGGGPTT---------------------
D.yakuba/B            NCVNKEYFMRTPPAAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSTVGGGPTT---------------------
D.erecta/X1           NCVNKEYFMRTPPAAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSTVAGGPTA---------------------
D.ananassae/X1        NCVNKEYFMRTPPTAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHEGPHSTTNLAAAAAA---------------AAALAPPPPG
D.persimilis/X1       NCVNKEYFMRTPPTAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSSSSGS--------------SAMAQPPPPN
D.pseudoobscura/X1    NCVNKEYFMRTPPTAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSSSSGS--------------SAMAQPPPPN
D.willistoni/X1       NCVNKEYFMRTPQTAAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTSHDGPHSTTNIGSSS----------------SSALA-PPPPN
D.mojavensis/X1       NCVNKEYFMRTPPTTAAGGVAAAVYPEGLNGGYYTGALQHRAHSTTPHDGPHSTTNLGSGSGSGVVVGGIGSSGNSASSAALIPPPPT
D.virilis/X1          NCVNKEYFMRTPPTTAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSSSSSG---GG-------TTSAALIPPPA-
D.grimshawi/X1        NCVNKEYFMRTPPTTAAGGVAAAVYPEGLNGGYYTGALHHRAHSTTPHDGPHSTTNLGSSTS----------------SALVQPPN--

D.melanogaster/PH     -----VA-QAAAAAVAAAAAAA-LLPSPVP------------SHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSS-SIKTEVKAESTEAP*
D.simulans/X1         -----VA-QAAAAAVAAAAAAA-LLPSPVP------------PHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSS-SIKTEVKAESTEAP*
D.sechellia/X1        -----VA-QAAAAAVAAAAAAA-LLPSPVP------------PHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSS-SIKTEVKAESTEAP*
D.yakuba/B            -----VA-QAAAAAVAAAAAAAALLPSPVP------------PHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSS-SIKTEVKAESTEAP*
D.erecta/X1           -----VA-QAAAAAAAAVAA---LLPSPVP------------PHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSS-SIKTEVKAESTEAP*
D.ananassae/X1        SNSSSSSVAAAAAAVAAAAAAAAALLPVPSPVPQVPSNSVPSVPHLSPQVAGLPLGIVSSQSPLSAPDPNNHLSSSNIKTEVKTESTEAP*
D.persimilis/X1       SV---AAAAAA------------LLPSPVPPTA-----VPTVPHLSPQVGGLPLGIVSSQSPLSAPDPSNHVLS-SIKAEVKAESTETP*
D.pseudoobscura/X1    SI---AAAAAA------------LLPSPVPPTA-----VPTVPHLSPQVGGLPLGIVSSQSPLSAPDPSNHVLS-SIKAEVKAESTETP*
D.willistoni/X1       S----AAAAAAVVAA--------LLPPSVPPFAEQ-----TVPHLSPQVGGLPLGIVSSQSPLSAPDPNNHLSS-SIKTEVKAESSEAP*
D.mojavensis/X1       ATGNNSM-TAAVAVAAAAAAAAALLPSTVPPQNAA---VPTVPHLSPQVGGLPLGIVSSQSHLSAPDPNNHLSS-SIKSEVKVESTETP*
D.virilis/X1          AGSNTVA-ATAVAVAAAAAAAAALLPSPVPPPNAAA--VPTVPHLSPQIGGLPLGIVSSQSHLSAPDPNNHLSS-SIKSEVKVESTETP*
D.grimshawi/X1        AGTNTMA-AAAAAVAAAAAAA--LLPSPVPP-NAAA--VPTVPHLSPQIGGLPLGIVSSQSHLSAPDPNNHLTS-SIKSEVKVESTETP*
```

Figure 4.2. Long and short CtBP isoforms are expressed in Drosophila. A) Phylogenetic relationship of the Drosophila species used. These species diverged from their last common ancestor ~40 Ma. **B)** Alignment of the CTDs of the longest Drosophila CtBP(L) sequences. Alignment begins with NCVN, which is the end of the conserved core. In this and subsequent figures, blue highlighting is used for conservation of a residue in >50% of species, gold for chemically conserved residues, and army green for conservation of a second residue in 25–50% of species. Orange horizontal bars highlight variable regions rich in proline and alanine. Specific isoform letters or numbers are indicated after the species name, and the final asterisk indicates the STOP codon. **C)** The presence (X) of various alternative CtBP(S) terminal sequences. The SNQEK version is created by suppressing a splice donor site and extending the ORF into the intron. APECARP, DNTAR, and AKK are created through exon skipping and alternative splicing. All species express an APECARP short version and all species have the ability to encode the SNQEK version, but it is not always detected in cDNA sequences. For this figure and subsequent figures, phylogenetic trees were generated by phyloT.

rich in prolines, followed by a short stretch of N/H and acidic residues in a conserved PExSEVH/Q terminus. It is apparent that the Drosophila C-terminal sequence noted above (ESTEAP) is a derived feature within this genus (also found in the closely related *Scaptodrosophila lebanonensis* sequence; **Figure S4.3A**), although it shares the acidic character with the consensus SEVH ending found across Diptera. Among this set of sequences from Brachycera, Drosophila also stands out for the central block of polyalanine repeats not present in other species (**Figure 4.3B**). Within specific species, variants exist in which the YSEGINGGYY, AHSTTPHD, and VSSKS motifs are alternatively spliced out, as was observed in Drosophila (data not shown).

Within Nematocera, and specifically Culicidae (mosquitoes), the terminal sequences of the CtBP(L) isoforms are highly variable—much more so than seen within Brachycera. Across the three mosquito genera we sampled (Aedes, Culex, and Anopheles), we note many genus-specific sequences, as well as some that resemble those found in Brachycera (**Figure 4.3B**, **S4.3D**). Depending on the species, the CtBP gene can give rise to up to six potential long-CtBP isoforms and five potential short CtBP isoforms, all created through alternative splicing (**Figure S4.3B**). We hypothesize that these diverse protein isoforms may serve tissue- or temporal-specific functions.

In most of the Brachycera, we find that the CtBP(S) isoforms SNQEK and APECARP (conserved in drosophilids) are also expressed, with the predominant short form ending in APECARP (**Figure 4.3C**). Interestingly, we find that splice donor site suppression occurs in *Bactrocera oleae* to form an SNQEK-like ending, as seen with *D. melanogaster*. The APECARP endings in the other Diptera are also created through alternative splicing. In contrast, only four of the sampled mosquitoes report short CtBP isoforms, all within the Anopheles genus (**Figure 4.3A, S4.3B**). These do not resemble the conserved brachyceran SNQEK or APECARP variants, but instead have one or more of seven different short variants (**Figure S4.3B, C**). In summary, the production of short- and long-CtBP isoforms is found in Diptera, with certain sequences of the long forms showing strong conservation.

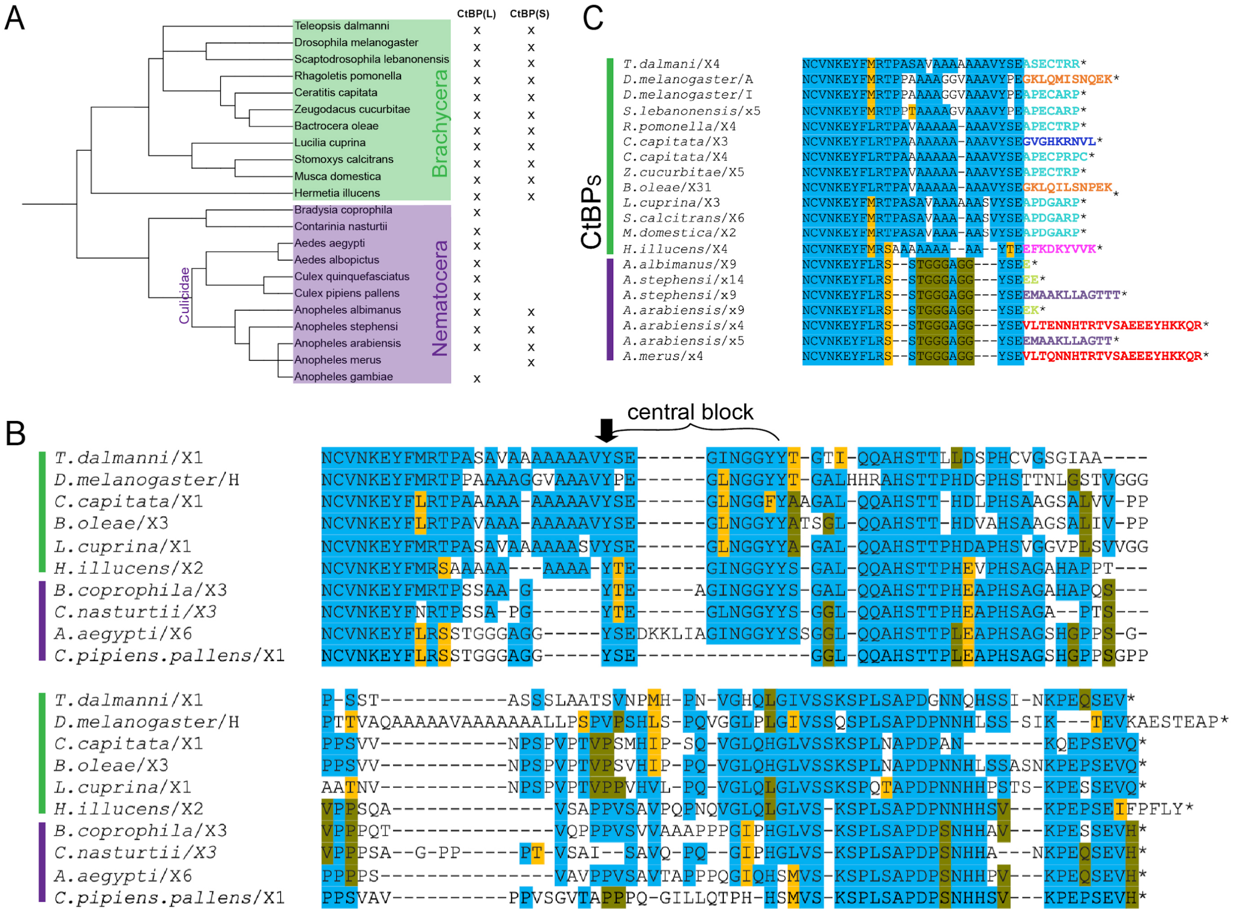


Figure 4.3. Dipterans express both long and short isoforms, with a diversity of short forms. A) Phylogenetic tree showing evolutionary relationship of two Diptera groups: Brachycera and Nematocera. Presence (X) of long or short isoforms is indicated. **B)** Alignment of CtBP(L) in representative Diptera. Sequences from only two mosquito genomes are presented in this alignment, as isoforms from most other species encode novel sequences at the very C terminus (**Figure S4.3D**). The "central block" is indicated, as well as the conserved aromatic Y (black arrow). **C)** Alignment of CtBP(S) isoforms in Diptera. Brachycera encode the conserved APECARP and SNQEK, whereas Nematocera express short forms not seen in their close relatives. Colors for terminal residues indicate the diversity of short endings observed in Diptera (i.e. the variant ending in APECARP is in nine of the species, colored in light blue). Variants that are found in more than one species are colored the same.

Deep Conservation of Arthropod CtBP Structure, With Lineage-specific Modifications

We compared CTD sequences from CtBP genes across representative insect orders as well as from springtails, a related hexapod (**Figure 4.4A**). Hexapod CTD sequences exhibit a deeply conserved central block including the YPEGINGGYY and AHSTTPHD motifs, as well as proline-rich terminal region ending in SEVH (**Figure 4.4B**). The ancestral SEVH-like terminal region is conserved across all sampled insect orders other than Hymenoptera, which instead feature a glycine- and proline-rich terminal sequence unique to this order. Interestingly, the springtail (*Folsomia candida*) CTD terminates just beyond the conserved central block, highlighting two lineages in the hexapods for which the terminal regions have been remodeled. Lineage-specific "spacers" rich in alanine, glycine, and proline separate the conserved central block from more N- and C-terminal residues in Blattodea, Hemiptera, Lepidoptera, Diptera, and Hymenoptera (**Figure 4.4B**). CtBP(S) isoforms are not unique to Diptera; hymenopteran isoforms also encode putative short variants (**Figure 4.4C**). Within Hymenoptera, alternative splicing produces the conserved order-specific RLSSRC short terminal sequence. The production of short variants appears to have arisen independently in these two orders.

A comparison of conserved hexapod sequences with those of crustaceans, myriapods, and chelicerates, which altogether make up the arthropod phylum, reveals that the central and terminal conserved regions noted in hexapods are generally conserved across arthropods (**Figure 4.5**). Sequences from representative species from these four groups demonstrate that four key motifs (NCVNKEY followed by an aromatic, ΨNGGYY (central block), AHSTT, and PEPSEVH) are present in all lineages, indicating that they are derived from an ancestral CtBP. From the two myriapod genomes available, no large deviations from the consensus are found. However, in crustaceans (shrimp, barnacle, and planktonic crustaceans), considerable variation is found in terminal sequence regions for all three classes analyzed (**Figure 4.5B**, **S4.4B**). The ancestral SEVH terminus is found only in *Pollicipes pollicipes* (Gooseneck barnacle).

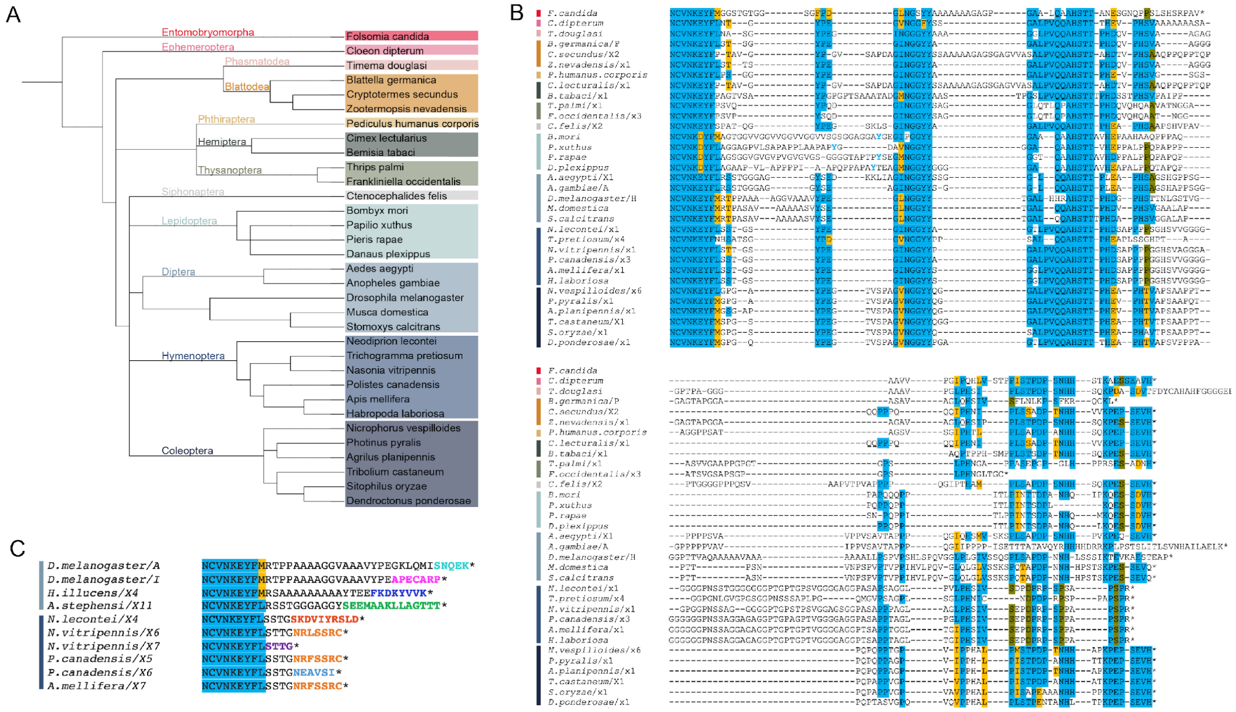


Figure 4.4. Hexapods express long forms of CtBP, with certain lineages producing short variants. A) Phylogenetic tree of all species analyzed in Hexapoda. Colored vertical bars in B and C correspond to orders indicated in A. **B)** Alignment of long CTDs from all Hexapod species analyzed. Certain motifs are conserved across all orders. In Hymenoptera, YG/S/TE residues within a variable region are highlighted in blue lettering to indicate presumed conservation. The *Timema douglasi* sequence extends another 47 residues past what is shown. **C)** Alignment of short CTDs from Diptera and Hymenoptera indicates that within Hymenoptera, some short endings are conserved, but are distinct from sequences of Diptera. Other Insecta orders do not have short endings. Variants that are found in more than one species are colored the same (i.e. the variant ending in SSRC is found in both *Nasonia vitripennis* and *Apis mellifera*, colored in orange).

An interesting finding comes from consideration of chelicerate CtBP sequences. Most CtBP CTD sequences from this subphylum, which includes mites, ticks, scorpions, spiders, and horseshoe crabs, have clearly alignable motifs in central and terminal regions (**Figure S4D, 5B**). All the species sampled (**Figure S4.4C**) have only one long version of the CTD, with no indication that short isoforms are produced. For most species, very few differences in the CTD sequences are present; the conserved blocks are not separated by repeat expansions noted in some insect orders, and sequences terminate with the same SEVH motif observed in the hexapods (**Figure 4.5B**).

Three species from the order Mesostigmata, which includes predatory and parasitic mites, share a CTD that is entirely dissimilar to other arthropod sequences (**Figure S4.4E**). The proline/alanine-rich CTDs of *Varroa destructor*, *Varroa jacobsoni*, and *Galendromus occidentalis* do not show compelling similarity to other chelicerate sequences, including those of more distantly related ticks (Ixodida) and dust mites (Sarcoptiformes). Thus, it is evident that the CTD of CtBP has undergone a wholesale replacement in the Mesostigmata lineage, which diverged from the order Ixodida ~300 Ma (Mans *et al*. 2016). The novel CTD is likely to be similarly disordered, based on sequence composition, but functional properties may have changed.

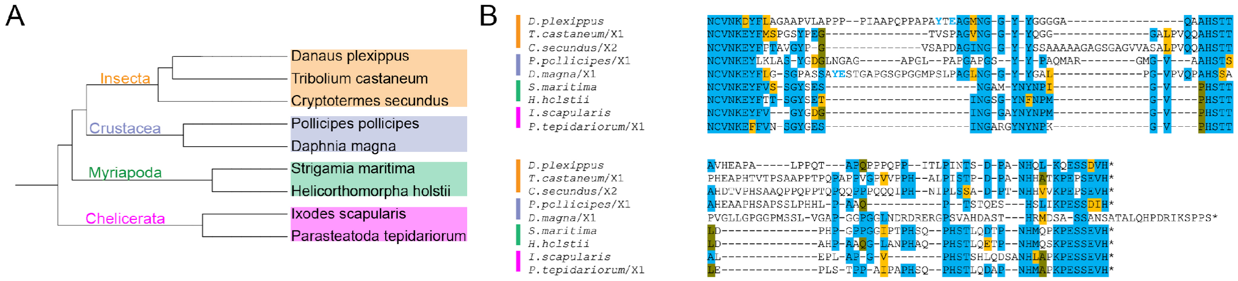

A
Insecta
Crustacea
Myriapoda
Chelicerata
Danaus plexippus
Tribolium castaneum
Cryptotermes secundus
Pollicipes pollicipes
Daphnia magna
Strigamia maritima
Helicorthomorpha holstii
Ixodes scapularis
Parasteatoda tepidariorum
B
D.plexippus NCVNKDYFLAGAAPVLAPPP--PIAAPQPPAPAYTEAGMNG-G-Y-YGGGGA----------------QAAHSTT
T.castaneum/X1 NCVNKEYFMSPGSYPEG--------------TVSPAGVNG-G-Y-YQGG-------------GALPVQQAHSTT
C.secundus/X2 NCVNKEYFPTAVGYP-G-------------VSAPDAGING-G-Y-YSSAAAAAGAGSGAGVVASALPVQQAHSTT
P.pcllicipes/X1 NCVNKEYLKLAS-YGDGLNGAG------APGL--PAPGAPGS--Y-Y-PAQMAR--------GMG-V---AAHSTS
D.magna/X1 NCVNKEYFLG-SGPASSAYESTGAPGSGPGGMPSLPAGLNG-G-Y-YGAL-------------PG-VPVQPAHSSA
S.maritima NCVNKEYFVS-SGYSES---------------------NGAM-YNYNPI-------------G-V----PHSTT
H.hclstii NCVNKEYFTT-SGYSET--------------------INGSG-YNFNPM-------------G-V----PHSTT
I.scapularis NCVNKEYFV---GYGDG--------------------ING-GAYNYNPM-------------G-V----PHSTT
P.tepidariorum/X1 NCVNKEFFVN-SGYGES--------------------INGARGYNYNPK-------------G-V----PHSTT
D.plexippus AVHEAPA-----LPPQT---APQPPPQPP--ITLPINTS-D-PA-NHQL-KQESSDVH*
T.castaneum/X1 PHEAPHTVTPSAAPPTPQPAPPVGPVVPPH--ALPISTP-D-PA-NHHATKPEPSEVH*
C.secundus/X2 AHDTVPHSAAQPPQPPTQPQQPPPQQQIPH--NIPLSSA-D-PT-NHHVVKPEPSEVH*
P.pcllicipes/X1 AHEAAPHSAPSSLPHHL-P-AAQ-----------P--TSTQES---HSLIKPESSDIH*
D.magna/X1 PVGLLGPGGPMSSL-VGAP-GGPGGLNDRDRERGPSVAHDAST---HRMDSA-SSANSATALQHPDRIKSPPS*
S.maritima LD--------------PHP-GPPGGIPTPHSQ--PHSTLQDTP--NHMQPKPESSEVH*
H.hclstii LD--------------AHP-AAQG-LANPHAQ--PHSTLQETP--NHMQSKPESSEVH*
I.scapularis AL--------------EPL-AP-G-V--------PHSTSHLQDSANHLAPKPESSEVH*
P.tepidariorum/X1 LE--------------PLS-TPP-AIPAPHSQ--PHSTLQDAP--NHMAPKPESSEVH*

Figure 4.5. Arthropod CTDs contain conserved motifs including ancestral ending. A) Phylogenetic tree of representatives of the four major arthropod groups. **B)** Alignment of representative species illustrates that motifs seen across Diptera are conserved within these groups of arthropods. Vertical bars on the left represent the lineages in A. The tyrosines in light blue (found in *Danaus plexippus* and *Daphnia magna*) indicate there is a conserved aromatic residue found between the NCVN and NGGYY motifs, but spaced slightly differently in these two species.

Diversification in Protostomia

Within other ecdysozoan lineages, the CtBP CTD of the velvet worm (Onychophora) was substantially similar to the consensus arthropod sequence (**Figure 4.6A, B**). Similarly, the CTD from the priapulid *Priapulus caudatus* provides another example of a non-arthropod ecdysozoan with highly similar CTD (**Figure 4.6A**, **B**). However, wholesale changes were found for the tardigrade CTD, where clear homology ends just after the start of the conserved central block.

This alternative CTD features poly-asparagine and multiple polyalanine stretches to generate a sequence slightly longer than those in many arthropods (**Figure 4.6A, B**). In Nematoda (roundworm), multiple lineage- specific forms of the CTD were identified that bore no close similarity to the previously identified conserved elements in arthropods (**Figure S4.5, 6A**). Notably, the NAD- binding core of these proteins showed high conservation with invertebrate sequences (~60%), indicating that evolutionary changes are focused on the CTD. Sequence alignments from ten roundworm species from the orders Rhabditida and Trichinnelida showed at least three distinct primary structures (**Figure S4.5**). In the nematodes, aside from the Caenorhabditis worms, an aromatic residue (F) universally found near the N-terminal portion of the CTD is also present. Although the Caenorhabditis worms lack this feature, they have an LNMGF motif that is present at approximately the same position as the conserved central block in arthropods. In short sequence blocks, some level of similarity is present in Rhabditida, especially within Caenorhabditis (**Figure S4.5B**). Only CtBP(L) isoforms were identified in the nematodes, with the Trichinella species having the longest CTDs (when compared with all other Ecdysozoa), ranging from 200 to 720 amino acids, in some cases virtually doubling the size of the CtBP protein.

To better understand what structural features of the CtBP CTD may be generally conserved in protostomes, we examined CtBP sequences from the morphologically diverse clade Spiralia, including mollusks, annelids, flatworms, and other taxa (**Figure S4.6A**). Species from six selected phyla, excepting Platyhelminthes, share core conserved motifs found in the consensus ecdysozoan CTD (**Figure S4.6D**). A striking exception was found in certain annelids; the leeches (Hirudinea) lack a C-terminal extension entirely (**Figure S4.6E**). This represents the only animal lineage that appears to lack a long form of CtBP. Polychaete annelids express CtBP with a CTD containing homology to the central block and terminal core conserved motifs, whereas earthworms

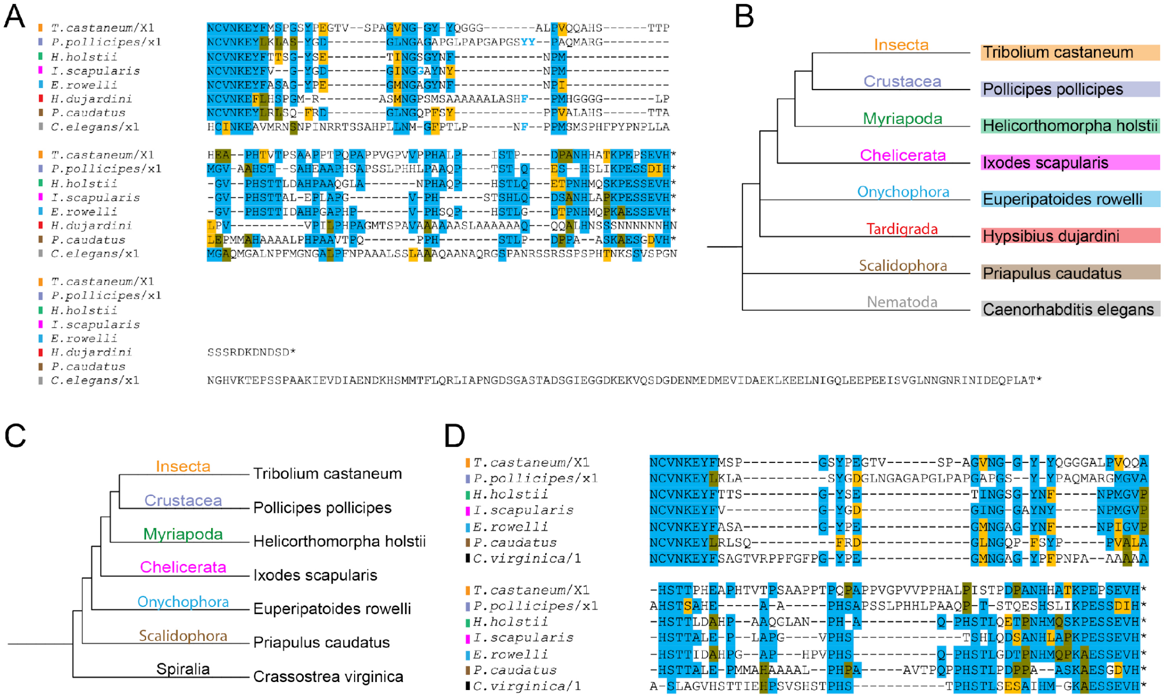


Figure 4.6. Comparative CtBP CTD alignments across Ecdysozoa and Protostomia. A) Alignment of CtBP sequences from representative Ecdysozoa shows conservation of central block and C-terminal sequences in most lineages. Unique and completely divergent sequences are found in tardigrades and nematodes. The residues in light blue indicate there is a conserved NGGYY-like sequence, but spaced slightly differently in these species. **B)** Phylogenetic tree of representative ecdysozoan species used in panel A. **C)** Phylogenetic tree of representative species from Protostomia (Ecdysozoa and Spiralia) that have a canonical CTD with conserved motifs. **D)** Alignment of representative protostomes shown in C illustrates the conservation of particular motifs across these invertebrates, including the central block and most C-terminal portion of the CTD.

(*Lumbricus rubellus* and *Eisenia fetida*) and the oligochaete *Olavius algarvensis* bear shorter CTD sequences with small regions of sequence similarity to the protostome consensus (**Figure S4.6E**). Representatives of Nemertea (*Notospermus geniculatus*; ribbon worm) and Phoronida (*Phoronis australis*; horseshoe worm) showed strong conservation in central and terminal sequences, with minor variations (**Figure S4.6D**). Interestingly, the rotifer (*Adineta vaga*) CTD has recognizable homologies through the central block, and then is sharply divergent from other protostomes, a pattern of variation resembling that of the order Hymenoptera in insects (**Figure S4.6D, 4.4B**). The Platyhelminthes have unique proline and polyalanine-rich CTD sequences that do not resemble those of other species. Interestingly, there is considerable diversity within the Platyhelminthes phylum; there are weakly alignable blocks within trematode, cestode, and monogeneid CTDs, whereas CTD sequences of triclad planaria form a separate homology set (**Figure S4.6B, C**). Platyhelminth CTDs range in size from 150 to 550 residues, formed by addition of novel residues to the terminus.

Overall, deep conservation of the CTD of CtBP within Protostomia is punctuated by rapid evolution in this domain in certain lineages (**Figure 4.6A, D**). The chemical nature and size of the typical CTD sequence are generally conserved (**Figure S4.10A, B**). Only the leech appears to have done away with the CTD entirely, but various arthropods have devised splice variants that presumably allow for facultative expression of a short form, as in Drosophila. Despite occasional bursts of evolution lengthening the CTD, the CtBP proteins are clear homologs and ~80% of the protein (~400 positions) is alignable across diverse protostomes (**Supplementary file 3, online**). Interestingly, even when only alignable positions are considered, the sequence of the nematode CtBP shows greater divergence from sequences of other protostomes suggesting that a lengthened CTD may change functional constraint across the protein (**Figure S4.15**). The significance of these massive alternative CTDs (>500 residues) remains obscure.

Conservation of CTD Sequences Between Protostomes and Deuterostomes

In contrast to a single CtBP gene found across Protostomia, mammals have two CtBP paralogs, CtBP1 and CtBP2, which have both overlapping and unique functions in development (Katsanis and Fisher 1998; Hildebrand and Soriano 2002). CtBP1 null mice are viable but have developmental phenotypes, whereas CtBP2 null mice are embryonic lethal (Hildebrand and

Soriano 2002). The human CtBP1 and CtBP2 proteins exhibit ~90% conservation in the dehydrogenase core, with most of the remaining 10% reflecting chemically conserved substitutions. Interestingly, the CTD itself has more variation, with only 50% of the primary sequence being conserved between the two human paralogs (**Figure S4.7A**). We can conclude, however, that the CTD sequences are derived from a common ancestor. Specific motifs in the CTD show stronger conservation, such as a central PELNGAxYRY motif and the aromatic residue (W) situated near the N-terminal region of the CTD, both of which are conserved in the protostomes (**Figure S4.7A, 4.6D**). The charged residues (one basic/two acidic) at the very terminus also appear to represent conserved features, whereas various alignable prolines are less compelling as evidence of homology for these overall proline-rich sequences. Deeply conserved AHSTT and PHS–PHS motifs located between the central motif and terminus in many protostomes are not conserved in the human CTDs (**Figure S4.7B**), but the overall length of the CTDs (~100 residues) is similar to those of representative protostomes (**Figure S4.10A**). Overall, the similarities argue for a common CTD sequence shared by the last common ancestor of protostomes and deuterostomes, a feature that was not apparent when only a few CtBP genes were available such as the *C. elegans* CtBP with its highly derived CTD.

To better understand evolutionary processes in deuterostomes, we turned to genomes of species representing echinoderms, acorn worms (hemichordates), and non-vertebrate chordates, including tunicates (urochordates) and lancelets (cephalochordates) (**Figure S4.7D**). In contrast to mammals, only a single CtBP gene is found in these species, as in protostomes. These CtBP CTDs are clearly homologous; they share a lone tryptophan toward the N-terminus, adjacent to the central block PELNGxYRY (similar to central block from protostomes), more lineage-specific blocks, and a highly conserved terminus with one basic and two acid residues in conserved spacing

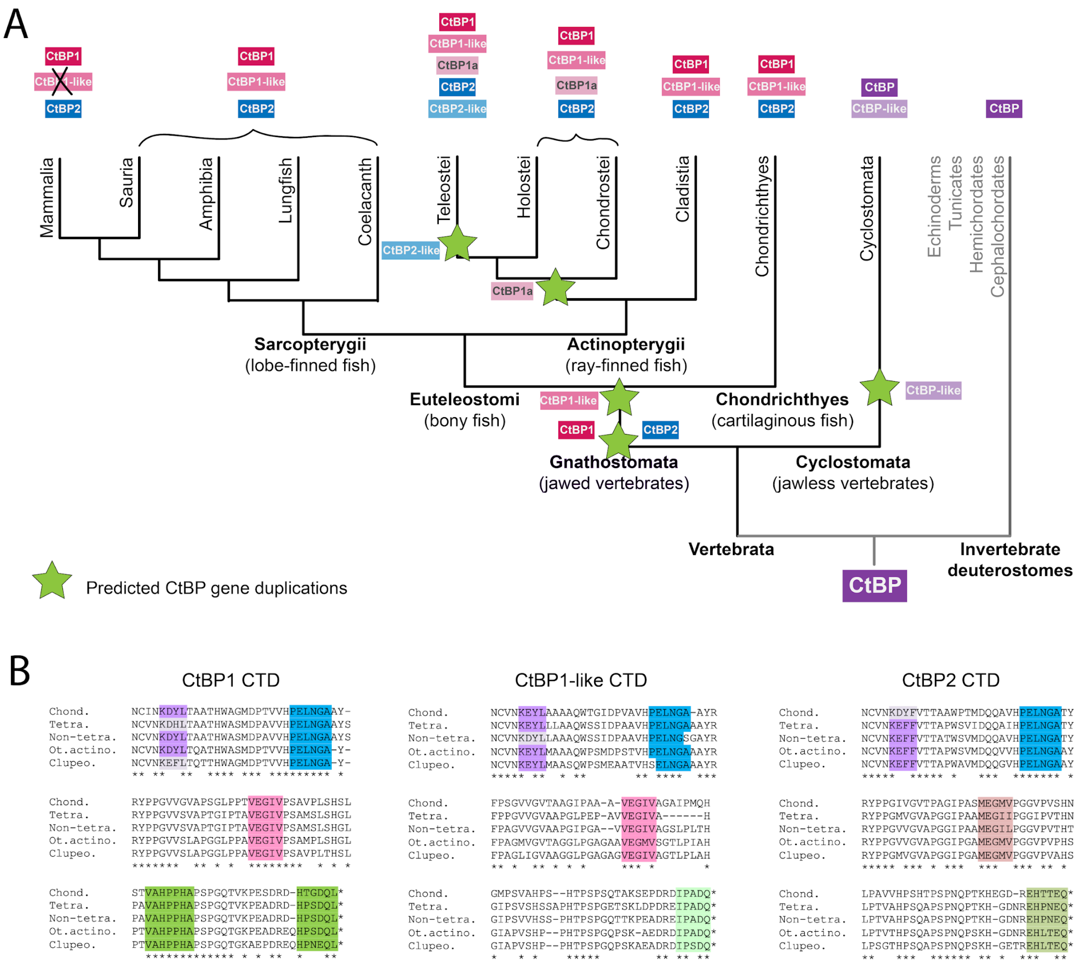
(**Figure S4.7C**). Insertions between these more conserved regions are present in the two tunicate CtBP sequences, generating a longer CTD.

Vertebrates Encode Multiple CtBP Genes

To better understand the molecular transformations that occurred as vertebrates diversified from the last common ancestor of other deuterostomes, we analyzed CtBP isoforms in Vertebrata, where multiple rounds of whole-genome duplications have increased the number of many paralogous genes (**Figure 4.7A**). In vertebrate genome annotations, paralogous genes are based on presumed similarities to mammalian CtBP1 or CtBP2 paralogs; however, we find that these designations are in some cases inaccurate, based on the presence of highly conserved residues characteristic of one or the other paralog. We prepared a systematic set of criteria to reliably designate a paralog CtBP1-like, CtBP1a, or CtBP2-like (see Materials and Methods; **Figure 4.7B**).

In the lamprey (Cyclostomata, a jawless fish), two paralogs are found which we have named CtBP and CtBP-like (**Figure 4.8**). The lamprey CtBP is more similar to the vertebrate CtBP1 and CtBP2 than to its own paralog, showing ≥85% identity to the vertebrate proteins across the dehydrogenase core. The lamprey CtBP CTD and vertebrate CTD sequences are likewise very similar. In contrast, the lamprey CtBP-like CTD, whereas clearly derived from the canonical ancestral sequence, is less similar, and contains insertions between conserved blocks. This evidence suggests that CtBP-like is derived from an independent duplication of the single CtBP gene in jawless fish.

In contrast, species of cartilaginous fish (Chondrichthyes) encode CtBP1, CtBP1-like, and CtBP2. The CtBP paralogs in this ancient fish lineage may have originated during basal whole-genome duplication events in the jawed fish (Gnathostomata) (**Figure 4.7A, 4.9**). The CTDs from Chondrichthyes are similar to the lamprey CtBP CTD, but differ greatly from the lamprey CtBP-



CtBP1 CTD

```
Chond.        NCINKDYLTAATHWAGMDPTVVHPELNGAAY-
Tetra.        NCVNKDHLTAATHWASMDPAVVHPELNGAAYS
Non-tetra.    NCVNKDYLTAATHWASMDPAVVHPELNGAAYS
Ot.actino.    NCVNKDYLTQATHWASMDPTVVHPELNGA-Y-
Clupeo.       NCVNKEFLTQTTHWAGMDPTVVHPELNGA-Y-
              ** **  **  **** *** ********* *

Chond.        RYPPGVVGVAPSGLPPTVEGIVPSAVPLSHSL
Tetra.        RYPPGVVSVAPTGIPTAVEGIVPSAMSLSHGL
Non-tetra.    RYPPGVVSVAPTGIPTAVEGIVPSAMSLSHGL
Ot.actino.    RYPPGVVSLAPGGLPPAVEGIVPSAMPLSHGL
Clupeo.       RYPPGVVSLAPGGLPPAVEGIVPSAVPLTHSL
              *******  ** * *  ********  * * *

Chond.        STVAHPPHAPSPGQTVKPESDRD-HTGDQL*
Tetra.        PAVAHPPHAPSPGQTVKPEADRD-HPSDQL*
Non-tetra.    PAVAHPPHAPSPGQTVKPEADRD-HPSDQL*
Ot.actino.    PTVAHPPHAPSPGQTGKPEADREQHPSDQL*
Clupeo.       PTVAHPPHAPSPGQTGKAEPDREQHPNEQL*
                ************* * * **  *   **
```

CtBP1-like CTD

```
Chond.        NCVNKEYLAAAAQWTGIDPVAVHPELNGA-AYR
Tetra.        NCVNKEYLLLAAQWSSIDPAAVHPELNGAAAYR
Non-tetra.    NCVNKDYLLAAAQWSSIDPAAVHPELNGSGAYR
Ot.actino.    NCVNKEYLMAAAQWPSMDPSTVHPELNGAAAYR
Clupeo.       NCVNKEYLMAASQWPSMEAATVHSELNGAAAYR
              ***** **  * **       ** ****  ***

Chond.        FPSGVVGVTAAGIPAA-A-VEGIVAGAIPMQH
Tetra.        FPPGVVGVAAPGLPEP-AVVEGIVA------H
Non-tetra.    FPAGVVGVAAPGIPGA--VVEGIVAGSLPLTH
Ot.actino.    FPAGMVGVTAGGLPGAGAAVEGMVSGTLPIAH
Clupeo.       FPAGLIGVAAGGLPGAGAGVEGIVAGTLPLAH
              ** *  ** * * *     *** *       *

Chond.        GMPSVAHPS--HTPSPSQTAKSEPDRDIPADQ*
Tetra.        GIPSVSHSSAPHTPSPGETSKLDPDREIPADQ*
Non-tetra.    GIPSVVHSS-PHTPSPGQTTKPEADREIPADQ*
Ot.actino.    GIAPVSHP--PHTPSPGQPSK-AEDRDIPADQ*
Clupeo.       GIAPVSHP--PHTPSPGQPSKAEADRDIPSDQ*
              *   * *    *****        ** ** **
```

CtBP2 CTD

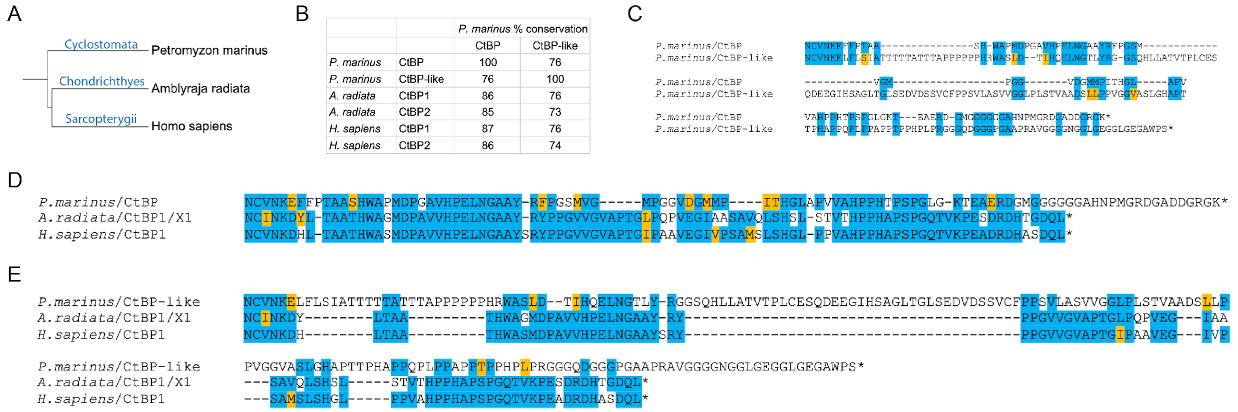
```
Chond.        NCVNKDYFVTTAAWPTMDQQAVHPELNGATY
Tetra.        NCVNKEFFVTTAPWSVIDQQAIHPELNGATY
Non-tetra.    NCVNKEFFVTTATWSVMDQQAVHPELNGATY
Ot.actino.    NCVNKEFFVTTAPWSVMDQQVVHPELNGAAY
Clupeo.       NCVNKEFFVTTAPWAVMDQQGVHPELNGAAY
              *****  ***** *   ***  ******* *

Chond.        RYPPGIVGVTPAGIPASMEGMVPGGVPVSHN
Tetra.        RYPPGMVGVAPGGIPAAMEGIIPGGIPVTHN
Non-tetra.    RYPPGVVGVAPGGIPAAMEGILPGGVPVTHT
Ot.actino.    RYPPGVVGVAPGGIPGAMEGMVPGGVPVPHS
Clupeo.       RYPPGMVGVAPGGIPGAMEGMVPGGVPVAHS
              ***** *** * ***  ***  *** ** *

Chond.        LPAVVHPSHTPSPNQPTKHEGD-REHTTEQ*
Tetra.        LPTVAHPSQAPSPNQPTKH-GDNREHPNEQ*
Non-tetra.    LPTVAHPSQAPSPNQPTKH-GDNREHPNEQ*
Ot.actino.    LPTVTHPSQAPSPNQPSKH-GDNREHLTEQ*
Clupeo.       LPSGTHPSQAPSPNQPSKH-GETREHLTEQ*
              **   ***  ****** ** *   ***  **
```

Figure 4.7. CtBP paralogs in vertebrates. A) All vertebrates express two or more CtBP genes. Based on sequence similarities, an independent duplication (green star) is suggested to have happened in Cyclostomata, whereas three paralogs (CtBP1, CtBP1-like, and CtBP2) originated in an ancestor to jawed vertebrates, possibly generated through whole-genome or independent gene duplication events. Loss of the CtBP1-like gene occurred only in mammals, whereas additional gene duplications occurred at different times in ray-finned fish. **B)** Alignment of representative Gnathostomata sequences of the CtBP1, CtBP1-like, and CtBP2 CTDs. The sequences do not represent those of a particular species, but rather a consensus that illustrates the representative CTD for that particular clade. For all alignments: Chond. (Chondrichthyes), Tetra. (Tetrapods), Non-tetra. (Non-tetrapod sarcopterygians, including lungfish and coelacanth), Ot. actino (Other actinopterygii, includes Cladistia, Chondrostei, Holostei and non-Clupeocephala Teleostei), Cupleo. (Clupeocephala, includes some Teleostei like zebrafish, pufferfish, and northern pike). Asterisks on the bottom indicate complete conservation of a particular residue. Purple highlighting indicates a region characteristic of a particular paralog grouping, and light purple indicates a lineage-specific derivation. Blue highlight indicates a sequence that's conserved across all clades, across all paralogs. Pink highlight indicates a motif unique to CtBP1 and 1-like, but that differs in CtBP2. Green highlight indicates a motif that is highly conserved within each protein family, and is representative of that protein, but not of the other paralogs.

like CTD (**Figure 4.8D, E**). Chondrichthyes are the first lineage in which we find expression of three different CtBP proteins, with conservation of the third, CtBP1-like, across the selected species. Interestingly, short isoforms of CtBP1 and CtBP2 are reported in some of the species, suggesting that formation of a CtBP(S) isoform arose independently in these vertebrates, similar to what was observed in certain insect orders (**Figure 4.9B**).



		P. marinus % conservation	
		CtBP	CtBP-like
P. marinus	CtBP	100	76
P. marinus	CtBP-like	76	100
A. radiata	CtBP1	86	76
A. radiata	CtBP2	85	73
H. sapiens	CtBP1	87	76
H. sapiens	CtBP2	86	74

```
P.marinus/CtBP        NCVNKEFFPTAA---------------SH-WAPMDPGAVHPELNGAAYRFPGSM------------
P.marinus/CtBP-like   NCVNKELFLSIATTTTTATTTAPPPPPPHRWASLD--TIHQELNGTLYRG-GSQHLLATVTPLCES

P.marinus/CtBP        -----------VGM-----------------PGG-------VDGMMPITHGL-----APV
P.marinus/CtBP-like   QDEEGIHSAGLTGLSEDVDSSVCFPPSVLASVVGGLPLSTVAADSLLPPVGGVASLGHAPT

P.marinus/CtBP        VAHPPHTPSPGLGKT---EAERD-GMGGGGGAHNPMGRDGADDGRGK*
P.marinus/CtBP-like   TPHAPPQPLPPAPPTPPHPLPRGGGQDGGGPGAAPRAVGGGGNGGLGEGGLGEGAWPS*
```

```
P.marinus/CtBP        NCVNKEFFPTAASHWAPMDPGAVHPELNGAAY-RFPGSMVG-----MPGGVDGMMP----ITHGLAPVVAHPPHTPSPGLG-KTEAERDGMGGGGGGAHNPMGRDGADDGRGK*
A.radiata/CtBP1/X1    NCINKDYL-TAATHWAGMDPAVVHPELNGAAY-RYPPGVVGVAPTGLPQPVEGIAASAVQLSHSL-STVTHPPHAPSPGQTVKPESDRDHTGDQL*
H.sapiens/CtBP1       NCVNKDHL-TAATHWASMDPAVVHPELNGAAYSRYPPGVVGVAPTGIPAAVEGIVPSAMSLSHGL-PPVAHPPHAPSPGQTVKPEADRDHASDQL*
```

```
P.marinus/CtBP-like   NCVNKELFLSIATTTTTATTTAPPPPPPHRWASLD--TIHQELNGTLY-RGGSQHLLATVTPLCESQDEEGIHSAGLTGLSEDVDSSVCFPPSVLASVVGGLPLSTVAADSLLP
A.radiata/CtBP1/X1    NCINKDY--------LTAA---------THWAGMDPAVVHPELNGAAY-RY--------------------------------------PPGVVGVAPTGLPQPVEG---IAA
H.sapiens/CtBP1       NCVNKDH--------LTAA---------THWASMDPAVVHPELNGAAYSRY--------------------------------------PPGVVGVAPTGIPAAVEG---IVP

P.marinus/CtBP-like   PVGGVASLGHAPTTPHAPPQPLPPAPPTPPHPLPRGGGQDGGGPGAAPRAVGGGGNGGLGEGGLGEGAWPS*
A.radiata/CtBP1/X1    ---SAVQLSHSL-----STVTHPPHAPSPGQTVKPESDRDHTGDQL*
H.sapiens/CtBP1       ---SAMSLSHGL-----PPVAHPPHAPSPGQTVKPEADRDHASDQL*
```

Figure 4.8. Cyclostomata CtBP CTD sequences differ from those of other deuterostomes. A) Phylogenetic tree representing the relationship between the lamprey (Cyclostomata, basal vertebrates), thorny skate (Chondrichthyes), and human (Sarcopterygii). **B)** Comparison of the percent conservation of the dehydrogenase core of the lamprey (*Petromyzon marinus*) CtBP and CtBP-like to CtBP1 and CtBP2 from a representative Chondrichthyes (*Amblyraja radiata*) and Sarcopterygii (*Homo sapiens*). Numbers indicate percentage of completely conserved residues including and between the RPLVALL and NCVN motifs. The lamprey CtBP has a higher degree of similarity to vertebrate CtBP1 and CtBP2 than does the lamprey CtBP-like paralog. CtBP-like may have originated as a duplication specific to the lamprey, and then diverged within this lineage. **C)** Alignment of the lamprey CtBP and CtBP-like CTDs indicates low conservation, and differences in CTD length. **D)** Alignment of the lamprey CtBP CTD with that of representative jawed vertebrates' CtBP1 CTD. The terminal residues of lamprey CtBP show a derived extension, with otherwise high level of similarity. (E) Alignment of the lamprey CtBP-like CTD with representative vertebrates' CtBP1 CTD. Residues more C-terminal to the central block motif constitute a much longer sequence that appears to be derived in this lineage.

An examination of representative species from the two groups of bony fish (Euteleostomi) reveals additional changes in CtBP gene copy number. In the lobe-finned fish (Sarcopterygii), extant species have homologs of CtBP1, CtBP1-like, and CtBP2 as found in the ancestral

Chondrichthyes, with retention in most tetrapods. The CtBP1-like paralog is lost solely in mammals (**Figure 4.7A**). In ray-finned fish (Actinopterygii), additional CtBP1a and CtBP2-like genes are found. Teleost-specific gene duplication events are associated with up to five CtBP genes in certain lineages (**Figure 4.7**).

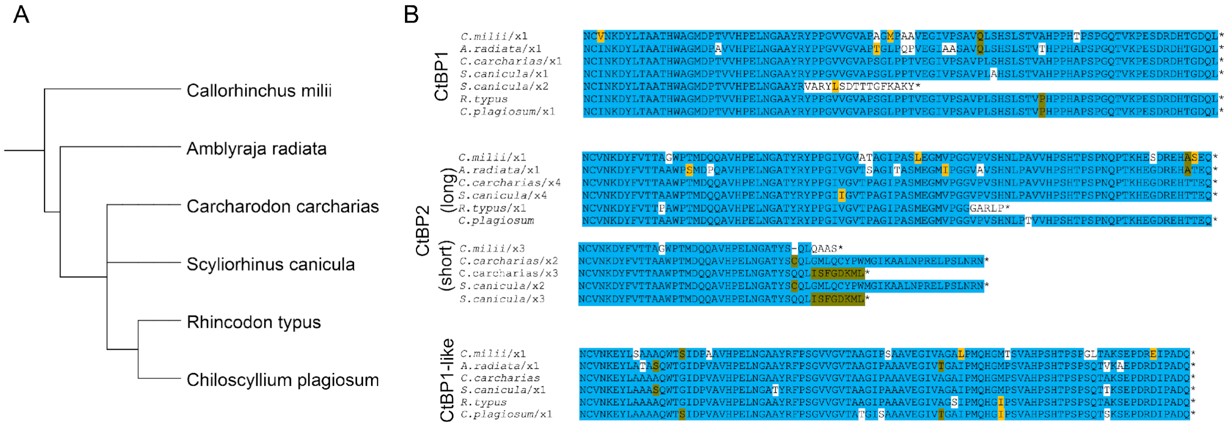

A
Callorhinchus milii
Amblyraja radiata
Carcharodon carcharias
Scyliorhinus canicula
Rhincodon typus
Chiloscyllium plagiosum
B
CtBP1
C.milii/x1 NCVNKDYLTAATHWAGMDPTVVHPELNGAAYRYPPGVVGVAPAGMPAAVEGIVPSAVQLSHSLSTVAHPPHTPSPGQTVKPESDRDHTGDQL*
A.radiata/x1 NCINKDYLTAATHWAGMDPAVVHPELNGAAYRYPPGVVGVAPTGLPQPVEGIAASAVQLSHSLSTVTHPPHAPSPGQTVKPESDRDHTGDQL*
C.carcharias/x1 NCINKDYLTAATHWAGMDPTVVHPELNGAAYRYPPGVVGVAPSGLPPTVEGIVPSAVPLSHSLSTVAHPPHAPSPGQTVKPESDRDHTGDQL*
S.canicula/x1 NCINKDYLTAATHWAGMDPTVVHPELNGAAYRYPPGVVGVAPSGLPPTVEGIVPSAVPLAHSLSTVAHPPHAPSPGQTVKPESDRDHTGDQL*
S.canicula/x2 NCINKDYLTAATHWAGMDPTVVHPELNGAAYRVARYLSDTTTGFKAKY*
R.typus NCINKDYLTAATHWAGMDPTVVHPELNGAAYRYPPGVVGVAPSGLPPTVEGIVPSAVPLSHSLSTVPHPPHAPSPGQTVKPESDRDHTGDQL*
C.plagiosum/x1 NCINKDYLTAATHWAGMDPTVVHPELNGAAYRYPPGVVGVAPSGLPPTVEGIVPSAVPLSHSLSTVPHPPHAPSPGQTVKPESDRDHTGDQL*
CtBP2
(long)
C.milii/x1 NCVNKDYFVTTAGWPTMDQQAVHPELNGATYRYPPGIVGVATAGIPASLEGMVPGGVPVSHNLPAVVHPSHTPSPNQPTKHESDREHASEQ*
A.radiata/x1 NCVNKDYFVTTAAWPSMDPQAVHPELNGATYRYPPGIVGVTSAGITASMEGMIPGGVAVSHNLPAVVHPSHTPSPNQPTKHEGDREHATEQ*
C.carcharias/x4 NCVNKDYFVTTAAWPTMDQQAVHPELNGATYRYPPGIVGVTPAGIPASMEGMVPGGVPVSHNLPAVVHPSHTPSPNQPTKHEGDREHTTEQ*
S.canicula/x4 NCVNKDYFVTTAAWPTMDQQAVHPELNGATYRYPPGIIGVTPAGIPASMEGMVPGGVPVSHNLPAVVHPSHTPSPNQPTKHEGDREHTTEQ*
R.typus/x1 NCVNKDYFVTTPAWPTMDQQAVHPELNGATYRYPPGIVGVTPAGIPASMEGMVPGGGARLP*
C.plagiosum NCVNKDYFVTTAAWPTMDQQAVHPELNGATYRYPPGIVGVTPAGIPASMEGMVPGGVPVSHNLPTVVHPSHTPSPNQPTKHEGDREHTTEQ*
(short)
C.milii/x3 NCVNKDYFVTTAGWPTMDQQAVHPELNGATYS-QLQAAS*
C.carcharias/x2 NCVNKDYFVTTAAWPTMDQQAVHPELNGATYSCQLGMLQCYPWMGIKAALNPRELPSLNRN*
C.carcharias/x3 NCVNKDYFVTTAAWPTMDQQAVHPELNGATYSQQLISFGDKML*
S.canicula/x2 NCVNKDYFVTTAAWPTMDQQAVHPELNGATYSCQLGMLQCYPWMGIKAALNPRELPSLNRN*
S.canicula/x3 NCVNKDYFVTTAAWPTMDQQAVHPELNGATYSQQLISFGDKML*
CtBP1-like
C.milii/x1 NCVNKEYLSAAAQWTSIDPAAVHPELNGAAYRFPSGVVGVTAAGIPSAAVEGIVAGALPMQHGMTSVAHPSHTPSPGLTAKSEPDREIPADQ*
A.radiata/x1 NCVNKEYLATASQWTGIDPVAVHPELNGAAYRFPSGVVGVTAAGIPAAAVEGIVTGAIPMQHGMPSVAHPSHTPSPSQTVKAEPDRDIPADQ*
C.carcharias NCVNKEYLAAAAQWTGIDPVAVHPELNGAAYRFPSGVVGVTAAGIPAAAVEGIVAGAIPMQHGMPSVAHPSHTPSPSQTAKSEPDRDIPADQ*
S.canicula/x1 NCVNKEYLAAASQWTGIDPVAVHPELNGATYRFPSGVVGVTAAGIPAAAVEGIVAGAIPMQHGMPSVAHPSHTPSPSQTTKSEPDRDIPADQ*
R.typus NCVNKEYLAAAAQWTGIDPVAVHPELNGAAYRFPSGVVGVTAAGIPAAAVEGIVAGSIPMQHGIPSVAHPSHTPSPSQTAKSEPDRDIPADQ*
C.plagiosum/x1 NCVNKEYLAAAAQWTSIDPVAVHPELNGAAYRFPSGVVGVTATGISAAAVEGIVTGAIPMQHGIPSVAHPSHTPSPSQTSKSEPDRDIPADQ*

Figure 4.9. Chondrichthyes encode three CtBP genes. A) Phylogenetic tree of selected Chondrichthyes. **B)** Alignment of representative isoforms from each of six cartilaginous fish indicates that the CTDs are very highly conserved within CtBP1, CtBP2, and CtBP1-like, with short isoforms appearing in both CtBP1 and CtBP2.

Characteristic residues present in CtBP1, CtBP1-like, and CtBP2 across Gnathostomata reveal particular segments of the CTD that have undergone modifications at different evolutionary times. For instance, more recent derivations are represented by a tetrapod-specific change in the more N-terminal portion of the CtBP1 CTD from the ancestral KDYL to KDHL, whereas the same region underwent a conversion from KDYL to KEFL within Clupeocephala, a specific clade within Actinopterygii (purple highlight; **Figure 4.7B**). A more ancient derivation is observed in a comparable location in the CtBP2 CTD, where a KDYF motif is found in Chondrichthyes and a KEFF motif in all other bony fish and tetrapods. Certain motifs are unique to the specific CtBP paralogs, and are completely conserved across species; these include the very C-terminal sequences, which were also found to be highly conserved in protostomes (green highlight; **Figure 4.7B**). It is likely that these distinct motifs represent variations that arose relatively soon after CtBP gene duplication. Examples of motifs that are common to CtBP1 family paralogs include central VEGIV motifs, that are clearly related to, but distinct from, the somewhat less conserved CtBP2 MEGMV motif (pink highlight; **Figure 4.7B**). More ancient motifs such as PELNGA, appearing just N-terminal to deeply conserved aromatic residues (W) of the central block, appear to have been present in the last common ancestor of vertebrates and echinoderms (blue highlight; **Figure 4.7B, S4.7C**). To infer a phylogenetic history of the CtBP sequences, we assembled an alignment of homologous CtBP sequences from representative deuterostome and protostome species. Much like in protostomes, the deuterostome sequences are clearly alignable (>80% sites), despite the presence of regions with length variation (**Supplementary file S3, online**). We then inferred a maximum-likelihood phylogeny using the best-fit model of protein evolution (**Figure S4.15**). From this phylogeny, we inferred the timing of the gene duplications that created the paralogs found in modern vertebrate genomes. The gene duplications on the phylogeny clearly show when the paralogs originated on the vertebrate phylogeny, and are consistent with our proposed model of duplications (**Figure 4.7**). One deviation from the expected species tree was observed with the Cyclostomata CtBP sequences, which can be explained by two different evolutionary scenarios (**Figure S4.15**). The two sequences from lamprey are likely difficult to place because they are the only two sequences obtained from jawless vertebrates.

Conservation of CtBP Paralogs Across Sarcopterygii

An examination of CtBP paralogs in these vertebrate species reveals very different levels of evolutionary variation. The super class Sarcopterygii comprises the more basal lungfish (Dipnoi) and coelacanth (Coelacanthimorpha), as well as more recently derived tetrapods including mammals, birds, reptiles, and amphibians (**Figure 4.10A**). The sequences of both the CtBP1 and CtBP2 CTDs are very highly conserved (**Figure 4.10B, C, S4.8A, B**). We find evidence of some substitutions at specific sites, with the length and sequence having high conservation across all Sarcopterygii sampled, and much more within each particular class. Intriguingly, we found that the CtBP2 of amphibians diversified in the length and sequence (**Figure 4.10C, S4.8B**). Conserved truncated versions of the CTD observed in some amphibians terminate immediately C-terminal to the central block motif, suggesting that the first portion of the CTD, which includes highly conserved aromatic/hydrophobic residues, may possess a function that is conserved even in these variants (**Figure S4.8B**). These amphibians are the only Sarcopterygii to have modified the CtBP2 tail, whereas some Sauria also produce a second short variant (data not shown). In mammals, the only major deviation from the canonical sequence was found in bats (Chiroptera), which are the only order to have short isoforms of both CtBP1 and CtBP2 (**Figure S4.8D**). The third gene in Sarcopterygii, CtBP1-like, which was lost solely in mammals, has more variation than the other paralogs (**Figure 4.10D, S4.8C**). Short variants of CtBP1-like exist in Sauria as well (data not shown). Over the course of ~400 My of evolution of Sarcopterygii, we find very high conservation of the CtBP CTD, suggesting that in Sarcopterygii, conservation of this sequence is critical for function.

Actinopterygii Express up to Five CtBP Genes

Actinopterygii are the second and most speciose branch of the Euteleostomi clade of bony vertebrates. In Actinopterygii, additional CtBP paralogs have arisen, likely through the whole-genome duplications documented in fish, including the Teleost-specific Genome Duplication that occurred 225–333 Ma (Berthelot *et al.* 2014). We selected sixteen fish that cover all major groups of ray-finned fishes, including many teleost fish and some more basal, ancient fish such as the bichir (*Polypterus senegalus*), paddlefish (*Polyodon spathula*), and gar (*Lepisosteus oculatus*). Up to five unique CtBP proteins were found in select species including some Teleostei (**Figure 4.11A**, **B**). All of these ray-finned fishes express the canonical CtBP1 and CtBP2 paralogs, which differ slightly in their CTDs, but are very highly conserved (**Figure 4.11C**). We found evidence for short isoforms in some Actinopterygii; the CtBP1-short CTDs are highly conserved, whereas the CtBP2-short CTD sequences are characterized by a greater degree of variation (**Figure 4.11D**). Core motifs, such as the NCVNKEY at the beginning of the CTD and the conserved central block, are present in all isoforms analyzed.

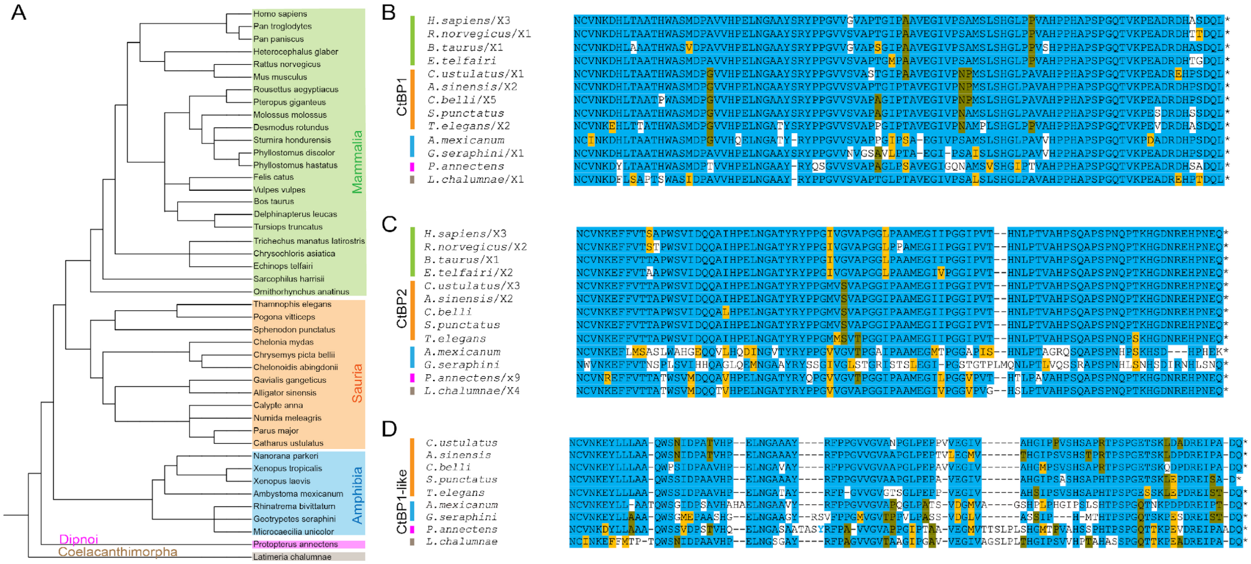
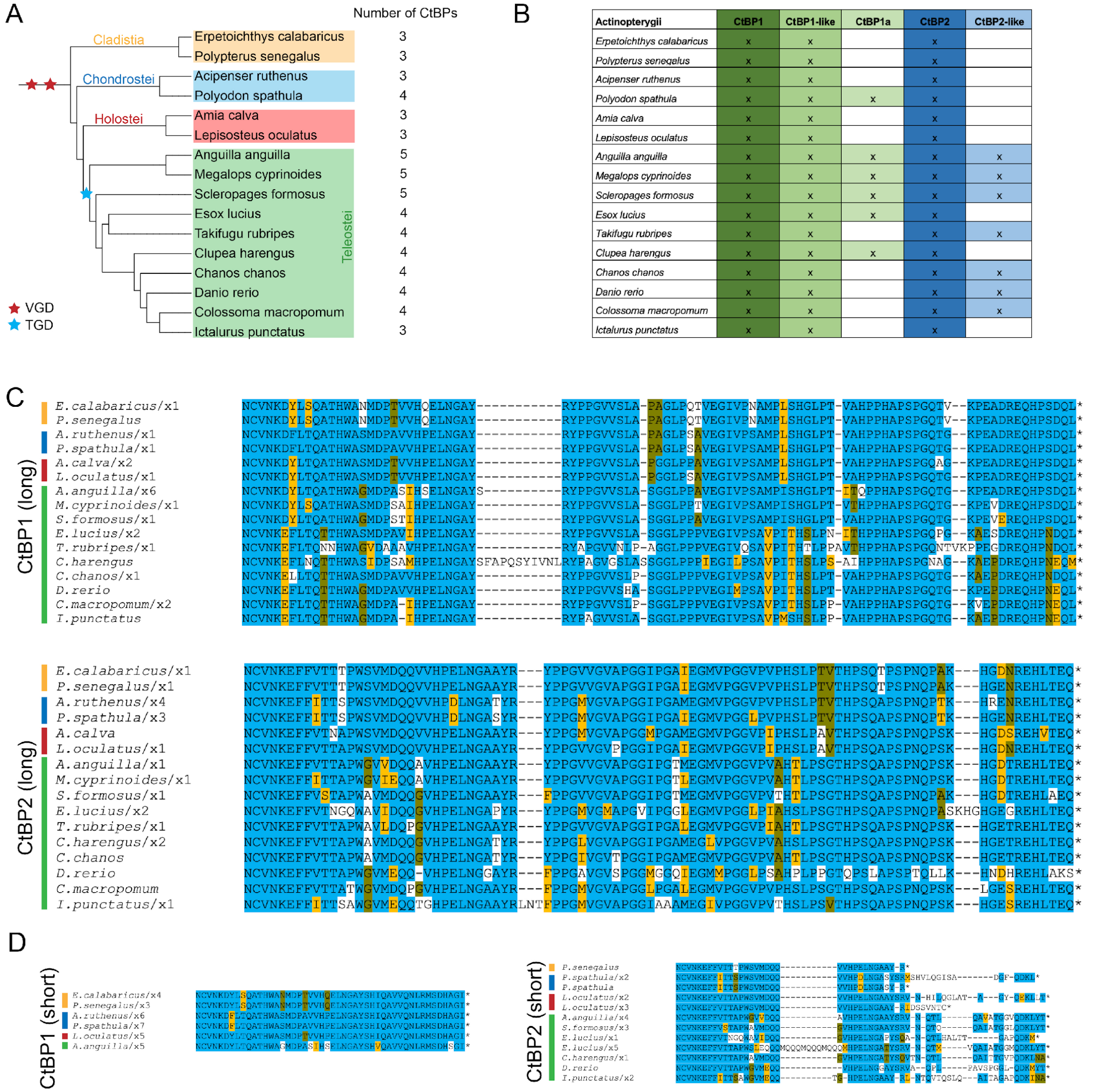

A
Homo sapiens
Pan troglodytes
Pan paniscus
Heterocephalus glaber
Rattus norvegicus
Mus musculus
Rousettus aegyptiacus
Pteropus giganteus
Molossus molossus
Desmodus rotundus
Sturnira hondurensis
Phyllostomus discolor
Phyllostomus hastatus
Felis catus
Vulpes vulpes
Bos taurus
Delphinapterus leucas
Tursiops truncatus
Trichechus manatus latirostris
Chrysochloris asiatica
Echinops telfairi
Sarcophilus harrisii
Ornithorhynchus anatinus
Mammalia
Thamnophis elegans
Pogona vitticeps
Sphenodon punctatus
Chelonia mydas
Chrysemys picta bellii
Chelonoidis abingdonii
Gavialis gangeticus
Alligator sinensis
Calypte anna
Numida meleagris
Parus major
Catharus ustulatus
Sauria
Nanorana parkeri
Xenopus tropicalis
Xenopus laevis
Ambystoma mexicanum
Rhinatrema bivittatum
Geotrypetes seraphini
Microcaecilia unicolor
Amphibia
Dipnoi
Protopterus annectens
Coelacanthimorpha
Latimeria chalumnae
B
CtBP1
H.sapiens/X3 NCVNKDHLTAATHWASMDPAVVHPELNGAAYSRYPPGVVGVAPTGIPAAVEGIVPSAMSLSHGLPPVAHPPHAPSPGQTVKPEADRDHASDQL*
R.norvegicus/X1 NCVNKDHLTAATHWASMDPAVVHPELNGAAYSRYPPGVVSVAPTGIPAAVEGIVPSAMSLSHGLPPVAHPPHAPSPGQTVKPEADRDHTTDQL*
B.taurus/X1 NCVNKDHLAAATHWASVDPAVVHPELNGAAYSRYPPGVVGVAPSGIPAAVEGIVPSAMSLSHGLPPVSHPPHAPSPGQTVKPEADRDHASDQL*
E.telfairi NCVNKDHLTAATHWASMDPAVVHPELNGAAYSRYPPGVVSVAPTGMPAAVEGIVPSAMSLSHGLPPVAHPPHAPSPGQTVKPEADRDHTGDQL*
C.ustulatus/X1 NCVNKDHLTAATHWASMDPGVVHPELNGAAYSRYPPGVVSVASTGIPAAVEGIVPNPMSLSHGLPAVAHPPHAPSPGQTVKPEADREHPSDQL*
A.sinensis/X2 NCVNKDHLTAATHWASMDPGVVHPELNGAAYSRYPPGVVSVAPTGIPTAVEGIVPNPMSLSHGLPAVAHPPHAPSPGQTVKPEADRDHPSDQL*
C.belli/X5 NCVNKDHLTAATPWASMDPGVVHPELNGAAYSRYPPGVVSVAPAGIPTAVEGIVPNPMSLSHGLPAVAHPPHAPSPGQTVKPEADRDHPSDQL*
S.punctatus NCVNKDHLTAATHWASMDPGVVHPELNGAAYSRYPPGVVSVAPAGIPTAVEGIVPNAMSLSHGLPAVAHPPHAPSPGQTVKPESDRDHSSDQL*
T.elegans/X2 NCVNKEHLTTATHWASMDPGVVHPELNGATYSRYPPGVVSVAPPGIPTAVEGIVPNAMPLSHGLPPVAHPPHAPSPGQTVKPEVDRDHASDQL*
A.mexicanum NCINKDHLTAATHWASMDPGVVHQELNGATY-RYPPGVVSVAPPGIPSA-EGIVPSAMSLSHGLPAVVHPPHAPSPGQTVKPDADRDHPSDQL*
G.seraphini/X1 NCVNKDHLTAATHWASMDPAVVHPELNGAAYSRYPPGVVNVGSAVLPTA-EGI-PSAISLSHGLPAVVHPPHAPSPGQTVKPEADRDHPSDQL*
P.annectens NCVNKDYLTAATHWASMDPTVVHPELNGAAY-RYQSGVVSVAPAGLPSAVEGIGQNAMSVSHGIPTVAHPPHAPSPGQTVKPEADRDHSADQL*
L.chalumnae/X1 NCVNKDFLSAPTSWASIDPAVVHPELNGAAY-RYPPGVVSVAPTGLPTAVEGIVPSALSLSHGLPAVAHPPHAPSPGQTVKPEADREHPTDQL*
C
CtBP2
H.sapiens/X3 NCVNKEFFVTSAPWSVIDQQAIHPELNGATYRYPPGIVGVAPGGLPAAMEGIIPGGIPVT--HNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
R.norvegicus/X2 NCVNKEFFVTSTPWSVIDQQAIHPELNGATYRYPPGIVGVAPGGLPPAMEGIIPGGIPVT--HNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
B.taurus/X1 NCVNKEFFVTTAPWSVIDQQAIHPELNGATYRYPPGIVGVAPGGLPAAMEGIIPGGIPVT--HNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
E.telfairi/X2 NCVNKEFFVTAAPWSVIDQQAIHPELNGATYRYPPGIVGVAPGGLPAAMEGIVPGGIPVT--HNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
C.ustulatus/X3 NCVNKEFFVTTAPWSVIDQQAIHPELNGATYRYPPGMVSVAPGGIPAAMEGIIPGGIPVT--HNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
A.sinensis/X2 NCVNKEFFVTTAPWSVIDQQAIHPELNGATYRYPPGMVSVAPGGIPAAMEGIIPGGIPVT--HNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
C.belli NCVNKEFFVTTAPWSVIDQQALHPELNGATYRYPPGMVSVAPGGIPAAMEGIIPGGIPVT--HNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
S.punctatus NCVNKEFFVTTAPWSVIDQQAIHPELNGATYRYPPGMVSVAPGGIPAAMEGIIPGGIPVT--HNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
T.elegans NCVNKEFFVTTAPWSVIDQQAIHPELNGATYRYPPGMMSVTPGGIPAAMEGIIPGGIPVT--HNLPTVAHPSQAPSPNQPSKHGDNREHPNEQ*
A.mexicanum NCVNKEFLMSASLWAHGEQQVLHQDINGVTYRYPPGVVGVTPGAIPAAMEGMTPGGAPIS--HNLPTAGRQSQAPSPNHPSKHSD---HPHEK*
G.seraphini NWVNKEFFVTNSPLSVIHHQAGLQEMNGAAYRYSSGIVGLSTGRISTSLEGI-PGSTGTPLMQNLPTLVQSSRAPSPNHLSNHSDIRNHLSNQ*
P.annectens/x9 NCVNREFFVTTATWSVMDQQAVHPELNGATYRYQPGVVGVTPGGIPAAMEGILPGGVPVT--HTLPAVAHPSQAPSPNQPTKHGDNREHPNEQ*
L.chalumnae/X4 NCVNKEFFVTTATWSVMDQQTVHPELNGATYRYPPGVVGVAPGGIPAAMEGIVPGGVPVG--HSLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
D
CtBP1-like
C.ustulatus NCVNKEYLLLAA-QWSNIDPATVHP--ELNGAAAY----RFPPGVVGVANPGLPEPPVVEGIV------AHGIPPVSHSAPRTPSPGETSKLDADREIPA-DQ*
A.sinensis NCVNKEYLLLAA-QWSNIDPATVHP--ELNGAAAY----RFPPGVVGVAAPGLPEPTVLEGMV------THGIPSVSHSTPRTPSPGETSKLDPDREIPA-DQ*
C.belli NCVNKEYLLLAA-QWPSIDPAAVHP--ELNGAVAY----RFPPGVVGVAAPGLPEPAVVEGIV------AHGMPSVSHSAPRTPSPGETSKQDPDREIPA-DQ*
S.punctatus NCVNKEYLLLAA-QWSSIDPAAVHP--ELNGAAAY----RFPPGVVGVAAPGLPEPA-VEGIV------AHGIPSASHSAPHTPSPGETSKLEPDREISA-D*
T.elegans NCVNKEYLLLAA-QWSSIDPAAVHP--ELNGATAY----RFP-GVVGVGTSGLPEPP-VEGIV------AHSIPSVSHSAPHTPSPGESSKLEPDREIST-DQ*
A.mexicanum NCVNKEYLL-AATQWSGIDPSAVHAHAELNGAAVY----RFPPGVVGVAPQGLPATS-VDGMVA----GSHPLPHGIPSLSHTPSPGQTNKPDPDREIPT-DQ*
G.seraphini NCVNKEYLLAAA-QWSGMEPAASHG--ELNGAAGY--RSVFPPGMVGVTPPVLPASS-VDGLV------ASSIP---H-MTHTPSPGQTSKPESDREIST-DQ*
P.annectens NCVNKDYLLAAA-QWSSVDPSTVHQ--ELNGASAATASYRFPA-VVGVAPPGIPTAA-VEGMVTTSLPLSHGMTPVAHSSPHTPSPGQTTKPEVDREIPAADQ*
L.chalumnae NCINKEFFMTP-TQWSNIDPAAVHP--ELNGSGAY----RFPAGVVGVTAAGIPGAV-VEGIVAGSLPLTHGIPSVVHPTAHASSPGQTTKPEADREIPA-DQ*

Figure 4.10. Conservation of CtBP CTD sequences in Sarcopterygii. A) Phylogenetic tree of Sarcopterygii species analyzed. Colored boxes indicate major classes of Sarcopterygii, including mammals, Sauria, amphibians, lungfish, and coelacanth. Vertical lines in B-D correspond to the groups shown in panel A. **B)** Representative alignment of CtBP1 isoforms indicates that the CtBP1 CTD is highly conserved among selected lobe-finned fishes. A handful of species also express a shorter version of CtBP1 (not shown), whereas all other species have only a long variant of CtBP1. **C)** Representative alignment of CtBP2 isoforms indicates that the CTD is also highly conserved but is less well conserved than CtBP1 among selected amphibian species. Some bats (**Figure S4.8D**) and Sauria have short versions of CtBP2. **D)** Representative alignment of CtBP1-like isoforms. CtBP1-like paralogs are absent in mammals.

All Actinopterygii have a CtBP1-like paralog, as was observed in most Sarcopterygii, and those with four or five CtBP paralogs express what we have named CtBP1a and CtBP2-like, determined based on the degree of similarity in the dehydrogenase core to CtBP1 or CtBP2 (see Materials and Methods; **Figure 4.11B**). Alignments of each of these additional CtBP proteins to one another across Actinopterygii confirm their high degree of conservation and paralog-specific sequences (**Figure S4.9**). In summary, we find that across 400 My of evolution of Actinopterygii, gene duplications played a big role in the diversification of this family, whereas retaining key features of CtBP seen in Sarcopterygii. We hypothesize that the additional CtBP paralogs may function in new roles distinct from transcription; these may include cytoplasmic functions in the Golgi and in synaptic vesicles and neurons, as found for the RIBEYE variant of CtBP2 (Schmitz *et al*. 2000; tom Dieck *et al*. 2005).

Structural Properties of CtBP C-termini

Our survey of the C-termini of CtBP forms across Bilateria indicates that many, but not all, lineages have retained primary structural elements that presumably reflect important functional properties. "Canonical" CTD structures include conserved hydrophobic/aromatic N-terminal residues, an aromatic (Y, F, or W) adjacent to the central block, and paralog-specific C-termini with similar arrangements of lysine and acidic residues. This general structure is found across deuterostomes, where most CTDs range in length from 90 to 100 residues (**Figure S4.10C-E**). In the vertebrates, the CTD of CtBP1 paralogs in most tetrapod lineages have few modifications, and remain 90–100 residues in length (**Figure S4.10C**), in contrast to the more dynamic length changes evident even within the Drosophila genus. As noted in the analysis of protostome CtBP structure, certain lineages have independently substituted canonical CTD sequences with novel structures, leading to a greater diversity of lengths (**Figure S4.10A**).



Actinopterygii	CtBP1	CtBP1-like	CtBP1a	CtBP2	CtBP2-like
Erpetoichthys calabaricus	x	x		x	
Polypterus senegalus	x	x		x	
Acipenser ruthenus	x	x		x	
Polyodon spathula	x	x	x	x	
Amia calva	x	x		x	
Lepisosteus oculatus	x	x		x	
Anguilla anguilla	x	x	x	x	x
Megalops cyprinoides	x	x	x	x	x
Scleropages formosus	x	x	x	x	x
Esox lucius	x	x	x	x	
Takifugu rubripes	x	x		x	x
Clupea harengus	x	x	x	x	
Chanos chanos	x	x		x	x
Danio rerio	x	x		x	x
Colossoma macropomum	x	x		x	x
Ictalurus punctatus	x	x		x	

Figure 4.11. Actinopterygii possess up to five CtBP genes. A) Phylogenetic tree of ray-finned fishes divided into four classes/infraclasses. Red stars represent Vertebrate Whole-Genome Duplication events (VGD) and the blue star represents a Teleost-specific Genome Duplication event (TGD). **B)** Chart representing CtBP isoforms encoded in each species' genome. **C)** Alignment of CtBP1-long and CtBP2-long isoforms in the fish indicates that the CTD is well conserved. **D)** Alignment of CtBP1 and CtBP2-short isoforms. Evidence for short isoforms is limited to a subset of species. The CTD sequences are much more conserved in CtBP1-short than in CtBP2-short.

Are there common properties of these diverse CTD sequences, which may reveal common functions? The CTD of CtBP is predicted to be unstructured; therefore, we focused on properties of intrinsically disordered regions (IDR). We measured hydrophobic content, proportion of charged residues, and proportion of disorder-promoting residues across Bilaterian CTDs. Considering overall amino acid composition, the occurrence of hydrophobic residues (M, I, V, L, F, Y, W) is around 21% in protostomes, whereas some lineages average closer to 10% (**Figure S4.10B**). Deuterostomes range from 16% to 27%, averaging ~25% across the paralogs (**Figure S4.10F**). This is lower than the frequency found in the CtBP structured dehydrogenase domain, which, consistent with folded, water-excluding structures, is ~33% hydrophobic in the fly CtBP and human CtBP1. In comparison to experimentally validated repressor domains in the human proteome, which average around 45% hydrophobic content, the CtBP CTD also has much lower hydrophobicity (A and P residues were included and V excluded, compared with our method; Soto *et al*. 2022).

We also calculated the proportion of positively charged (K and R) and negatively charged (D and E) residues (**Figure S4.11A-D**). There is some variability across Protostomia, with annelids having only 6% of the CTD composed of charged residues, whereas Nemertea have 18%. Across Deuterostomia, there is much less variability, with both CtBP1 and CtBP2 CTDs composed of just under 15% charged residues. Using CIDER (Holehouse *et al*. 2017), we find that based on the low hydrophobicity, and high content of charged residues, these CTDs are considered "weak polyampholytes and weak polyelectrolytes" (FCR <0.3, and NCPR <0.25). These properties suggest that the CTDs may form defined structures in a facultative manner.

IDRs are often enriched in proline, glycine, and alanine residues, which are considered structure-breaking residues (Habchi *et al*. 2014). Proline is the most disorder-promoting of all amino acids, with a disorder propensity score of 1.0, whereas alanine and glycine have scores of 0.45 and 0.43, respectively (Theillet *et al*. 2013). We analyzed the composition of the primary peptide sequences of CtBP CTDs across Metazoa, and found that the composition remains similar across protostomes and deuterostomes, with P, G, and A residues accounting for 35–45% of the entire CTD in most lineages regardless of the CTD length or primary sequence (**Figure S4.11E-H**). These values are much higher than in the dehydrogenase core, where P, G, and A only make up about 22% of the primary sequence in the fly CtBP and human CtBP1.

We also performed secondary structure predictions to determine whether any species have discernible CTD structures. Using PSIPRED and Robetta (Buchan and Jones 2019; Baek *et al*. 2021), we found that most protostome CtBP CTDs have predicted unstructured domains (**Figure 4.12A**), with predicted short alpha helices correlated to nonconserved alanine-rich insertions in some species, such as in Drosophila (data not shown). Interestingly, a much greater degree of predicted structure is found in the highly derived, lineage-specific CTD sequences, such as in certain mites, nematodes, tardigrades, and flatworms (**Figure 4.12B**). These predicted structures do not bear structural resemblance to each other, and may play specialized roles in these species. Vertebrate CtBP1 and CtBP2 CTDs were predicted to also be highly unstructured, with most having a beta turn or a small alpha helix (**Figure 4.13**). The predicted beta turns are found within the conserved central block motif, towards the N-terminus of the CTD, and this feature is conserved across representative protostomes and deuterostomes. This structured motif may be important for binding to cofactor.

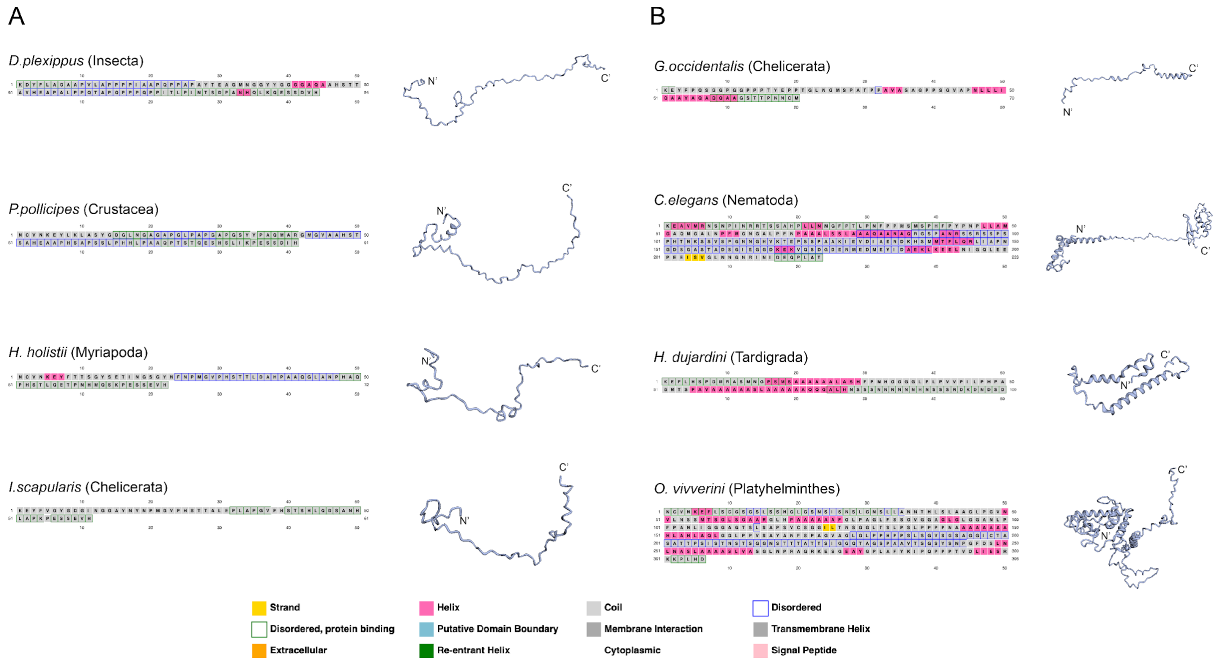


Figure 4.12. Secondary structure predictions of select protostome CtBP CTDs. PSIPRED (boxed amino acids) and Robetta (structures) predictions. The legend on the bottom indicates the significance of colored boxes. **A)** Representative protostomes with canonical CtBP CTD sequences were selected. In all the predicted structures, the CTD is highly disordered, with a beta turn toward the N-terminus, which maps to the central block motif. **B)** CTD sequences and structures from four species with derived CTDs are shown (specific mites [chelicerates], nematodes, tardigrades, and Platyhelminthes). These secondary structures were found to have less disordered regions and instead had a higher number of alpha helices predicted. These are the only CTDs that are predicted to have a distinct structure. The N- and C-termini are indicated.

Diversity of the N-terminal Domain of Vertebrate CtBP

Aside from the great variation seen in the C-terminal domain across bilaterians, variations in CtBP also exist in the N-terminal domain. This is particularly evident in Gnathostomata, the jawed vertebrates, who have diversified CtBP2 through usage of alternative TSS to create isoforms with extended NTDs. In mammals, this CtBP2 isoform with an NTD extension is termed RIBEYE, encoding a protein with a 572 residue extension (**Figure S4.12A**). RIBEYE localizes to synaptic vesicles in the retina and in sensory neurons (Schmitz *et al*. 2000; tom Dieck *et al*. 2005).

To determine where the RIBEYE isoform first arose outside mammals, and whether non-mammalian vertebrates have a conserved RIBEYE isoform, we analyzed NTD sequences of CtBP2 proteins across Vertebrata. There is no evidence of long NTD isoforms in the single Cyclostomata species analyzed. Most Gnathostomata express CtBP2 isoforms with extended RIBEYE-like NTDs, with lengths of 550–620 residues (**Figure S4.12B**). The lack of long NTDs observed in select species may be due to lack of expression in the tissues from which transcriptomic data were collected, or poor detection during sequencing, rather than a loss of long NTD isoforms in these vertebrates. Compared with the human RIBEYE sequence, mammals have 65–80% sequence identity, whereas other Sarcopterygii have 45–55%. Actinopterygii and Chondrichthyes also have 40–50% sequence identity, similar to birds, reptiles, and amphibians (**Figure S4.12B**). Although the levels of conservation among RIBEYE domains is lower than that found in the catalytic core and CTD, there are blocks of sequences that are highly conserved across Gnathostomata (**Figure S4.12C**). For instance, the MPVPS-like motif at the start of the NTD is conserved across the sampled gnathostomes, and there are additional 5–10mer motifs that are highly conserved and scattered across the RIBEYE sequence (**Figure S4.12C**). In species encoding CtBP2-like isoforms (found only in Teleostei), long NTD isoforms are sometimes present, confirming that CtBP2-like originated from a CtBP2 duplication event. Only a few teleosts have long CtBP2-like NTDs, which are either ~300 or 700 residues long. Those with 700-residue NTDs (*Colossoma macropomum* and *Chanos chanos*) have sequences that resemble the human RIBEYE, with about 40% primary sequence conservation in the NTD, similar to that seen with the CtBP2 of select Actinopterygii (data not shown). We also find conserved 10mer motifs scattered throughout, with insertions of polyQ tracts, which results in the longer observed lengths (data not shown). Taken together, these results indicate that the extended CtBP2 N-terminus originated in the last common ancestor of Gnathostomata, and that the extended NTD was retained in the CtBP2-like paralog after gene duplication.

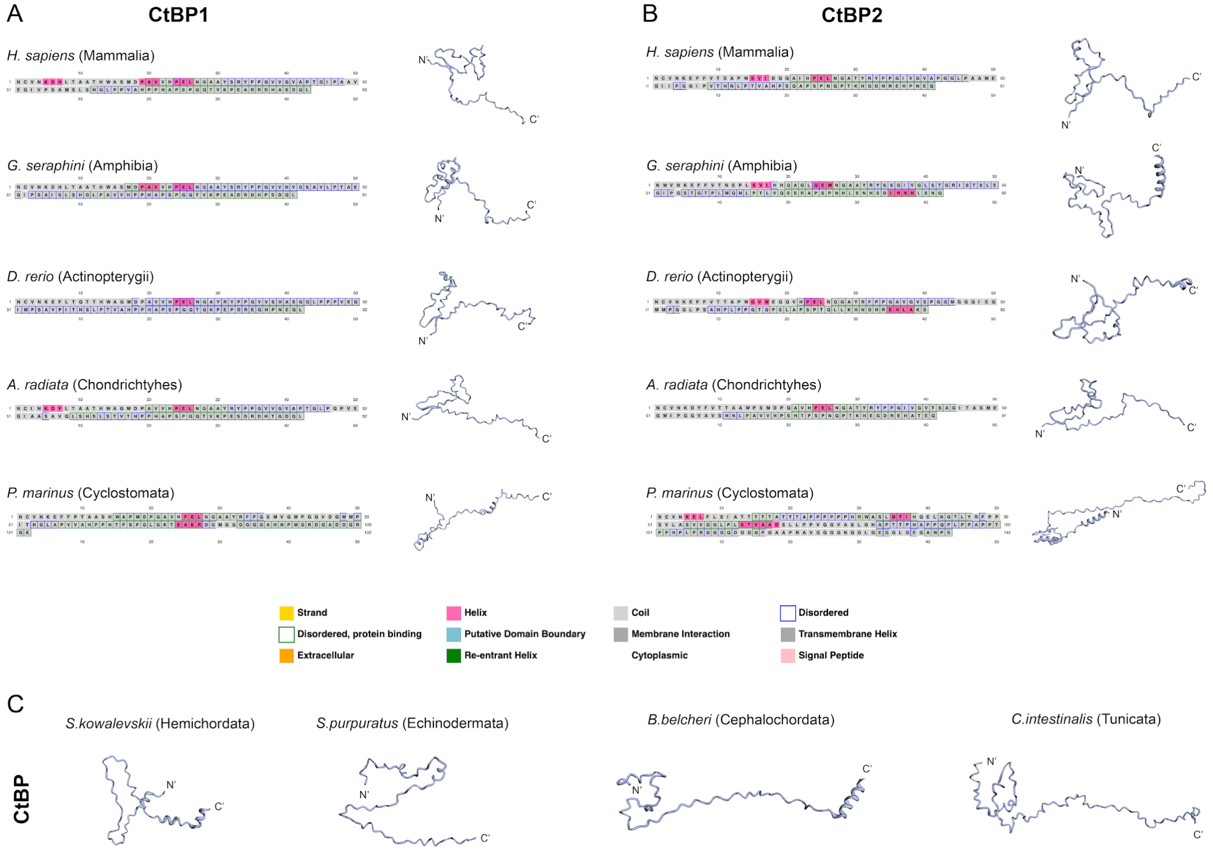

A
CtBP1
H. sapiens (Mammalia)
G. seraphini (Amphibia)
D. rerio (Actinopterygii)
A. radiata (Chondrichtyhes)
P. marinus (Cyclostomata)
B
CtBP2
H. sapiens (Mammalia)
G. seraphini (Amphibia)
D. rerio (Actinopterygii)
A. radiata (Chondrichtyhes)
P. marinus (Cyclostomata)
N'
C'
Strand
Helix
Coil
Disordered
Disordered, protein binding
Putative Domain Boundary
Membrane Interaction
Transmembrane Helix
Extracellular
Re-entrant Helix
Cytoplasmic
Signal Peptide
C
CtBP
S.kowalevskii (Hemichordata)
S.purpuratus (Echinodermata)
B.belcheri (Cephalochordata)
C.intestinalis (Tunicata)

Figure 4.13. Secondary structure predictions of select deuterostome CtBP CTDs. PSIPRED (boxed amino acids) and Robetta (structures) predictions. The legend on the bottom indicates the significance of colored boxes. **A)** Representative structures for the CtBP1 CTD from Sarcopterygii (Mammalia & Amphibia), Actinopterygii, cartilaginous fish (Chondrichthyes), and jawless vertebrates (Cyclostomata). **B)** Representative structures for the CtBP2 CTD from the same species shown in A. **C)** Predicted secondary structures for non-vertebrate deuterostome CTDs. These CTDs are predicted to be unstructured, with some small alpha helices on the C-terminal portion. The widely conserved central block motif is found to form a beta-turn in most deuterostome species. The N- and C-termini are indicated.

Modifications to the CtBP CTD may add an Additional Layer of Regulation

We have shown that over longer evolutionary times, novel forms of CtBP have developed at the gene level through wholesale adoption of unique CTD sequences, isoform production using alternative splicing, gene duplication, and alternative promoter usage. Not surprisingly, the CtBP CTD can undergo many PTMs, which is a common feature of IDRs because they are accessible to enzymes for modifications (Musselman and Kutateladze 2021). The CtBP1 CTD is phosphorylated at S422 by HIPK2, which triggers CtBP degradation and cell death, and is sumoylated at K428 by SUMO-1, which allows for its nuclear localization (Zhang *et al*. 2003, 2005; Wang *et al*. 2006; Kagey *et al*. 2003; Lin *et al*. 2003; **Figure S4.13A**). These residues are completely conserved across vertebrate CtBP1 and CtBP1-like, and also among some non-vertebrate deuterostomes, suggesting that the CtBP1 tail can be modified and regulated in a similar manner in these species (**Figure S4.13B**). CtBP2 is phosphorylated on residues S365, T414, and S428. HIPK2 phosphorylates S428, but the impact of this and other modifications have not been experimentally determined (Bian *et al*. 2014; Dewi *et al*. 2015; **Figure S4.13A**). Only T428 is conserved across vertebrates, whereas the other residues show lower conservation (**Figure S4.13C**).

To determine whether PTMs such as phosphorylation and sumoylation may involve conserved portions of the CTD in our selected species, we used predictive PTM software. We determined putative sumoylation sites using JASSA v4 (Beauclair *et al*. 2015). We find that many invertebrates with a canonical CTD sequence including insects, chelicerates, and some Spiralia, have high consensus SUMO motifs, usually in the extreme C-terminus. Many of the derived CTD sequences from Nematoda also have a predicted sumoylation site. The majority of Sarcopterygii CtBP1 sequences have a strong SUMO consensus motif in the extreme C-terminus, whereas

Actinopterygii have a weak motif. The CtBP2 CTDs lack SUMO motifs, suggesting a different form of regulation. We also predicted possible phosphorylation sites in the CTDs, as IDRs have been shown to be particularly enriched in phosphorylated residues (Habchi *et al.* 2014). Using NetPhos 3.1 (Blom *et al.* 1999), we find that the Y and S/T residues of the vertebrate central block are predicted phosphorylation sites, as are the same residues in the invertebrate central block motif. The high conservation of this motif across Bilateria, and its predicted phosphorylation status may point to an important role in regulation of CtBP activity. Additionally, protostome-specific motifs (AHSTTP and the terminal SEVH ending) are also predicted phosphorylation sites, again pointing to positive selection perhaps due to an important regulatory role.

Evolutionary Variation in the Conserved Dehydrogenase Core

The well-structured dehydrogenase domain of CtBP shows much higher sequence conservation than the CTD, and across longer evolutionary time (**Figure S4.1**). However, small variations in the core are found between species; in Diptera, a number of species generate alternative splice forms that affect the VFQ tripeptide motif, which is predicted to be in an unstructured loop on the surface of the protein (**Figure S4.14A**). VFQ is present across most insects, with some variations, and is found in some arthropods including crustaceans and myriapods, but the motif is not conserved across protostomes (**Figure S4.14A**). It presumably is only spliced out in Diptera, as there is no evidence that there are isoforms without VFQ in other insects or protostomes. Additional core variations are found more broadly in arthropods, such as a five-residue insertion in some splice isoforms of select insects and chelicerates, N-terminal to the start of the CTD (**Figure S4.14A**). Interestingly, this motif maps just C-terminal to the VFQ, also in a predicted unstructured portion of the protein on the surface of the structure, and away from the tetramerization interface. Among the protostomes, there are several spiralian and crustacean species that have 1–15 amino acid motifs that are inserted or deleted, which are unique to only those species, and presumably arose much later in their evolution since they are not alternatively spliced in other species (data not shown).

Among the deuterostomes, the only conserved alternatively spliced motif is SF, found ~50 residues N-terminal to the CtBP1 CTD (**Figure S4.14B**). SF is alternatively spliced in select Actinopterygii and Sarcopterygii, but not in all examined species. Interestingly, this motif also maps to an unstructured loop in the human CtBP1 protein, and overlaps the dipteran VFQ motif, suggesting that its alternative splicing event is significant, either because it was retained, or independently arose in these separate lineages.

The catalytic triad, which is emblematic of CtBP as an ancient dehydrogenase, is conserved in all bilaterians, aside from nematodes, which have lost one of the three residues (**Figure S4.1**). Interestingly, all metazoans retain these residues, but the *Arabidopsis thaliana* ANGUSTIFOLIA homolog does not, consistent with the divergent function of the plant protein in the cytoplasm. Tetramerization residues found in the core (S128, A129, R190, G216, and L221), which have recently been shown to be necessary for CtBP2's activity as a transcriptional repressor, are also highly conserved (Jecrois *et al*. 2021). Between the human CtBP1 and CtBP2, four of these are conserved (not S128; Raicu *et al*. 2021). In bilaterians, the R, G, and L residues are completely conserved, and SA is GY, GF, or GV. Perhaps tetramerization and a possible catalytic role are more broadly conserved structural features of these proteins.

DISCUSSION

Our comparative phylogenetic study demonstrates that CtBP is a bilaterian innovation, with virtually all orthologs possessing an unstructured C-terminus, usually of about 100 residues. Although initial observations of CtBP protein sequences suggested that the CTD was not conserved, here we demonstrate striking patterns of deep conservation (Kim *et al*. 2002; Nicholas *et al*. 2008). Across Metazoa, the CTD is highly conserved in length, in its propensity for disorder, and in certain blocks of sequence that are found in most species. The long C-terminus is found in virtually all lineages, with additional shorter isoforms arising through alternative splicing independently in a number of insects and vertebrates. Interestingly, there are lineages where the sequence and structure of the C-terminus has independently undergone radical transformations; in mites and tardigrades, the length is maintained but the sequence has diverged, whereas divergent flatworm and nematode CTD sequences extend to several hundred residues. In vertebrates, additional diversification of CtBP is found through gene duplication, with up to five unique genes encoded in certain fishes. Diversification of the CtBP CTD may have implications in gene regulatory networks, and more broadly in evolutionary transformations of bilaterians.

Viewed broadly, this analysis of CtBP evolution shows some parallels to previous studies of other components of the bilaterian transcriptional machinery, whereas raising some still unanswered questions. From pioneering work by Lewis and others, reverse engineering of transcriptional systems has uncovered important cis and trans variations in components of the transcriptional machinery that drive profound evolutionary transformations in the metazoan body plan (Lewis, 1978). Those variations affecting DNA-binding transcription factors, such as Hox proteins, provide some of the best-known cases (Pearson *et al.* 2005). On the other hand, the potential impact on morphological evolution stemming from variation in the core regulatory machinery that is responsible for expression of most genes is less well known. Indeed, initial biochemical studies of the basal transcriptional machinery, including RNA polymerase II and associated factors, emphasized the conservation of a largely invariant and nearly universal collection of components specific to eukaryotes, underlining the early emergence of these factors in the last common ancestor. However, more recent work has demonstrated lineage-specific features of this machinery, including the diversity of factors within the TFIID complex (TBP and TAFs) pointing to the specialization of even the pleiotropic core machinery (Li *et al.* 2009; Goodrich and Tjian 2010).

As we document for CtBP, a significant source of variation within the core transcriptional machinery is found in IDR. Overall, IDRs feature low sequence complexity, low hydrophobicity, and are heterogeneous in their conformation (Shukla *et al.* 2022). They can self-associate, adopt structured conformations in association with cofactors, or participate in flexible interaction surfaces (so-called "fuzzy" complexes; Shukla *et al.* 2022). These properties appear to lend IDRs a particularly active role in evolutionary change, as they can tolerate substitutions and still perform their diverse functions (Musselman and Kutateladze 2021; Pajkos and Dosztanyi 2021; Shukla *et al.* 2022). What functions might be associated with the CtBP CTD? Studies of a number of IDR-containing proteins point to a diversity of roles, including roles in regulation of DNA binding, cofactor recruitment, anchors for posttranslational modifications, homodimerization, and adopting defined structures in a larger complex. It is notable that structural studies of CtBP that have emphasized its unstructured CTD used purified protein, thus it is possible that the CTD is highly structured when combined into a complex of other interacting protein partners.

We think that the most attractive model for the CTD of CtBP is provided by a transcriptional cofactor that is derived from another class of hydroxyacid dehydrogenases, N-PAC (also known as GLYR1). This protein, which like CtBP can form homotetramers, possesses a disordered N-terminus. The IDR associates with the LSD2 demethylase, and a portion of its sequence adopts a defined structure to assist LSD2 to access histone tails (Marabelli *et al.* 2019). The long flexible N-terminal region has been suggested to allow the tetrameric complex to simultaneously span several nucleosomes, coordinating the action of this chromatin modifying complex. Similarly, those conserved portions of CtBP's CTD may form defined structures in the context of a larger complex, whereas also contributing to regulation via posttranslational modifications and possible long-range interactions. Ongoing advances in structural biology will likely deliver important information on such multiprotein complexes, which will generate important hypotheses relating to CtBP, such as the expected impact of CtBP-short forms lacking a CTD. How the lineage-specific variations impact gene expression remains a significant challenge that will require an integrated genomic and molecular genetic approach.

MATERIALS AND METHODS

cDNA and Peptide Sequences

cDNA and peptide sequences for *D. melanogaster* CtBP isoforms were downloaded from flybase (www.flybase.org; version FB2020_03 and FB2020_05; dm6; Gramates *et al*. 2022). *Drosophila melanogaster* sequences were used as a reference to retrieve cDNA and peptide sequences using NCBI blastn and blastp (https://blast.ncbi.nlm.nih.gov/Blast.cgi) for human CtBP1 and CtBP2, and most protostomes. The human CtBP1 and CtBP2 sequences were used as a reference to retrieve cDNA and peptide sequences for most deuterostomes. When peptide sequences were not available through NCBI, we translated the available cDNA sequences using the "Show translation" tool on bioinformatics.org (https://www.bioinformatics.org/sms/show_trans.html), selecting the translations for "reading frames 1 to 3". The open reading frame for *D. melanogaster* and *H. sapiens* CtBP sequences were used to determine the correct reading frames for other species. Most of the downloaded sequences were annotated as CtBP or CtBP-like in NCBI; sequences labeled as "dehydrogenase" with <40% identity, where another hit was labeled "CtBP", were not included in this analysis. For non-Bilaterians including Cnidarians, Porifera, and other non-Metazoans, we used the only hits labeled “dehydrogenase” with low sequence identity to perform our analysis. Platyhelminthes sequences were retrieved by selecting “Transcriptome Shotgun Assembly” on NCBI BLAST. *Adineta vaga* sequences were obtained from GENOSCOPE Adineta vaga genome browser (https://www.genoscope.cns.fr/adineta/cgi-bin/gbrowse/adineta/), *Strigamia maritima* from e! EnsemblMetazoa, *Protopterus annectens* from Marco Gerdol (Biscotti *et al.* 2016), *Euperipatoides rowelli* from https://datadryad.org/stash/dataset/doi:10.5061/dryad. bk3j9kdc0 and *Gyrodactylus salaris* from Paps and Holland 2018. All species used, their taxonomic ID, and genome version are listed in **Supplementary File 1** (online).

Multiple Sequence Alignments

Peptide sequences were aligned using the MAFFT multiple sequence alignment (https://www.ebi.ac.uk/Tools/msa/mafft/) using the ClustalW output format. All isoforms for a given species were aligned against one another to note differences between isoforms from the same species. A representative isoform from each species was included in the figures. Amino acids that were conserved in >50% of the species in an alignment were colored blue. Chemically conserved amino acids in the same position were colored orange, using the following conservation scheme: Hydrophobic aliphatic amino acids: M, V, I, L; hydrophobic aromatic amino acids: W, Y, F; acidic amino acids: D, E; basic amino acids: K, R; hydroxyl containing amino acids: S, T. Where there was conservation of a second amino acid but in only 25–50% of species, they were colored army green. Amino acids that were not chemically similar and were not conserved across many species in the alignment were left uncolored.

Phylogenetic Trees

We used NCBI to collect taxonomic IDs for all species used in this study. Tax IDs were inputted into phyloT v2 (https://phylot.biobyte.de/) which generates phylogenetic trees based on NCBI taxonomy, incorporating phylogenetic and taxonomic information from multiple types of sources, including sequencing data and morphological information (Federhen, 2012). Timetree (http://www. timetree.org/) was used to compare phylogenetic trees and determine estimated time of divergence between select species.

To infer a phylogenetic history of the CtBP sequences (**Figure S4.15**), we first created an alignment containing 35 representative sequences from protostomes and 105 sequences from deuterostomes, lacking the NTD. We sampled and aligned the sequences within each group separately and used trimAI software packages to remove highly gapped regions that were poorly alignable ("–gappyout" function; Capella-Gutierrez *et al.* 2009). The protostome and deuterostome sequences were then profile-aligned to each other to generate a multiple sequence alignment file containing all of the CtBP sequences using MUSCLE (Edgar 2004; **Supplementary file S3**, online). We next inferred a maximum-likelihood phylogeny from this sequence using the best-fit model of sequence evolution (Q.insect+Invariant Sites+Gamma), as determined by the "ModelFinder" algorithm implemented in IQtree2 (Minh *et al.* 2020, 2021). We also inferred a phylogeny using a commonly used model of protein evolution (JTT +Gamma) to ensure robustness of the uncovered relationships to model choice.

Analysis of CtBP CTD Properties

We collected CtBP CTD peptide sequences from >200 metazoan species, and used a script (**Supplementary File S2**, online) to determine the following for each CTD sequence: length (in Amino Acids, AA), proportion of A, P, G residues, percent hydrophobic residues (M, V, I, L, W, Y, F), and percent charged residues (K, R for positive and D, E for negative). For each property, a species' longest CTD was used, and properties were averaged by groups to display in the graphs (i.e. all the insects were averaged together, using a single CTD sequence from each of the selected species). Short CTDs were not used in the analysis aside from the leeches.

Secondary RNA and Protein Structure Predictions

Secondary structure predictions of RNA were made using RNAstructure (version 6.2) from the Mathews lab at University of Rochester Medical Center (Xu and Mathews 2016). Data were inputted using the Predict a Secondary Structure Web Server with default parameters, with temperature set to 293 K. The structure with the highest probability was used. Secondary structure predictions of CtBP CTD peptides were made using PSIPRED Workbench V3.2 (Buchan and Jones 2019). For data input, "sequence data" were selected under the "Select input data type" heading. PSIPRED 4.0 and DISOPRED3 were selected under the "Choose prediction methods" heading. Under "Submission details", the FASTA peptide sequence of interest was inputted and submitted. Secondary structure predictions of CtBP CTDs were also made into homology models using Robetta from the Baker lab at the University of Washington Institute for Protein Design (Baek *et al.* 2021). For data input, "Submit" was selected under the "Structure Prediction" heading. Under "Protein Sequence", the FASTA peptide sequence of interest was pasted. RoseTTAFold was used to create models.

Determination of CtBP Paralogs in Vertebrates

Several CtBP sequences from the vertebrates with more than two CtBP paralogs were misannotated in NCBI. We performed pairwise sequence alignments and determined the correct CtBP1-like, CtBP1a, and CtBP2-like sequences based on percent conservation to the dehydrogenase core of *H. sapiens* CtBP1 and CtBP2, and to a species' own CtBP1 and CtBP2. We also determined motifs in the CTD which were representative of CtBP1 or CtBP2 to accurately assign 1-like and 2-like names to the additional proteins. CtBP1-like and 1a sequences have higher conservation in the core and CTD to CtBP1, and the same for CtBP2-like with CtBP2. CtBP1-like and 1a are very similar to each other, but are not 100% conserved in the core within the same species. CtBP1-like/1a CTDs typically start and end with KEYL…PADQ. CtBP2-like CTDs typically start and end with KEFF…LTEQ. We additionally determined whether the cDNA sequences originate from different genomic locations, and found that for the Actinopterygii with up to five different genes (such as Anguilla anguilla), they originate from different chromosomes, indicating that they are five unique genes. Species where only one CtBP exists but was annotated with a variant name was re-assigned as CtBP (i.e. the non-vertebrate deuterostomes with a single CtBP). The two *P. marinus* paralogs were also renamed to CtBP and CtBP-like based on alignments to each other, and to other vertebrate sequences, as described in the text.

Classification of IDRs

CIDER (Classification of Intrinsically Disordered Ensemble Regions; http://pappulab.wustl.edu/CIDER/analysis/) was used to determine FCR (fraction of charged residues), NCPR (net charge per residue), and the Das-Pappu phase diagram position for representative CtBP CTDs from bilaterian species. CTD sequences were inputted in FASTA format.

PTM Predictions

To predict putative SUMO motifs in the CtBP CTDs, we used JASSA v4 (Joined Advanced SUMOylation site and SIM analyzer; http://www.jassa.fr/; Beauclair *et al.* 2015), which predicts sumoylated lysines based on the presence of a ψKxα _motif, or a variation of it (ψ=_hydrophobic residue; x=_any amino acid, α=_D or E). We inputted sequences from representative species across Bilateria in FASTA format, with the set parameters. To predict putative phosphorylated S, T, and Y residues, we used NetPhos—3.1 (https://services.healthtech.dtu.dk/service.php?

NetPhos-3.1; Blom *et al*. 1999) by inputting sequences from representative species across Bilateria in FASTA format.

AUTHOR CONTRIBUTIONS

D.N.A. conceived of the project, assisted with data interpretation, wrote and edited the manuscript. A.M.R. curated and analyzed the data, assisted with data interpretation, created figures, and wrote and edited the manuscript. D.K., M.N., A.J., Y.Y., K.B., and A.S. curated and analyzed the data, and assisted with creating the figures. K.M.B. performed secondary structure predictions and assisted with creating the figures. M.S. performed the phylogenetic analyses, assisted with data interpretation, and edited the manuscript. All authors reviewed the manuscript.

SUPPLEMENTARY MATERIAL (In Appendix)

Supplementary data are available at Molecular Biology and Evolution online. All data obtained for analysis is available from the resources listed in Materials & Methods.

ACKNOWLEDGEMENTS

This work was supported by the National Institute for General Medical Sciences (grant number R01GM124137 to D.N.A.), the National Institute of Child Health and Development (grant number F31HD105410 to A.M.R.), the MSU Biochemistry and Molecular Biology Undergraduate Research Fellowship to M.N., the Wielenga Research Fellowship to M.N, the MSU BEACON Luminaries grant to Y.Y., the Moore-Billman Scholarship Award in BMB to K.M.B, the College of Wooster APEX Fellowship to K.B., and the National Cancer Institute (grant number F32CA26111) and Michigan Life Sciences Fellowship to M.S. We would like to thank FlyBase, the researchers that supplied their genomic and transcriptomic data to NCBI, Diego Granados-Villanueva for his assistance with writing the script for the CTD analysis, and Andrew Thompson and Sandhya Payankaulam for helpful discussions.

REFERENCES

Achouri Y, Noel G, Van Schaftingen E. 2007. **2-Keto-4-methylthiobutyrate, an intermediate in the methionine salvage pathway, is a good substrate for CtBP1.** Biochem Biophys Res Commun. 352:903-906.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. 2021. **Accurate prediction of protein structures and interactions using a three-track neural network.** Science. 373:871–876.

Balasubramanian P, Zhao L-J, Chinnadurai G. 2003. **Nicotinamide adenine dinucleotide stimulates oligomerization, interaction with adenovirus E1A and an intrinsic dehydrogenase activity of CtBP.** FEBS Letters. 537(1-3):157-160.

Barroilhet L, Yang J, Hasselblatt K, Paranal RM, Ng SK, Rauh-Hain JA, Welch WR, Bradner JE, Berkowitz RS, Ng SW. 2013. **C-terminal binding protein-2 regulates response of epithelial ovarian cancer cells to histone deacetylase inhibitors.** Oncogene. 32(33):3896-903.

Beauclair G, Bridier-Nahmias A, Zagury J-F, Saib A, Zamborlini A. 2015. **JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMS.** Bioinformatics. 31(21):3483-3491.

Bellesis AG, Jecrois AM, Hayes JA, Schiffer CA, Royer WE, Jr. 2018. **Assembly of human C-terminal binding protein (CtBP) into tetramers.** J Biol Chem. 293(23):9101-9112.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A, et al. 2014. **The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates.** Nat Commun. 5:3657.

Bhambhani C, Chang JL, Akey DL, Cadigan KM. 2011. **The oligomeric state of CtBP determines its role as a transcriptional co-activator and co-repressor of Wingless targets.** EMBO J. 30(10):2031-43.

Bhasin H, Hulskamp M. 2017. **ANGUSTIFOLIA, a plant homolog of CtBP/BARS localizes to stress granules and regulates their formation.** Front Plant Sci. 8:1004.

Bi C, Meng F, Yang L, Cheng L, Wang P, Chen M, Fang M, Xie H. 2018. **CtBP represses Dpp signaling as a dimer.** Biochem Biophys Res Commun. 495(2):1980-1985.

Bian Y, Song C, Cheng K, Dong M, Wang F, Huang J, Sun D, Wang L, Ye M, Zou H. 2014. **An enzyme assisted RP-RPLC approach for in-depth analysis of human liver phosphoproteome.** J Proteomics. 96:253-62.

Biscotti MA, Gerdol M, Canapa A, Forconi M, Olmo E, Pallavicini A, Barucca M, Schartl M. 2016. **The lungfish transcriptome: a glimpse into molecular evolution events at the transition from water to land.** Sci Rep. 6:21571.

Blom N, Gammeltoft S, Brunak S. 1999. **Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.** Journal of Molecular Biology. 294:1351-1362.

Boyd JM, Subramanian T, Schaeper U, La Regina M, Bayley S, Chinnadurai G. 1993. **A region in the C-terminus of adenovirus 2/5 E1a protein is required for association with a cellular phosphoprotein and important for the negative modulation of T24-ras mediated transformation, tumorigenesis and metastasis.** The EMBO Journal. 12(20):469-478.

Buchan DWA, Jones DT. 2019**. The PSIPRED Protein Analysis Workbench: 20 years on.** Nucleic Acids Res. 47:W402-407.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.** Bioinformatics 2009;25(15):1972-3.

Chinnadurai G. 2002. **CtBP, an unconventional transcriptional corepressor in development and oncogenesis.** Molecular Cell. 9:213-224.

Chinnadurai G. 2007. **Transcriptional regulation by C-terminal binding proteins.** Int J Biochem Cell Biol. 39(9):1593-607.

Dcona MM, Morris BL, Ellis KC, Grossman SR. 2017. **CtBP-an emerging oncogene and novel small molecule drug target: Advances in the understanding of its oncogenic action and identification of therapeutic inhibitors.** Cancer Biol Ther. 18(6):379-391.

Deng H, Liu J, Deng Y, Han G, Shellman YG, Robinson SE, Tentler JJ, Robinson WA, Norris DA, Wang XJ, et al. 2013. **CtBP1 is expressed in melanoma and represses the transcription of p16INK4a and Brca1.** J Invest Dermatol. 133(5):1294-301.

Dewi V, Kwok A, Lee S, Lee MM, Tan YM, Nicholas HR, Isono K, Wienert B, Mak KS, Knights AJ, et al. 2015. **Phosphorylation of Kruppel-like factor 3 (KLF3/BKLF) and C-terminal binding protein 2 (CtBP2) by homeodomain-interacting protein kinase 2 (HIPK2) modulates KLF3 DNA binding and activity.** J Biol Chem. 290(13):8591-605.

Edgar RC. **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** Nucleic Acids Res. 2004;32(5):1792-7.

Erlandsen H, Jecrois AM, Nichols JC, Cole JL, Royer WE, Jr. 2022. **NADH/NAD(+) binding and linked tetrameric assembly of the oncogenic transcription factors CtBP1 and CtBP2.** FEBS Lett. 596(4):479-490.

Fang M, Li J, Blauwkamp T, Bhambhani C, Campbell N, Cadigan KM. 2006. **C-terminal-binding protein directly activates and represses Wnt transcriptional targets in Drosophila.** EMBO J. 25(12):2735-45.

Federhen S. 2012**. The NCBI Taxonomy database.** Nucleic Acids Res. 40:D136-D143.

Fjeld CC, Birdsong WT, Goodman RH. 2003. **Differential binding of NAD+ and NADH allows the transcriptional corepressor carboxyl-terminal binding protein to serve as a metabolic sensor.** Proc Natl Acad Sci U S A. 100(16):9202-7.

Folkers U, Kirik V, Schobinger U, Falk S, Krishnakumar S, Pollock MA, Oppenheimer DG, Day I, Reddy AR, Jurgens G, Hulskamp, M. 2002**. The cell morphogenesis gene ANGUSTIFOLIA encodes a CtBP/BARS-like protein and is involved in the control of the microtubule cytoskeleton.** The EMBO Journal. 21(6):1280-1288.

Goodrich JA, Tjian R. 2010. **Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation.** Nat Rev Genet. 11(8):549-58.

Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T, et al. 2022. **FlyBase: a guided tour of highlighted features.** Genetics. 220(4):1-12.

Grooteclaes M, Deveraux Q, Hildebrand J, Zhang Q, Goodman RH, Frisch SM. 2003. **C-terminal-binding protein corepresses epithelial and proapoptotic gene expression programs.** Proc Natl Acad Sci U S A. 100(8):4568-4573.

Habchi J, Tompa P, Longhi S, Uversky VN. 2014. **Introducing protein intrinsic disorder.** Chem Rev. 114(13):6561-6588.

Hildebrand JD, Soriano P. 2002**. Overlapping and unique roles for C-terminal binding protein 1 (CtBP1) and CtBP2 during mouse development.** Mol Cell Biol. 22(15):5296-5307.

Holehouse AS, Das RK, Ahad JN, Richardson MO, Pappu RV. 2017**. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins.** Biophys J. 112(1):16-21.

Jecrois AM, Dcona MM, Deng X, Bandyopadhyay D, Grossman SR, Schiffer CA, Royer WE, Jr. 2021. **Cryo-EM structure of CtBP2 confirms tetrameric architecture.** Structure. 29(4):310-319.

Jin W, Scotto KW, Hait WH, Yang J-M. 2007**. Involvement of CtBP1 in the transcriptional activation of the MDR1 gene in human multidrug resistant cancer cells.** Biochemical Pharmacology. 74(6):851-859.

Kagey MH, Melhuish TA, Wotton D. 2003. **The polycomb protein Pc2 is a SUMO E3.** Cell. 113:127–137.

Katsanis N, Fisher EMC. 1998. **A novel C-terminal Binding Protein (CTBP2) is closely related to CTBP1, an adenovirus E1A-binding protein, and maps to human chromosome 21q21.3.** Genomics. 47:294-299.

Kim G-T, Shoda K, Tsuge T, Cho K-H, Uchimiya H, Yokoyama R, Nishitani K, Tsukaya H. 2002. **The ANGUSTIFOLIA gene of Arabidopsis, a plant CtPB gene, regulates lead-cell expansion, the arrangement of cortical microtubules in leaf cells and expression of a gene involved in cell-wall formation.** The EMBO Journal. 21(6):1267-1279.

Kumar A, Carlson JE, Ohgi KA, Edwards TA, Rose DW, Escalante CR, Rosenfeld MG, Aggarwal AK. 2002. **Transcription corepressor CtBP is an NAD$^+$-regulated dehydrogenase.** Molecular Cell. 10:857-869.

Kuppuswamy M, Vijayalingam S, Zhao LJ, Zhou Y, Subramanian T, Ryerse J, Chinnadurai G. 2008. **Role of the PLDLS-binding cleft region of CtBP1 in recruitment of core and auxiliary components of the corepressor complex.** Mol Cell Biol. 28(1):269-281.

Lewis EB. 1978. **A gene complex controlling segmentation in Drosophila.** Nature. 276:565-570.

Li VC, Davis JC, Lenkov K, Bolival B, Fuller MT, Petrov DA. 2009. **Molecular evolution of the testis TAFs of Drosophila.** Mol Biol Evol. 26(5):1103-1116.

Lin X, Sun B, Liang M, Liang Y-Y, Gast A, Hildebrand J, Brunicardi FC, Melchior F, Feng X-H. 2003. **Opposed regulation of corepressor CtBP by SUMOylation and PDZ binding.** Molecular Cell. 11:1389-1396.

Madison DL, Wirz JA, Siess D, Lundblad JR. 2013. **Nicotinamide adenine dinucleotide-induced multimerization of the co-repressor CtBP1 relies on a switching tryptophan.** J Biol Chem. 288(39):27836-27848.

Mani-Telang P, Arnosti DN. 2007. **Developmental expression and phylogenetic conservation of alternatively spliced forms of the C-terminal binding protein corepressor.** Dev Genes Evol. 217(2):127-135.

Mans BJ, de Castro MH, Pienaar R, de Klerk D, Gaven P, Genu S, Latif AA. 2016. **Ancestral reconstruction of tick lineages.** Ticks Tick Borne Dis. 7(4):509-535.

Marabelli C, Marrocco B, Pilotto S, Chittori S, Picaud S, Marchese S, Ciossani G, Forneris F, Filippakopoulos P, Schoehn G, et al. 2019. **A tail-based mechanism drives nucleosome demethylation by the LSD2/NPAC multimeric complex.** Cell Rep. 27(2):387-399.

Minh BQ, Dang CC, Vinh LS, Lanfear R. **QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution.** Syst Biol 2021;70(5):1046-1060.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. **IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.** Mol Biol Evol 2020;37(5):1530-1534.

Musselman CA, Kutateladze TG. 2021. **Characterization of functional disordered regions within chromatin-associated proteins.** iScience. 24(2):102070.

Nadauld LD, Phelps R, Moore BC, Eisinger A, Sandoval IT, Chidester S, Peterson PW, Manos EJ, Sklow B, Burt RW, et al. 2006. **Adenomatous polyposis coli control of C-terminal binding protein-1 stability regulates expression of intestinal retinol dehydrogenases.** J Biol Chem. 281(49):37828-37835.

Nardini M, Svergun D, Konarev PV, Spano S, Fasano M, Bracco C, Pesce A, Donadini A, Cericola C, Secundo F, et al. 2006. **The C-terminal domain of the transcriptional corepressor CtBP is intrinsically unstructured.** Protein Sci. 15(5):1042-1050.

Nibu Y, Zhang H, Levine M. 1998. **Interaction of short-range repressors with Drosophila CtBP in the embryo.** Science. 280(5360):101-104.

Nicholas HR, Lowry JA, Wu T, Crossley M. 2008. **The Caenorhabditis elegans protein CTBP-1 defines a new group of THAP domain-containing CtBP corepressors.** J Mol Biol. 375(1):1-11.

Pajkos M, Dosztanyi Z. 2021. **Functions of intrinsically disordered proteins through evolutionary lenses.** Prog Mol Biol Transl Sci. 183:45-74.

Paliwal S, Ho N, Parker D, Grossman SR. 2012. **CtBP2 promotes human cancer cell migration by transcriptional activation of Tiam1.** Genes Cancer. 3(7-8):481-90.

Paps J, Holland PWH. 2018. **Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty.** Nat Commun. 9(1):1730.

Pearson JC, Lemons D, McGinnis W. 2005. **Modulating Hox gene functions during animal body patterning.** Nat Rev Genet. 6(12):893-904.

Poortinga G, Watanabe M, Parkhurst SM. 1998. **Drosophila CtBP: a Hairy-interacting protein required for embryonic segmentation and Hairy-mediated transcriptional repression.** The EMBO Journal. 117(7):2067-2078.

Quinlan KG, Nardini M, Verger A, Francescato P, Yaswen P, Corda D, Bolognesi M, Crossley M. 2006. **Specific recognition of ZNF217 and other zinc finger proteins at a surface groove of C-terminal binding proteins.** Mol Cell Biol. 26(21):8159-8172.

Raicu AM, Bird KM, Arnosti DN. 2021. **Tête-à-tête with CtBP dimers.** Structure. 29(4):307-309.

Raker VA, Mironov AA, Gelfand MS, Pervouchine DD. 2009. **Modulation of alternative splicing by long-range RNA structures in Drosophila.** Nucleic Acids Res. 37(14):4533-44.

Ray SK, Li HJ, Leiter AB. 2017. **Oligomeric form of C-terminal-binding protein coactivates NeuroD1-mediated transcription.** FEBS Lett. 591(1):205-212.

Schaeper U, Boyd, J.M., Verma, S., Uhlmann, E., Subramanian, T., Chinnadurai, G. 1995. **Molecular cloning and characterization of a cellular phosphoprotein that interacts with a conserved C-terminal domain of adenovirus E1A involved in negative modulation of oncogenic transformation.** Proc Natl Acad Sci U S A. 92:10467-10471.

Schmitz F, Konigstorfer A, Sudhof TC. 2000. **RIBEYE, a component of synaptic ribbons: a protein's journey through evolution provides insight into synaptic ribbon function.** Neuron. 28:857-872.

Shi Y, Sawada, J, Sui G, Affar EB, Whetstine JR, Lan F, Ogawa H, Luke MPS, Nakatani Y, Shi Y. 2003. **Coordinated histone modifications mediated by a CtBP co-repressor complex.** Nature. 422(6933):735-738.

Shukla S, Agarwal P, Kumar A. 2022. **Disordered regions tune order in chromatin organization and function.** Biophys Chem. 281:106716.

Soto LF, Li Z, Santoso CS, Berenson A, Ho I, Shen VX, Yuan S, Fuxman Bass JI. 2022. **Compendium of human transcription factor effector domains.** Mol Cell. 82(3):514-526.

Stankiewicz TR, Gray JJ, Winter AN, Linseman DA. 2014. **C-terminal binding proteins: central players in development and disease.** Biomol Concepts. 5(6):489-511.

Stern MD, Aihara H, Cho KH, Kim GT, Horiguchi G, Roccaro GA, Guevara E, Sun HH, Negeri D, Tsukaya H and others. 2007. **Structurally related Arabidopsis ANGUSTIFOLIA is functionally distinct from the transcriptional corepressor CtBP.** Dev Genes Evol. 217(11-12):759-769.

Sutrias-Grau M, Arnosti DN. 2004. **CtBP contributes quantitatively to Knirps repression activity in an NAD binding-dependent manner.** Mol Cell Biol. 24(13):5953-5966.

Theillet FX, Kalmar L, Tompa P, Han KH, Selenko P, Dunker AK, Daughdrill GW, Uversky VN. 2013. **The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins.** Intrinsically Disord Proteins. 1(1):e24360.

tom Dieck S, Altrock WD, Kessels MM, Qualmann B, Regus H, Brauner D, Fejtova A, Bracko O, Gundelfinger ED, Brandstatter JH. 2005. **Molecular dissection of the photoreceptor ribbon synapse: physical interaction of Bassoon and RIBEYE is essential for the assembly of the ribbon complex.** J Cell Biol. 168(5):825-836.

Turner J, Crossley M. 2001. **The CtBP family: enigmatic and enzymatic transcriptional co-repressors.** BioEssays. 23:683-690.

Wang SY, Iordanov M, Zhang Q. 2006. **c-Jun NH2-terminal kinase promotes apoptosis by down-regulating the transcriptional co-repressor CtBP.** J Biol Chem. 281(46):34810-34815.

Xie M, Zhang J, Yao T, Bryan AC, Pu Y, Labbe J, Pelletier DA, Engle N, Morrell-Falvey JL, Schmutz J, et al. 2020. **Arabidopsis C-terminal binding protein ANGUSTIFOLIA modulates transcriptional co-regulation of MYB46 and WRKY33.** New Phytol. 228(5):1627-1639.

Xu ZZ, Mathews DH. 2016. **Secondary structure prediction of single sequences using RNAstructure.** Methods Mol Bio. 1490:15-34.

Zhang YW, Arnosti DN. 2011. **Conserved catalytic and C-terminal regulatory domains of the C-terminal binding protein corepressor fine-tune the transcriptional response in development.** Mol Cell Biol 31(2):375-384.

Zhang Q, Nottke A, Goodman RH. 2005. **Homeodomain-interacting protein kinase-2 mediates CtBP phosphorylation and degradation in UV-triggered apoptosis.** Proc Natl Acad Sci U S A. 102(8):2802-2807.

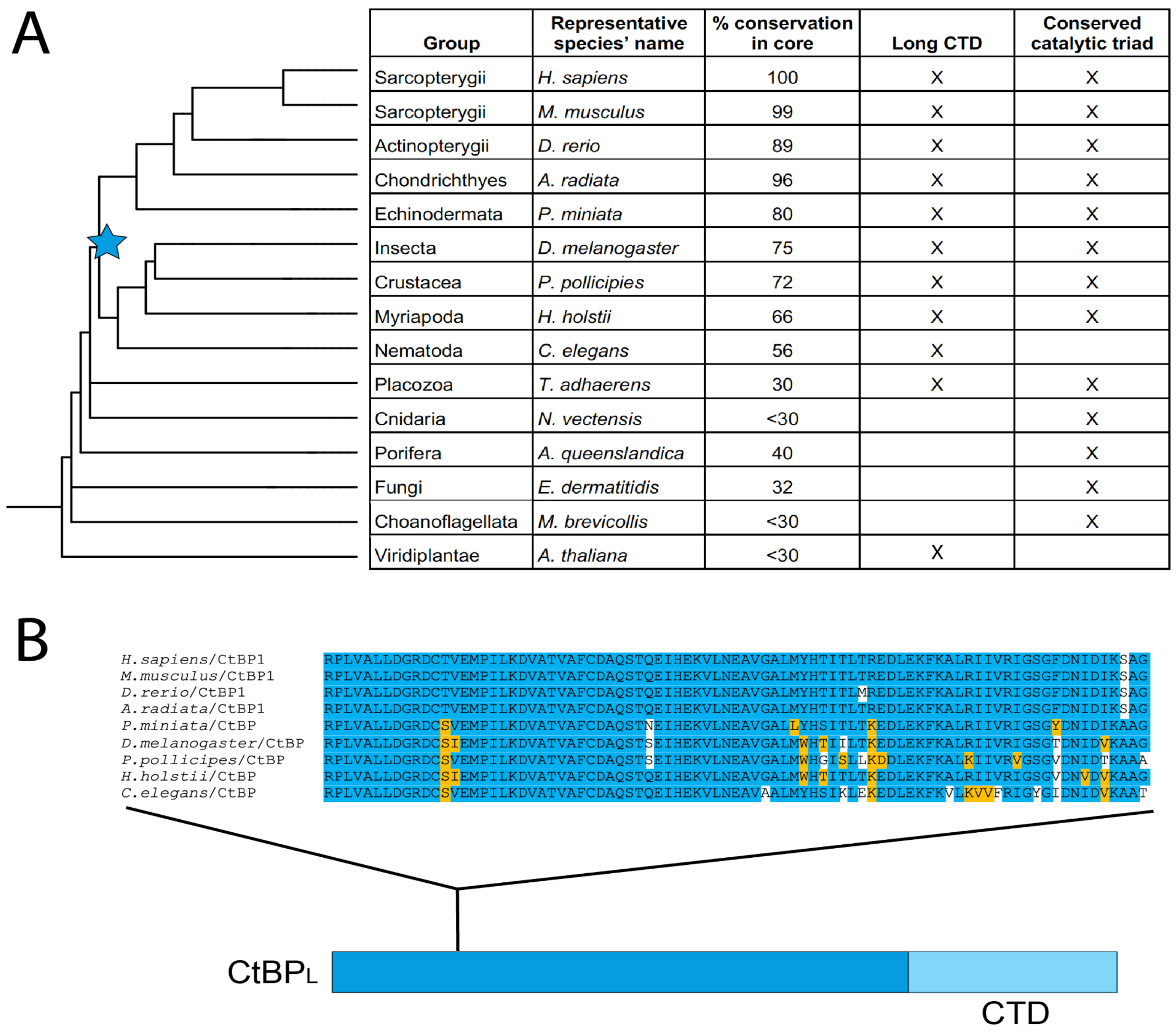
Zhang Q, Piston DW, Goodman RH. 2002. **Regulation of corepressor function by nuclear NADH**. Science. 295(5561):1895-1897.

Zhang Q, Yoshimatsu Y, Hildebrand J, Frisch SM, Goodman RH. 2003. **Homeodomain interacting protein kinase 2 promotes apoptosis by downregulating the transcriptional corepressor CtBP.** Cell. 115:177–186.

APPENDIX

This work was published as “Supplementary material” in the following manuscript:

Raicu, A.M., Kadiyala, D., Niblock, M., Jain, A., Yang, Y., Bird, K.M., Bertholf, K., Seenivasan, A., Siddiq, M., Arnosti, D.N. **The Cynosure of CtBP: Evolution of a Bilaterian Transcriptional Corepressor.** Mol Biol Evol. 2023 Feb 3;40(2):msad003. doi: 10.1093/molbev/msad003. PMID: 36625090.



Group	Representative species' name	% conservation in core	Long CTD	Conserved catalytic triad
Sarcopterygii	*H. sapiens*	100	X	X
Sarcopterygii	*M. musculus*	99	X	X
Actinopterygii	*D. rerio*	89	X	X
Chondrichthyes	*A. radiata*	96	X	X
Echinodermata	*P. miniata*	80	X	X
Insecta	*D. melanogaster*	75	X	X
Crustacea	*P. pollicipies*	72	X	X
Myriapoda	*H. holstii*	66	X	X
Nematoda	*C. elegans*	56	X	
Placozoa	*T. adhaerens*	30	X	X
Cnidaria	*N. vectensis*	<30		X
Porifera	*A. queenslandica*	40		X
Fungi	*E. dermatitidis*	32		X
Choanoflagellata	*M. brevicollis*	<30		X
Viridiplantae	*A. thaliana*	<30	X	

```
H.sapiens/CtBP1        RPLVALLDGRDCTVEMPILKDVATVAFCDAQSTQEIHEKVLNEAVGALMYHTITLTREDLEKFKALRIIVRIGSGFDNIDIKSAG
M.musculus/CtBP1       RPLVALLDGRDCTVEMPILKDVATVAFCDAQSTQEIHEKVLNEAVGALMYHTITLTREDLEKFKALRIIVRIGSGFDNIDIKSAG
D.rerio/CtBP1          RPLVALLDGRDCTVEMPILKDVATVAFCDAQSTQEIHEKVLNEAVGALMYHTITLMREDLEKFKALRIIVRIGSGFDNIDIKSAG
A.radiata/CtBP1        RPLVALLDGRDCTVEMPILKDVATVAFCDAQSTQEIHEKVLNEAVGALMYHTITLTREDLEKFKALRIIVRIGSGFDNIDIKSAG
P.miniata/CtBP         RPLVALLDGRDCSVEMPILKDVATVAFCDAQSTNEIHEKVLNEAVGALLYHSITLTKEDLEKFKALRIIVRIGSGYDNIDIKAAG
D.melanogaster/CtBP    RPLVALLDGRDCSIEMPILKDVATVAFCDAQSTSEIHEKVLNEAVGALMWHTIILTKEDLEKFKALRIIVRIGSGTDNIDVKAAG
P.pollicipes/CtBP      RPLVALLDGRDCSVEMPILKDVATVAFCDAQSTSEIHEKVLNEAVGALMWHGISLLKDDLEKFKALKIIVRVGSGVDNIDTKAAA
H.holstii/CtBP         RPLVALLDGRDCSIEMPILKDVATVAFCDAQSTQEIHEKVLNEAVGALMWHTITLTKEDLEKFKALRIIVRIGSGVDNVDVKAAG
C.elegans/CtBP         RPLVALLDGRDCSVEMPILKDVATVAFCDAQSTQEIHEKVLNEAVAALMYHSIKLEKEDLEKFKVLKVVFRIGYGIDNIDVKAAT
```

Figure S4.1. Conservation of CtBP features in metazoans and other eukaryotes. A) Representative sequences from vertebrates, invertebrates, and non-metazoan eukaryotes were aligned to the human CtBP1 sequence. Percent conservation in the core was calculated as the proportion of conserved residues between the region flanked by the NTD and CTD in humans (everything including and between the RPLVALL and NCVN motifs). We indicate (X) the presence of a long CTD (>80 residues), and the catalytic triad (REH). Bilaterians, indicated by the blue star, have canonical CtBP sequences, resembling the known human co-repressor. No CtBP homolog was identified in a representative Ctenophore (*M. leydi*, not shown). **B)** Alignment of bilaterians shown in panel A indicates a very high level of conservation in the dehydrogenase core. A sample alignment of the first 90 amino acids at the start of the dehydrogenase core illustrates this high level of conservation. Blue highlighting is used for conservation of a residue in >50% of species, and gold for chemically conserved residues.

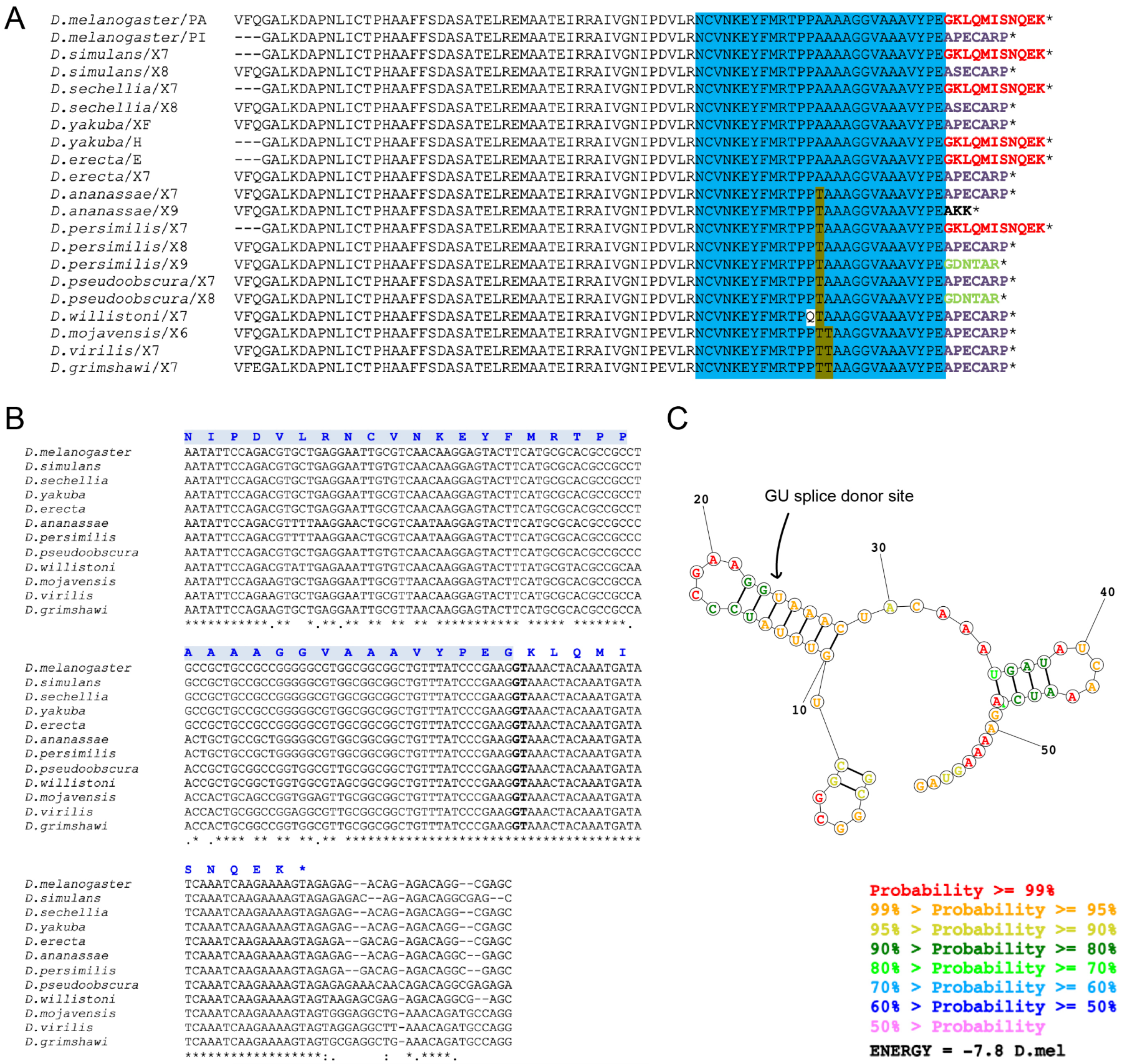
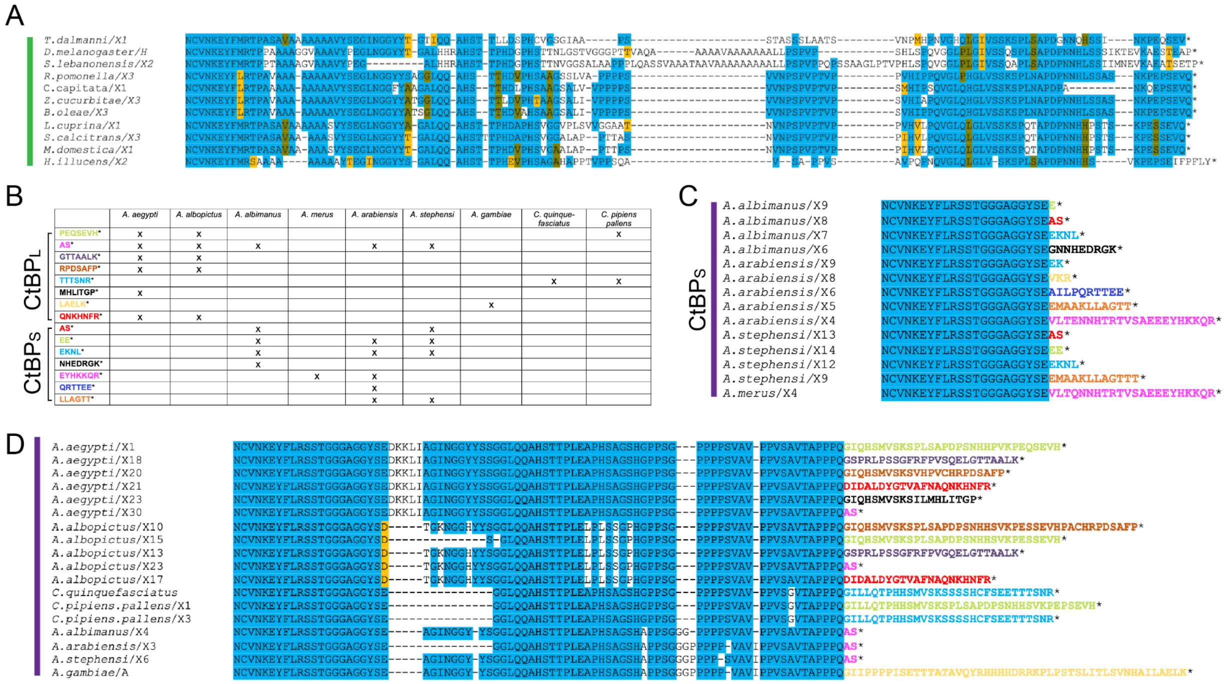


Figure S4.2. Formation of CtBP(S) through predicted RNA hairpin structures. A) Alignment of CtBP(S) from 12 Drosophila species. SNQEK-like endings are reported for half of the species, while APECARP-like are found in all. Each CTD variant is shown in a different color for the terminal residues. **B)** Alignment of the 3' end of protein coding exon 5 (light blue highlight), and 5' end of adjacent intron among the 12 Drosophila species. The single letter amino acid code indicates the corresponding peptide sequence for *D. melanogaster* for the SNQEK short isoform. This isoform is created through reading through the GT splice donor site (bold) and continuing until a STOP (TAG) codon is reached. We observe 100% conservation (asterisks below alignment) of ~50 nucleotides around the region where splicing normally creates the CtBP(L) isoform through use of the donor site. **C)** RNA structure prediction of the region starting with the triple alanine codons (panel B) highlights formation of a possible hairpin structure incorporating the GU splice donor site. Predicted probability of the structure is indicated in the bottom right legend.


A
T.dalmanni/X1
D.melanogaster/H
S.lebanonensis/X2
R.pomonella/X3
C.capitata/X1
Z.cucurbitae/X3
B.oleae/X3
L.cuprina/X1
S.calcitrans/X3
M.domestica/X1
H.illucens/X2
B
C
CtBPs
A.albimanus/X9 NCVNKEYFLRSSTGGGAGGYSEE*
A.albimanus/X8 NCVNKEYFLRSSTGGGAGGYSEAS*
A.albimanus/X7 NCVNKEYFLRSSTGGGAGGYSEEKNL*
A.albimanus/X6 NCVNKEYFLRSSTGGGAGGYSEGNNHEDRGK*
A.arabiensis/X9 NCVNKEYFLRSSTGGGAGGYSEEK*
A.arabiensis/X8 NCVNKEYFLRSSTGGGAGGYSEVKR*
A.arabiensis/X6 NCVNKEYFLRSSTGGGAGGYSEAILPQRTTEE*
A.arabiensis/X5 NCVNKEYFLRSSTGGGAGGYSEEMAAKLLAGTT*
A.arabiensis/X4 NCVNKEYFLRSSTGGGAGGYSEVLTENNHTRTVSAEEEYHKKQR*
A.stephensi/X13 NCVNKEYFLRSSTGGGAGGYSEAS*
A.stephensi/X14 NCVNKEYFLRSSTGGGAGGYSEEE*
A.stephensi/X12 NCVNKEYFLRSSTGGGAGGYSEEKNL*
A.stephensi/X9 NCVNKEYFLRSSTGGGAGGYSEEMAAKLLAGTTT*
A.merus/X4 NCVNKEYFLRSSTGGGAGGYSEVLTQNNHTRTVSAEEEYHKKQR*
D
A.aegypti/X1
A.aegypti/X18
A.aegypti/X20
A.aegypti/X21
A.aegypti/X23
A.aegypti/X30
A.albopictus/X10
A.albopictus/X15
A.albopictus/X13
A.albopictus/X23
A.albopictus/X17
C.quinquefasciatus
C.pipiens.pallens/X1
C.pipiens.pallens/X3
A.albimanus/X4
A.arabiensis/X3
A.stephensi/X6
A.gambiae/A

		A. aegypti	A. albopictus	A. albimanus	A. merus	A. arabiensis	A. stephensi	A. gambiae	C. quinque-fasciatus	C. pipiens pallens
CtBPL	PEQSEVH*	x	x							x
	AS*	x	x	x		x	x			
	GTTAALK*	x	x							
	RPDSAFP*	x	x							
	TTTSNR*								x	x
	MHLITGP*	x								
	LAELK*							x		
	QNKHNFR*	x	x							
CtBPs	AS*			x			x			
	EE*			x		x	x			
	EKNL*			x		x	x			
	NHEDRGK*			x						
	EYHKKQR*				x	x				
	QRTTEE*					x				
	LLAGTT*					x	x			

Figure S4.3. Culicidae (mosquitoes) have unique CTDs with both long and short isoforms. A) Alignment of long isoforms from all Brachycera examined. For all alignments, Brachycera are indicated by the vertical green line on the left, and Nematocera by the vertical purple line. **B)** Chart of the long and short CTD variants found in the selected mosquitoes (variants are indicated by the very C-terminal sequences e.g. PEQSEVH). All species except A. merus have one or more long CTDs, and some species also have a diversity of short CTDs. **C)** Alignment of short isoforms from mosquitoes. Short isoforms are only found in the Anopheles genus, and some species have up to five variants. Variants that are found in more than one species are colored the same (i.e the variant ending in KKQR is found in both *A. arabiensis* and *A. merus*, colored in pink.) **D)** Alignment of long isoforms from Culicidae examined indicates that there is high conservation until the terminal PPQ, after which many variations exist across the genera.

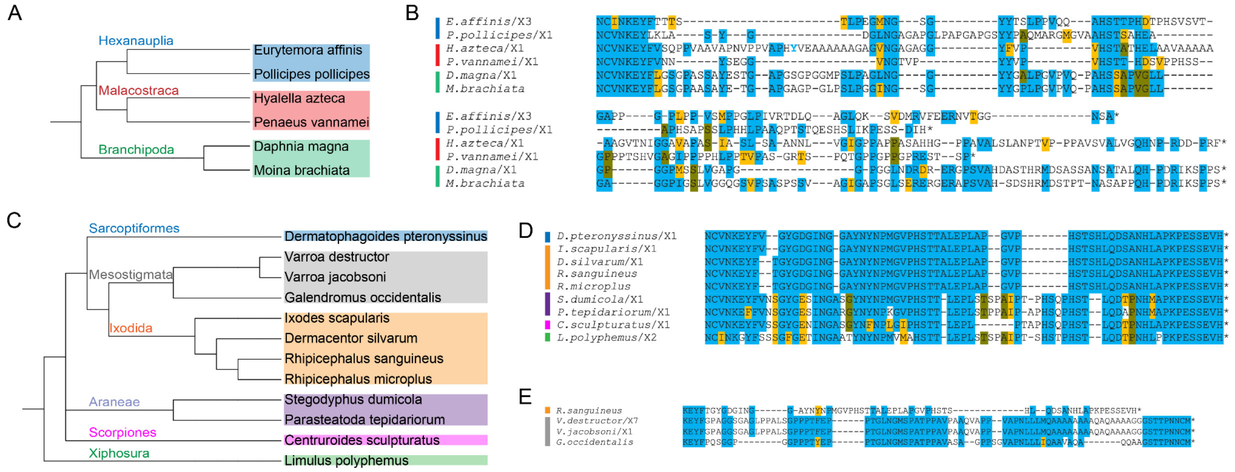


Figure S4.4. Alignment of crustacean and chelicerate CtBP CTDs. A) Phylogenetic tree of three Crustacean groups. **B)** Alignment of six crustaceans indicates that this subphylum experienced diversification particularly affecting the C-terminal sequences after the central block, while some species (*P. pollicipes*) retain presumed ancestral C-terminal sequences (e.g. SDIH). Vertical lines represent species shown in panel A. The tyrosine in light blue (*H. azteca*) is a conserved aromatic residue spaced differently than the other species, but conserved between the NCVN and GLNG--YY motifs. **C)** Phylogenetic tree of select chelicerates including mites, ticks, spiders, scorpions, and horseshoe crab. **D)** Alignment of chelicerate CTDs indicates very high conservation ending with the ancestral SEVH. Vertical bars on left follow the labels in panel C. **E)** Alignment of divergent CTDs from three Mesostigmata mites with the *R. sanguineus* tick, which represents an ancestral sequence.

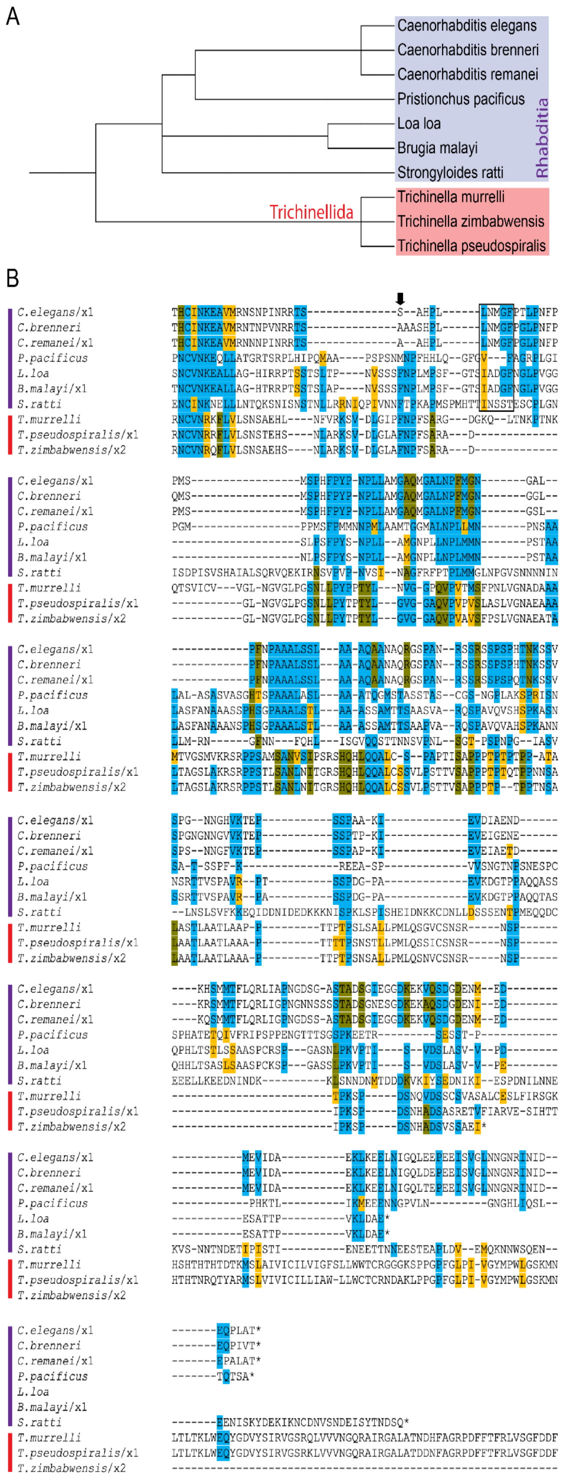


Figure S4.5. Nematodes have lineage-specific derived CTD sequences. A) Phylogenetic tree of select roundworms. **B)** Alignment of CtBP CTDs from nematode species covering several genera. Within the same genera (for instance, Caenorhabditis), many features are conserved, but from one genus to the next, there are few similarities in primary sequence. Only portions of the *T. murrelli* and *T. pseudospiralis* sequences are shown; an extension of an additional ~400 amino acids is predicted to complete these CTDs.

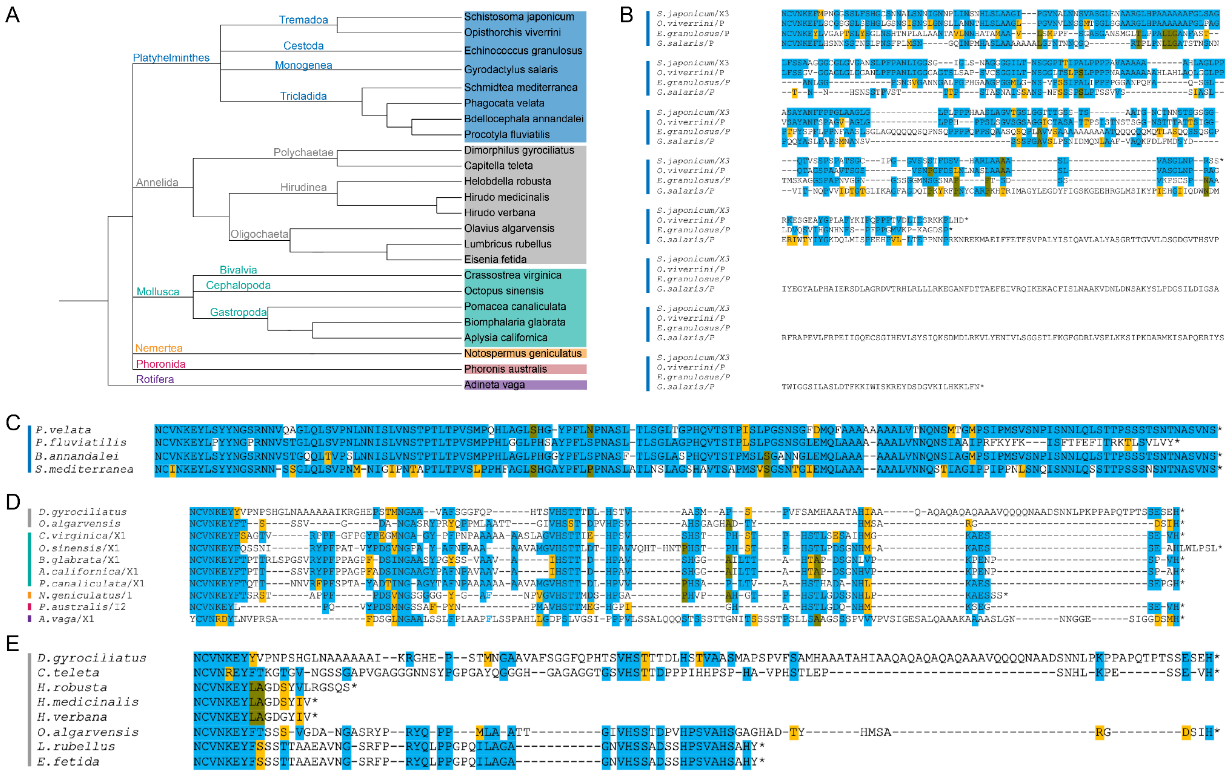


Figure S4.6. Spiralia are diverse in their CTDs. A) Phylogenetic tree of diverse Spiralia species encompassing six phyla. Colored vertical bars in B-E correspond to phyla indicated in A. **B)** Alignment of CTD sequences from trematodes, cestodes, and monogeneids indicates some similarities among these groups, but the sequences are distinct from those found in Triclad Platyhelminthes and the other spiralian species. **C)** Alignment of CTD from four flatworms (Tricladida) showing strong conservation of unique, derived CTD sequences, unlike those found in other spiralians. **D)** Alignment of species within five phyla illustrates the conservation of distinct CtBP CTD motifs, similar to conserved ecdysozoan CTDs. Conserved aromatic (F) residues that are spaced differently in the central block are indicated in blue font. **E)** Alignment of diverse annelids, including earthworms and leeches. The leeches (Hirudinea) are the only species found to express only CtBP(S). The Oligochaeta annelids have truncated CTDs that still resemble some of the protostome motifs, such as the central block (VNGSRY/F motif before the highlighted ILAGA).

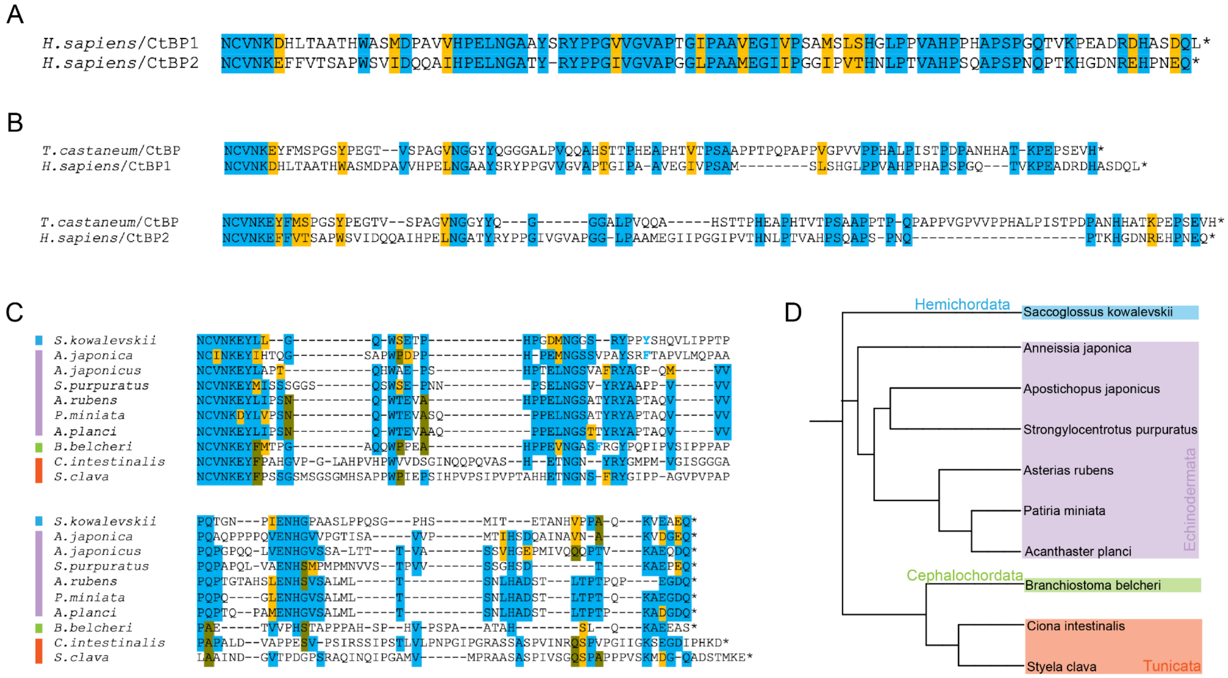

A
H.sapiens/CtBP1 NCVNKDHLTAATHWASMDPAVVHPELNGAAYSRYPPGVVGVAPTGIPAAVEGIVPSAMSLSHGLPPVAHPPHAPSPGQTVKPEADRDHASDQL*
H.sapiens/CtBP2 NCVNKEFFVTSAPWSVIDQQAIHPELNGATY-RYPPGIVGVAPGGLPAAMEGIIPGGIPVTHNLPTVAHPSQAPSPNQPTKHGDNREHPNEQ*
B
T.castaneum/CtBP NCVNKEYFMSPGSYPEGT--VSPAGVNGGYYQGGGALPVQQAHSTTPHEAPHTVTPSAAPPTPQPAPPVGPVVPPHALPISTPDPANHHAT-KPEPSEVH*
H.sapiens/CtBP1 NCVNKDHLTAATHWASMDPAVVHPELNGAAYSRYPPGVVGVAPTGIPA-AVEGIVPSAM--------SLSHGLPPVAHPPHAPSPGQ---TVKPEADRDHASDQL*
T.castaneum/CtBP NCVNKEYFMSPGSYPEGTV--SPAGVNGGYYQ---G------GGALPVQQA-----HSTTPHEAPHTVTPSAAPPTP-QPAPPVGPVVPPHALPISTPDPANHHATKPEPSEVH*
H.sapiens/CtBP2 NCVNKEFFVTSAPWSVIDQQAIHPELNGATYRYPPGIVGVAPGG-LPAAMEGIIPGGIPVTHNLPTVAHPSQAPS-PNQ-------------------PTKHGDNREHPNEQ*
C
S.kowalevskii NCVNKEYLL--G----------Q-WSETP------------HPGDMNGGS--RYPPYSHQVLIPPTP
A.japonica NCINKEYIHTQG---------SAPWPDPP------------H-PEMNGSSVPAYSRFTAPVLMQPAA
A.japonicus NCVNKEYLAPT-----------QHWAE-PS-----------HPTELNGSVAFRYAGP-QM-----VV
S.purpuratus NCVNKEYMISSSGGS-------QSWSE-PNN-----------PSELNGSV-YRYAAPP-V-----VV
A.rubens NCVNKEYLIPSN----------Q-WTEVA------------HPPELNGSATYRYAPTAQV-----VV
P.miniata NCVNKDYLVPSN----------Q-WTEVASQ-----------PPELNGSATYRYAPTAQV-----VV
A.planci NCVNKEYLIPSN----------Q-WTEVAAQ-----------PPELNGSTTYRYAPTAQV-----VV
B.belcheri NCVNKEYFMTPG---------AQQWPPEA------------HPPEVNGASFRGYPQPIPVSIPPPAP
C.intestinalis NCVNKEYFPAHGVP-G-LAHPVHPWVVDSGINQQPQVAS--H--ETNGN--YRYGMPM-VGISGGGA
S.clava NCVNKEYFPSSGSMSGSGMHSAPPWPIEPSIHPVPSIPVPTAHHETNGNS-FRYGIPP-AGVPVPAP
S.kowalevskii PQTGN---PIENHGPAASLPPQSG---PHS------MIT---ETANHVPPA-Q---KVEAEQ*
A.japonica PQAQPPPPQVENHGVVPGTISA-----VVP------MTIHSDQAINAVN-A-----KVDGEQ*
A.japonicus PQPGPQQ-LVENHGVSSA-LTT---T-VA-------SSVHGEPMIVQQQPTV----KAEQDQ*
S.purpuratus PQPAPQL-VAENHSMPMPMNVVS--TPVV-------SSGHSD--------T-----KAEPEQ*
A.rubens PQPTGTAHSLENHSVSALML-----T----------SNLHADST---LTPTPQP----EGDQ*
P.miniata PQPQ----GLENHGVSALML-----T----------SNLHADST---LTPT-Q---KAEGDQ*
A.planci PQPTQ--PAMENHGVSALML-----T----------SNLHADST---LTPTP----KADGDQ*
B.belcheri PAE-----TVVPHSTAPPPAH-SP--HV-PSPA---ATAH--------SL--Q---KAEEAS*
C.intestinalis PAPALD--VAPPESV-PSIRSSIPSTLVLPNPGIPGRASSASPVINRQSPVPGIIGKSEGDIPHKD*
S.clava LAAIND--GVTPDGPSRAQINQIPGAMV------MPRAASASPIVSGQSPAPPPVSKMDG-QADSTMKE*
D
Hemichordata
Saccoglossus kowalevskii
Anneissia japonica
Apostichopus japonicus
Strongylocentrotus purpuratus
Asterias rubens
Patiria miniata
Acanthaster planci
Echinodermata
Cephalochordata
Branchiostoma belcheri
Ciona intestinalis
Styela clava
Tunicata

Figure S4.7. Evidence for homology among deuterostome CTD sequences. A) Human CtBP1 vs. CtBP2 CTD alignment. This portion is 50% conserved. **B)** Representative protostome (*T. castaneum* beetle) compared to human CtBP1 CTD and CtBP2 CTD. Some motifs are conserved in protostomes and deuterostomes, such as the central block. **C)** Alignment of non-vertebrate deuterostome CtBP CTD sequences. These species have a single CtBP protein. Aromatic-containing central block sequences are conserved, as are terminal amino acids. Conserved Y and F residues that are spaced differently in the central block are indicated in blue font. **D)** Phylogenetic tree of the four (sub)phyla in Deuterostomia that harbor a single CtBP gene.

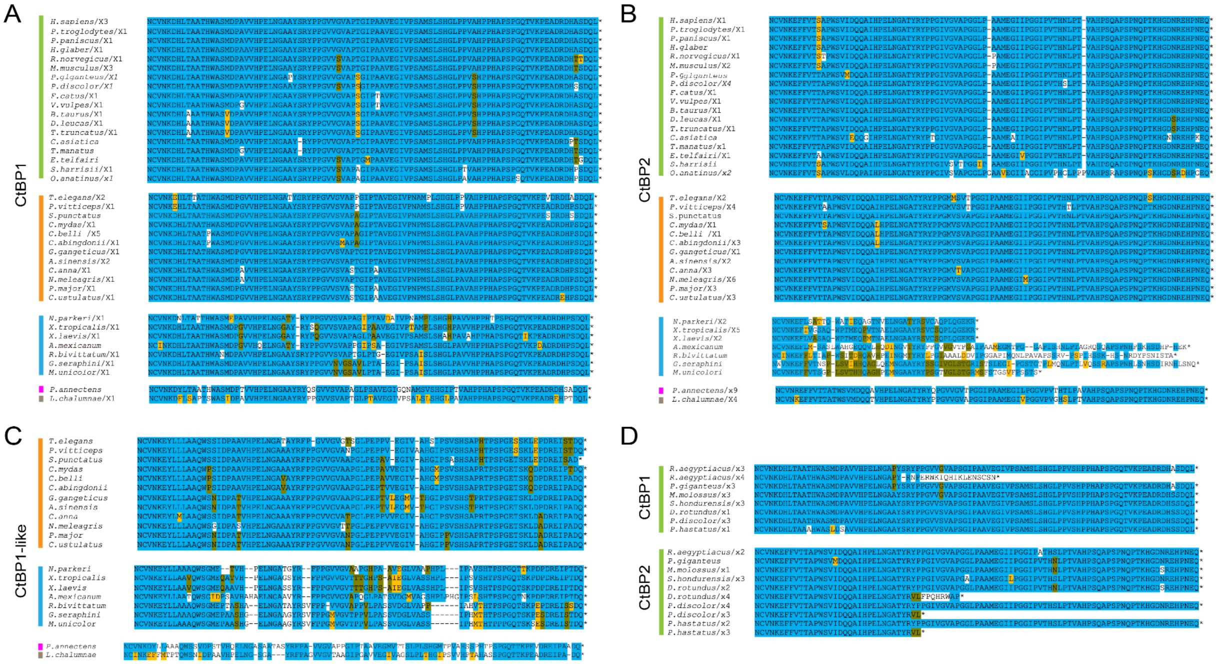


Figure S4.8. Sarcopterygii have conserved CtBP1 and CtBP2 CTDs, with variations in amphibian CtBP2. A) CtBP1 CTD alignment of all Sarcopterygii species analyzed. For all panels, vertical colored bars indicate the same clades as in Figure 10. **B)** CtBP2 CTD alignment of all Sarcopterygii species analyzed indicates high conservation of the CTD, aside from amphibia, where shorter and more divergent variants are observed. **C)** Alignment of CtBP1-like sequences, which are found in all Sarcopterygii other than mammals. **D)** Bats are the only mammals found to have short and long versions of CtBP1 and CtBP2 CTDs.

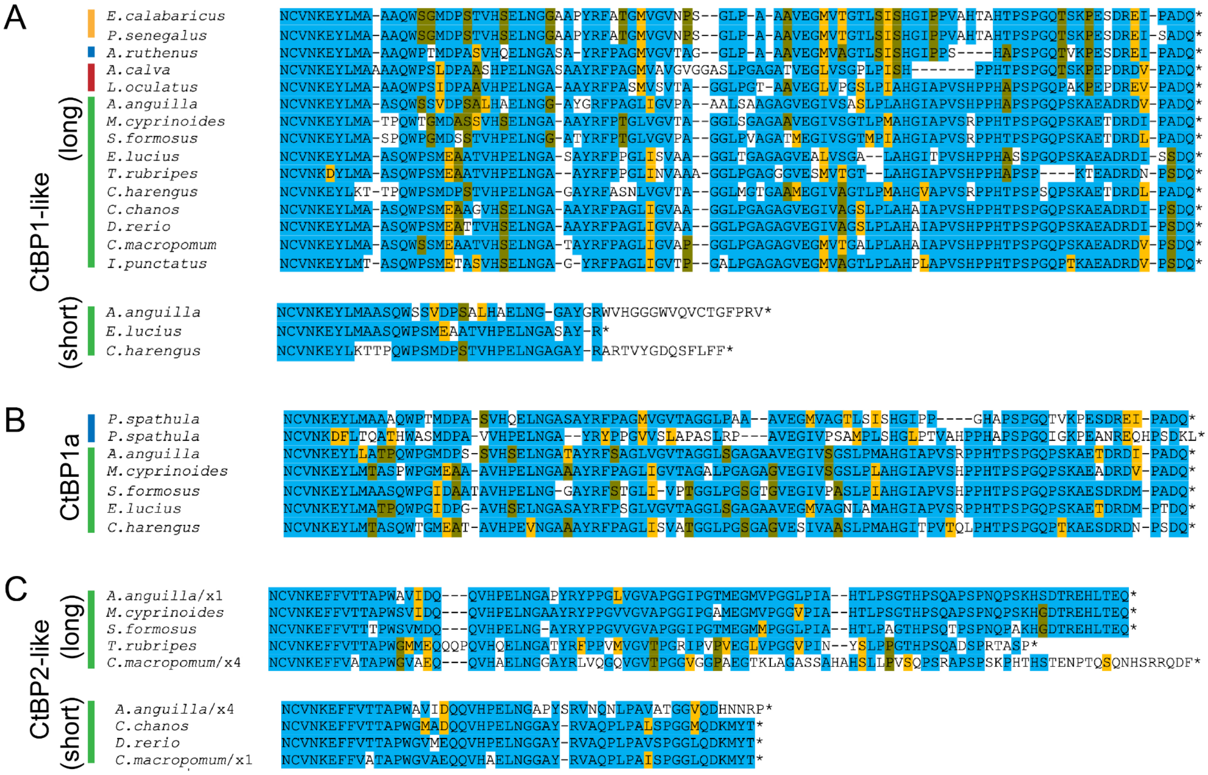

A
CtBP1-like
(long)
E.calabaricus NCVNKEYLMA-AAQWSGMDPSTVHSELNGGAAPYRFATGMVGVNPS--GLP-A-AAVEGMVTGTLSISHGIPPVAHTAHTPSPGQTSKPESDREI-PADQ*
P.senegalus NCVNKEYLMA-AAQWSGMDPSTVHSELNGGAAPYRFATGMVGVNPS--GLP-A-AAVEGMVTGTLSISHGIPPVAHTAHTPSPGQTSKPESDREI-SADQ*
A.ruthenus NCVNKEYLMA-AAQWPTMDPASVHQELNGASA-YRFPAGMVGVTAG--GLP-A-AAVEGMVAGTLSISHGIPPS----HAPSPGQTVKPESDREI-PADQ*
A.calva NCVNKEYLMAAAAQWPSLDPAASHPELNGASAAYRFPAGMVAVGVGGASLPGAGATVEGLVSGPLPISH-------PPHTPSPGQTSKPEPDRDV-PADQ*
L.oculatus NCVNKEYLMA-AAQWPSIDPAAVHPELNGA-AAYRFPASMVSVTA--GGLPGT-AAVEGLVPGSLPIAHGIAPVSHPPHAPSPGQPAKPEPDREV-PADQ*
A.anguilla NCVNKEYLMA-ASQWSSVDPSALHAELNGG-AYGRFPAGLIGVPA--AALSAAGAGVEGIVSASLPLAHGIAPVSHPPHAPSPGQPSKAEADRDV-PADQ*
M.cyprinoides NCVNKEYLMA-TPQWTGMDASSVHSELNGA-AAYRFPTGLVGVTA--GGLSGAGAAVEGIVSGTLPMAHGIAPVSRPPHTPSPGQPSKAETDRDI-PADQ*
S.formosus NCVNKEYLMA-SPQWPGMDSSTVHPELNGG-ATYRFPTGLVGVPA--GGLPVAGATMEGIVSGTMPIAHGIAPVSRPPHTPSPGQPSKAETDRDL-PADQ*
E.lucius NCVNKEYLMA-ASQWPSMEAATVHPELNGA-SAYRFPPGLISVAA--GGLTGAGAGVEALVSGA--LAHGITPVSHPPHASSPGQPSKAEADRDI-SSDQ*
T.rubripes NCVNKDYLMA-ASQWPSMEAATVHPELNGA--AYRFPPGLINVAAA-GGLPGAGGGVESMVTGT--LAHGIAPVSHPPHAPSP----KTEADRDN-PSDQ*
C.harengus NCVNKEYLKT-TPQWPSMDPSTVHPELNGA-GAYRFASNLVGVTA--GGLMGTGAAMEGIVAGTLPMAHGVAPVSRPPHTPSPSQPSKAETDRDL-PADQ*
C.chanos NCVNKEYLMA-ASQWPSMEAAGVHSELNGA-AAYRFPAGLIGVAA--GGLPGAGAGVEGIVAGSLPLAHAIAPVSHPPHTPSPGQPSKAEADRDI-PSDQ*
D.rerio NCVNKEYLMA-ASQWPSMEATTVHSELNGA-AAYRFPAGLIGVAA--GGLPGAGAGVEGIVAGSLPLAHAIAPVSHPPHTPSPGQPSKAEADRDI-PSDQ*
C.macropomum NCVNKEYLMA-ASQWSSMEAATVHSELNGA-TAYRFPAGLIGVAP--GGLPGAGAGVEGMVTGALPLAHAIAPVSHPPHTPSPGQPSKAEADRDV-PSDQ*
I.punctatus NCVNKEYLMT-ASQWPSMETASVHSELNGA-G-YRFPAGLIGVTP--GALPGAGAGVEGMVAGTLPLAHPLAPVSHPPHTPSPGQPTKAEADRDV-PSDQ*
(short)
A.anguilla NCVNKEYLMAASQWSSVDPSALHAELNG-GAYGRWVHGGGWVQVCTGFPRV*
E.lucius NCVNKEYLMAASQWPSMEAATVHPELNGASAY-R*
C.harengus NCVNKEYLKTTPQWPSMDPSTVHPELNGAGAY-RARTVYGDQSFLFF*
B
CtBP1a
P.spathula NCVNKEYLMAAAQWPTMDPA-SVHQELNGASAYRFPAGMVGVTAGGLPAA--AVEGMVAGTLSISHGIPP----GHAPSPGQTVKPESDREI-PADQ*
P.spathula NCVNKDFLTQATHWASMDPA-VVHPELNGA--YRYPPGVVSLAPASLRP---AVEGIVPSAMPLSHGLPTVAHPPHAPSPGQIGKPEANREQHPSDKL*
A.anguilla NCVNKEYLLATPQWPGMDPS-SVHSELNGATAYRFSAGLVGVTAGGLSGAGAAVEGIVSGSLPMAHGIAPVSRPPHTPSPGQPSKAETDRDI-PADQ*
M.cyprinoides NCVNKEYLMTASPWPGMEAA-AVHPELNGAAAYRFPAGLIGVTAGALPGAGAGVEGIVSGSLPLAHGIAPVSHPPHTPSPGQPSKAEADRDV-PADQ*
S.formosus NCVNKEYLMAASQWPGIDAATAVHPELNG-GAYRFSTGLI-VPTGGLPGSGTGVEGIVPASLPIAHGIAPVSHPPHTPSPGQPSKAESDRDM-PADQ*
E.lucius NCVNKEYLMATPQWPGIDPG-AVHSELNGASAYRFPSGLVGVTAGGLSGAGAAVEGMVAGNLAMAHGIAPVSRPPHTPSPGQPSKAETDRDM-PTDQ*
C.harengus NCVNKEYLMTASQWTGMEAT-AVHPEVNGAAAYRFPAGLISVATGGLPGSGAGVESIVAASLPMAHGITPVTQLPHTPSPGQPTKAESDRDN-PSDQ*
C
CtBP2-like
(long)
A.anguilla/x1 NCVNKEFFVTTAPWAVIDQ---QVHPELNGAPYRYPPGLVGVAPGGIPGTMEGMVPGGLPIA--HTLPSGTHPSQAPSPNQPSKHSDTREHLTEQ*
M.cyprinoides NCVNKEFFVTTAPWSVIDQ---QVHPELNGAAYRYPPGVVGVAPGGIPGAMEGMVPGGVPIA--HTLPSGTHPSQAPSPNQPSKHGDTREHLTEQ*
S.formosus NCVNKEFFVTTTPWSVMDQ---QVHPELNG-AYRYPPGVVGVAPGGIPGTMEGMMPGGLPIA--HTLPAGTHPSQTPSPNQPAKHGDTREHLTEQ*
T.rubripes NCVNKEFFVTTAPWGMMEQQQPQVHQELNGATYRFPPVMVGVTPGRIPVPVEGLVPGGVPIN--YSLPPGTHPSQADSPRTASP*
C.macropomum/x4 NCVNKEFFVATAPWGVAEQ---QVHAELNGGAYRLVQGQVGVTPGGVGGPAEGTKLAGASSAHAHSLLPVSQPSRAPSPSKPHTHSTENPTQSQNHSRRQDF*
(short)
A.anguilla/x4 NCVNKEFFVTTAPWAVIDQQVHPELNGAPYSRVNQNLPAVATGGVQDHNNRP*
C.chanos NCVNKEFFVTTAPWGMADQQVHPELNGGAY-RVAQPLPALSPGGMQDKMYT*
D.rerio NCVNKEFFVTTAPWGVMEQQVHPELNGGAY-RVAQPLPAVSPGGLQDKMYT*
C.macropomum/x1 NCVNKEFFVATAPWGVAEQQVHAELNGGAY-RVAQPLPAISPGGLQDKMYT*

Figure S4.9. Some Actinopterygii encode two additional CtBP paralogs. A) Alignment of long isoforms of CtBP1-like. This protein is found in all Actinopterygii sampled. Short isoforms are present in certain Teleostei. **B)** Alignment of long isoforms of CtBP1a indicates this protein is conserved only in select Teleostei and Chondrostei. *P. spathula* seems to be the only species to have two versions of what we termed CtBP1a; it is unclear if one of them is indeed CtBP1-like. **C)** Alignment of CtBP2-like, which is found only in select Teleostei. Both long and short isoforms are encoded, and conserved across species. Vertical bars on the left correspond to the clades highlighted in Figure 11A. Due to mis-annotation of some of the genes in Actinopterygii, we renamed sequences from certain species (see Materials and Methods).

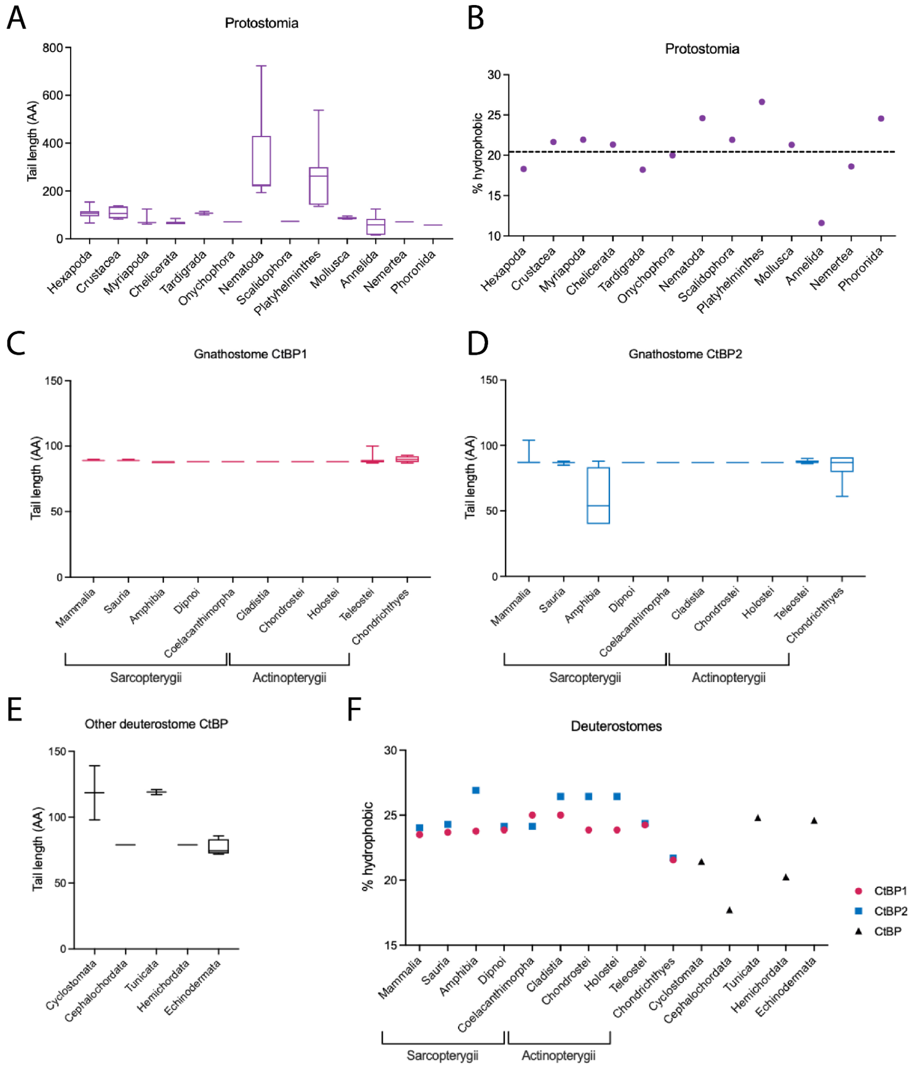


Figure S4.10. Properties of CtBP CTDs in Bilateria. A) CTD length (in Amino Acids, AA) across Protostomia. Average CTD lengths are similar except in divergent nematode and platyhelminth lineages, where some are over 500 AA. The leech family within annelids are the only phylum to have lost the long CTD. Box and whisker plots are used to indicate the middle 50% of values (box) and range (whiskers from lowest to highest value). Horizontal line indicates the median value. For all plots, tail properties were determined by averaging the longest CTD of all species within each clade. **B)** Hydrophobic content of CTD in protostomes (average ~21%, indicated by dashed line). **C)** Length of CtBP1 CTD in gnathostomes spanning Sarcopterygii, Actinopterygii, and Chondrichthyes. The length has remained mostly unchanged. **D)** Length of CtBP2 CTD in gnathostomes indicates that amphibians experienced some diversification in tail length. **E)** Length of CtBP CTD in non-Gnathostome deuterostomes also averages ~100 residues in length, but is more variable. **F)** Proportion of hydrophobic residues in vertebrate CtBP1 and CtBP2 CTDs and invertebrate deuterostome CtBP CTDs. Most classes have an average of about 24% hydrophobic residues, while some experience diversification, such as those within the non-vertebrate phyla. Percent hydrophobicity was determined by calculating the number of M, I, V, L, F, Y, and W residues in the longest tail from each species, and averaging the hydrophobic content across each clade. Clades include 1 to 18 species in total.

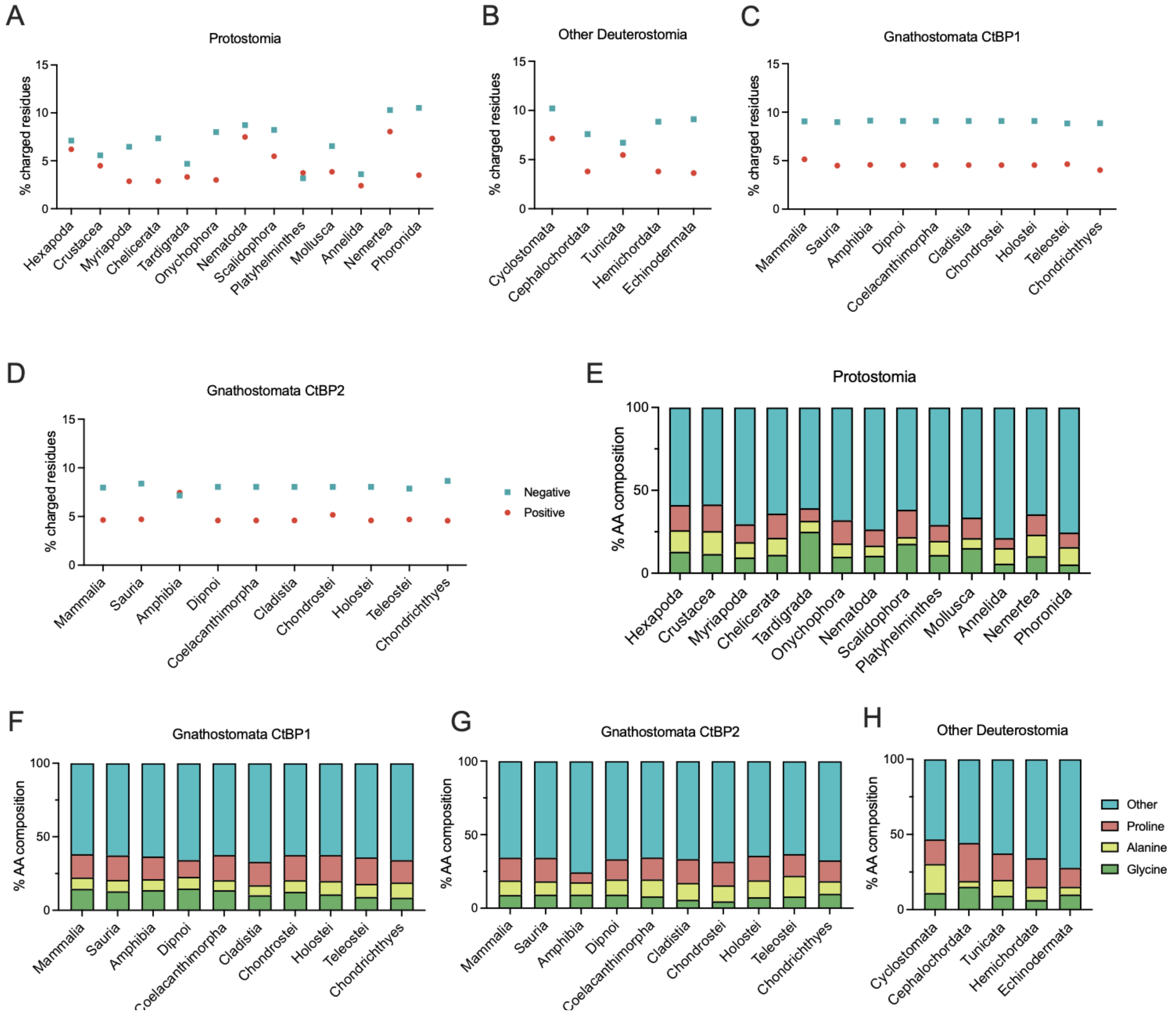
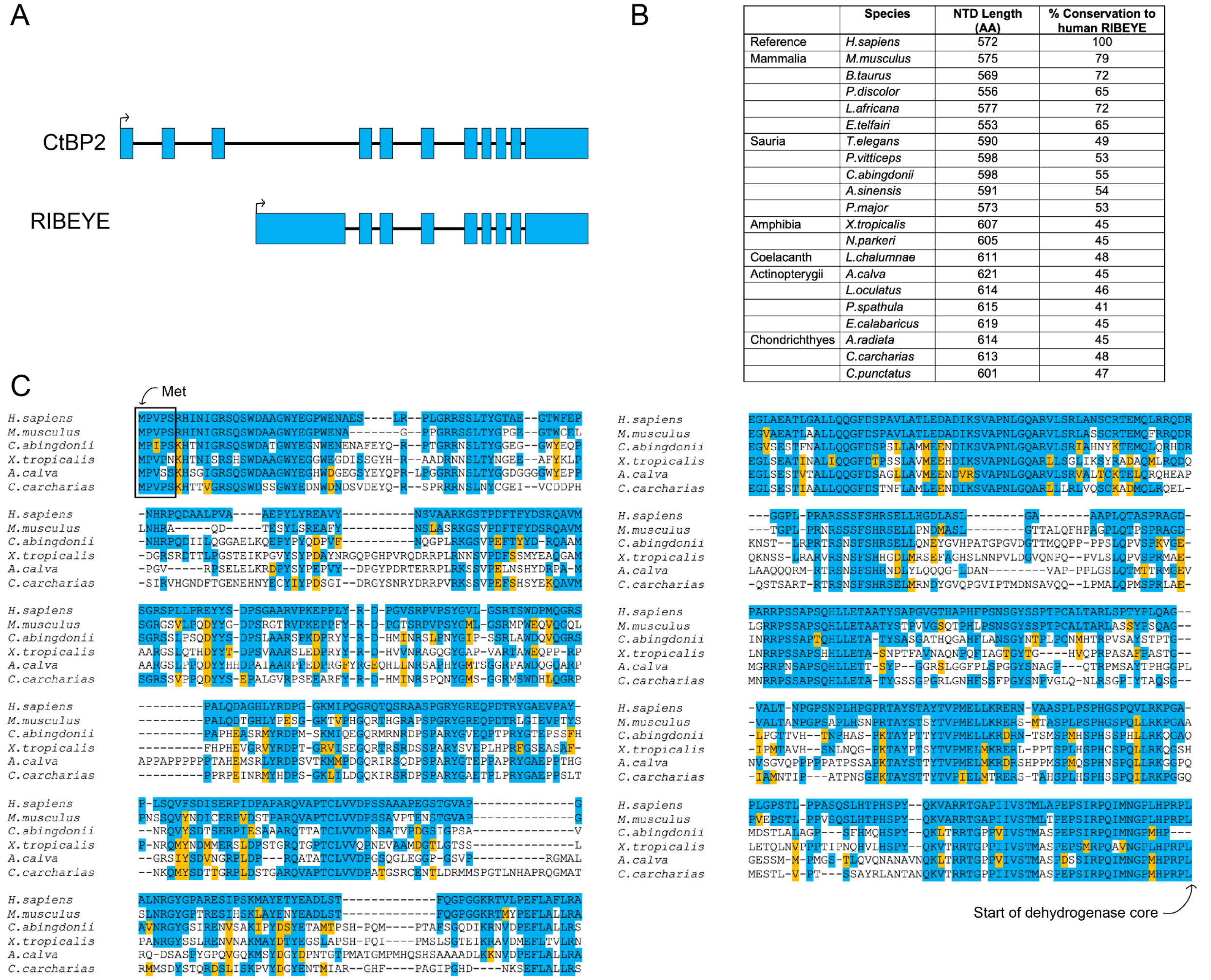
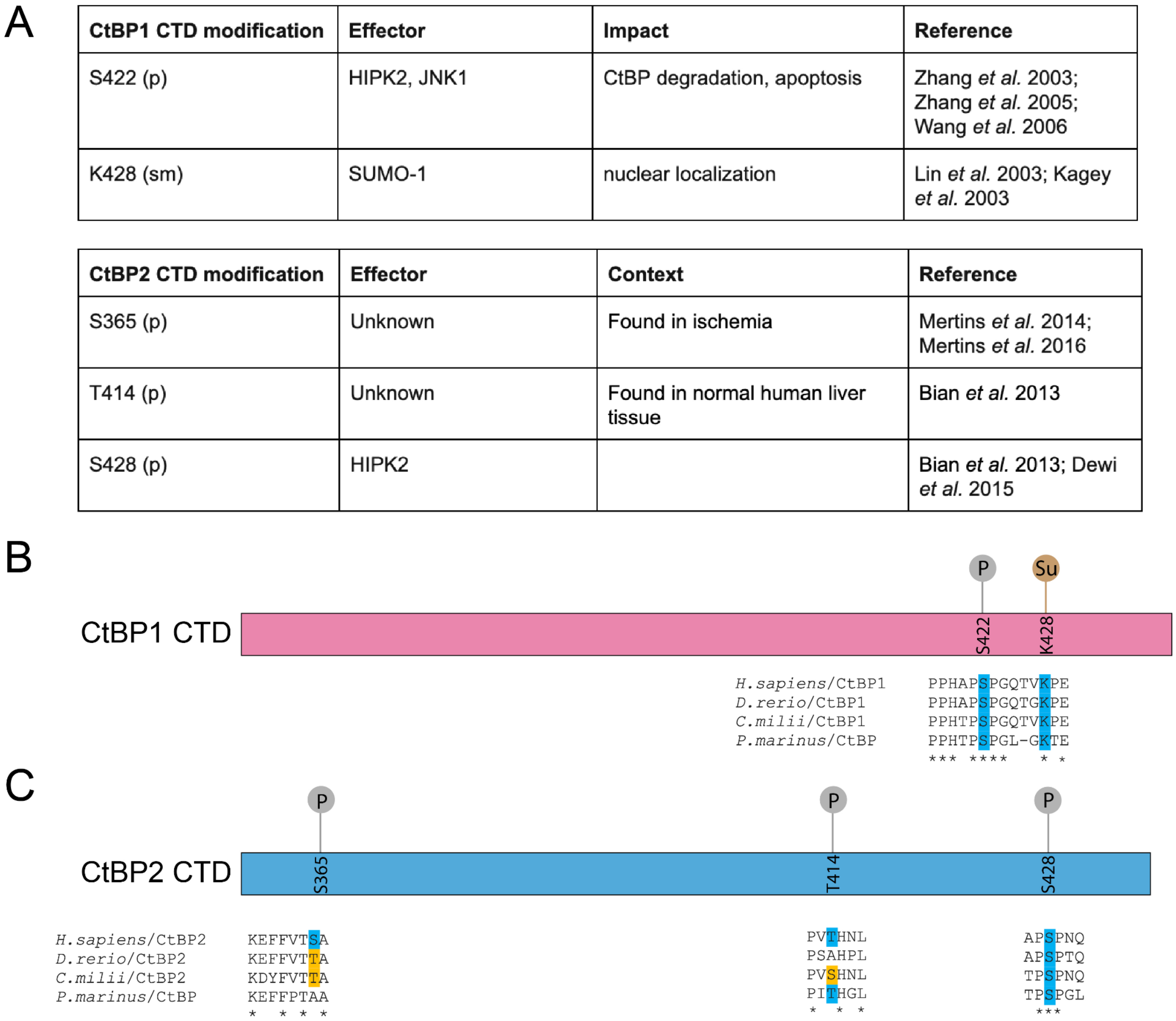


Figure S4.11. Composition of amino acid residues in the CtBP CTD across Metazoa. Composition of charged residues (positive: K, R; negative: D, E) in CtBP CTDs across **A)** Protostomia, **B)** non-gnathostome Deuterostomia, **C)** Gnathostomata CtBP1, and **D)** Gnathostomata CtBP2. Negatively charged residues are indicated by blue squares, and positively charged residues are indicated by red circles. We find that in deuterostomes, the charged residues are consistently making up under 15% of the CTD, while in protostomes there's more variability (6% in Annelida to 18% in Nemertea). AA composition of the CTD in **E)** Protostomia, **F)** Gnathostomata CtBP1, **G)** Gnathostomata CtBP2, and **H)** non-gnathostome Deuterostomia. Glycines are indicated in green, alanines in yellow, prolines in pink and all other residues in blue. We find that across Metazoa, most species have about 40% of their CTD made up of P, A, G residues, which are disorder-promoting. These analyses were performed by selecting one isoform of longest length from each species and averaging within each clade.



	Species	NTD Length (AA)	% Conservation to human RIBEYE
Reference	*H.sapiens*	572	100
Mammalia	*M.musculus*	575	79
	B.taurus	569	72
	P.discolor	556	65
	L.africana	577	72
	E.telfairi	553	65
Sauria	*T.elegans*	590	49
	P.vitticeps	598	53
	C.abingdonii	598	55
	A.sinensis	591	54
	P.major	573	53
Amphibia	*X.tropicalis*	607	45
	N.parkeri	605	45
Coelacanth	*L.chalumnae*	611	48
Actinopterygii	*A.calva*	621	45
	L.oculatus	614	46
	P.spathula	615	41
	E.calabaricus	619	45
Chondrichthyes	*A.radiata*	614	45
	C.carcharias	613	48
	C.punctatus	601	47

Figure S4.12. N-terminal variations in CtBP2 are conserved. A) Schematic of human CtBP2 and the RIBEYE variant with an extended NTD. Different transcriptional start sites are used to create a unique NTD. Blue boxes represent all exons (not to scale). **B)** Table indicating percent identity relative to the human sequence of the CtBP2 NTD in select gnathostomes. NTD lengths across Gnathostomata are consistent with the human RIBEYE variant length. Cyclostomata (not shown) do not have RIBEYE isoforms. **C)** Alignment of the NTD sequences from select gnathostomes. Blue highlight was used for the human RIBEYE sequence and for any residues that were completely conserved in the other species as compared to *H. sapiens*. Orange was used for chemical conservation compared to the human sequence. We find that the translational start site is highly conserved (boxed), and there are many blocks of conservation across the NTD. The sequence terminates with the RPL sequence, which is the start of the conserved dehydrogenase core.



CtBP1 CTD modification	Effector	Impact	Reference

S422 (p)	HIPK2, JNK1	CtBP degradation, apoptosis	Zhang *et al.* 2003;
			Zhang *et al.* 2005;
			Wang *et al.* 2006

K428 (sm)	SUMO-1	nuclear localization	Lin *et al.* 2003; Kagey *et al.* 2003

CtBP2 CTD modification	Effector	Context	Reference
S365 (p)	Unknown	Found in ischemia	Mertins *et al.* 2014; Mertins *et al.* 2016
T414 (p)	Unknown	Found in normal human liver tissue	Bian *et al.* 2013
S428 (p)	HIPK2		Bian *et al.* 2013; Dewi *et al.* 2015

B
P
Su
CtBP1 CTD
S422
K428
H.sapiens/CtBP1 PPHAPSPGQTVKPE
D.rerio/CtBP1 PPHAPSPGQTGKPE
C.milii/CtBP1 PPHTPSPGQTVKPE
P.marinus/CtBP PPHTPSPGL-GKTE
*** **** * *
C
P
P
P
CtBP2 CTD
S365
T414
S428
H.sapiens/CtBP2 KEFFVTSA PVTHNL APSPNQ
D.rerio/CtBP2 KEFFVTTA PSAHPL APSPTQ
C.milii/CtBP2 KDYFVTTA PVSHNL TPSPNQ
P.marinus/CtBP KEFFPTAA PITHGL TPSPGL
* * * * * * * ***

Figure S4.13. Post-translational modifications of the human CtBP CTDs. A) Chart summarizing experimentally validated PTMs on the *H. sapiens* CtBP1 and CtBP2 CTDs. **B)** Schematic of the *H. sapiens* CtBP1 CTD (89 residues). S422 and K428 are empirically determined phosphorylation and sumoylation targets, respectively. Both residues are conserved across vertebrates (blue highlight in alignment). **C)** Schematic of the *H. sapiens* CtBP2 CTD (87 residues). S365, T414 and S428 are empirically determined phosphorylation targets. Only S428 is conserved across vertebrates, while S365 and T414 show lower conservation.

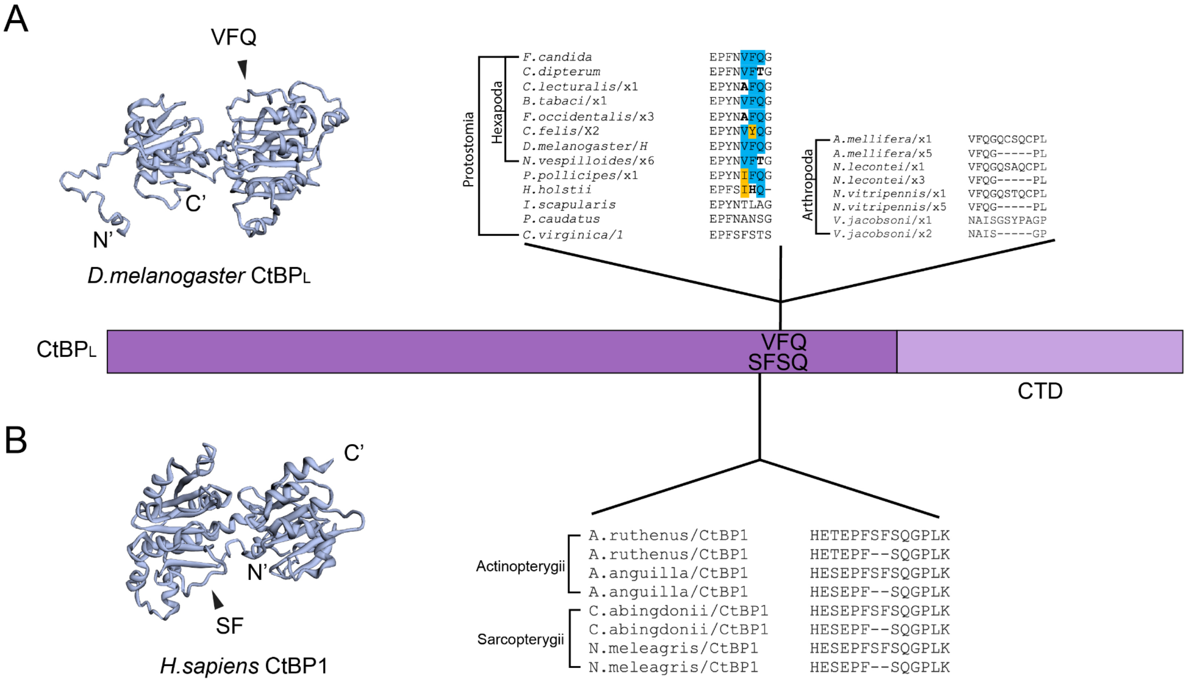

A
VFQ
C'
N'
D.melanogaster CtBPL
Protostomia
Hexapoda
F.candida EPFNVFQG
C.dipterum EPFNVFTG
C.lecturalis/x1 EPYNAFQG
B.tabaci/x1 EPYNVFQG
F.occidentalis/x3 EPYNAFQG
C.felis/X2 EPYNVYQG
D.melanogaster/H EPYNVFQG
N.vespilloides/x6 EPYNVFTG
P.pollicipes/x1 EPYNIFQG
H.holstii EPFSIHQ-
I.scapularis EPYNTLAG
P.caudatus EPFNANSG
C.virginica/1 EPFSFSTS
Arthropoda
A.mellifera/x1 VFQGQCSQCPL
A.mellifera/x5 VFQG-----PL
N.lecontei/x1 VFQGQSAQCPL
N.lecontei/x3 VFQG-----PL
N.vitripennis/x1 VFQGQSTQCPL
N.vitripennis/x5 VFQG-----PL
V.jacobsoni/x1 NAISGSYPAGP
V.jacobsoni/x2 NAIS-----GP
CtBPL
VFQ
SFSQ
CTD
B
C'
N'
SF
H.sapiens CtBP1
Actinopterygii
Sarcopterygii
A.ruthenus/CtBP1 HETEPFSFSQGPLK
A.ruthenus/CtBP1 HETEPF--SQGPLK
A.anguilla/CtBP1 HESEPFSFSQGPLK
A.anguilla/CtBP1 HESEPF--SQGPLK
C.abingdonii/CtBP1 HESEPFSFSQGPLK
C.abingdonii/CtBP1 HESEPF--SQGPLK
N.meleagris/CtBP1 HESEPFSFSQGPLK
N.meleagris/CtBP1 HESEPF--SQGPLK

Figure S4.14. Little variation in the CtBP dehydrogenase core. A) Variation in the dehydrogenase core in Protostomia is limited to alternative splicing of the VFQ tripeptide in Diptera and alternative splicing of a 5mer in Arthropoda (alignments shown on right). This site is indicated on the Drosophila predicted structure on the left, and maps to an unstructured loop on the surface of the protein. Schematic of CtBP(L) is shown in purple, below. **B)** Variation in the dehydrogenase core in Deuterostomia is limited to alternative splicing of the SF motif in Actinopterygii and Sarcopterygii CtBP1 (alignment shown on right). This site is indicated on the human predicted structure on the left, and maps to an unstructured loop on the surface of the protein. These conserved alternative splicing events map to the same region of the core, near the CTD.

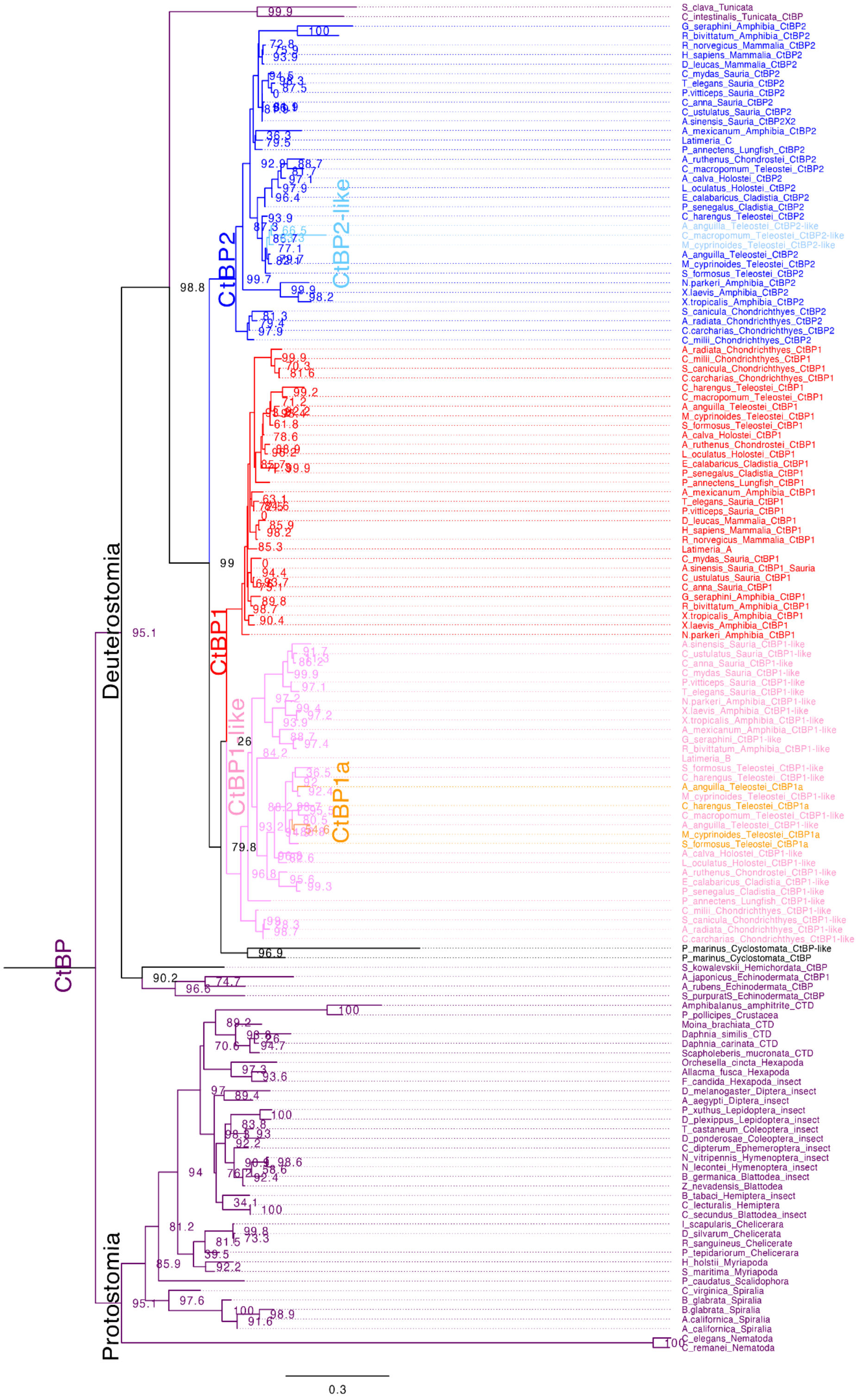

S_clava_Tunicata
C_intestinalis_Tunicata_CtBP
G_seraphini_Amphibia_CtBP2
R_bivittatum_Amphibia_CtBP2
R_norvegicus_Mammalia_CtBP2
H_sapiens_Mammalia_CtBP2
D_leucas_Mammalia_CtBP2
C_mydas_Sauria_CtBP2
T_elegans_Sauria_CtBP2
P.vitticeps_Sauria_CtBP2
C_anna_Sauria_CtBP2
C_ustulatus_Sauria_CtBP2
A.sinensis_Sauria_CtBP2X2
A_mexicanum_Amphibia_CtBP2
Latimeria_C
P_annectens_Lungfish_CtBP2
A_ruthenus_Chondrostei_CtBP2
C_macropomum_Teleostei_CtBP2
A_calva_Holostei_CtBP2
L_oculatus_Holostei_CtBP2
E_calabaricus_Cladistia_CtBP2
P_senegalus_Cladistia_CtBP2
C_harengus_Teleostei_CtBP2
A_anguilla_Teleostei_CtBP2-like
C_macropomum_Teleostei_CtBP2-like
M_cyprinoides_Teleostei_CtBP2-like
A_anguilla_Teleostei_CtBP2
M_cyprinoides_Teleostei_CtBP2
S_formosus_Teleostei_CtBP2
N.parkeri_Amphibia_CtBP2
X.laevis_Amphibia_CtBP2
X.tropicalis_Amphibia_CtBP2
S_canicula_Chondrichthyes_CtBP2
A_radiata_Chondrichthyes_CtBP2
C.carcharias_Chondrichthyes_CtBP2
C_milii_Chondrichthyes_CtBP2
A_radiata_Chondrichthyes_CtBP1
C_milii_Chondrichthyes_CtBP1
S_canicula_Chondrichthyes_CtBP1
C.carcharias_Chondrichthyes_CtBP1
C_harengus_Teleostei_CtBP1
C_macropomum_Teleostei_CtBP1
A_anguilla_Teleostei_CtBP1
M_cyprinoides_Teleostei_CtBP1
S_formosus_Teleostei_CtBP1
A_calva_Holostei_CtBP1
A_ruthenus_Chondrostei_CtBP1
L_oculatus_Holostei_CtBP1
E_calabaricus_Cladistia_CtBP1
P_senegalus_Cladistia_CtBP1
P_annectens_Lungfish_CtBP1
A_mexicanum_Amphibia_CtBP1
T_elegans_Sauria_CtBP1
P.vitticeps_Sauria_CtBP1
D_leucas_Mammalia_CtBP1
H_sapiens_Mammalia_CtBP1
R_norvegicus_Mammalia_CtBP1
Latimeria_A
C_mydas_Sauria_CtBP1
A.sinensis_Sauria_CtBP1_Sauria
C_ustulatus_Sauria_CtBP1
C_anna_Sauria_CtBP1
G_seraphini_Amphibia_CtBP1
R_bivittatum_Amphibia_CtBP1
X.tropicalis_Amphibia_CtBP1
X.laevis_Amphibia_CtBP1
N.parkeri_Amphibia_CtBP1
A.sinensis_Sauria_CtBP1-like
C_ustulatus_Sauria_CtBP1-like
C_anna_Sauria_CtBP1-like
C_mydas_Sauria_CtBP1-like
P.vitticeps_Sauria_CtBP1-like
T_elegans_Sauria_CtBP1-like
N.parkeri_Amphibia_CtBP1-like
X.laevis_Amphibia_CtBP1-like
X.tropicalis_Amphibia_CtBP1-like
A_mexicanum_Amphibia_CtBP1-like
G_seraphini_CtBP1-like
R_bivittatum_Amphibia_CtBP1-like
Latimeria_B
S_formosus_Teleostei_CtBP1-like
C_harengus_Teleostei_CtBP1-like
A_anguilla_Teleostei_CtBP1a
M_cyprinoides_Teleostei_CtBP1-like
C_harengus_Teleostei_CtBP1a
C_macropomum_Teleostei_CtBP1-like
A_anguilla_Teleostei_CtBP1-like
M_cyprinoides_Teleostei_CtBP1a
S_formosus_Teleostei_CtBP1a
A_calva_Holostei_CtBP1-like
L_oculatus_Holostei_CtBP1-like
A_ruthenus_Chondrostei_CtBP1-like
E_calabaricus_Cladistia_CtBP1-like
P_senegalus_Cladistia_CtBP1-like
P_annectens_Lungfish_CtBP1-like
C_milii_Chondrichthyes_CtBP1-like
S_canicula_Chondrichthyes_CtBP1-like
A_radiata_Chondrichthyes_CtBP1-like
C.carcharias_Chondrichthyes_CtBP1-like
P_marinus_Cyclostomata_CtBP-like
P_marinus_Cyclostomata_CtBP
S_kowalevskii_Hemichordata_CtBP
A_japonicus_Echinodermata_CtBP1
A_rubens_Echinodermata_CtBP
S_purpuratS_Echinodermata_CtBP
Amphibalanus_amphitrite_CTD
P_pollicipes_Crustacea
Moina_brachiata_CTD
Daphnia_similis_CTD
Daphnia_carinata_CTD
Scapholeberis_mucronata_CTD
Orchesella_cincta_Hexapoda
Allacma_fusca_Hexapoda
F_candida_Hexapoda_insect
D_melanogaster_Diptera_insect
A_aegypti_Diptera_insect
P_xuthus_Lepidoptera_insect
D_plexippus_Lepidoptera_insect
T_castaneum_Coleoptera_insect
D_ponderosae_Coleoptera_insect
C_dipterum_Ephemeroptera_insect
N_vitripennis_Hymenoptera_insect
N_lecontei_Hymenoptera_insect
B_germanica_Blattodea_insect
Z_nevadensis_Blattodea
B_tabaci_Hemiptera_insect
C_lecturalis_Hemiptera
C_secundus_Blattodea_insect
I_scapularis_Chelicerara
D_silvarum_Chelicerata
R_sanguineus_Chelicerate
P_tepidariorum_Chelicerara
H_holstii_Myriapoda
S_maritima_Myriapoda
P_caudatus_Scalidophora
C_virginica_Spiralia
B_glabrata_Spiralia
B.glabrata_Spiralia
A.californica_Spiralia
A_californica_Spiralia
C_elegans_Nematoda
C_remanei_Nematoda
CtBP2
CtBP2-like
CtBP1
CtBP1-like
CtBP1a
Deuterostomia
Protostomia
CtBP
99.9
100
98.8
99
26
79.8
95.1
90.2
96.6
74.7
96.9
89.2
70.6
97.3
93.6
97
89.4
94
81.2
85.9
95.1
97.6
100
98.9
91.6
81.5
39.5
92.2
99.8
73.3
34.1
100
0.3

Figure S4.15. The evolutionary history of CtBP sequences across Bilateria. Maximum likelihood phylogeny of the CtBP sequences inferred with the best-fit model of sequence evolution and parameters (Q.insect + F + Gamma) as determined by BIC (see Materials & Methods). The root of the phylogeny is placed at the divergence of protostomes and deuterostomes. Colors correspond to different CtBP homologs: Dark purple (CtBP), red (CtBP1), pink (CtBP1-like),

Figure S4.15 (cont'd)
orange (CtBP1a), blue (CtBP2), and light blue (CtBP2-like). Node support values are shown for major gene duplication events and represent SH-aLRT from 1000 replicates. The inferred phylogenetic tree shows all vertebrates to encode multiple CtBP paralogs, whereas protostomes and invertebrate deuterostomes encode a single CtBP. In the vertebrates, a first duplication created the CtBP1 and CtBP2 paralogs, and a subsequent duplication created CtBP1-like. Two other gene duplication events occurred within the Actinopterygii, and Cyclostomata experienced their own duplication event. These two lamprey paralogs branch together; because no other cyclostome sequences exist with which this branch can be better resolved, two different evolutionary scenarios are possible: 1) lamprey-specific duplication of CtBP after divergence of jawed and jawless vertebrates or 2) retention of an earlier duplication shared by all vertebrates but in which paralogs resemble each other because of gene conversion. Certain sequence patterns support scenario one, as illustrated in Figure 8. The second scenario alleviates a need for a cyclostome-specific duplication; however, this scenario is less likely to be the case because the duplication event that formed CtBP1 and CtBP2 is likely more ancient based on structural features of the molecules. We also inferred a phylogeny using a commonly used model of protein evolution (JTT + Gamma) to ensure robustness of the uncovered relationships to model choice (data not shown). The major discrepancy between the two models is in the placement of Cyclostomata. In the JTT model, the Cyclostomata is placed sister to CtBP1-like.

CHAPTER 5: EVIDENCE FOR A REGULATORY ROLE OF THE UNSTRUCTURED C-TERMINAL DOMAIN OF CTBP

This work was published in the following preprint and adapted here:

Raicu, A.M., Suresh, M., & Arnosti, D. N. (2023). **A regulatory role for the unstructured C-terminal domain of the CtBP transcriptional corepressor.** *bioRxiv*. 2023.05.19.541472; doi: https://doi.org/10.1101/2023.05.19.541472

ABSTRACT

The C-terminal Binding Protein (CtBP) is a transcriptional corepressor that plays critical roles in development, tumorigenesis, and cell fate. CtBP proteins are structurally similar to alpha hydroxyacid dehydrogenases and additionally feature an unstructured C-terminal domain (CTD). The role of a possible dehydrogenase activity has been postulated for the corepressor, although *in vivo* substrates are unknown, but the functional significance of the CTD is unclear. In the mammalian system, CtBP proteins lacking the CTD are able to function as transcriptional regulators and oligomerize, putting into question the significance of the CTD for gene regulation. Yet, the presence of an unstructured CTD of ~100 residues, including some short motifs, is conserved across Bilateria, indicating the importance of this domain. To study the *in vivo* functional significance of the CTD, we turned to the *Drosophila melanogaster* system, which naturally expresses isoforms with the CTD (CtBP(L)), and isoforms lacking the CTD (CtBP(S)). We used the CRISPRi system to test dCas9-CtBP(S) and dCas9-CtBP(L) on diverse endogenous genes, to directly compare their transcriptional impacts *in vivo*. Interestingly, CtBP(S) was able to significantly repress transcription of the *E2F2* and *Mpp6* genes, while CtBP(L) had minimal impact, suggesting that the long CTD modulates CtBP's repression activity. In contrast, in cell culture, the isoforms behaved similarly on a transfected *Mpp6* reporter. Thus, we have identified context-specific effects of these two developmentally-regulated isoforms, and propose that differential expression of CtBP(S) and CtBP(L) may provide a spectrum of repression activity suitable for developmental programs.

INTRODUCTION

Eukaryotic transcription factors and cofactors are rich in unstructured domains; these proteins have a higher percentage of predicted intrinsically disordered regions (IDR) than the average protein (Uversky, 2016). Some unstructured domains have been shown to participate in specific transcriptional processes, such as the C-terminal domain (CTD) of RNA polymerase II, which is a platform for association of factors involved in capping, splicing, and polyadenylation (Harlen and Churchman, 2017). However, the roles of many IDRs present in these factors are still unknown. Transcriptional regulators can take on a diversity of roles in the cell, and unstructured domains may not necessarily play a role specific to gene regulation; yet, it has been speculated that these IDRs can assist with protein-protein interactions, or in phase separation of transcription condensates.

The C-terminal binding protein (CtBP) is a highly conserved transcriptional corepressor that contains a prominent IDR in its CTD. The CtBP CTD of about 100 amino acids is conserved across Bilateria, and despite overall lower sequence conservation than other parts of the protein, it retains certain properties, such as the predicted unstructured nature of the domain (Raicu *et al.* 2023). A few lineages, such as roundworms and flatworms, have novel, derived CTD sequences that are predicted to form structures. However, the conservation in primary sequence, length, and unstructured property of the CTD in bilaterians suggests that this IDR plays an important role, perhaps in gene regulation.

Mammalian genomes encode the CtBP1 and CtBP2 paralogs, which play overlapping and non-redundant roles in regulating expression of genes involved in apoptosis, the epithelial to mesenchymal transition, and cell differentiation (Grooteclaes *et al.* 2003; Fang *et al.* 2006; Jin *et al.* 2007; Paliwal *et al.* 2012). The CtBP1 and CtBP2 CTDs exhibit 50% sequence conservation, which is much lower than that of the central core dehydrogenase domain (Raicu *et al.* 2023). This core domain contains residues critical for oligomerization of CtBP monomers and for NADH binding, as well as *in vitro* dehydrogenase activity (Madison *et al.* 2013). CtBP can oligomerize and repress genes without the CTD, putting into question the significance of the CTD in gene regulation (Kumar *et al.* 2002; Madison *et al.* 2013). Interestingly, CtBP isoforms without the CTD exist in certain tetrapods such as birds and amphibians (Raicu *et al.* 2023). Additionally, the single *Drosophila melanogaster* CtBP locus encodes short isoforms that lack the CTD (CtBP(S)) and another which retains the long CTD (CtBP(L); Mani-Telang and Arnosti, 2007). Thus, *D. melanogaster* is an appropriate model system to test for a possible role of the CtBP CTD in gene regulation.

Previous work using GAL4-CtBP fusions in the Drosophila embryo demonstrated that the two isoforms have similar repressive effects on even-skipped-lacZ reporters, and both isoforms individually rescue a CtBP null fly, albeit with some phenotypes (Sutrias-Grau and Arnosti, 2004; Zhang and Arnosti, 2011). Thus, the CTD does not seem to play an essential role in developmental programs. The expression pattern of the two isoforms exhibit developmentally distinct profiles; CtBP(S) is expressed throughout development, while CtBP(L) is highly expressed in the embryonic stage (Mani-Telang and Arnosti, 2007). The fact that short isoforms have been independently derived in other insects, such as Hymenoptera, and in other lineages in Bilateria suggests that expression of both isoforms is somehow important (Raicu *et al.* 2023). The conflicting evidence compelled us to test how the two CtBP isoforms regulate gene expression *in vivo*.

Here, we have made use of precise genetic tools in Drosophila to probe the function of the fly CtBP isoforms, CtBP(L) and CtBP(S), for their ability to regulate gene expression in a developing fly. Specifically, we used the CRISPRi system to assess the function of chimeric dCas9-CtBP proteins targeted to gene promoters *in vivo*. This method allowed us to compare the activity of the long and short isoforms in the same context, in both fly wing tissue and in cell culture. We found that CtBP(S) is a more potent repressor of the *E2F2/Mpp6* bidirectional promoter than CtBP(L), but that this difference in repression ability is not observed on a transiently transfected *Mpp6*-luciferase reporter. Thus, in some contexts the CTD seems to provide a regulatory function, but the difference observed between endogenous gene regulation and transient transfections raises the possibility that the effect may be chromatin-dependent. Additionally, gene promoters targeted here had differential sensitivity to CtBP recruitment, indicating a further level of regulatory specificity, in accord with recent high-throughput assays (Jacobs *et al.* 2022, *bioRxiv*).

RESULTS

Creation of dCas9-CtBP chimeras to regulate gene expression

To investigate differences in gene regulation by the CtBP(L) and CtBP(S) isoforms in Drosophila, we employed CRISPRi (Reviewed in Kampmann *et al.* 2018). We fused the coding sequence of each CtBP isoform to a nuclease dead Cas9 (dCas9) enzyme to recruit CtBP corepressors to target promoters using gene-specific guide RNAs (gRNA; **Figure 5.1A**). dCas9-CtBP(L) and dCas9-CtBP(S) are expressed at similar levels in S2 cells, according to western blot (**Figure S5.1**).

We expressed the chimeric proteins in L3 wing discs using the *nubbin*-GAL4 driver. Flies homozygous for both *nub*-GAL4 and UAS:dCas9-CtBP were crossed to lines obtained from Harvard TRiP expressing two tandem gRNAs targeting a gene's proximal promoter (**Figure 5.1B;** Zirin *et al.* 2022). We previously tested dCas9-Rb chimeras in L3 discs, where we observed gene-specific effects after targeting ~30 different gene promoters; here, we targeted many of the same promoters (Raicu, Castanheira, Arnosti. *bioRxiv*; **Table S5.1**).

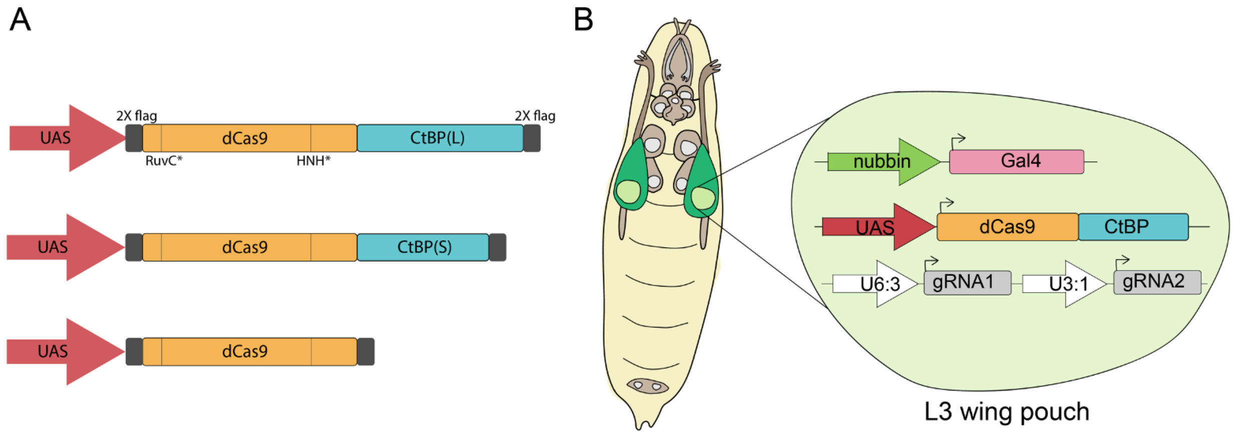


Figure 5.1. An *in vivo* system for targeting CtBP isoforms to gene promoters using CRISPRi. A) The fly CtBP(L) and CtBP(S) FLAG-tagged coding sequences were fused to the C-terminus of the S. pyogenes nuclease dead Cas9 (dCas9; D10A mutation in RuvC catalytic domain and H840A mutation in HNH catalytic domain), and placed under UAS expression. FLAG-tagged dCas9 was used as a negative control. Vertical lines in dCas9 represent the inactivating mutations. **B)** *Drosophila melanogaster* expressing three transgenes were generated for tissue-specific expression of dCas9-CtBP effectors using GAL4-UAS. Flies express dCas9-CtBP chimeras in the *nubbin* expression pattern (wing pouch of L3 wing discs), with ubiquitous expression of two tandem gRNAs designed to target a single gene's promoter. Flies used in experiments express one copy of each of the three transgenes. gRNA flies were designed by Harvard TRiP (Zirin *et al.* 2022).

The epithelial cells of the developing wing are a highly sensitive tissue that has been used to measure developmental perturbation of a number of regulatory pathways. To screen for genetic effects, we allowed the flies expressing the three transgenes to grow to adulthood, and then assessed adult wing phenotypes from targeting each promoter. We note that the *nub*-GAL4>UAS:dCas9-CtBP flies crossed to a non-targeting gRNA control fly line (QUAS) produced mild wing phenotypes, consisting chiefly of supernumerary bristles (**Figure 5.2A**). We presume that ectopic CtBP, even when fused to dCas9, may interact with diverse endogenous CtBP binding sites on the genome. The control gRNAs used here did not produce phenotypes with dCas9-Rb corepressor fusions tested previously, so the effect here is CtBP-specific (Raicu, Castanheira, Arnosti. *bioRxiv*).

CtBP isoforms have diverse effects on gene promoters

We recruited CtBP(L) and CtBP(S) to a number of gene promoters, with specific effects observed only on a few (**Table S5.1**). Here, we detail the effects of targeting the *E2F2/Mpp6* bidirectional promoter, the insulin receptor (*InR*) promoter, and the promoter of *Acf*, a nucleosome remodeling subunit (**Figure 5.2**). Targeting CtBP(S) to the divergent *E2F2/Mpp6* promoter produced small wings with severe morphological defects, similar to that seen with dCas9-Rb proteins (**Figure 5.2B**; Raicu, Castanheira, Arnosti. *bioRxiv*). Intriguingly, CtBP(L) did not produce this phenotype, but instead produced milder effects, including wings with ectopic veins and supernumerary bristles (**Figure 5.2B**). dCas9 alone did not produce any phenotypic effect, indicating that the observed phenotypes are CtBP-specific. The clear difference between the long and short isoforms on this promoter suggests that the long CTD may inhibit CtBP's gene regulatory activities.

Interestingly, the strong CtBP(S) effect is only seen when using two gRNAs; recruitment using the individual gRNAs produced milder effects, including the ectopic veins seen with the CtBP(L) isoform when both gRNA were used (**Figure S5.2**). Interestingly, the number of wings with supernumerary bristles was less than that observed for the non-targeting control QUAS gRNA; we speculate that nonspecific CtBP overexpression effects are suppressed by targeting the chimeric protein to specific DNA locations.

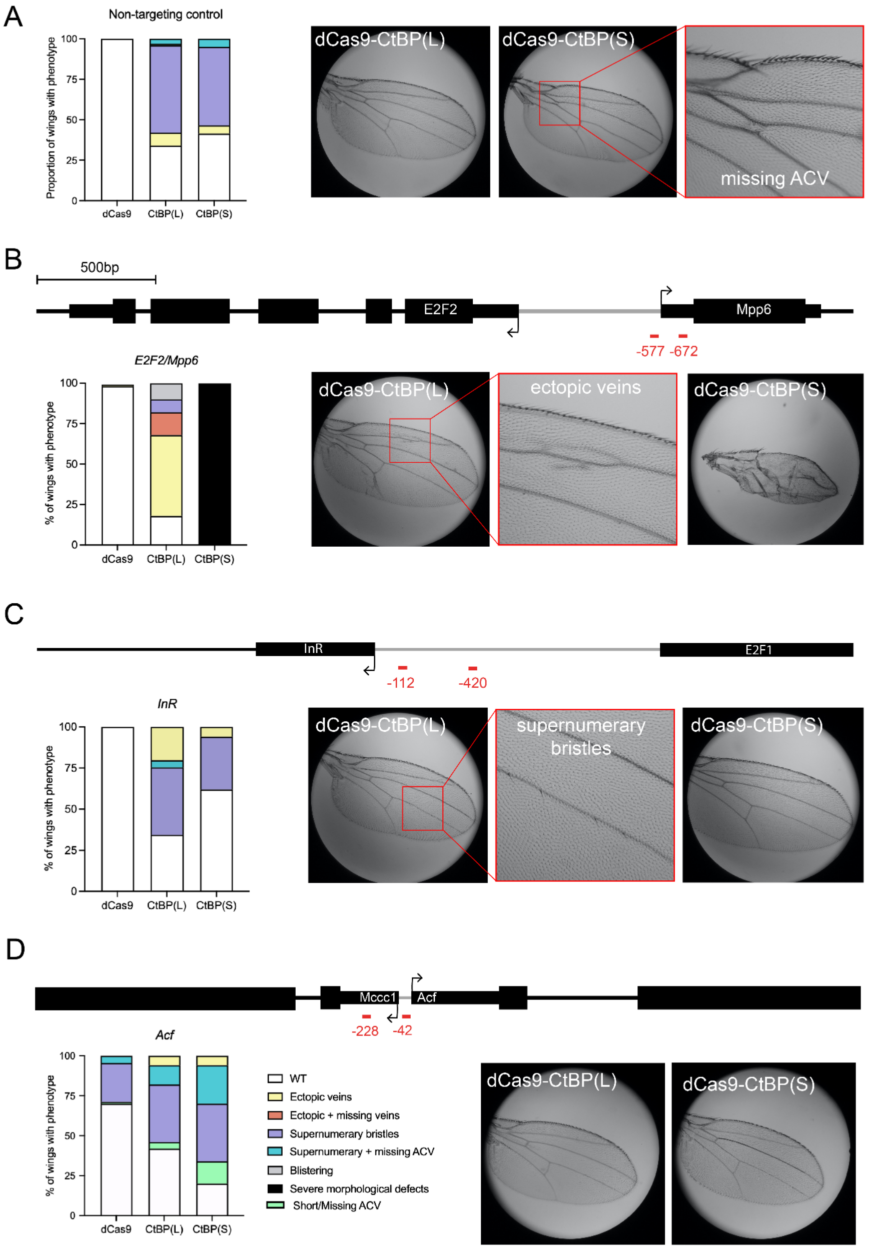


Figure 5.2. Targeting CtBP(S) and CtBP(L) to gene promoters leads to diverse phenotypic effects. For all crosses, ~100 wings from ~50 adults were used for analysis. Black arrows indicate the TSS, and red lines indicate gRNA binding sites relative to the target gene's TSS. **A)** Using a non-targeting control gRNA (QUAS), expression of one copy of dCas9-CtBP effectors leads to >50% of adult wings with a phenotype, such as supernumerary bristles. Legend is in panel D. **B)** Targeting the *E2F2/Mpp6* bidirectional promoter leads to severe morphological defects observed only from CtBP(S) targeting, with milder effects caused by CtBP(L). gRNA positions are relative to the *E2F2* TSS. **C)** Targeting the *InR* promoter leads to phenotypes similar to the QUAS non-targeting control, suggesting little or no specific effect on this promoter. **D)** Targeting the *Acf* promoter leads to mild phenotypes, some of which are also observed with dCas9 alone, at lower frequency. CtBP isoforms lead to a higher penetrance of phenotypes than dCas9.

Targeting the *InR* promoter produced adult wings with mild phenotypes, similar to those produced with the non-targeting QUAS gRNA control, so this effect is difficult to distinguish from a mild overexpression phenotype rather than specific *InR* targeting (**Figure 5.2C**). Clearly, positioning the CtBP chimeras near the transcriptional start site does not strongly affect the wing, although we know that positioning dCas9-Rb chimeras at this promoter does impact development and transcription (Raicu, Castanheira, Arnosti. *bioRxiv*). This distinct effect is consistent with CtBP promoter selectivity, a property illustrated from recent high-throughput assays (Jacobs et al. 2022, *bioRxiv*).

Recruitment to the *Acf* promoter region generated a different spectrum of phenotypes. In this case, a significant proportion of wings from the dCas9 control cross showed supernumerary bristles, evidence that dCas9 alone can disrupt gene function in certain locations. Notably, the position of one of the gRNAs used here is 3' of the initiation site for the divergently transcribed *Mccc1* gene, a position from which transcriptional inhibition is possible by dCas9 (Qi et al. 2013). Despite this dCas9 effect, the CtBP fusions had specific effects, with CtBP(S) causing a larger proportion of wings to be affected (80%) than CtBP(L) (60%; **Figure 5.2D**). Results from these targeted promoters indicate that CtBP exhibits gene-specific effects, and that in some cases, CtBP(S) has a more pronounced effect than CtBP(L).

CtBP(S) is a more potent transcriptional repressor than CtBP(L) on *E2F2/Mpp6*

Given the noticeable differences in phenotypes as a result of targeting the two CtBP isoforms to the *E2F2/Mpp6* shared promoter, we measured transcript levels of both of these genes in the wing disc using RT-qPCR. The two gRNAs used here bind at -577 and -672 relative to the *E2F2* TSS, and at -18 and +57 relative to the *Mpp6* TSS (**Figure 5.3A**). CtBP(S) showed specific repression of the *Mpp6* gene, whereas CtBP(L) effects were indistinguishable from those of dCas9 alone (**Figure 5.3C**). Effects on *E2F2* were more modest, with no apparent change for CtBP(L), and a small but significant reduction of similar magnitude for both dCas9 and CtBP(S) (**Figure 5.3B**). Interestingly, although CtBP(L) had a weaker effect on transcription than dCas9 alone at the time point measured (late L3 larval stage), it clearly showed more pronounced phenotypic effects in the adult stage. This may be due to gene regulatory effects later in development, where we did not measure transcriptional impacts. Taken together, these results indicate that both of the dCas9-CtBP corepressors do have specific effects, and at least CtBP(S) can be found to demonstrate classical repression effects.

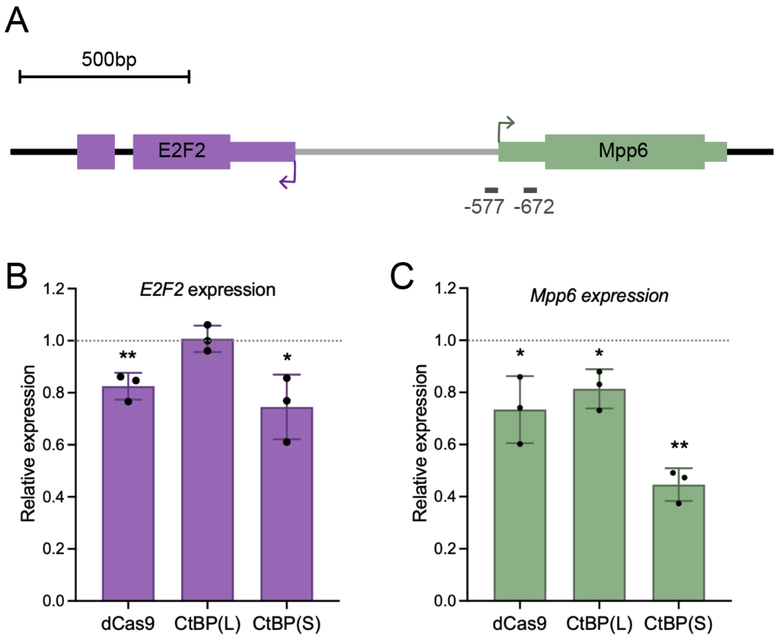


Figure 5.3. CtBP(S) is a more potent repressor of *Mpp6* than CtBP(L) in wing discs. A) Schematic of the *E2F2/Mpp6* bidirectional promoter, with the two tandem gRNAs indicated in gray. **B)** Targeting dCas9-CtBP(S) led to repression of *E2F2* by about 25%, similar to the effect of dCas9 alone. dCas9-CtBP(L) recruitment to the same sites did not lead to any measurable repression. **C)** Targeting dCas9-CtBP(S) led to significant repression of *Mpp6* (~50%), and this repression is greater than effects by dCas9 alone. dCas9 alone and dCas9-CtBP(L) led to about 20-25% repression. * p<0.05, ** p<.0.

Position-sensitive CtBP repression in cell culture

Many tests of CtBP function have relied on transiently transfected reporter genes; however, few studies have directly compared repression activity on the same genes in their endogenous chromosomal location. To further assess CtBP(L) and CtBP(S) function, we expressed the dCas9 chimeras in S2 cells, using an *Mpp6* reporter, which we have previously demonstrated is susceptible to repression by dCas9-Rb proteins (Raicu, Castanheira, Arnosti. *bioRxiv*). Here, we employed seven different gRNAs to test for possible position effects on this 1 kbp promoter region (**Figure 5.4A**). Both CtBP(S) and CtBP(L) showed strongest effects with gRNA 2 and 5; dCas9 alone did not mediate significant repression from the gRNA 2 position, but did from gRNA 5, likely due to steric effects (**Figure 5.4B-D**). The dCas9 control did not mediate repression from any other site, clearly different from the CtBP effects with gRNAs 1, 2, and 3. A simple distance effect, with stronger repression proximal to the transcriptional start site, was not evident. Additionally, CtBP(S) appeared to be more effective at the more distal gRNA 1 and B positions than near the TSS, at 4. Overall, it is striking that CtBP(L) performed similarly to CtBP(S) on this reporter, given the clear differences *in vivo*.

DISCUSSION

Our study of CtBP(L) and CtBP(S) isoforms using a CRISPRi approach has revealed that these repressors do exhibit different functional potential, and that CtBP itself shows promoter selectivity, consistent with the findings of the Stark laboratory (Jacobs *et al.* 2022, *bioRxiv*). Our data suggest that CtBP proteins are involved in selective modulation of their gene targets, consistent with a “soft repression” form of regulation that may characterize many repressive interactions in the cell (Mitra, Raicu *et al.* 2021).

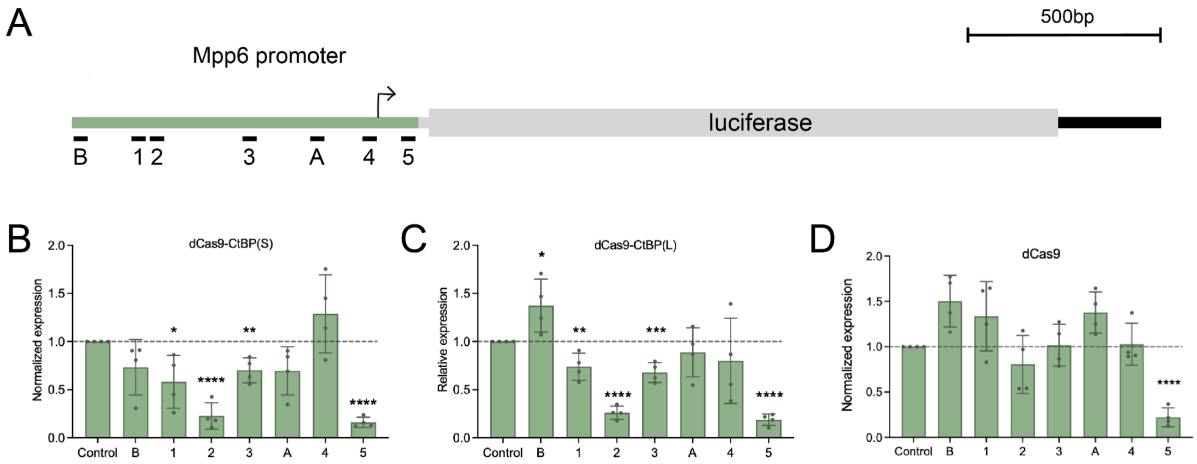


Figure 5.4. Testing range of action in S2 cells. S2 cells were transfected with actin-GAL4, the *Mpp6*-luciferase reporter, one of the dCas9 effectors, and a single gRNA. **A)** Schematic of luciferase reporter that was designed to be regulated by the *Mpp6* promoter, with gRNA positions indicated below. **B)** dCas9-CtBP(S) has position-specific effects. Position 2 led to the most severe repression. Position 5 caused the same level of repression as dCas9 alone, suggesting steric hindrance. **C)** dCas9-CtBP(L) has position-specific effects, which are similar to those of CtBP(S). **D)** The dCas9 control did not lead to significant repression, aside from position 5. The dCas9 results were presented in Raicu, Castanheira, Arnosti. *bioRxiv.*

Evolutionary conservation of the CTD of CtBP indicates that this portion of the corepressor must be of importance, yet most assays employed in previous studies have not identified a difference in function at the transcriptional level (Kumar *et al.* 2002; Madison *et al.* 2013). One possible explanation is that the domain is involved in other aspects of CtBP biology, such as turnover or intracellular targeting, which may be overlooked in overexpression assays. Alternatively, its function in gene regulation may not have been identified yet, as the context in which CtBP has been assayed is limited; even the recent high throughput assessment of GAL4-CtBP was carried out with transient transfections and effects of the CTD were not assessed (Jacobs *et al.* 2022, *bioRxiv*).

Few studies have tested the impact of CtBP proteins with or without the conserved, long CTD on expression of endogenous genes, with the exception of genomic rescue experiments that demonstrated that viability is possible with either a CtBP(S) or CtBP(L) rescue construct (Zhang and Arnosti, 2011). However, the survivors from genomic rescues employing single isoforms showed a variety of phenotypes, including elevated embryonic lethality and aberrant wing development, indicating that limiting expression to one isoform alone does not fully satisfy developmental demands. Here, by directly testing CtBP isoforms in a CRISPRi setting on endogenous genes, we uncovered a striking difference between CtBP(L) and CtBP(S). On the *E2F2/Mpp6* bidirectional promoter, CtBP(S) was a potent repressor of gene expression and caused a severe wing phenotype, while CtBP(L) was much milder in its transcriptional and phenotypic effects. What might be the molecular action of the CTD on CtBP itself? Biochemical assays have shown that this intrinsically disordered domain is not required for NAD(H) binding or oligomerization, which are required for *in vivo* functionality (Kumar et al. 2002; Bellesis et al. 2018; Jecrois et al. 2021). The CTD of mammalian CtBP has been shown to be a target of post-translational modifications, which may affect conformation or protein-protein interactions of this domain. Our CRISPRi system ensures targeting to the promoter, thus the CTD regulatory impact is likely to be at the level of transcriptional action, rather than promoter binding. It is interesting that a different eukaryotic dehydrogenase-like corepressor, NPAC/GLYR1, similar to CtBP, forms tetramers and possesses an IDR that is involved in functional contacts with histone-modifying lysine demethylases (Marabelli et al. 2019). Our finding that the CtBP(L) isoform is less active only on the chromatinized endogenous *E2F2/Mpp6* regulatory region, but not when this element is tested in a transient reporter assay, provides support for the notion that the CTD regulation is chromatin-related, but deeper understanding will require further biochemical and molecular genetic studies.

MATERIALS AND METHODS

Plasmids used in this study

To create UAS:dCas9-CtBP constructs, the FLAG-tagged (DYKDDDDK) coding sequences for CtBP(L) and CtBP(S) were used, as described previously (Sutrias-Grau and Arnosti, 2004). These coding sequences were amplified from their parent vector using 5' PacI and 3' XbaI sites, and inserted in place of Rbf1 in the UAS:dCas9-Rbf1 plasmid described previously (Raicu, Castanheira, Arnosti. *bioRxiv*). CtBP(L) is isoform F and CtBP(S) is a combination of isoform E and J, based on Flybase nomenclature. The *Mpp6*-luciferase reporter construct uses the *Mpp6* promoter, which includes the *Mpp6* 5'UTR, to drive luciferase expression, as was described previously (Raicu, Castanheira, Arnosti. *bioRxiv*). The gRNA plasmids used in transfections were described previously, and target different sites of the *E2F2/Mpp6* bidirectional promoter (Raicu, Castanheira, Arnosti. *bioRxiv*).

Transgenic flies

Flies were fed on standard lab food (molasses, yeast, corn meal) and kept at RT in the lab, under normal dark-light conditions. The *nubbin*-GAL4 fly line was obtained from the Bloomington Drosophila Stock Center (BDSC; #25754) and was maintained as a homozygous line with a Chr 3 balancer obtained from BDSC #3704 (*w[1118]/Dp(1; Y)y[+]; CyO/Bl[1]; TM2, e/TM6B, e, Tb[1]*). Homozygous UAS:dCas9-CtBP flies were generated by using the ϕC31 integrase service at Rainbow Transgenic Flies Inc. #24749 embryos were injected with each dCas9-CtBP construct to integrate into Chr 3, landing site 86Fb. Successful transgenic flies were selected through the mini-white selectable marker expression in-house, and maintained as a homozygous line with Chr 2 balancer (from BDSC #3704). *nub*-GAL4 and UAS:dCas9-CtBP homozygous flies were crossed to generate double homozygotes (*nub*-GAL4>UAS:dCas9-CtBP), using the Chr 2 and Chr 3 balancers (from #3704). sgRNA fly lines were obtained from the BDSC (fly line numbers indicated in **Table S5.1**). Single gRNA flies (-577 and -672) were previously described (Raicu, Castanheira, Arnosti. *bioRxiv*). Homozygous *nub*-GAL4>UAS:dCas9-CtBP flies were crossed to homozygous gRNA flies to generate triple heterozygotes (-/-; *nubbin*-GAL4/sgRNA; UAS:dCas9-CtBP/+) that are used for all fly experiments described here.

Genotyping flies

All flies generated in this study were genotyped at the adult stage. Flies of each genotype were homogenized (1 fly/tube) in squish buffer (1M Tris pH 8.0, 0.5M EDTA, 5M NaCl with 1µl of 10mg/mL Proteinase K for each fly). Tubes were set at 37C for 30 minutes, 95C for 2 mins, centrifuged at 14,000RPM for 7 minutes, and stored at 4C. Following PCR amplification, amplicons were cleaned using Wizard SV-Gel and PCR Clean-Up System and sent for Sanger sequencing.

Imaging adult wings

Adult wings were collected from ~50 male and female 1-3 day-old adults. They were stored in 200 proof ethanol in -20C until mounted. Wings were removed, mounted onto Asi non-charged microscope slides using Permount, and photographed with a Canon PowerShot A95 camera mounted onto a Leica DMLB microscope. Images were all taken at 10X magnification and using the same software settings.

Wing disc dissections and RT-qPCR

50 third instar wing discs were dissected from L3 larvae and placed in 200µl Trizol (ambion TRIzol Reagent) and stored in -80C until use. RNA was extracted using chloroform and the QIAGEN maXtract High Density kit, and stored in -80C. cDNA synthesis was performed using applied biosystems High Capacity cDNA Reverse Transcription Kit. RT-qPCR was performed

using SYBR green (PerfeCTa SYBR Green FastMix Low ROX by Quantabio) and measured using the QuantStudio 3 machine by applied biosystems. Three control genes were averaged (Rp49, RpS13, CG8636) for all samples with control obtained from crossing dCas9 to a non-targeting gRNA (QUAS). Primers used were described previously (Raicu, Castanheira, Arnosti. *bioRxiv*). RT-qPCR was performed on 3 biological replicates with two technical duplicates. Student's t-test (two tailed, $p<0.05$) was used to measure statistical significance. Error bars indicate SEM.

Luciferase reporter assays

Reporter assays were performed as described previously, but with dCas9-CtBP(L) and dCas9-CtBP(S) effectors here (Raicu, Castanheira, Arnosti. *bioRxiv*).

Western blot

Western blot was performed as described previously for S2 cells (Raicu, Castanheira, Arnosti. *bioRxiv*).

ACKNOWLEDGEMENTS

We would like to thank members of the Arnosti lab for their thoughtful suggestions, and the Michigan State University RTSF and plasmidsaurus for sequencing plasmids.

FUNDING

This work was supported by the National Institute for General Medical Sciences grant (number R01GM124137) to D.N.A., the National Institute of Child Health and Development grant (number F31HD105410) to A.M.R., and the BEACON luminaries grant to M.S.

REFERENCES

Bellesis, A. G., Jecrois, A. M., Hayes, J. A., Schiffer, C. A., & Royer, W. E. (2018). **Assembly of human C-terminal binding protein (CtBP) into tetramers**. Journal of Biological Chemistry, 293(23), 9101–9112. https://doi.org/10.1074/jbc.RA118.002514

Fang, M., Li, J., Blauwkamp, T., Bhambhani, C., Campbell, N., & Cadigan, K. M. (2006). **C-terminal-binding protein directly activates and represses Wnt transcriptional targets in Drosophila.** The EMBO Journal, 25(12), 2735–2745. https://doi.org/10.1038/sj.emboj.7601153

Grooteclaes, M., Deveraux, Q., Hildebrand, J., Zhang, Q., Goodman, R. H., & Frisch, S. M. (2003). **C-terminal-binding protein corepresses epithelial and proapoptotic gene expression programs.** Proceedings of the National Academy of Sciences, 100(8), 4568–4573. https://doi.org/10.1073/pnas.0830998100

Harlen, K. M., & Churchman, L. S. (2017). **The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain.** Nature Reviews Molecular Cell Biology, 18(4), 263–273. https://doi.org/10.1038/nrm.2017.10

Jacobs, J., Pagani, M., Wenzl, C., & Stark, A. (2022). **Widespread regulatory specificities between transcriptional corepressors and enhancers in Drosophila** [Preprint]. https://doi.org/10.1101/2022.11.07.515017

Jecrois, A. M., Dcona, M. M., Deng, X., Bandyopadhyay, D., Grossman, S. R., Schiffer, C. A., & Royer, W. E. (2021). **Cryo-EM structure of CtBP2 confirms tetrameric architecture.** Structure, 29(4), 310-319.e5. https://doi.org/10.1016/j.str.2020.11.008

Jin, W., Scotto, K. W., Hait, W. N., & Yang, J.-M. (2007). **Involvement of CtBP1 in the transcriptional activation of the MDR1 gene in human multidrug resistant cancer cells.** Biochemical Pharmacology, 74(6), 851–859. https://doi.org/10.1016/j.bcp.2007.06.017

Kampmann, M. (2018). **CRISPRi and CRISPRa Screens in Mammalian Cells for Precision Biology and Medicine.** ACS Chemical Biology, 13(2), 406–416. https://doi.org/10.1021/acschembio.7b00657

Kumar, V., Carlson, J. E., Ohgi, K. A., Edwards, T. A., Rose, D. W., Escalante, C. R., Rosenfeld, M. G., & Aggarwal, A. K. (2002). **Transcription Corepressor CtBP Is an NAD+-Regulated Dehydrogenase.** Molecular Cell, 10(4), 857–869. https://doi.org/10.1016/S1097-2765(02)00650-0

Madison, D. L., Wirz, J. A., Siess, D., & Lundblad, J. R. (2013). **Nicotinamide Adenine Dinucleotide-induced Multimerization of the Co-repressor CtBP1 Relies on a Switching Tryptophan.** Journal of Biological Chemistry, 288(39), 27836–27848. https://doi.org/10.1074/jbc.M113.493569

Mani-Telang, P., & Arnosti, D. N. (2007). **Developmental expression and phylogenetic conservation of alternatively spliced forms of the C-terminal binding protein corepressor.** Development Genes and Evolution, 217(2), 127–135. https://doi.org/10.1007/s00427-006-0121-4

Marabelli, C., Marrocco, B., Pilotto, S., Chittori, S., Picaud, S., Marchese, S., Ciossani, G., Forneris, F., Filippakopoulos, P., Schoehn, G., Rhodes, D., Subramaniam, S., & Mattevi, A. (2019). **A Tail-Based Mechanism Drives Nucleosome Demethylation by the LSD2/NPAC Multimeric Complex.** Cell Reports, 27(2), 387-399.e7. https://doi.org/10.1016/j.celrep.2019.03.061

Mitra, A., Raicu, A.M., Hickey, S. L., Pile, L. A., & Arnosti, D. N. (2021). **Soft repression: Subtle transcriptional regulation with global impact.** BioEssays, 43(2), 2000231. https://doi.org/10.1002/bies.202000231

Paliwal, S., Ho, N., Parker, D., & Grossman, S. R. (2012). **CtBP2 Promotes Human Cancer Cell Migration by Transcriptional Activation of Tiam1.** Genes & Cancer, 1947601912463695. https://doi.org/10.1177/1947601912463695

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). **Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression.** Cell, 152(5), 1173–1183. https://doi.org/10.1016/j.cell.2013.02.022

Raicu, A.M., Kadiyala, D., Niblock, M., Jain, A., Yang, Y., Bird, K. M., Bertholf, K., Seenivasan, A., Siddiq, M., & Arnosti, D. N. (2023). **The Cynosure of CtBP: Evolution of a Bilaterian Transcriptional Corepressor.** Molecular Biology and Evolution, 40(2), msad003. https://doi.org/10.1093/molbev/msad003

Raicu, A.M., Castanheira, P. H., & Arnosti, D. N. **Retinoblastoma protein activity revealed by CRISPRi study of divergent Rbf1 and Rbf2 paralogs.** [Preprint] bioRxiv.

Sutrias-Grau, M., & Arnosti, D. N. (2004). **CtBP Contributes Quantitatively to Knirps Repression Activity in an NAD Binding-Dependent Manner.** Molecular and Cellular Biology, 24(13), 5953–5966. https://doi.org/10.1128/MCB.24.13.5953-5966.2004

Uversky, V. N. (2016). **Paradoxes and wonders of intrinsic disorder: Complexity of simplicity.** Intrinsically Disordered Proteins, 4(1), e1135015. https://doi.org/10.1080/21690707.2015.1135015

Zhang, Y. W., & Arnosti, D. N. (2011). **Conserved Catalytic and C-Terminal Regulatory Domains of the C-Terminal Binding Protein Corepressor Fine-Tune the Transcriptional Response in Development.** Molecular and Cellular Biology, 31(2), 375–384. https://doi.org/10.1128/MCB.00772-10

Zirin, J., Bosch, J., Viswanatha, R., Mohr, S. E., & Perrimon, N. (2022). **State-of-the-art CRISPR for in vivo and cell-based studies in Drosophila.** Trends in Genetics, 38(5), 437–453. https://doi.org/10.1016/j.tig.2021.11.006

APPENDIX

This work was published as supplementary material in the following preprint:

Raicu, A.M., Suresh, M., & Arnosti, D. N. (2023). **A regulatory role for the unstructured C-terminal domain of the CtBP transcriptional corepressor.** *bioRxiv*. 2023.05.19.541472; doi: https://doi.org/10.1101/2023.05.19.541472

Gene	Function	gRNA distance from TSS
E2F2	Cell cycle regulator	-577, -672
Mpp6	M-phase phosphoprotein	-18, +57
Acf	Subunit of two ATP dependent nucleosome remodeling complexes.	-42, -228
InR	Insulin receptor	-112, -420
Atx2	Involved in eye development	-476, -213
Rbf1	Transcriptional repressor	-65, -442
Atf3	Activating transcriptional factor	-112, -283
p53	Transcription factor	-211, -327
Vang	Establishes planar polarity in epithelia	-69, -332
Dad	Inhibitory SMAD in dpp pathway	-294, -398
Cad99c	Cell-cell adhesion	-36, -135
Arm	Cell adhesion and wingless signaling	-376, -95
Dtg	gastrulation	-257, -133
Mip40	Critical regulator of the cell cycle	-428, -315
Ap	Transcription factor	-378, -148
Caz	Locomotion and eye development	-455, -334

Table S5.1. Table indicating genes that were targeted by dCas9-CtBP(S), and the positions of the gRNA binding sites. Targeting most genes did not cause a severe phenotype.

Table S5.1 (cont'd)

Cyc	Transcription of circadian clock genes.	-249, -93
Hh	Signaling pathway ligand	-163, -114
Wwox	Oxidoreductase	-191, -359
DNApola	Catalytic subunit of DNAP	-57, -202
Sta	Ribosomal protein	-72, -302
mRps22	Mitochondrial ribosomal protein	-476, -364
GstE13	Glutathione	-471, -244
Spen	Regulator of wnt signaling	-424, -355
wg	Encodes a ligand of the Wnt/Wg signaling pathway	-314, -366
Mcm6	Subunit of the hetero-hexameric mcm complex	-438, -383
dpp	Ligand of the transforming growth factor β signaling pathway	-450, -3
Rbf2	Cell and development regulator	-410, -187
elob	Control of wing cell identity.	-374, -266

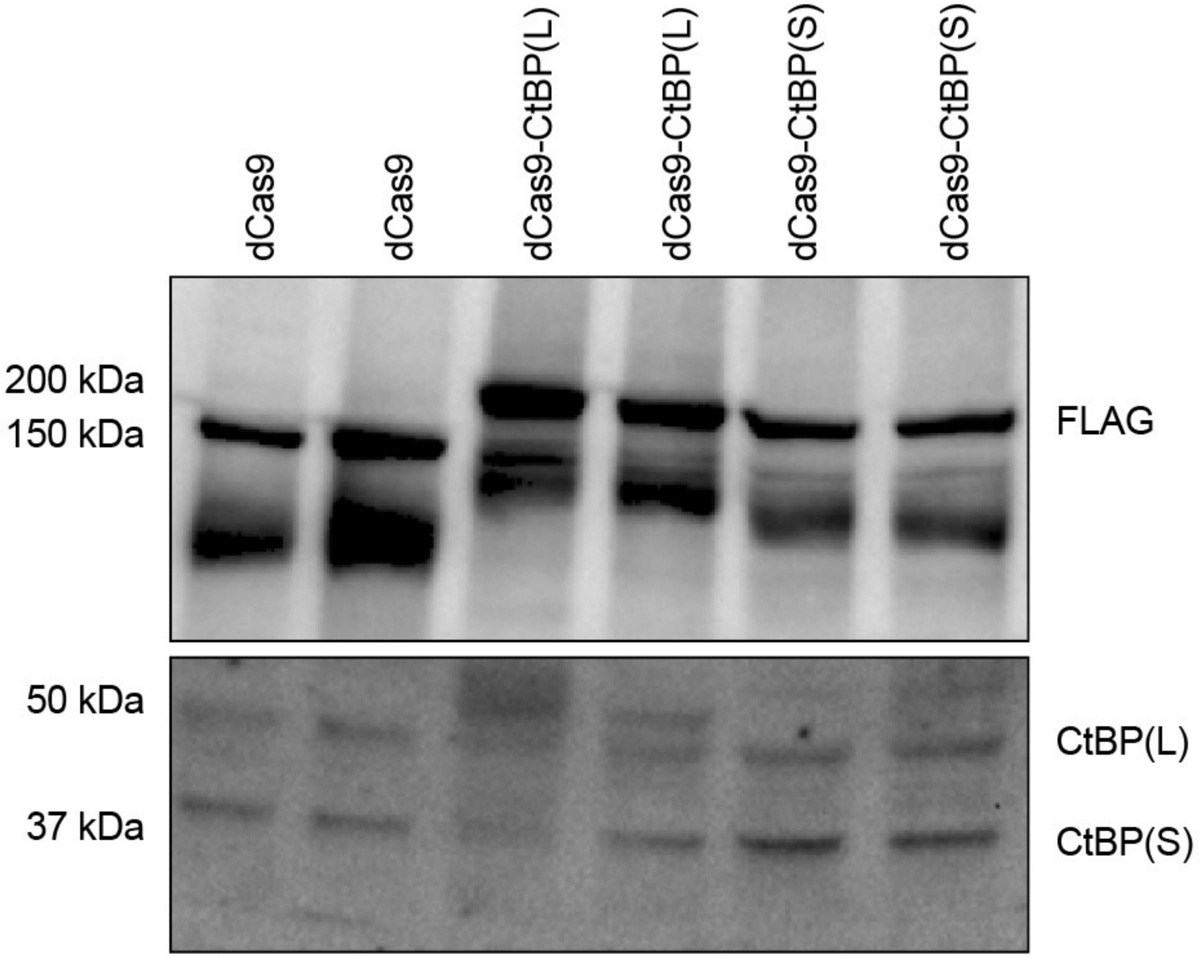


Figure S5.1. dCas9-CtBP effectors are expressed in S2 cells at similar levels. dCas9, dCas9-CtBP(L) and dCas9-CtBP(S) were co-transfected in S2 cells with actin-GAL4 in duplicate. Effectors run at expected sizes of ~150 kDa for dCas9 alone, and <200 kDa for the CtBP effectors. dCas9-CtBP(S) runs faster than dCas9-CtBP(L), consistent with a size difference of ~10 kDa.

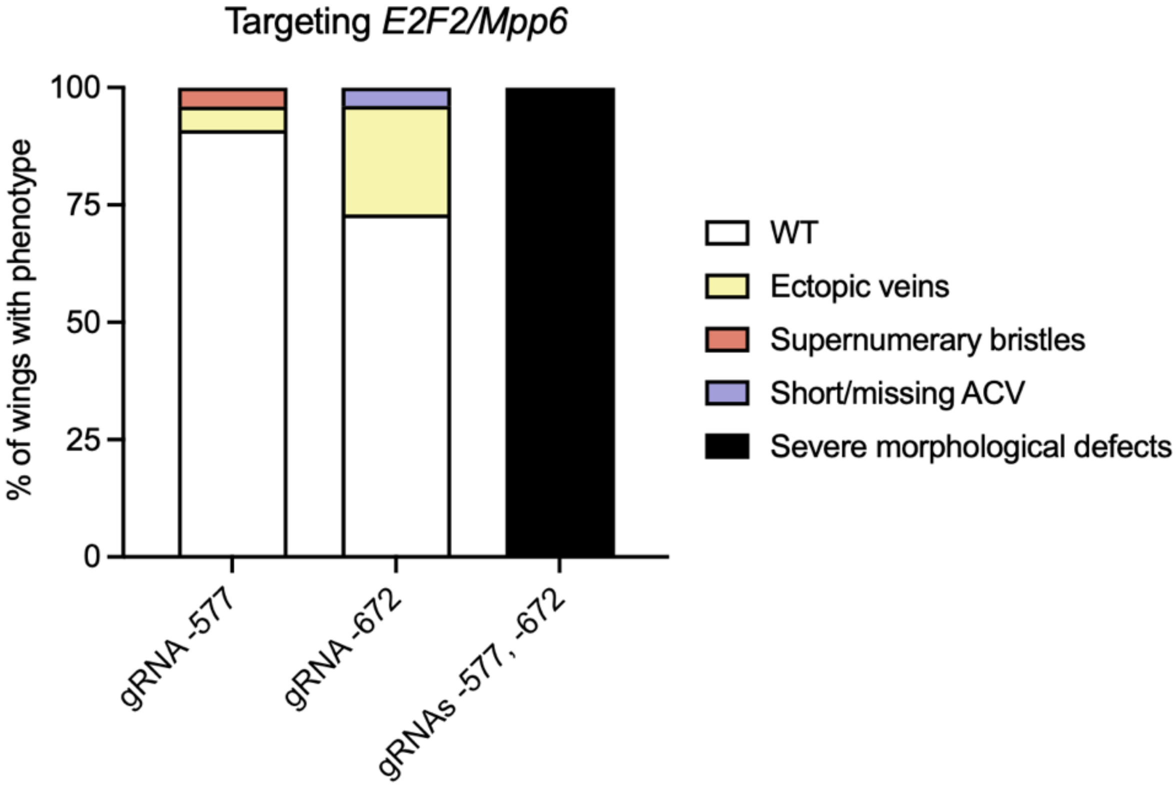


Figure S5.2. Single gRNAs produce milder CtBP(S) effects. Recruiting CtBP(S) with individual gRNAs at -577 or -672 led to much milder effects than when used together in tandem, suggesting a possible cooperative effect when two dCas9-CtBP molecules are brought together.

CHAPTER 6: TÊTE-À-TÊTE WITH CTBP DIMERS

This work was published in the following manuscript:

Raicu, A.M., Bird, K.M., & Arnosti, D.N. (2021) **Tête-à-tête with CtBP dimers**. Structure, Apr 1; 29(4):307-309. doi: 10.1016/j.str.2021.03.006

Jecrois *et al.* (2020) use cryoelectron microscopy to illuminate the tetrameric conformation of the CtBP2 transcriptional corepressor, a protein frequently overexpressed in human cancers. The *in vivo* functional characterization of tetramer-destabilizing mutants indicates that tetramerization is a physiologically important process, critical for CtBP control of gene regulation and cell migration.

The paralogous CtBP1 and CtBP2 proteins are transcriptional regulators involved in cell fate, apoptosis, and the epithelial-to-mesenchymal transition. CtBP proteins are misregulated in many human cancers including breast, colon, and ovarian cancers, leading to inhibition of apoptosis and promotion of metastasis. CtBP was first identified by the Chinnadurai lab as a cofactor binding the C-terminus of the adenoviral E1A oncoprotein (Boyd *et al.* 1993), and has since been recognized as a transcriptional scaffold that binds histone modifiers and chromatin remodelers. Unique among transcriptional coregulators, CtBP structurally resembles D-2-hydroxyacid dehydrogenases; the protein binds NAD(H), and this has been suggested to permit the sensing of the cell's metabolic state. Notwithstanding, the catalytic activity has remained an enigma, and an *in vivo* CtBP substrate has yet to be identified. Early studies focused on determining the link between the proposed dehydrogenase activity, NAD(H) binding, and transcriptional regulation by CtBP. Through biochemical and structural analysis, it was determined that NAD(H) binding leads to dimerization of CtBP, which was thought to be the physiologically relevant form of the protein (**Figure 6.1A**; Kumar *et al.* 2002). In conditions of low NAD(H) levels, CtBP remains monomeric; thus, a shift in NAD(H) levels may drive dimerization and activity under conditions of hypoxia. These findings provided clues to the significance of the unique features of CtBP.

Further structural analysis of this important protein has revealed additional, previously unsuspected properties. Using *in vitro* characterization of purified recombinant protein, the Royer laboratory provided evidence of a CtBP1 and CtBP2 tetrameric state using size exclusion chromatography as well as X-ray crystallography (Bellesis *et al*. 2018). The biological significance of this higher-order structure, also reported by Lundblad and colleagues (Madison *et al*. 2013), remained unclear until now.

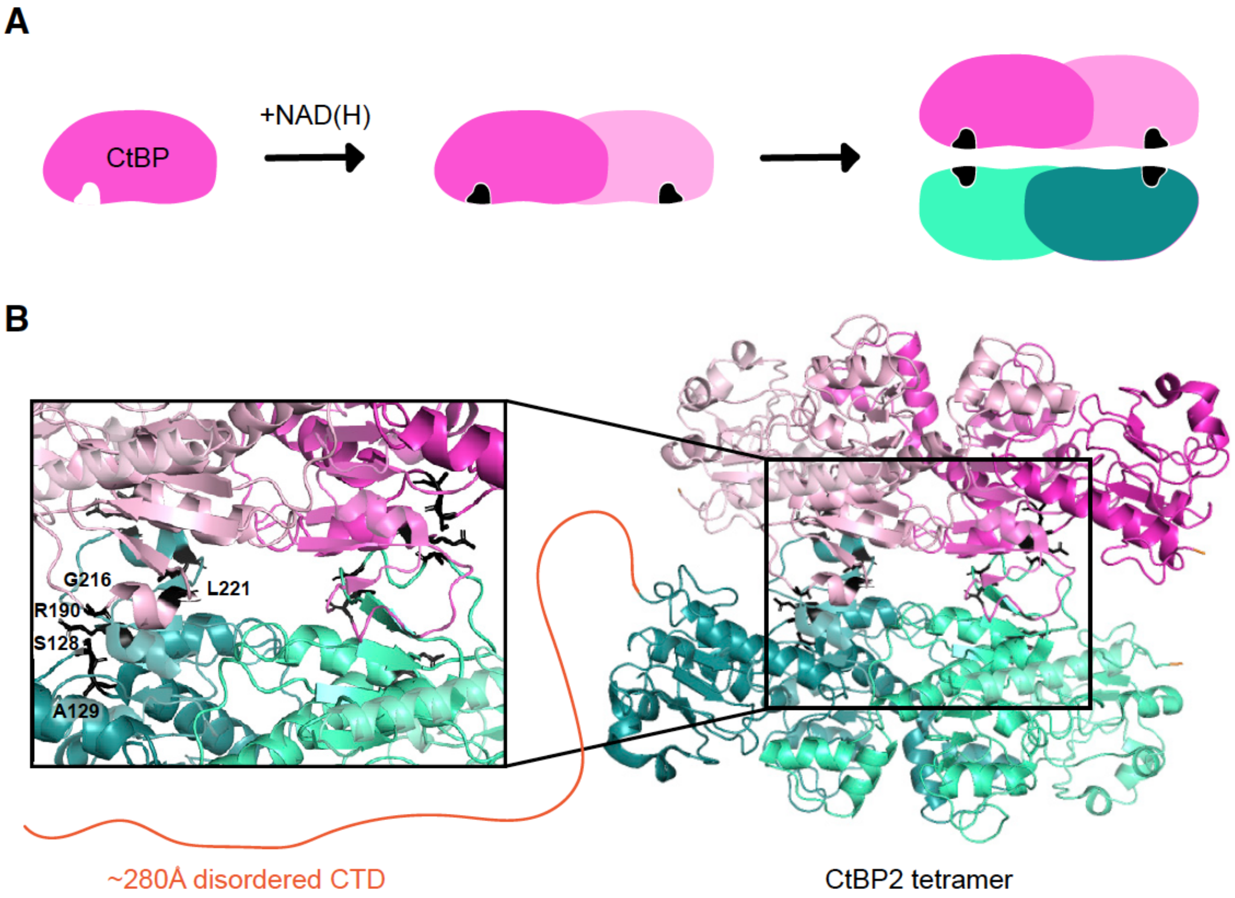


Figure 6.1. CtBP proteins tetramerize through NAD(H) binding, and several residues at the interdimer interface are required for tetramerization. A) NAD(H) allows for CtBP monomers to dimerize and also form tetramers. **B)** CtBP2 cryo-EM structure shows a tetramer with five residues involved in tetramerization highlighted in the inset. The C-terminal domain, which is unstructured and not resolved, may stretch out to 280 A° (structure from Jecrois *et al*. 2020; PDB: 6WKW).

In this issue of Structure, Jecrois *et al*. (2020) use high-resolution cryo-electron microscopy (cryo-EM) to validate the important residues mediating inter-subunit contacts also seen in X-ray studies (Bellesis *et al*. 2018; Jecrois *et al*. 2020). Importantly, they exploited these molecular insights to generate point mutants that are still competent for dimerization, but abrogate tetramerization *in vitro* (**Figure 6.1B**). These mutants were then assessed for biological function in cell culture (Jecrois *et al*. 2020). Colon cancer cells null for CtBP2 (HCT116) were transfected with the five CtBP2 mutant forms and the effects on gene expression and cell migratory activity were tested. Strikingly, although the mutant proteins were expressed at similar levels to the wildtype protein, they were unable to activate TIAM1 expression or repress CHD1 expression, both cancer-related genes that are direct CtBP2 targets (Jecrois *et al*. 2020). These results indicate that CtBP2 mutants that are unable to tetramerize lose their transcriptional co-regulatory activities in cell culture. The mutant forms were also unable to support heightened cell migration. Taken together, these results provide strong evidence of the functional relevance of CtBP protein tetramerization.

Jecrois *et al*. (2020) also determined the structure of CtBP2 with its flexible C-terminal domain (CTD), although at low resolution. Unfortunately, electron density maps could not confidently assign a structure to the CTD, but the full-length protein formed a similar tetrameric structure to the truncated protein. The cryo-EM structures confirm that the CTD is not required for single-particle oligomerization. These results contrast with those of Lundblad and colleagues, who suggested that the CTD is required for tetramerization of CtBP1 *in vitro* (Madison *et al*. 2013). Yet Royer and colleagues determined that tetrameric forms of CtBP1 and CtBP2 lacking the entire C-terminal region can be obtained *in vitro*, although the presence of the CTD yields more tetramer formation (Bellesis *et al*. 2018). In fact, in some organisms, CtBP isoforms are expressed that entirely lack the CTD, which may impact tetramer formation (Mani-Telang and Arnosti, 2007). Thus, it is still unclear whether this C-terminal extension is required *in vivo* for stabilization of tetramers.

The five residues in CtBP2 that were found to be required for tetramerization are embedded in a larger block of 100 amino acids that are highly conserved: three (R190, G216, and L221) are absolutely conserved across CtBP2 proteins in vertebrates, and the other two (S128 and A129) are very highly conserved (data not shown). Similar or identical residues are found in mammalian CtBP1, which can also form tetramers. In fact, the tetramer-promoting R,G,L residues are also absolutely conserved across arthropods, indicating that tetramerization may be an ancient structural property of CtBP.

This study raises additional interesting questions. Here, the CtBP2 mutants were tested in only one cellular context through overexpression; testing these tetramer-destabilizing mutants in a developmental context, through overexpression or CRISPR-mediated mutagenesis, would provide additional clues to the physiological significance of tetramerization. For instance, a catalytic domain residue, H315, was previously shown to be unnecessary in mouse embryonic fibroblasts; however, a similar genomic rescue construct containing this catalytic site mutation failed to fully rescue CtBP mutations in Drosophila (Grooteclaes *et al*. 2003; Zhang and Arnosti, 2011). A similar assay with the CtBP2 tetramer-destabilizing mutants in a developmental system would unequivocally determine the physiological importance of tetramerization outside of overexpression in select cell lines.

Additionally, whether tetramer assembly is a regulated and reversible process is not currently known. Early studies indicated that CtBP binds to NAD+ and NAD(H) with different affinities, leading to differences in binding to other proteins (Zhang *et al*. 2002; Fjeld *et al*. 2003), yet this difference in affinity was disputed by other studies (Kumar *et al*. 2002; Madison *et al*. 2013). Still, CtBP may function as a redox sensor, being activated through varying levels of NAD(H), impacting oligomerization of the protein and its interactions with cofactors. Regulation of tetramer assembly by endogenous processes such as post-translational modifications also remains to be determined. Regardless of whether tetramerization is normally a regulated process, the assembly of CtBP tetramers might be exploited in cancer therapeutics to control CtBP activity in certain contexts and reduce its oncogenic properties. The structural insights of CtBP will continue to reveal surprising aspects of this central player in cellular processes and disease.

ACKNOWLEDGEMENTS

This work is supported by the National Institutes of Health grant R01GM124137 to D.N.A., the Michigan State University Dissertation Continuation Fellowship to A.M.R., and the Moore-Billman Scholarship in Biochemistry and Molecular Biology to K.M.B.

REFERENCES

Bellesis, A.G., Jecrois, A.M., Hayes, J.A., Schiffer, C.A., and Royer, W.E., Jr. (2018). **Assembly of human C-terminal binding protein (CtBP) into tetramers.** J. Biol. Chem. 293, 9101–9112.

Boyd, J.M., Subramanian, T., Schaeper, U., La Regina, M., Bayley, S., and Chinnadurai, G. (1993). **A region in the C-terminus of adenovirus 2/5 E1a protein is required for association with a cellular phosphoprotein and important for the negative modulation of T24-ras mediated transformation, tumorigenesis and metastasis.** EMBO J. 12, 469–478.

Fjeld, C.C., Birdsong, W.T., and Goodman, R.H. (2003). **Differential binding of NAD+ and NADH allows the transcriptional corepressor carboxyl terminal binding protein to serve as a metabolic sensor.** Proc. Natl. Acad. Sci. USA 100, 9202–9207.

Grooteclaes, M., Deveraux, Q., Hildebrand, J., Zhang, Q., Goodman, R.H., and Frisch, S.M. (2003). **C-terminal-binding protein corepresses epithelial and proapoptotic gene expression programs.** Proc. Natl. Acad. Sci. USA 100, 4568–4573.

Jecrois, A.M., Dcona, M.M., Deng, X., Bandyopadhyay, D., Grossman, S.R., Schiffer, C.A., and Royer, W.E., Jr. (2020). **Cryo-EM structure of CtBP2 confirms tetrameric architecture.** Structure 29, this issue, 310–319.

Kumar, V., Carlson, J.E., Ohgi, K.A., Edwards, T.A., Rose, D.W., Escalante, C.R., Rosenfeld, M.G., and Aggarwal, A.K. (2002). **Transcription corepressor CtBP is an NAD(+)-regulated dehydrogenase.** Mol. Cell 10, 857–869.

Madison, D.L., Wirz, J.A., Siess, D., and Lundblad, J.R. (2013). **Nicotinamide adenine dinucleotide induced multimerization of the co-repressor CtBP1 relies on a switching tryptophan.** J. Biol. Chem. 288, 27836–27848.

Mani-Telang, P., and Arnosti, D.N. (2007). **Developmental expression and phylogenetic conservation of alternatively spliced forms of the C-terminal binding protein corepressor.** Dev. Genes Evol. 217, 127–135.

Zhang, Y.W., and Arnosti, D.N. (2011). **Conserved catalytic and C-terminal regulatory domains of the C-terminal binding protein corepressor finetune the transcriptional response in development.** Mol. Cell. Biol. 31, 375–384.

Zhang, Q., Piston, D.W., and Goodman, R.H. (2002). **Regulation of corepressor function by nuclear NADH.** Science 295, 1895–1897.

CHAPTER 7: OFF THE DEEP END: WHAT CAN DEEP LEARNING DO FOR THE GENE EXPRESSION FIELD?

This work was published in the following manuscript and adapted here:

Raicu, A.M., Fay, J.C., Rohner, N., Zeitlinger, J., & Arnosti, D.N. **Off the deep end: What can deep learning do for the gene expression field?** J Biol Chem. 2023 Jan;299(1):102760. Doi: 10.1016/j.jbc.2022.102760.

After a COVID-related hiatus, the fifth biennial symposium on Evolution and Core Processes in Gene Regulation met at the Stowers Institute in Kansas City, Missouri July 21 to 24, 2022. This symposium, sponsored by the American Society for Biochemistry and Molecular Biology (ASBMB), featured experts in gene regulation and evolutionary biology. Topic areas covered enhancer evolution, the cis-regulatory code, and regulatory variation, with an overall focus on bringing the power of deep learning (DL) to decipher DNA sequence information. DL is a machine learning method that uses neural networks to learn complex rules that make predictions about diverse types of data. When DL models are trained to predict genomic data from DNA sequence information, their high prediction accuracy allows the identification of impactful genetic variants within and across species. In addition, the learned sequence rules can be extracted from the model and provide important clues about the mechanistic underpinnings of the cis-regulatory code.

Interpreting the cis-regulatory sequence rules to obtain a mechanistic understanding of gene regulation

A sought-after goal of the gene regulation field is to decode enhancer grammar (Zeitlinger, 2020). How do transcription factor (TF) binding motifs within an enhancer combine to generate its unique activity? Can we learn the enhancer grammar to create synthetic spatially or temporally regulated enhancers? The great advantage of deep learning (DL) models over traditional methods is that they learn complex cis-regulatory rules in a precise and unbiased manner, allowing for new sequence rules to be discovered. Identifying these rules is done after model training and is not trivial; yet, a variety of interpretation tools already exist to obtain the important sequence features and their rules of interactions. In this manner, DL models reliably reveal the binding motifs of TFs and provide important clues as to how the motifs combine to produce an experimental outcome.

Interpreting DL models can therefore reveal novel mechanistic insights that can then be tested experimentally. For example, Alexander Stark (IMP, Vienna) discussed his laboratory's recent application of DL to their STARR-seq method to predict enhancer activity from DNA sequence (de Almeida *et al*. 2022). Deep-STARR can be successfully used to design synthetic enhancers with desired activities. Julia Zeitlinger (Stowers Institute) described her laboratory's work on applying DL to understand the cis-regulatory rules of enhancers during Drosophila embryogenesis. Using BPNet (Avsec *et al*. 2020) and chromBPNet as DL models, they uncovered TF cooperativity and sequences critical to opening chromatin, revealing a different effect for high versus low affinity motifs. Shaun Mahony (Pennsylvania State University) also explored the mechanistic basis of chromatin accessibility by studying the evolutionary impacts of FOX gene paralogs (Srivastava *et al*. 2021). Their DL approach specifically modeled chromatin state and DNA sequence to predict whether individual paralogs required prior chromatin accessibility for binding.

Several talks focused on further improving our ability to extract cis-regulatory information from DL models. Sara Mostafavi (University of Washington) talked about her approaches to identify sequence features important for determining chromatin states in diverse human immune cells. She discussed approaches to unlock elements in the DL algorithms that were informative and reproducible, including identifying the number of active motifs found at diverse enhancers (Novakovsky *et al*. 2022). Vivekanandan Ramalingam (Kundaje laboratory; Stanford University) illustrated how the Kundaje laboratory extracts important DNA sequences from a DL model to identify the distance rules by which TF motifs cooperate in binding and opening chromatin (Avsec *et al*. 2020). He also shared dynseq, a browser tool for visually exploring the sequences that were learned by a DL model (Nair *et al*. 2022). These examples highlight the importance of further tool development for motif identification and functional contributions to enhancer activity.

A strength of DL models is that they can learn sequence rules in an unbiased manner. Reassuringly, the learned rules can often be matched to known processes involved in gene regulation. For example, the learned TF binding motifs and their affinities correspond remarkably well to biophysical models of TF binding. Therefore, an important goal is to combine DL with biophysical models of transcription. Along this line, Justin Kinney (Cold Spring Harbor Laboratory) discussed Mave-NN, a computational framework for integration of diverse gene expression data to make DL accessible to a biological user base (Tareen *et al*. 2022). These biophysical models will illuminate the functional properties of gene switches such as those studied using optogenetic technology by Hernan Garcia (UC Berkeley) in the Drosophila embryo (Zhao *et al*. 2022).

Placing extracted sequences into gene regulatory networks and developmental processes

Since sequence information plays a central role in the fields of gene regulation, development, and evolution, sequence-centered DL models are an excellent way to promote cross-fertilization between these fields, which was the overarching theme for this American Society for Biochemistry and Molecular Biology (ASBMB)-sponsored conference. Evolution is a crucial trove of knowledge for understanding gene regulation, and conversely, an understanding of gene regulation is a key to unlocking evolutionary processes. Deep learning can therefore have a significant impact toward accelerating this dialogue between fields. One path forward is to integrate the sequence information extracted from DL models with other data modalities to construct gene regulatory networks (GRNs). For example, to characterize key players in zebrafish inner ear regeneration, Erin Jimenez (Shawn Burgess laboratory; NIH) used single-cell ATAC-seq and RNA-seq to identify activated enhancers during regeneration. Using DL, she uncovered a role for the Sox and Six TFs in coordinately regulating ear regeneration (Jimenez *et al*. 2022). This and other studies illustrate how DL approaches can facilitate the molecular identification of GRNs and increase the power of traditional genetic and genomic studies for biomedical research.

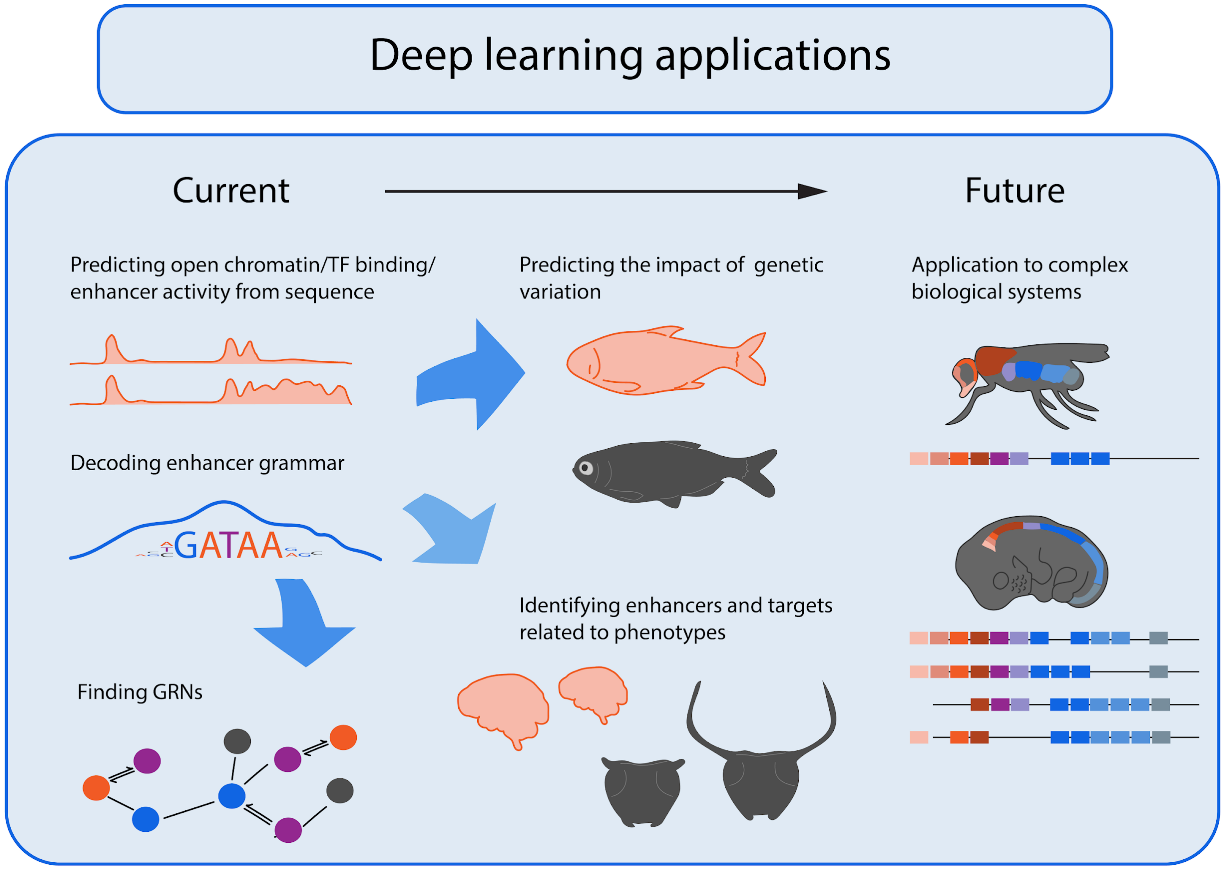

Deep learning applications
Current
Future
Predicting open chromatin/TF binding/
enhancer activity from sequence
Predicting the impact of genetic
variation
Application to complex
biological systems
Decoding enhancer grammar
GATAA
Identifying enhancers and targets
related to phenotypes
Finding GRNs

Figure 7.1. Applications of deep learning to studying the complexity of gene regulation. Current deep learning models enable predictions of open chromatin sites, transcription factor (TF) binding, or enhancer activity data from DNA sequence, which can then be interpreted to uncover enhancer grammar and key players of a gene regulatory network (GRN). The prediction capacity is excellent for identifying the impact of genetic variation at a species and population level, such as the difference between cave and surface fish morphs. Altogether, deep learning models are poised to help the identification of enhancers and gene targets involved in evolving traits, e.g., brain size or the phenotypic plasticity of dung beetles in response to nutritional cues. Understanding the activity of complex, highly context-dependent biological systems, such as HOX genes, may represent a future frontier for deep learning.

Predicting the effect of genetic variation using DL models

The high prediction accuracy of DL models can also be leveraged without understanding the learned sequence rules. When DL models are trained to predict a readout such as ATAC-seq accessibility, the high prediction accuracy holds for similar sequences, including genetic variants within a population or across related species. This makes DL models ideal for the identification of causal genetic variants and has the potential to identify genes and alleles underlying the evolution of complex phenotypes such as mammalian brain size. This extremely challenging problem was tackled by Irene Kaplow (Andreas Pfenning laboratory; Carnegie Mellon University) by training a DL model on ATAC-seq data from several well-characterized mammalian brains (Kaplow *et al*. 2022). Her model, TACIT, allowed the identification of several motor cortex enhancers that are associated with the evolution of brain size relative to body size. This work paves the way for using deep learning to identify enhancers and candidate genes involved in complex traits that are subject to evolutionary selection.

Species- and population-level variation was also the subject of DL analysis by Michael Wilson (Hospital for Sick Children, Toronto) who applied the BPNet model to mouse liver TFs and showed how it could predict TF binding profiles in other mammalian species. In addition, his laboratory applied this model to interpret variations related to disease-causing alleles involved in blood coagulation and lipid regulation.

Such use of DL models is poised to influence and complement the current genetic approaches used to map and understand the impact of cis-regulatory sequence variation. Excellent examples were the talks from Drs Brem, Vierbuchen, Wunderlich, Fay, and Wittkopp. Using a mouse fibroblast senescence model, Rachel Brem's laboratory at UC Berkeley used a classic F1 hybrid approach to identify cis-regulatory changes between two mouse species that explains their differential response to irradiation. Their analysis highlighted the TF USF2, which may play a role in senescence decision-making. Similarly, Thomas Vierbuchen (Memorial Sloan Kettering Cancer Center) described the use of mouse hybrid cells to link cis-regulatory variation with TF binding and enhancer function in the context of mouse embryonic and brain development. Zeba Wunderlich (Boston University) studied distinct populations and hybrids of *Drosophila melanogaster* to understand how population-level variation impacts the innate immune response to infection. Her laboratory found that trans-acting alleles dominate the response to a Gram-negative infection, while cis-acting effects dominate in a Gram-positive infection. A similar population variation approach was used by the laboratory of Justin Fay (University of Rochester) to identify protein-based variation linked to thermal tolerance in different species of yeast. They identified numerous differences in protein stability and also an important role of the hybrid cellular environment. Patricia Wittkopp (University of Michigan) described her laboratory's classical genetic approaches to empirically test the assumption that trans-acting variants are more pleiotropic than cis-acting variants in Saccharomyces cerevisiae. By comparing the impact of cis- and trans-acting mutations on fitness and gene expression, their highly quantitative assays revealed differences in the effects of these two classes of mutations that support the hypothesis that trans-regulatory mutations are more pleiotropic than cis (Vande Zande *et al*. 2022).

The influence of genetic variation can also be studied at the organismal level. Nicolas Rohner (Stowers Institute) described work in his laboratory using Mexican cavefish, which independently underwent metabolic adaptation to the cave environment multiple times. His team combined 'omics datasets from livers of surface and cave morphs to identify putative cis-regulatory changes that alter target genes and pathways directly involved in cave adaptation (Krishnan *et al*. 2022). Such a dissection of cis-regulatory evolution was also explored by Phillip

Davidson (Armin Moczek laboratory; Indiana University Bloomington) using dung beetles. These beetles have sexually dimorphic horn development, and in some species, males develop horns in response to nutritional cues. Davidson explored the cis-regulatory basis of this developmental plasticity using ‘omics approaches and identified enhancers that may be responsible for nutritional and sex-responsive differential development. In these studies, identification of relevant cis-regulatory changes relied on current molecular biological tools including ATAC-seq. It is intriguing to consider how DL approaches may complement this objective in the future.

Using DL models to obtain insights into the mechanisms of gene regulation should be a welcome addition to the current purely experimental approaches. For example, Evgeny Kvon (UC Irvine) used chromatin conformation capture technology to map enhancer-promoter interactions for thousands of validated mouse enhancers (Chen *et al.* 2022). They concluded that most enhancer-promoter loops are tissue specific and are significantly stronger when enhancers are active. Similarly, Tathagata Biswas (Nicolas Rohner laboratory; Stowers Institute) shared his work looking at global chromosomal architecture. These studies make inferences about critical 3D genome interactions that may differ between populations of cavefish. Since DL models can be trained to predict Hi-C data from sequence, these are areas that could benefit from an integrative approach using both DL and targeted experiments.

Several talks at the meeting leveraged the interplay between evolutionary changes and mechanistic insights into gene regulation. By taking evolutionary changes as a starting point, they used classical experimental approaches to uncover specific functions of enhancers, insulators, histone proteins, and TFs, sometimes at the level of a single locus. For instance, Mark Rebeiz (University of Pittsburgh) described elegant experiments dissecting evolutionary transformations at the *ebony* locus, where silencers have been systematically reshaped to impact pigmentation in specific Drosophila species. Likewise, Nicolas Gompel (Ludwig Maximilians University, Munich) used the Drosophila pigmentation system to revisit the notion of enhancer modularity at the *yellow* locus. Through analysis of wing spot pigmentation in a number of species, his laboratory showed how regulatory regions of this gene exhibit multifunctionality and partial redundancy that evolved over time. Dimple Notani (National Centre for Biological Sciences, Bangalore) discussed her laboratory's studies on estrogen-driven gene regulation, where clusters of enhancers appear to act in a cooperative fashion to drive gene expression. Some elements are prebound to the estrogen receptor prior to signaling. Others are induced and appear to require "driver" enhancers for activity. Notably, these elements are not functionally distinguishable based on previously measured chromatin properties.

Exploring evolutionary variation at the protein level, talks by Pravrutha Raman (Harmit Malik laboratory; Fred Hutchinson Cancer Center), David Arnosti (Michigan State University), and Pinar Onal (Shelby Blythe laboratory; Northwestern University) focused on particular TFs and their evolution. Pravrutha Raman's work examines the variability in histone proteins over evolutionary time. She found that ancestral histone variants H2A.X and H2A.Z are found to have fused to form a composite gene, H2A.V in many Diptera. Interestingly, some Drosophila have duplicated the H2A.V variant, with the duplicates expressed in males, indicating that evolutionary innovations in histone proteins may drive biological novelties. David Arnosti's talk about evolution of the C-terminal Binding Protein also investigated how this core component of the transcriptional apparatus has evolved in eukaryotes (Raicu *et al.* 2022). Deep phylogenetic analysis demonstrated that the intrinsically disordered C terminus bears a surprising level of conservation of short linear motifs, dating back to its earliest last common bilaterian ancestor. Evolution of the Bicoid TF was discussed by Pinar Onal, who specifically focused on the role of the Bicoid DNA-binding domain. In testing ancestral forms of this protein in the Drosophila embryo, she was able to replay the evolutionary history of this domain as Bicoid duplicated and evolved (Onar *et al.* 2021).

These focused individual studies demonstrate that learning general rules for gene regulation using DL models should inspire, but not replace, the focus on specific biological problems. Like individual works of art, biological systems need to be considered in their own right. Many aspects by which they operate represent highly specialized solutions to specific organismal challenges. As a result, highly complex combinations of molecular components, such as the HOX gene cluster, are often unique and may only be represented once in a genome (**Figure 7.1**). Thus, while DL innovations are ripe for application to the fields of gene regulation, development, and evolution, they do not replace the unique perspective that individual studies bring.

Perspectives

The meeting brought together experimental and computational biologists, whose goal is to uncover how gene expression is regulated in the context of evolution. In particular, there was a focus on how DL models can impact the study of gene regulation, development, and evolution (**Figure 7.1**). DL models can be interpreted to identify complex sequence rules that underlie TF binding, enhancer function, open chromatin regions, global chromosomal architecture, and key players in a GRN. These rules can then be tested and explored further using targeted experiments. DL models can also be exploited to make accurate predictions about genetic variation, which together with experimental approaches can help uncover enhancers and target genes involved in specific biological processes and complex phenotypes. Thus, DL models have the potential to become important tools among experimentalists, thereby accelerating unique insights into biological systems.

ACKNOWLEDGEMENTS

This work was supported by the National Institute for General Medical Sciences (grant number R01GM124137 to D.N.A), the National Institute of Child Health and Development (grant number F31HD105410 to A.M.R) and funding from the Stowers Institute for Medical Research to J.Z. We would like to thank Sara Jackson (Stowers Institute) and Megan Reder and other staff at ASBMB for excellent support of the meeting. We apologize to the colleagues whose work we could not highlight due to space constraints.

AUTHOR CONTRIBUTIONS

A. M. R. methodology; A. M. R., J. Z., and D. N. A. writing–original draft; A. M. R., J. Z., D. N. A., J. C. F., and N. R. writing–review and editing.

REFERENCES

[preprint] Avsec, Z., Weilert, M., Shrikumar, A., Alexandari, A., Krueger, S., Dalal, K., et al. (2020) **Deep learning at base-resolution reveals motif syntax of the cis-regulatory code.** BioRxiv. https://doi.org/10.1101/737981

[preprint] Chen, Z., Snetkova, V., Bower, G., Jacinto, S., Clock, B., Barozzi, I., et al. (2022) **Widespread increase in enhancer-promoter interactions during developmental enhancer activation in Mammals.** BioRxiv. https://doi.org/10.1101/2022.11.18.516017

[preprint] Raicu, A. M., Kadiyala, D., Niblock, M., Jain, A., Yang, Y., Bird, K. M., et al. (2022) **The cynosure of CtBP: Evolution of a bilaterian transcriptional corepressor.** BioRxiv. https://doi.org/10.1101/2022.06.23.497424

[preprint] Zhao, J., Lammers, N. C., Alamos, S., Kim, Y. J., Martini, G., and Garcia, H. G. (2022) **Optogenetic dissection of transcriptional repression in a multicellular organism.** BioRxiv. https://doi.org/10.1101/2022.11.20.517211 2022.100170

de Almeida, B. P., Reiter, F., Pagani, M., and Stark, A. (2022) **DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers.** Nat. Genet. 54, 613–624

Jimenez, E., Slevin, C. C., Song, W., Chen, Z., Frederickson, S. C., Gildea, D., et al. (2022) **A regulatory network of Sox and Six transcription factors initiate a cell fate transformation during hearing regeneration in adult zebrafish.** Cell Genomics. https://doi.org/10.1016/j.xgen.

Kaplow, I. M., Schäffer, D. E., Wirthlin, M. E., Lawler, A. J., Brown, A. R., Kleyman, M., et al. (2022) **Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin.** BMC Genomics 23, 291

Krishnan, J., Seidel, C. W., Zhang, N., Singh, N. P., VanCampen, J., Peuß, R., et al. (2022) **Genome-wide analysis of cis-regulatory changes underlying metabolic adaptation in cavefish.** Nat. Genet. 54, 684–693

Nair, S., Barrett, A., Li, D., Raney, B. J., Lee, B. T., Kerpedjiev, P., et al. (2022) **The dynseq browser track shows context-specific features at nucleotide resolution.** Nat. Genet. 54, 1581–1583

Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., and Mostafavi, S. (2022) **Obtaining genetics insights from deep learning via explainable artificial intelligence.** Nat. Rev. Genet. https://doi.org/10.1038/s41576-022-00532-2

Onar, P., Gunasinghe, H. I., Umezawa, K. Y., Zheng, M., Ling, J., Azeez, L., et al. (2021) **Suboptimal intermediates underlie evolution of the bicoid homeodomain.** Mol. Biol. Evol. 38, 2179–2190

Srivastava, D., Aydin, B., Mazzoni, E. O., and Mahony, S. (2021) **An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced transcription factor binding.** Genome Biol. 22, 20

Tareen, A., Kooshkbaghi, M., Posfai, A., Ireland, W. T., McCandlish, D. M., and Kinney, J. B. (2022) **MAVE-NN: Learning genotype phenotype maps from multiplex assays of variant effect.** Genome Biol. 23, 98

Vande Zande, P., Hill, M. S., and Wittkopp, P. J. (2022) **Pleiotropic effects of trans-regulatory mutations on fitness and gene expression.** Science 377, 105–109

Zeitlinger, J. (2020) **Seven myths of how transcription factors read the cis-regulatory code.** Curr. Opin. Syst. Biol. 23, 22–31

CHAPTER 8: CONCLUSIONS

What have we learned about Rb family proteins?

Traditionally, Rb proteins have been studied through genetic manipulations such as gene ablation or overexpression. These complementary approaches have been useful in determining genome-wide regulation of gene expression by Rb proteins, developmental impact of loss or gain of Rb, and for uncovering differences between paralogs. For instance, in human cells, Rb paralog ablation using RNAi followed by microarray and ChIP-seq identified classes of genes that are misexpressed only by certain paralogs, indicating paralog-specific regulation (Chicas *et al.* 2010). In the fly, *rbf1* RNAi has identified regulation of polarity genes and of myogenesis, among other processes (Payankaulam *et al.* 2016; Zappia *et al.* 2019). *rbf2* KO flies have been used to study the significance of Rbf2 in the ovary and testis, which cannot be done with *rbf1* KO, due to its lethal phenotype (de Oliveira, unpublished results; Payankaulam, unpublished results; Du and Dyson, 1999; Steveaux *et al.* 2002). We and others have overexpressed Rb paralogs and mutants in developing flies, followed by phenotypic analyses or RNA-seq, and uncovered the relative importance of certain domains or residues in development, as well as the differences in Rb paralog function in terms of genome-wide transcriptional effects (Acharya *et al.* 2010; Wei *et al.* 2015; Mouawad *et al.* 2019). These studies have been important in furthering the Rb field, and illustrating the use of Drosophila as a model for understanding Rb biology. Yet, these genetic manipulations may trigger complex effects, which are both direct and indirect, making it difficult for us to unravel which effects are Rb-specific, and which are due to non-specific effects. Additionally, RNA-seq and related tools are often used to identify genes that are significantly misregulated (with a cut-off of more than two-fold), removing any genes whose expression is modestly affected. Therefore, Rb-specific modest changes have mostly been overlooked, even though they may be physiologically relevant, as we discuss in Chapter 3.

To overcome the issue of indirect effects from genetic manipulations and to measure possible soft repression by Rb proteins, I developed an Rb-specific CRISPRi system as described in Chapter 2. I tested dCas9 fusions to Rb proteins in a living organism for the first time, using this tool for precise targeting of a single genomic locus, followed by phenotypic and transcriptional assays to measure the impact of targeting single genes at a time. Another motivation to use this tool was to directly compare Rbf1 and Rbf2 transcriptional impact in the same context, which has not been done so precisely and comprehensively before. I developed new dCas9-Rb constructs, generated novel fly lines expressing the dCas9 effectors in a tissue-specific manner, created gRNAs targeting different sites on a gene's promoter, compared Rbf1 and Rbf2 activity on ~30 promoters in the developing wing, and compared results to those obtained from transient transfections in S2 cells. I found that 1) Rb paralogs can have differential effects on targeted genes, 2) some genes are more sensitive to Rb recruitment than others, 3) loss of certain domains previously deemed necessary for repression does not affect Rbf1 ability to repress a target gene, indicating a role for promoter association rather than transcriptional repression per se, and 4) the effects of targeting genes *in vivo* do not always reflect results on transient reporters in cell culture. The implications of these findings are manyfold, as described below.

As we discuss in Chapter 2, we found paralog-specific effects. For instance, the phenotypic effects of targeting *InR* and *Pex2* are more pronounced for Rbf2, with little effect of Rbf1 targeting. We also see a specific Rbf2 effect using gRNA 4 or 5 alone on the *E2F2/Mpp6* promoter, while no phenotypic effect is observed with Rbf1 using those gRNAs. Finally, Rbf1 repression using gRNA 4 + 5 in the wing disc and ovary is more potent than Rbf2. These results all indicate that while on some genes the effect is similar (*wg*, *Acf*), we do observe and measure paralog-specific transcriptional impact.

The fact that some genes are more sensitive to Rb recruitment than others is not surprising, but is meaningful because this indicates that Rb proteins have promoter-specific effects, and are not reproducibly interfering with some common aspect of all regulatory regions. For instance, Rb's effects may include a change in nucleosome positioning that may differentially impact regulatory factors and the basal transcriptional machinery, depending on the composition of the basal promoter sequence, intrinsic nucleosome positioning sequences, and nearby transcription factors. My studies complement a plethora of work which have identified Rb effects only on particular sets of genes, such as cell cycle genes and ribosomal protein genes, but not on others. Out of the 28 promoters that we targeted *in vivo*, only one-third (10/28) showed a wing phenotype after targeting by Rbf1 or Rbf2, and six were impacted by both Rbf1 or Rbf2, with mostly mild effects. These observations are in line with our categorization of Rb as a "soft repressor" in Chapter 3. We infer that these genes are modestly repressed, as we saw for *E2F2*, although we did not measure expression changes for all of these genes.

Our third finding was the most surprising, as the Instability Element and pocket domain were previously deemed as necessary for Rb-mediated repression in both the fly and mammalian systems (Bremner *et al.* 1995; Acharya *et al.* 2010; Sengupta *et al.* 2015). We previously identified that Rbf1 IE and pocket mutants are unable to repress reporter genes in cell culture very well (Acharya *et al.* 2010; Raj *et al.* 2012a). Thus, our expectation that recruitment of Rbf1$^{\Delta IE}$ and Rbf1$^{\Delta pocket}$ to gene promoters would not lead to measurable repression was incorrect. In Chapter 2, we show that when tethered to dCas9, both Rbf1$^{\Delta IE}$ and Rbf1$^{\Delta pocket}$ are able to significantly repress *E2F2* and *Mpp6*. Therefore, the IE may only be required for recruitment to gene promoters, and that is why in overexpression assays, it does not repress very well, as it lacks full E2F binding. The result with the Rbf1$^{\Delta pocket}$ mutant is more surprising and difficult to explain. We propose that because the NTD resembles the pocket's cyclin fold structure, and because the CTD is known to interact with various factors including E2Fs, it is possible that the Rbf1$^{\Delta pocket}$ mutant, which retains NTD and CTD, may interact with cofactors using these remaining domains, to create a repression complex.

Finally, in Chapter 2 we show that an identical regulatory region responded somewhat differently to Rb factors when assayed in its native chromosomal environment versus as a reporter gene in cell culture. While position 4 and 5 were optimal for repression of *Mpp6 in vivo*, it was position 2 and 3 that were optimal in cell culture. These results point to differences in the chromatin environment *in vivo* versus on a reporter plasmid. A possibly related finding is that Rbf2 responsive genes differ between the embryo, female ovaries, and S2 cells, although in this case, cell-type differences in binding or expression of other TF may play a role (Stevaux *et al.* 2005). Our results suggest that the reliance on reporter gene assays for Rb studies in the mammalian system may overlook aspects of Rb activity in different tissue types and in development. Thus, my work in the developing fly is important for understanding how Rb paralog recruitment to endogenous genes impact tissue systems.

Future directions for understanding Rb activity

An obvious next step for the study of Rb impact on the *E2F2/Mpp6* promoter is testing the impact of these corepressors on the chromatin environment. We alluded to the significance of the chromatin environment on Rb's relative impact on this locus, but did not test how the basal transcriptional machinery was impacted, or whether histone tails were altered through Rb recruitment. Based on our finding that both *E2F2* and *Mpp6* are best repressed from near their respective TSSs in the larval wing disc, we expect that Rb is somehow impacting the basal transcriptional machinery, possibly through directly inhibiting the TFs or TAFs that recruit

RNAPII. Using ChIP-qPCR, a next step in measuring possible impacts on assembly of the preinitiation complex would include testing the binding of TBP, various TAFs, and RNAPII to the TSS before and after targeting of dCas9-Rb. Additionally, using antibodies against activating or repressing histone marks on the putative ~1 kb promoter will reveal how histone marks are impacted by Rb recruitment. Comparing the results from recruiting Rbf1 versus Rbf2 would convincingly show us whether these Rb paralogs have different effects on the chromatin environment on an Rb-sensitive promoter.

Future studies on Rb paralog proteins in Drosophila would also benefit from a more detailed study of the divergent domains of these proteins, which may provide them with different intrinsic repressive activities. As we have described, Rbf1 and Rbf2 differ predominantly in their N- and C-terminal domains. Performing a swap of these domains and testing their effects *in vivo* will uncover how evolution of these domains impact their activity, shedding light on how evolution has impacted Rb paralogy and transcription. This has been performed on the mammalian Rb and p107 proteins, whose pocket domains were swapped to create Rb-p107 chimeras that were tested as GAL4 fusions for repressive activity (Chow *et al.* 1996). We previously created chimeric proteins in which the Rbf1 and Rbf2 CTDs were swapped for their paralog's C-terminal domain (Yiliang Wei, unpublished data). These chimeras were then used in S2 transfection assays to test their ability to repress a *PCNA*-luciferase reporter gene. Rbf1 with the Rbf2 CTD was termed Rbf1-Rbf2C, and Rbf2 with the Rbf1 CTD was termed Rbf2-Rbf1C. Compared to WT Rbf1, Rbf1-Rbf2C was a slightly worse repressor of a *PCNA*-luciferase reporter in cell culture, indicating that the Rbf2 CTD somehow makes Rbf1 a weaker repressor. Perhaps this CTD alters the types of protein interactions Rbf1 can engage in. Compared to WT Rbf2, Rbf2-Rbf1C was a slightly better repressor of the *PCNA*-luciferase reporter. Thus, the Rbf1 CTD, with its IE, is superior to the Rbf2

CTD. This sequence may be optimal for creating a potent repression complex, or increase its affinity for cofactors, allowing for increased repression. These preliminary results are very interesting because they indicate that the CTD sequence truly does matter in repression potency. Perhaps divergence of the Rb paralog's C-terminal sequences is what led them to have different abilities to repress different promoters.

As a follow-up to this experiment, and to my work in Chapter 2, I propose testing these chimeric proteins through dCas9 fusions. Such an experiment, performed *in vivo* on some of the same gene targets I tested in the wing will shed light on the significance of these divergent CTDs. Testing them in the wing, rather than S2 cells, would give us a clue about their relative importance in a developing organism. Additionally, I propose creating chimeras where the Rbf1 CTD is swapped out with other Drosophila Rbf1 CTDs, and doing the same for Rbf2, to evaluate how evolutionary changes among different species of the same genus of fruit flies impacts Rb repression ability. For instance, taking the Rbf1 CTD from a closely related species like Drosophila simulans versus one from a more distantly related species, such as Drosophila grimshawi, and comparing effects on the same locus would illustrate how minute differences may play an important role in transcriptional regulation.

Another proposed experiment involves a high-throughput analysis of Rb paralog impact on gene expression, using the CRISPRi system. Such an experiment would allow us to uncover the rules governing each paralog's ability to repress a certain promoter. To do this, I propose performing a CRISPRi screen in S2 cells. S2 cells would be transduced with a lentiviral library of gRNAs targeting the promoter region of all Drosophila genes. Each cell would receive either dCas9-Rbf1 or dCas9-Rbf2, and one gRNA targeting one promoter (with 3-5 gRNAs designed for each promoter, in case some gRNAs are ineffective). This would be followed by single cell RNA-seq to determine whether the particular targeting event impacted expression of the gene target. Next, we would have to determine 1) if gRNA targeting led specifically to the intended gene's misexpression, 2) how many genes in total were misexpressed after the targeting, and 3) whether Rbf1 and Rbf2 impacted the same genes to the same magnitude, or if different categories of genes were impacted by each paralog. This high-throughput screen of Rb paralog impact on gene expression has never been done in any system, with such specificity. The results of this proposed experiment will be very meaningful in determining the significance of Drosophila encoding multiple Rb proteins. Such a screen using the mammalian Rb proteins in mammalian cell culture would also be of great significance to the field.

What have we learned about the elusive CtBP CTD?

In Chapter 4, we undertook a comparative phylogenetic study of the CtBP corepressor protein, focusing specifically on its unstructured C-terminal domain. Necessitated by the COVID-19 lockdown, I formed a group of junior researchers and performed a phylogenetic analysis of the CtBP CTD across 200 species. We performed hundreds of alignments of CtBP protein sequences to uncover the evolution of the CtBP CTD. Until our study, evolutionary studies of the CTD were limited to observations that the human, mouse, fly, and worm CTDs differed greatly (Nichols *et al.* 2008). While the mammalian and fly CTDs resemble one another, the worm CTD was known to be several fold longer, and not have comparable features. Until then, it was believed that the CTD was highly variable across Metazoa, and because of experimental work indicating that the CTD is not required for oligomerization or repression, its function was largely untested and unknown (Sutrias-Grau and Arnosti, 2004; Madison *et al.* 2013; Bellesis *et al.* 2018). Our comprehensive analysis uncovered several key characteristics of the CTD which were previously unappreciated: 1) the unstructured CTD of about 100 amino acids is found across all Bilateria, 2) specific motifs are conserved from protostomes to deuterostomes, 3) alternative splicing to produce different isoforms of the CTD is found outside of Insecta, including in ray-finned and lobe-finned fishes, and 4) a few lineages have completely altered their CTD, such as the loss of the CTD in leeches and gain of a large, presumably structured CTD in flatworms and roundworms.

It was previously known that the fly CtBP transcript is alternatively spliced to create two isoforms, one retaining and one lacking the CTD, and that this is a feature of other insect CtBP orthologs (Mani-Telang and Arnosti, 2007). Here, we have shown that this feature is found across Insecta, as many Diptera and Hymenoptera produce short isoforms through alternative splicing. Additionally, we found many instances of short isoforms being produced in deuterostomes such as tetrapods and various water-dwelling fishes. Thus, formation of CtBP isoforms which lack the unstructured CTD seems to have evolved independently several times, suggesting an important role of this isoform.

Additionally, we found that CtBP as a transcriptional repressor is a bilaterian innovation, stemming from our observations in Chapter 4, as well as experimental data from animal and plant studies that indicate the Arabidopsis protein is largely active in the cytoplasm. Notably, this CtBP homolog, ANGUSTIFOLIA, is only 30% similar in sequence to the human CtBP, with a divergent CTD. Thus, we suggest that non-bilaterian Metazoans, such as Cnidaria and Porifera, which also have proteins with a low (~30%) sequence similarity to the human CtBP, are likely to be alpha hydroxy acid dehydrogenases that are not transcriptional corepressors. The CtBP proteins with >50% sequence conservation in the dehydrogenase domain, a “long” CTD, and conserved tetramerization residues are likely to exhibit a function as transcriptional corepressors.

Finally, we find that CtBP diversified through multiple duplications in deuterostomes. The expression of only two CtBP paralogs in mammals, while other tetrapods have three, appears to be reflective of a mammal-specific loss of the third paralog, CtBP1-like. Ray-finned fish encode up to five paralogs, which was not previously appreciated. We hypothesize that duplication events must have happened at various points in deuterostome evolution. With the help of our collaborator, Dr. Siddiq, we substantiated this hypothesis by inferring a maximum-likelihood phylogeny from select CTD sequences using the best-fit model of sequence evolution.

Along with the computational approach to studying the CtBP CTD in Chapter 4, in Chapter 5, I used the CRISPRi method to compare Drosophila CtBP isoforms *in vivo*. This was the first time that dCas9 fusions have been created with CtBP(S) and CtBP(L) isoforms, and the first time CtBP has been studied in such a way in a developing organism. Like the dCas9-Rb fusions, we used the developing wing and targeted diverse gene promoters. Expression of dCas9-CtBP without a targeting gRNA was sufficient to induce a mild phenotype in adult wings, suggesting that perturbation of CtBP levels alone can induce mild developmental effects (primarily bristle phenotypes, suggestive of an impact on cell fate specification). Whether acting alone or in complexes with native CtBP, the dCas9-CtBP proteins may interfere with or augment the amount of repression mediated by CtBP *in vivo*. Despite this background effect, we observed pronounced CtBP- and gene-specific effects when co-expressing gene-specific gRNAs. The *E2F2/Mpp6* locus produced the most striking result, where CtBP(S) recruitment led to severe morphological defects, while CtBP(L) recruitment led to milder effects. This was the first report of CtBP CTD isoforms having such distinct effects when targeted to a promoter. In comparison, using GAL4 fusions, we previously reported that the two isoforms have similar repressive effects on a lacZ reporter in the embryo (Sutrias-Grau and Arnosti, 2004).

Thus, the results presented in Chapter 5 indicate a possible difference in repression ability by these two isoforms, shedding light on the importance of expressing both isoforms during

Drosophila development. On the *E2F2/Mpp6* endogenous locus, the unstructured domain seems to somehow restrain CtBP, and prevent it from having a severe impact on gene expression. Yet, when testing these isoforms on an *Mpp6*-luciferase reporter in cell culture, we do not observe a major difference in repressive ability by the two isoforms. This difference in relative activities may reflect the chromatin environment, which is likely to be very different on endogenous genes versus transfected reporter genes tested in cell culture, an effect also noted for the dCas9-Rb chimeras. My work using dCas9-CtBP effectors was an important starting point in delineating the significance of the CTD in transcriptional regulation, as an addition to the computational approach attempting to answer a similar question.

Future directions for understanding CtBP activity

To better understand how promoter context affects repression ability, future studies of CtBP in Drosophila would benefit from testing dCas9-CtBP effectors on additional gene promoters and in different tissue types. Only a few of the genes we targeted were sensitive to CtBP recruitment. These promoters were chosen based on the fact that we previously targeted dCas9-Rb chimeras to them, and we decided to compare the effect of CtBP recruitment to the same sites. However, the lack of effect at the phenotypic level may be because many of these genes are known Rb targets, and may not be regulated by endogenous CtBP. The lack of ChIP-seq for the Drosophila CtBP proteins makes it difficult for us to identify what kinds of genes to target, but we can make inferences based on mammalian ChIP-seq to test additional genes. For instance, genes involved in the epithelial to mesenchymal transition, apoptosis, or cell differentiation, which CtBP is known to regulate, would be obvious next targets. Additionally, testing the isoforms in the embryo versus the adult may be important in uncovering relative importance of each isoform at different developmental stages. CtBP(L) is known to be highly expressed in the embryonic stage, but expressed at lower levels later on in development, while CtBP(S) is expressed throughout all stages (Mani-Telang and Arnosti, 2007). In Chapter 5, we tested the isoforms in the larval wing disc, which may not be a tissue or time point in which the relative contribution of each isoform may be easily uncovered. A further context in which we have yet to test CtBP action is the possible regulation from distal enhancer sequences. Unlike Rb, many transcription factors such as Knirps and Kruppel that recruit CtBP operate on cis regulatory elements that lie kilobases away from the transcriptional initiation site. A possible effect of the CtBP CTD at these more distal elements remains to be studied.

Another way to test the relative contributions of CtBP(S) and CtBP(L) to regulation of gene expression is by making use of traditional CRISPR/Cas9 approaches to edit the endogenous CtBP locus to create flies expressing only one isoform or the other. I propose the generation of Drosophila only encoding CtBP(S) by designing gRNAs to remove the terminal exon that encodes for the long CTD, and generating Drosophila only encoding CtBP(L) by removing the intron that is typically read through to produce CtBP(S) (Raicu *et al.* 2023). Such an approach may require the use of Homology Directed Repair, so that a repair template with the correct CtBP open reading frame is used to make the intended edits in the genome. Our previous genomic rescue experiments indicate that the organism is able to survive with some degree of perturbation to the wild-type isoform expression, so I expect these flies to survive and be used for further experiments (Zhang and Arnosti, 2011). Once these isoform-specific flies are generated, I would use the embryos for ChIP-seq, to identify the genomic targets of each isoform and compare the binding profiles. I would also use these flies and compare viability, fertility, lifespan, and other adult phenotypes to identify the potential impact that loss of one of the isoforms has on development.

Finally, to combine the evolutionary and functional findings in Chapter 4 and Chapter 5, I propose swapping out the CtBP CTD from *Drosophila melanogaster* with the CTDs of closely related Drosophilids, insects, and other protostomes. The chimeric proteins would then be fused to dCas9 and tested on the same gene promoters as the WT proteins to identify possible functional impacts of evolved CTD sequences. I propose testing these dCas9-CtBP chimeras in the embryo, where we know both of the Drosophila isoforms are highly expressed.

Final thoughts about transcriptional corepressors

The work described in this dissertation has focused on two highly conserved transcriptional corepressors, Rb and CtBP. By using genetic tools in Drosophila, I was able to probe how these corepressors function to regulate transcription in a developing organism. I made use of the CRISPRi system, which is not well-developed in Drosophila, and demonstrated that fusions of dCas9 to Rb paralogs and CtBP isoforms provide an excellent basis for studying these corepressors' activity. CRISPRi is traditionally used as a method for turning off gene expression, with most studies aimed at identifying the best repressor and best gRNA combination, as a tool for the complete silencing of gene transcription. Here, we used CRISPRi for a different goal, of probing repression activity and uncovering repression mechanisms. We compared Rb paralogs to one another, considered effects on different promoters, measured impact from different positions on the same promoter, and tested Rb mutants. For CtBP, we also compared isoforms in different contexts. To our knowledge, such a dissection of corepressor activity has never before been performed using a heterologous DNA-binding protein such as GAL4 or dCas9.

Studies of corepressors have been challenging due to their diversity in mechanisms of repression. The same corepressor, such as Rb, can function as a strong or soft repressor depending on the gene promoter. CtBP can be recruited either by long-range or short-range repressors to impact transcription in different ways. Genetic manipulations such as knockout, knockdown, and overexpression have suffered from indirect effects, which the CRISPRi system here aimed to overcome. By using a combination of the powerful CRISPRi technique, analysis of extant ‘omics data, and investigation of corepressor evolution and diversification using a comparative phylogenetic approach, we have uncovered details of Rb and CtBP biology that were previously unknown. As both Rb and CtBP are associated with a diversity of human cancers, understanding the basic biology of these corepressors’ functions through an evolutionary lens is critical for subsequent studies. The results presented here may eventually help inform design of cancer therapeutics and personalized medicine.

REFERENCES

Acharya, P., Raj, N., Buckley, M. S., Zhang, L., Duperon, S., Williams, G., Henry, R. W., & Arnosti, D. N. (2010). **Paradoxical Instability–Activity Relationship Defines a Novel Regulatory Pathway for Retinoblastoma Proteins.** Molecular Biology of the Cell, 21(22), 3890–3901. https://doi.org/10.1091/mbc.e10-06-0520

Bellesis, A. G., Jecrois, A. M., Hayes, J. A., Schiffer, C. A., & Royer, W. E. (2018). **Assembly of human C-terminal binding protein (CtBP) into tetramers**. Journal of Biological Chemistry, 293(23), 9101–9112. https://doi.org/10.1074/jbc.RA118.002514

Chicas, A., Wang, X., Zhang, C., McCurrach, M., Zhao, Z., Mert, O., Dickins, R. A., Narita, M., Zhang, M., & Lowe, S. W. (2010). **Dissecting the Unique Role of the Retinoblastoma Tumor Suppressor during Cellular Senescence.** Cancer Cell, 17(4), 376–387. https://doi.org/10.1016/j.ccr.2010.01.023

Chow, K. N. B., Starostik, P., & Dean, D. C. (1996). **The Rb Family Contains a Conserved Cyclin-Dependent-Kinase-Regulated Transcriptional Repressor Motif.** Molecular and Cellular Biology, 16(12), 7173–7181. https://doi.org/10.1128/MCB.16.12.7173

Du, W., & Dyson, N. (1999). **The role of RBF in the introduction of G1 regulation during Drosophila embryogenesis.** The EMBO Journal, 18(4), 916–925. https://doi.org/10.1093/emboj/18.4.916

Madison, D. L., Wirz, J. A., Siess, D., & Lundblad, J. R. (2013). **Nicotinamide Adenine Dinucleotide-induced Multimerization of the Co-repressor CtBP1 Relies on a Switching Tryptophan.** Journal of Biological Chemistry, 288(39), 27836–27848. https://doi.org/10.1074/jbc.M113.493569

Mani-Telang, P., & Arnosti, D. N. (2007). **Developmental expression and phylogenetic conservation of alternatively spliced forms of the C-terminal binding protein corepressor.** Development Genes and Evolution, 217(2), 127–135. https://doi.org/10.1007/s00427-006-0121-4

Mouawad, R., Prasad, J., Thorley, D., Himadewi, P., Kadiyala, D., Wilson, N., Kapranov, P., & Arnosti, D. N. (2019). **Diversification of Retinoblastoma Protein Function Associated with Cis and Trans Adaptations.** Molecular Biology and Evolution, 36(12), 2790–2804. https://doi.org/10.1093/molbev/msz187

Nicholas, H. R., Lowry, J. A., Wu, T., & Crossley, M. (2008). **The Caenorhabditis elegans Protein CTBP-1 Defines a New Group of THAP Domain-Containing CtBP Corepressors.** Journal of Molecular Biology, 375(1), 1–11. https://doi.org/10.1016/j.jmb.2007.10.041

Payankaulam, S., Yeung, K., McNeill, H., Henry, R. W., & Arnosti, D. N. (2016). **Regulation of cell polarity determinants by the Retinoblastoma tumor suppressor protein.** Scientific Reports, 6(1). https://doi.org/10.1038/srep22879

Raj, N., Zhang, L., Wei, Y., Arnosti, D. N., & Henry, R. W. (2012). **Ubiquitination of**

Retinoblastoma Family Protein 1 Potentiates Gene-specific Repression Function. Journal of Biological Chemistry, 287(50), 41835–41843. https://doi.org/10.1074/jbc.M112.422428

Raicu, A.M., Kadiyala, D., Niblock, M., Jain, A., Yang, Y., Bird, K. M., Bertholf, K., Seenivasan, A., Siddiq, M., & Arnosti, D. N. (2023). **The Cynosure of CtBP: Evolution of a Bilaterian Transcriptional Corepressor.** Molecular Biology and Evolution, 40(2), msad003. https://doi.org/10.1093/molbev/msad003

Sutrias-Grau, M., & Arnosti, D. N. (2004). **CtBP Contributes Quantitatively to Knirps Repression Activity in an NAD Binding-Dependent Manner.** Molecular and Cellular Biology, 24(13), 5953–5966. https://doi.org/10.1128/MCB.24.13.5953-5966.2004

Stevaux, O. (2002). **Distinct mechanisms of E2F regulation by Drosophila RBF1 and RBF2.** The EMBO Journal, 21(18), 4927–4937. https://doi.org/10.1093/emboj/cdf501

Stevaux, O., Dimova, D. K., Ji, J.-Y., Moon, N. S., Frolov, M. V., & Dyson, N. J. (2005). **Retinoblastoma Family 2 is Required In Vivo for the Tissue-Specific Repression of dE2F2 target Genes.** Cell Cycle, 4(9), 1272–1280. https://doi.org/10.4161/cc.4.9.1982

Wei, Y., Mondal, S. S., Mouawad, R., Wilczyński, B., Henry, R. W., & Arnosti, D. N. (2015). **Genome-Wide Analysis of Drosophila RBf2 Protein Highlights the Diversity of RB Family Targets and Possible Role in Regulation of Ribosome Biosynthesis.** G3 Genes|Genomes|Genetics, 5(7), 1503–1515. https://doi.org/10.1534/g3.115.019166

Zappia, M. P., Rogers, A., Islam, A. B. M. M. K., & Frolov, M. V. (2019). **Rbf Activates the Myogenic Transcriptional Program to Promote Skeletal Muscle Differentiation.** Cell Reports, 26(3), 702-719.e6. https://doi.org/10.1016/j.celrep.2018.12.080

Zhang, Y. W., & Arnosti, D. N. (2011). **Conserved Catalytic and C-Terminal Regulatory Domains of the C-Terminal Binding Protein Corepressor Fine-Tune the Transcriptional Response in Development.** Molecular and Cellular Biology, 31(2), 375–384. https://doi.org/10.1128/MCB.00772-10