# Understanding the Regulatory Mechanisms of Drosophila and Human TGF-β pathways

by

Samantha M. Daly

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved December 2021 by the
Graduate Supervisory Committee:

Stuart J. Newfeld, Chair
David G. Capco
Tatiana P. Ugarova

ARIZONA STATE UNIVERSITY

May 2022

ABSTRACT

Proper regulation of the Transforming Growth Factor-beta (TGF-β) pathway is important for maintaining homeostasis and development in various tissues across vertebrates and invertebrates. When TGF-β pathway signaling is disrupted it leads to tumor growth, birth defects, and other diseases. The identification and study of the various regulatory methods utilized within TGF-β pathway signaling is important to aid the understanding of disease prognosis and prevention. In the TGF-β pathway in Drosophila, dCORL functions in the dActivin subpathway and acts as a regulator of dSmad2 in the larval brain. dCORL is encoded by a gene on the fourth chromosome, in Drosophila. To learn more about dCORL's role in the pathway, two fourth chromosomes were created that allow clonal analysis to be conducted. Clonal analysis is needed to determine dCORL's role in TGF-β regulation in the adult brain. In my first project, both chromosomes were successfully created. Though, the importance of understanding regulatory mechanisms goes past one protein. In my second project, multiple conserved prodomain cysteines were identified in human amino acid alignments of 33 TGF-β family proteins across the three TGF-β subfamilies. Database mining identified conserved prodomain cysteine mutations in 10 proteins and their mutant phenotypes. Common phenotypes for conserved cysteine mutations suggest new heterodimer pairs. The most frequent mutant phenotypes associated with new heterodimers were tumors. Conserved prodomain cysteine mutations were connected to cysteine mutations in known regulatory partner proteins by mutant phenotype, yielding numerous new regulatory interactions. The most frequent mutant phenotypes connecting new regulatory interactions between TGF-β proteins and regulatory partners proteins were tumors. Together, my projects expand knowledge of regulatory mechanisms within the TGF-β pathway in Drosophila and humans, while providing hypotheses for further investigation.

ACKNOWLEDGMENTS

This work would not have been possible without the mentorship of Dr. Stuart Newfeld. I would also like to thank Nancy Tran, Mike Stinchfield and Samuel Goldsmith for their time spent training me, discussing different aspects of experiments, and developing our understanding of complex concepts in the lab. During my early years in the Newfeld lab, I was supported by an NIH Research Supplement to Promote Diversity in Health-Related Research. Finally, I want to thank my friends and family for always supporting me and lifting my spirits through the many ups and downs that have occurred throughout my four years in the Newfeld lab. My thesis truly wouldn't have been possible without your contributions, and I will be forever grateful for your patience, support, and guidance.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Multicellular organisms use intracellular signaling to govern homeostasis and development. Intracellular signaling requires signaling molecules, receptors on a cell's surface, and a multitude of signaling molecules, though the number involved depends on the pathway. The proteins involved in intracellular signaling are created from the genes in an organism's DNA. Studying genes allows for experiments that target one gene and thereby one protein at a time. Knowing this, the role, or roles of a specific protein within a pathway can be studied by altering the gene that encodes for it. Common methods used in a genetic approach to studying cell signaling include gene knock outs, overexpression experiments, RNAi experiments, the use of reporter genes, and the creation of clones. These methods allow for the function of specific proteins to be studied either across an entire organism or with in specific tissues.

One common consequence of gene knock outs is the negative systemic affects the removal of a single protein often yields, which can lead to premature death of the organism. If one is trying to determine the role of a protein in a specific tissue or tissue type, the use of clones is preferable to gene knockouts. Clones yield mutated cells surrounded by wild type cells. Thus, clones reduce the negative side effects of a gene being knocked out throughout the entire organism and allow for mutant cells to be studied next to control cells. All methods of producing clones utilize recombination to yield one mutant daughter cell and one wild-type cell from a heterozygous parent cell. The requirement of recombination for clones has prevented the study of genes on the fourth chromosome via clones, as the fourth chromosome in Drosophila as does not naturally undergo recombination.

In addition to clonal analysis, scientists have often turned to computational methods to expand their understanding of genes. Aligning the amino acid sequences for a protein of

interest with proteins that have known functions, can lead to the identification of conserved structural regions and yield predictions for the function of proteins. Other common methods used to computationally study cell signaling include database mining, analysis of mutations and variants, and mutant phenotype comparisons. These methods allow for evolutionary comparisons and the function and regulation of proteins can be predicted based on sequence similarity.

Common forms of regulation depend on proteins binding to one another through disulfide bonds at specific cysteines. Disulfide bonds created at specific cysteines are often used in the formation of protein tertiary structure, homodimers, heterodimers, and to support partner protein binding. Homodimers are composed of two identical proteins, while heterodimers are composed of two different proteins. Most proteins do not act on their own, partner proteins typically bind to a protein and regulate that protein's ability to perform its role. The role of these specific disulfide bonds also depends on their location within a protein. If a protein is lacking a cysteine where one normally is, due to a substitution or deletion, it likely will affect that protein's structure or its ability to form a dimer, or its capability to bind to a partner protein. Alignments are often used to identify the positions of conserved cysteines. Knowing the position of a conserved cysteine is then useful for extrapolating the regulation of that protein by disulfide bonds.

I utilized a combination of lab-based techniques and computational methods to study cell signaling from a genetics perspective and expand knowledge of regulation in the TGF-β pathway.

**The Transforming Growth Factor-β Pathway**

The Transforming Growth Factor-β (TGF-β) family initiates developmental pathways found in vertebrates and invertebrates. These highly conserved pathways have many roles in development (cell differentiation and proliferation) and often lead to disease

when mutated (cancer and fibrosis). In *Drosophila melanogaster* the TGF-β family is comprised of two branches: Activin (dActivin) and Decapentaplegic (Dpp) (Figure 1).

In the dActivin pathway, a dActivin ligand first binds to the type II receptor [Punt or Wishful Thinking (Wit)]. The dActivin ligand binding to the type II receptor recruits the type I Receptor Baboon (Babo). The constitutively active kinases Punt or Wit will then phosphorylate Babo and form an activated heterodimer complex. Once phosphorylated, the Babo kinase domain is activated and phosphorylates dSmad2. Then dSmad2 recruits Medea. Together dSmad2 and Medea form a complex that translocates into the nucleus where it acts as a transcription factor. The Dpp pathway signals in a similar manner with its own signal transducers (Figure 1). A Dpp ligand will bind to the type II receptor Punt and will recruit its type I receptor [Saxophone (Sax) or Thickveins (Tkv)]. The activated heterodimer complex will then phosphorylate Mothers Against Dpp (Mad). Mad then binds to Medea and moves into the nucleus where the Mad-Medea multi-Smad complex functions as a transcription factor. The dSmad2/Medea and Mad/Medea multi-Smad complexes will then stimulate or repress different genes in the nucleus (Kahlem and Newfeld, 2009).

My two thesis projects focus on understanding regulation of the TGF-β pathway at two stages. My first project aims to understand the role of a protein (dCORL) within the TGF-β pathway signal cascade of a receptor cell in Drosophila. While my second project aims to identify new interactions between human TGF-β proteins within a secreting cell prior to their secretion.

**The Regulation of Smads and Prodomain Cysteines**

The identification of regulatory proteins within the TGF-β pathway and the understanding of various regulatory methods employed within TGF-β pathway signaling is important for the expansion of our understanding of protein regulation and cell signaling.

dSno is a protein that regulates TGF-β acting as a pathway switch and increasing dActivin signaling in Drosophila. dSno acts in both the dActivin and Dpp subpathways. When bound to the Co-Smad Medea dSno has reduced affinity for Mad and increased affinity for dSmad2 yielding increased dSmad2 mediated signaling while also antagonizing Dpp signaling (Takaesu et al., 2006). dCORL was first identified in the phylogenetic analysis of dSno and is encoded by a gene on the long arm of the Drosophila fourth chromosome. The fourth chromosome is difficult to work with as it contains only 3.5 percent of the Drosophila genome and does not naturally undergo recombination (Locke and McDermid, 1993). The lack of recombination prevents the creation of clones for any gene on the fourth chromosome. dCORL is placed in the dActivin pathway as it is required for the activation of Ecdysone Receptor Beta1 (EcR-B1). EcR-B1 is a nuclear receptor downstream of Babo and dSmad2 that is expressed in the Mushroom Body (MB) of the brain in Drosophila. dCORL regulates EcR-B1 as a dosage dependent and tissue specific co-factor for dSmad2 (Takaesu et al., 2012).

The TGF-β family of proteins function within the TGF-β signaling pathway, and is a large family comprised of proteins with similar structures what interact with TGF-β receptors. In humans, there are 33 members of the TGF-β protein family. Each of the 33 proteins fits into to one of three subfamilies based on structural similarity, which increases within each subfamily. Eight of the 33 human TGF-β family proteins belong to the TGF-β subfamily, eight belong to the Activin subfamily, and 17 belong to Bone Morphogenetic Protein (BMP) subfamily. Proteins in all three TGF-β subfamilies are secreted as dimers with both prodomains covalently bonded to the two ligand domains prior to secretion. This is achieved by the preproprotein monomers first forming a dimer, the prodomains of this dimer are then cleaved, then they are non-covalently bound to the ligand dimer. The dimer is then secreted either in this conformation or in a conformation with one cysteine from each prodomain in the dimer forming a disulfide bond with two cysteines in a binding

partner protein. For this project, binding partners refer only to proteins known to bind with TGF-β family proteins through disulfide bonds. The prodomain dimer, in this conformation, blocks the ligand dimer from interacting with its receptor, preventing TGF-β family proteins from unnecessary signaling. Binding partners aid in regulating TGF-β signaling, by binding to this conformation and sequestering it to the extracellular matrix of the secreting cell. Only one of the 33 human TGF-β family proteins has had its cysteine involved in forming the disulfide bonds with a binding partner: TGFB1's prodomain Cys33. The study of TGF-β family proteins has been focused on the ligand domain, leaving the prodomain largely understudied. Recent work has identified 31 conserved cysteines in two regions of the prodomain of 33 mouse TGF-β family proteins. The Association (Assn) region contains two positions where conserved cysteines are found while the β8 element contains three positions where conserved cysteines have been identified. One such conserved cysteine is Cys33, in the association domain of TGFB1. Looking into the β8 element and Assn region of conserved cysteines in Humans would expand the limited knowledge currently held on TGF-β prodomain cysteines.

**Summary**

In Chapter 2, two new fourth chromosomes were created that will allow Mosaic Analysis with a Repressible Cell Marker (MARCM) clones to be produced in Drosophila brains (Lee and Lou, 1999). One new fourth chromosome contained a $w^+$FRT, $y^+$GAL80, and normal dCORL present and the second contained the same $w^+$FRT present with a seven-nucleotide deletion in dCORL. The lab will employ these chromosomes to create MARCM clones in neurons of adult Drosophila brains and identify the role of dCORL in Drosophila insulin-like peptide 2 (dILP2) secretion. Eight insulin like peptides have been identified in Drosophila and only three are expressed in Drosophila adult brains: dILP2, dILP3, and dILP5 (Grönke et al., 2010). A previous study found that the removal of

dCORL in dILP2 neurons yields a loss of neurons that express dILP2, while neurons that coexpress dILP2 and Drifter were unaffected (Tran et al., 2018b). Drifter (Drf) is a transcription factor that is expressed in Drosophila brains, within a subset of dCORL and dILP2 coexpressing neurons and in several non-dILP2 expressing neurons (Tran et al., 2018b). The exact role of dCORL in dILP2 transcription is currently unknown. dCORL is currently placed in the dActivin side of the TGF-β pathway. Evidence suggests that dCORL works below dSmad2 but before the signaling gets into the nucleus. My hypothesis is that dCORL may act as a regulator of dILP2 transcription. An alternative hypothesis is that the lack of dCORL in dILP2 cells leads to apoptosis. Identifying dCORL's role in dILP2 transcription will expand our knowledge of TGF-β signaling by adding to our understanding of dCORL function and identifying if it contributes to regulatory mechanisms.

In Chapter 3 alignments of the β8 element and Assn region for all 33 human TGF-β proteins were created. Within these two alignments, a total of 61 conserved prodomain cysteines were identified. Only one of the 33 human TGF-β family proteins have had the cysteine involved with partner protein binding biochemically validated: TGFB1 prodomain Cys33. It is through the disulfide bond that this cysteine creates that partner proteins regulate TGF-β signaling. Following the identification of 61 conserved prodomain cysteines, mutant phenotypes accosted with mutations in those conserved cysteines were identified in both the TGF-β proteins and their known binding partners. Shared mutant phenotypes caused by mutations in conserved prodomain cysteines within TGF-β family proteins suggest heterodimer pairs. While identification of shared mutant phenotypes caused by mutations in conserved prodomain cysteines in both TGF-β family proteins and known binding partners indicates a disruption in TGF-β/binding partner pairing via disulfide bonds. Analysis focused on identification of heterodimer pairs and binding partners has historically been limited to TGFB1. My work broadens the search to 33 human

TGF-β proteins and shared mutant phenotype data suggests 12 potential heterodimer pairs and 23 binding partner interactions, to be tested by the larger community. If any of these predicted interactions are confirmed it will expand the understanding of TGF-β signaling and regulation, including new examples of extra-subfamily heterodimer formation.

Together, my projects expand knowledge of regulatory mechanisms within the TGF-β pathway in the receptor cell in Drosophila and within the secreting cell in humans, while providing hypotheses for further investigation.

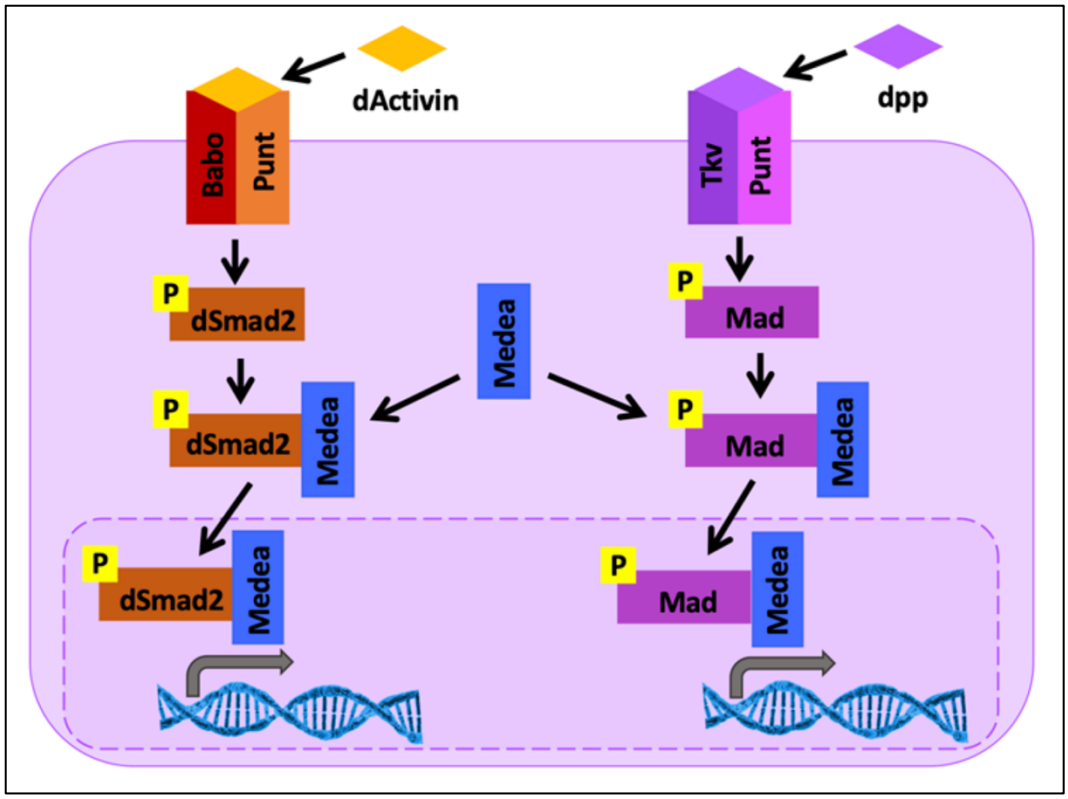


**Figure 1: The TGF-β Signaling Pathway with its Two Branches in Drosophila: dActivin and Dpp.** The cell is defined by the solid purple line and the nucleus is the dotted purple line. On the left is dActivin representing its subfamily while on the right is Dpp representing its subfamily. Each respective ligand binds to their type II receptor that will then recruit a branch specific type I receptor and phosphorylate the type I receptor. Once the two receptors form a heterodimer complex it phosphorylates a branch specific receptor associated Smad: dSmad2 or Mad. Phosphorylation is shown via the yellow squares with "P". The phosphorylated receptor associated Smads are then in their active form and recruit the shared common-mediator Smad Medea. Then the receptor associated Smad and common-mediator Smad complex moves into the nucleus to function as a transcription factor.

# CHAPTER 2

# GENETIC TECHNOLOGY ADVANCEMENTS TO IDENTIFY THE ROLE OF dCORL IN THE TGF-β PATHWAY

## INTRODUCTION

dCORL is encoded by a gene on the fourth chromosome and has been studied in the lab since its identification in Drosophila. dCORL is a protein in the dActivin pathway and the exact function of dCORL is currently unknown. dCORL shares sequence similarity with dSno indicating that dCORL may, like dSno, act in a regulatory manner. dSno binds to Medea and increases dActivin signaling. There is no evidence that dCORL also performs as a pathway switch. Rather, evidence suggests that it works with or beneath dSmad2 and prior to EcR-B1's transcription. dCORL acts as a model of TGF-β regulation. dCORL clones will allow for the identification of dCORL's function within the dActivin pathway. The regulation of pathways is important to ensure proper signaling control, especially for a timing and tissue specific pathway such as the TGF-β pathway.

### dCORL Functions in the TGF-β Pathway

dCORL was first identified during the analysis of the protein dSno, a Ski-related novel oncogene in Drosophila. The analysis showed that dSno functions in both signaling branches of the TGF-β pathway, acting as a novel pathway switch. dSno when bound to Medea has reduced affinity for Mad and increased affinity for dSmad2. This has the effect of increased dSmad2 mediated signaling and simultaneously antagonizing Dpp signaling. This is how dSno acts as a pathway switch (Takaesu et al., 2006).

During the analysis of dSno, a phylogenetic tree of Sno/Ski family proteins was constructed. The phylogeny included the amino acid sequences of the closely related protein sub-families Sno/Ski, and Dachshund (Dac). The CORL protein sub-family and its

predicted gene CG11093 were identified as related to the Sno/Ski/Dac protein sub-families. The Drosophila homolog of mouse CORL2 (Stinchfield et al., 2019), CG11093, was named Drosophila CORL (dCORL). CORL was first identified in mice as a Co-repressor for the homeodomain transcription factor Lbx1 (Mizuhara et al., 2005). The CORL protein family includes mouse CORL1, mouse CORL2, human SKOR1, human SKOR2, and Drosophila CORL (Takaesu et al., 2012). In specific domains the amino acid sequences for all three protein sub-families (Sno/Ski, Dac, CORL) have high levels of conservation between their human, mouse and Drosophila homologs. Sno/Ski proteins have roles in the TGF-β pathway during development (Liu et al., 2001). After finding that dCORL has high sequence homology to dSno the Newfeld lab hypothesized that it could function within the TGF-β signaling pathway.

In 2012, dCORL was demonstrated to be required for the activation of Ecdysone Receptor Beta1 (EcR-B1). EcR-B1 is a nuclear receptor downstream of Babo and dSmad2 that is expressed in the Mushroom Body of the brain in Drosophila. To study the function of dCORL the mutant fly line *Df(4)dCORL* was employed. *Df(4)dCORL* is a mutation that was created via FLP-FRT (recombinase flippase – flippase recognition target) induced intrachromosomal recombination, which deleted four genes (GluRA, CG32016, dCORL and Sphinx) on the fourth chromosome. *Df(4)dCORL* and dCORL RNAi experiments both showed disrupted EcR-B1 signaling in the Mushroom Body, which indicated dCORL's role in the dActivin subfamily of TGF-β signaling. Additional data suggested that dCORL acts as a dosage dependent and tissue specific co-factor for dSmad2 that could potentially have a role in determining Mushroom Body cell fate (Takaesu et al., 2012).

To better understand dCORL regulation a set of dCORL reporter lines were created and their expression patterns were analyzed. Immunofluorescence (IF) showed that AH-lacZ is the reporter whose expression pattern captures the largest fraction of dCORL expression at all stages (Tran et al., 2018a). AH-lacZ was expressed in the insulin

producing cells (IPCs) of the Drosophila brain in larvae and adults. In wild type larval AH-lacZ brains: between eight and10 neurons were observed to co-express Drf and AH-lacZ and all dILP2 expressing neurons (12-16) coexpressed AH-lacZ. In wild type adult AH-lacZ brains expression was observed in an average of 18.4 neurons, Drf in an average of 9.8 neurons and dILP2 in an average of 16.7 neurons (Tran et al., 2018b).

Following the discovery that AH-lacZ coexpresses with Drf and dILP2 in IPC's, experiments were conducted to determine the extent of Drf and dILP2 coexpression in wild type IPCs. Imaging of larval brains showed that Drf and dILP2 coexpress in 10 neurons of the brain while six neurons expressed dILP2 alone. The experiment was then repeated in *Df(4)dCORL* larvae to determine if dCORL has any effect on the expression of Drf or dILP2 in the brain. *Df(4)dCORL* staining demonstrated that all dILP2 neurons not expressing Drf were missing. dILP2 neurons expressing Drf were unaffected (Tran et al., 2018b).

Subsequently, analysis of wild type adult brains showed an apical monolayer of IPC's expressing Drf only as well as Drf and dILP2 coexpressing neurons. dILP2 only neurons were observed in an inverted pyramid shape medial the Drf monolayer. The experiment was then repeated in *Df(4)dCORL* adult brains. The coexpression of Drf and dILP2 is unaffected in the *Df(4)dCORL* adult brains but there is a reduction of 35% in dILP2 neurons to an average of 11.4 cells. As in larvae all dILP2 only neurons were missing in *Df(4)dCORL* (Tran et al., 2018b).

A dCORL RNAi experiment was then conducted in an attempt to phenocopy the *Df(4)dCORL* mutant phenotype in adults and show that the reduction in dILP2 cells resulted specifically from the deletion of dCORL. They employed a GAL4 driver that mimicked AH-lacZ expression in all dILP2 cells of the adult brain. This experiment revealed that dCORL RNAi induced the loss of dILP2 neurons that did not also express

Drf, supporting the hypothesis that the *Df(4)dCORL* IPC phenotype is due to loss of dCORL (Tran et al., 2018b).

The loss of dILP2 expressing cells in the IPC of *Df(4)dCORL* adult brains has not been explained to date. I hypothesize that it results from dILP2 not being transcribed rather than the death of that neuron. To test these hypotheses, Mosaic Analysis with a Repressible Cell Marker (MARCM) clones of a dCORL CRISPR mutation will be created. MARCM will allow me to understand how dCORL affects dILP2 transcription by comparing mutant cells to neighboring wild type cells.

**Genetic Systems Utilized to Conduct MARCM with dCORL**

Previous work from the Newfeld lab suggests the hypothesis that dCORL has a role in Drosophila insulin-like peptide 2 (dILP2) neuronal expression. My proposal is to create MARCM clones to determine if dCORL has a role in the neuronal expression of dILP2 or if dCORL has a role in expressing programmed cell death. I will investigate this question through loss of function experiments using a dCORL CRISPR mutation. Here a CRISPR/Cas9 derived mutation will be used in place of *Df(4)dCORL*. *Df(4)dCORL* was used in the initial experiment and required additional experiments to ascertain what extent of the phenotype resulted from the deletion of dCORL alone. Such additional steps will no longer be required with the use of a CRISPR mutation.

Conducting this experiment will require the use of multiple genetic systems including FLP/FRT, GAL4/UAS, GAL4/GAL80, and CRISPR/Cas9 (Lee and Lou, 1999). Genetic schemes will utilize these systems to create the genotypes necessary for MARCM clones. The MARCM clones will be dCORL CRISPR mutant neurons marked with GFP that are surrounded by wild type neurons lacking GFP. Having marked mutants allows for easier identification of mutant cells, a feature that is especially useful when looking for clones in the brain. Clones yield mutant cells surrounded by wild type cells, an arrangement

that removes the effects of the mutation outside the cell of interest. The effects of the mutation can be rigorously connected to a phenotype. The reliability of the MARCM clone method for the analysis of a loss of function phenotype in a single cell leads it to be widely used in Drosophila. One exception is chromosome four.

To conduct this MARCM experiment a fourth chromosome mutation in dCORL and a fourth chromosome expressing GAL80 and bearing an FRT proximal to dCORL are required. In nature, Drosophila fourth chromosomes do not recombine, so a unique method was required to get GAL80 and FRT transgenes onto the same fourth chromosome. Under normal conditions *Bloom syndrome helicase* (*Blm*) is a protein that regulates meiosis crossover patterning (crossover assurance, crossover interference and the centromere effect) resulting in accurate segregation. A double trans-heterozygous mutant genotype for *Blm* and a second gene called *recombination defect*ive (*rec*) yields higher crossover rates on all chromosomes and includes crossover events on the fourth chromosome. The loss of normal crossover patterning in the *Blm rec* double mutant also results in nondisjunction and aneuploidy during cell division (Hatkevich et al., 2017). Testing for the segregation of GAL80 and FRT away from one another on all potentially recombined fourth chromosomes will act as a control for any nondisjunction and ploidy issues.

The FLP-FRT (Flip recombinase – Flippase Recognition Target) system is one of the required aspects of MARCM. For a MARCM experiment an FRT must be present on homologous fourth chromosomes. One of these chromosomes will contain an FRT and a GAL80 transgene. The other will contain an FRT and a dCORL CRISPR mutant. The FLP-FRT system uses a yeast site-specific recombinase called flippase (FLP) to incite recombination events between two flippase recognition target (FRT) sites. See Figure 2. Recombination can occur between two FRT's on homologous chromosomes resulting in recombination or between two FRT's on the same chromosome resulting in a deletion of the genes between the FRT sites. The FLP in this system is induced via a heat shock

promotor, adding temperature dependent control to the system (Chou and Perrimon, 1996; Takaesu et al., 2012; Germani et al., 2018).

The GAL4/UAS system and GAL4/GAL80 system are also required components of MARCM. GAL4 and GAL80 are transcription factors initially found in yeast that are commonly applied to Drosophila. The GAL4 protein will bind to an upstream activation sequence (UAS) and drive the expression of the gene downstream of the UAS. The GAL4 gives the system specificity as it can be expressed in a time or tissue specific pattern. Genes downstream of the UAS are only activated when GAL4 is activated resulting in cells that are only marked (commonly with lacZ or GFP) where the GAL4/UAS interaction occurs (Lue et al., 1987).

The GAL80 protein in a transcriptional repressor of GAL4. When both proteins are present in a cell, GAL80 will bind to the GAL4 promotor repressing it and disrupting the GAL4/UAS system (Lue et al., 1987). In MARCM experiments, recombination at the FRT between the homologues noted above creates unequal daughter cells where one cell no longer expresses GAL80. In this daughter cell Gal4 expression is restored; Figure 3.

To create a dCORL mutant on an FRT chromosome for MARCM, CRISPR/Cas9 will be deployed. CRISPR/Cas9 is a system capable of creating germline mutations. This occurs when the Cas9 nuclease is directed to the dCORL gene by a dCORL-specific 20-nucleotide guide RNA via Watson-Crick base pairing. Once bound to the gene, Cas9 stimulates a double stand break at the site where the guide RNA hybridizes to the chromosome. After cleavage is complete, mutations can be formed by errors in DNA repair mechanisms: nonhomologous end joining or homology-directed repair (Ran et al., 2013).

Clones of mutant cells in an otherwise wild-type individual have been used to effectively study how mutations affect an organism without the burden of organismal death starting with twin-spot clones in the 1930's (Germani et al., 2018). Typically, the clones would be unmarked in an individual expressing a cell autonomous marker such as GFP

ubiquitously. The situation is reversed with Mosaic Analysis with a Repressible Cell Marker (MARCM). This is a method of creating marked clones in an unmarked background. MARCM is conducted in an organism that is heterozygous for the gene of interest with a GAL4/UAS driven GFP, FLP and two FRT's on homologous chromosomes (Lee and Lou, 1999). One chromosome contains an FRT proximal to a mutant gene of interest and the other contains an FRT proximal to a GAL80 and a wild-type copy of the gene of interest; Figure 3. The heat shock induced homologous recombination by the FLP-FRT system results in homozygous mutant cells that are marked with GFP. Though the FRT can create recombination between any two of the four fourth chromatids present during mitoses, the crossover shown in Figure 3B is one of two events that will result in genetically distinct daughter cells (Chou and Perrimon, 1996). Cells where the GAL80 is present will not express GFP, as GAL80 is a repressor of GAL4, and cells not containing the GAL80 will be marked clones (Lee and Lou, 1999).

MARCM has been conducted in Drosophila, with a fourth chromosome gene before, though it used a highly complicated method. In 2012, Sousa-Neves and Schinaman created MARCM clones with a fourth chromosome gene by translocating part of chromosome four onto chromosome two and part of chromosome two onto chromosome four. This method was necessary due to a lack of a fourth chromosome with both an FRT and GAL80 present (Sousa-Neves and Schinaman, 2012). My thesis project will be the first time MARCM can be completed with the FRT, GAL80 and gene of interest all on the fourth chromosome.

MARCM requires a fourth chromosome with the FRT, GAL80 and the gene of interest *in cis*. Before my project to create a fourth chromosome with an FRT and a GAL80 *in cis* with dCORL could occur, a fourth chromosome with a GAL80 was created. A previous project in the lab concluded with the jump of a transposon carrying $y^{+}$GAL80 onto the fourth chromosome. *y w*; P[$y^{+}$; GAL80-MW1] resulted from a jump off the X

chromosome onto a wild-type fourth chromosome. This jump was conducted with MARCM as the end goal. The functionality of *y w*; P[$y^+$; GAL80-MW1] was tested by comparing its ability to a tubulin promoter driven GAL80 in the suppression of three different GAL4 lines driving UAS.GFP (Figure 4). Experiments with MARCM clones on the X chromosome have used tub-GAL80 as suppressor of GAL4. Figure 4 demonstrates *y w*; P[$y^+$; GAL80-MW1]'s suppression of GAL4 to be as good as tub-GAL80 in neurons and glia of larval brains.

The schemes for the creation of both fourth chromosomes employ the FLP-FRT, GAL4/UAS, and GAL4/GAL80 systems to generate components of the MARCM method of creating clones. Together, with a dCORL CRISPR mutant, MARCM will allow me to test my hypothesis in single mutant neurons. Analyses of that neuron for dILP2 expression or caspase-3 (a marker of programmed cell death) will reveal the function of dCORL in dILP2 expressing neurons.

## MATERIALS AND METHODS

### Drosophila Stocks

The Bloomington Drosophila Stock Center (BDSC) provided the following stocks for the *Blm rec* double mutant recombination: #1836 *y w*; Pb[$w^+$; FRT-CG2316$^{f00836}$], #6782 w; P[$w^+$] *pum$^{bcm}$/Tm6 Ubx$^1$*. The BDSC also provided the #6420 *yw* P{ry[+t7.2]=70FLP}3F stock that was used in the creation of y- wing clones.

Other stocks utilized to generate the *Blm rec* double mutant flies include: *w*; *dpp$^{ho}$* map/*CyO*;*Blm$^{N1}$ rec$^2$* P[$w^+$;UASp.Blm]/*Tm6b, y*; *Blm$^{D2}$ ry rec$^1$ Ubx$^{by34e}$* P[$w^+$; mat α–GAL4]/*Tm6b* (Hatkevich et al., 2017). The Kondo lab (National Institute of Genetics, Mishima, Japan) provided the following stocks used in the generation of the dCORL CRISPR mutant fly lines: *y w*; P[y+] attP2 [nos.Cas9] and y cho v; P[y+] attP40 [U6-gRNAv2-dCORL].

Newfeld stocks utilized in the generation of the *Blm rec* double mutant flies include: 1) *y w*; *D gl/Tm3 Sb Ser,* 2) *y w*; Pb[w$^{+}$] *Df(4)dCORL/ci$^{D}$,* 3) *y w*; P[y$^{+}$; GAL80-MW1], 4) *y w*; *lgs/ci$^{D}$*, and 5) *y w*. Newfeld stocks utilized in the generation the dCORL CRISPR mutants include: 1) *w*; *Dr/TM3 Tb Sb; unc-13/ci$^{D}$*, 2) *w; Sco/CyO; unc-13/ci$^{D}$*, 3) *y w*; *Gla/SM6a*, 4) *y w*; PB[w+; FRT-CG2316$^{f00836}$], and 5) *y w*; *lgs/ci$^{D}$.* Newfeld stocks utilized to generate y- wing clones include: 1) *y w* hs-FLP; PB[w+; FRT-CG2316$^{f00836}$], and 2) *y w*; *y w*; PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1], (lines 332, 343, and 360).

**Creation of a Fourth Chromosome with an FRT and GAL80 for MARCM**

Creation of a fourth chromosome with both an FRT and GAL80 began with two fly lines: one with a w$^{+}$FRT at CG2316 and one with the newly jumped *y w*; P[y$^{+}$; GAL80-MW1] located distally. Females capable of recombination on the fourth chromosome were *Bloom syndrome helicase, recombination defective* (*Blm rec*) double mutants with the genotype: *y w*; *Blm$^{N1}$ rec$^{2}$* P[w$^{+}$;UAS.Blm]/ *Blm$^{D2}$ ry rec$^{1}$ Ubx$^{by34e}$* P[w$^{+}$;mat-α]; P[y$^{+}$; GAL80-MW1]/ PB[w+; FRT-CG2316$^{f00836}$]. These females had X chromosomes with mutations for the *yellow* and *white* genes allow tracking of the FRT transgene (marked with the white gene) and the GAL80 transgene (marked with the yellow gene). Their third chromosome is where the recombination stimulating mutations *Blm$^{N1}$ rec$^{2}$* P[w$^{+}$;UAS.Blm] / *Blm$^{D2}$ ry rec$^{1}$ Ubx$^{by34e}$* P [w$^{+}$; mat α– GAL4] are located. Their fourth chromosome had the transgenes P[y$^{+}$; GAL80-MW1]/ PB[w+; FRT-CG2316$^{f00836}$]. A summary of the scheme used to create a fourth chromosome capable of conducting MARCM clonal analysis is depicted in Figure 5 and took 21 months to complete.

**Creation of a dCORL CRISPR Mutation on the PB[w+; FRT-CG2316$^{f00836}$] Fourth Chromosome**

Creating dCORL CRISPR mutations utilized two fly lines, one containing nos-Cas9 and the other a guide RNA for dCORL. These lines were obtained from the O'Connor lab (Univ. Minnesota). The guide RNA contained a 20-nucleotide sequence (corresponding to nucleotides 499-468) from within the dCORL open reading frame. The nos promotor drives Cas9 in the female germline. The genotype *y w*; P[y+] attP40 [U6-gRNAv2-dCORL]; P[y+] attP2 [nos.Cas9]; PB[w+; FRT-CG2316$^{f00836}$]

allows CRISPR to occur in the ovaries of females (Figure 7, line two). Fifteen stocks with potential CRISPR mutations in dCORL, on the fourth chromosome with PB[w+; FRT-CG2316$^{f00836}$], were sent to the O'Connor lab for sequencing. Four dCORL mutants were identified *dCORL$^{A}$*, *dCORL$^{B}$*, *dCORL$^{F}$*, and *dCORL$^{J}$* (Table 2 and Figure 8).

**Heat Shock**

The function of the FRT at CG2316 was tested in two experiments. In the first, three FRT and GAL80 *in cis* lines were examined. In the second, one dCORL CRISPR line was examined. This was conducted to ensure that the function of PB[w+; FRT-CG2316$^{f00836}$] was not affected by the *Blm rec* double mutant or CRISPR. First, fly lines 332, 343, and 360 (*y w*; PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1]/ *ci$^{D}$*) were individually crossed to a heat shock FLP containing line *y w* hs-FLP; PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1]. If the FRT at CG2316 is still functional the heat shock was expected to yield y- wing clones in the y+ individuals. Each individual cross was placed in a water bath at 37$^{o}$C for one hour on days two, three, and four of the experiment. The 37$^{o}$C water bath activated the heat shock FLP promotor and FLP in-turn initiated recombination at the FRT of the PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1] chromosomes (Figure 2). When this occurs in developing larvae it results in two daughter

cells with different genotypes and phenotypes. One daughter cell has two copies of P[$y^+$; GAL80-MW1] yielding dark (y+) cells, the same color as the parental cell. The second daughter cell has no copies of P[$y^+$; GAL80-MW1] yielding light colored (y-) cells. Figure 2 displays the process of y- clone creation due to heat shock and Figure 6 displays the y- clones produced by this test.

The FRT in *y w*; PB[w+; FRT-CG2316$^{f00836}$]-*dCORL$^F$* / *ci$^D$* was then tested. This test followed the same steps as described above. Figure 9 displays the y- clones produced by this test.

**Adult Collection for Wing Clone Analysis**

After eclosing, adults were moved into empty vials and aged 24 hours to allow their wings to fully develop. At 24 hours post eclosion wings were dissected and mounted. After mounting, the slides were viewed under a light microscope for y- clones in a y+ background. Images were taken of wings with y-clones. All three lines yielded clones. One of the lines, 343, is shown with y- clones in a y+ background indicated by an arrowhead Figure 6C.

**RESULTS**

**Creation of a Fourth Chromosome with an FRT and GAL80 for MARCM**

Recombination-competent females with the genotype *y w*; *Blm$^{N1}$ rec$^2$* P[$w^+$;UAS.Blm]/ *Blm$^{D2}$ ry rec$^1$ ubx$^{by34e}$* P[$w^+$;mat-α]; P[$y^+$; GAL80-MW1] / PB[w+; FRT-CG2316$^{f00836}$] (Figure 5 line two) yielded nine candidate recombinant chromosomes, each in an independent fly line. Of the nine candidates two flies died without progeny and four had complicated segregation – possibly resulting from ploidy issues or illegitimate recombination. The remaining three candidate lines, 332, 343 and 360, displayed perfect segregation away from the fourth chromosome mutation *ci$^D$* (Table1). Perfect segregation

occurred when the dark bodies (from the y+ in P[$y^+$; GAL80-MW1]), and orange eyes (from the w+ in PB[w+; FRT-CG2316$^{f00836}$]) were always seen together and never with a wing vein truncation (from $ci^D$) in the progeny of *y w*; PB[w+; FRT-CG2316$^{f00836}$] P[$y^+$; GAL80-MW1] (332, 343, or 360) / $ci^D$ crossed to *y w* flies. See Figure 5 line four. The four candidate fly lines who were discarded gave progeny with only orange eyes or only dark bodies or both with $ci^D$ indicating that the PB[w+; FRT-CG2316$^{f00836}$] and P[$y^+$; GAL80-MW1] were on separate chromosomes or were together but not on the fourth chromosome.

Once lines 332, 343, and 360 were confirmed to result from perfect segregation they were individually tested to identify if the FRT on the fourth chromosome was still functional. Functionality of the FRT was conducted via heat shock. All three lines (332, 343, and 360) yielded y- wing clones in y+ organisms. Line 343 consistently produced the largest clones found in this analysis. Figure 6 displays images of the bristles located at the edge of the wing. A y- clone from line 343 is indicated by an arrowhead in Figure 6C.

The production of three fly lines whose progeny always segregate into *y w*; PB[w+; FRT-CG2316$^{f00836}$] P[$y^+$; GAL80-MW1] or *y w*; $ci^D$, I conclude that the *Blm rec* double mutant was successful in allowing recombination on the fourth chromosome. The increase in recombination yielded three fourth chromosomes with both PB[w+; FRT-CG2316$^{f00836}$] and P[$y^+$; GAL80-MW1] *in cis*. Successful clone generation demonstrated the continued functionality of the FRT in all three *y w*; PB[w+; FRT-CG2316$^{f00836}$] P[$y^+$; GAL80-MW1] lines. My three new fourth chromosomes containing the functional PB[w+; FRT-CG2316$^{f00836}$] and P[$y^+$; GAL80-MW1] will allow for MARCM clones to be created with fourth chromosome genes. The presence of wild type dCORL in all three lines will allow for any of them to be utilized when creating dCORL MARCM clones.

**Creation of a dCORL CRISPR Mutation on the PB[w+; FRT-CG2316$^{f00836}$] Fourth Chromosome**

Females with the genotype *y w*; P[y+] attP40 [U6-gRNAv2-dCORL]; P[y+] attP2 [nos.Cas9]; PB[w+; FRT-CG2316$^{f00836}$] had CRISPR occur in their oocytes (Figure 7, line two). Fifteen single female progeny in line three with potential mutations with potential CRISPR mutations in dCORL *in cis* with PB[w+; FRT-CG2316$^{f00836}$] were utilized to created 15 stocks with single *y w*; lgls / *ci$^{D}$* males. This cross is shown in Figure 7 line three. Sequencing confirmed four of the 15 stocks contained a unique CRISPR mutation in dCORL on a PB[w+; FRT-CG2316$^{f00836}$] containing fourth chromosome. The four unique dCORL mutants were identified as *dCORL$^{A}$*, *dCORL$^{B}$*, *dCORL$^{F}$*, and *dCORL$^{J}$* (Table 2 and Figure 8).

Table 2 and Figure 8 display the four mutations made in dCORL. In Figure 8 the blue box represents the three amino acid deletion in *dCORL$^{A}$*. The green arrow in Figure 8 indicates the single nucleotide deletion resulting in a frameshift starting at amino acid 153 for *dCORL$^{B}$*, the green arrow is also shown in Table 2. The red arrow in Figure 8 indicates the five-nucleotide deletion resulting in a frameshift starting at amino acid 150 in *dCORL$^{J}$*, the red arrow is also shown in Table 2. The yellow arrow in Figure 8 indicates the seven-nucleotide deletion resulting in a frameshift starting at amino acid 151 in *dCORL$^{F}$*, the green arrow is also shown in Table 2.

Before I could test if the FRTs in the four new CRISPR dCORL lines were still functional COVID-19 hit Arizona, and I had to move to conducting research from home due to my autoimmune disease. The test of the FRT on my four dCORL CRISPR lines was conduct by Samuel Goldsmith. All four unique dCORL CRISPR lines produced y- wing clones in y+ individuals. Successful clone generation demonstrated the continued functionality of the FRT in all four *y w*; PB[w+; FRT-CG2316$^{f00836}$]-dCORL / *ci$^{D}$* mutant lines. A y- clone from *dCORL$^{F}$* is indicated by an arrowhead in Figure 9C.

Sequencing data confirmed that the dCORL mutation is present on the PB[w+; FRT-CG2316$^{f00836}$] chromosome, and y- wing clones in the y+ background confirm that the FRT on dCORL mutant chromosome is functional. The four new fourth chromosomes containing the functional PB[w+; FRT-CG2316$^{f00836}$] and a dCORL CRISPR mutation will allow for dCORL MARCM clones to be created.

## DISCUSSION

Recombination-competent females (Figure 5 line two) yielded nine candidate recombinant chromosomes, believed to contain PB[w+; FRT-CG2316$^{f00836}$] and P[y$^{+}$; GAL80-MW1] *in cis* on the fourth chromosome. Four of the nine candidates had complicated segregation, possibly resulting from ploidy issues or illegitimate recombination. The *Blm rec* double trans-heterozygous mutant yields a higher crossover rate on all chromosomes and allows crossover events on the fourth chromosome. The loss of normal crossover patterning in the *Blm rec* double mutant is known to result in nondisjunction and aneuploidy during cell division (Hatkevich et al., 2017). The known potential for ploidy issues and illegitimate recombination events is why the nine candidate chromosomes were tested for perfect segregation (Figure 5 line four) and why the perfect segregation test was followed by an additional cross to *y w* (Figure 5 line five). The use of these steps and the sequencing that followed ensured that the fly lines that would be used in conducting MARCM clones contained PB[w+; FRT-CG2316$^{f00836}$] and P[y$^{+}$; GAL80-MW1] *in cis* on the fourth chromosome.

### Conducting MARCM Clones With a dCORL CRISPR Mutation

The creation of MARCM clones with a dCORL CRISPR mutation to determine if a phenotype found by a previous graduate student resulted from a disruption in the transcription of dILP2 or programmed cell death, will be completed by another graduate

student. Due to the COVID pandemic's appearance in Arizona in March of 2020, my thesis project switched from this lab biased project to a computational one, discussed in detail in Chapter 3. The fruition of this project will now be conducted by the Ph.D. student in the lab.

If I were to stay in the lab and complete this project, I would test my hypothesis that the loss of dILP2 expressing cells in the *Df(4)dCORL* mutant results from dILP2 not being expressed rather than cell death. The ability of MARCM to mark a single mutant cell with GFP will allow me to identify individual neurons and analyze if they are expressing caspase-3 or dILP2, due to the loss of dCORL. My prediction that dCORL CRISPR mutant MARCM clone neurons will not express dILP2 or Caspase-3, would also be tested. An alternative outcome could be that dCORL CRISPR mutant MARCM clone neurons express Caspase-3 but not dILP2. The latter result would support the alternative that dCORL has a role in cell death.

To produce MARCM clones and test my hypothesis, the two fly lines I created will be utilized. The MARCM experiment requires progeny with the following components: UAS.GFP, dCORL.GAL4, heat shock FLP, the PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1] fourth chromosome containing a wild type dCORL gene, and the PB[w+; FRT-CG2316$^{f00836}$] fourth chromosome with a dCORL CRISPR mutation (Figure 10). dCORL.GAL4 is necessary to drive UAS.GFP in the native expression pattern of dCORL. The dCORL.GAL4 transgenic line was designed by Dr. Newfeld and generated in the O'Connor lab (Univ. Minnesota).

The recombinant chromosome *y w*; PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1] line 343, will first be crossed with *y w* hs-FLP; PB[w+; FRT-CG2316$^{f00836}$] to add heat shock FLP. This will yield the male parent of the MARCM fly. The dCORL CRISPR *y w*; PB[w+; FRT-CG2316$^{f00836}$]-*dCORL$^{F7}$*, will be the female parent of the MARCM fly. These parents will yield a MARCM fly will have an X chromosome with UAS.GFP / heat

shock FLP, a second chromosome with dCORL.GAL4 and a fourth chromosome with PB[w+; FRT-CG2316$^{f00836}$]-*dCORL$^{F7}$* / PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1] line 343. Figure 10 displays the cross that will yield the MARCM fly.

Once the MARCM cross is set up, I would place the vials in a water bath at 37$^{o}$C for 1 hour on days two, three, and four of the experiment. The 37$^{o}$C water bath will activate the heat shock FLP promotor and the FLP in turn will initiate recombination at the FRTs of the PB[w+; FRT-CG2316$^{f00836}$]-*dCORL$^{F7}$* chromosome and the PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1] chromosome. The resulting daughter cells will either have the dCORL CRISPR mutation and be marked by GFP or will have the functional dCORL gene and not express GFP due to GAL80 repression.

I would then separate out one day post-eclosion female brains (the age and gender of the previous study; Tran et al., 2018b), dissect and fix them in a four percent formaldehyde solution, then rinse, and store them in methanol. Two groups of adult brains will then be stained in parallel: the control group of brains with GFP, dILP2, Drf and FasII and the experimental group of brains with GFP, dILP2, Caspase-3 and FasII. These two cocktails of antibodies would be used to test my hypothesis, as coexpression of Drf and dILP2 is expected in the control while coexpression of Caspase-3 and dILP2 is unexpected. The expectation results from Caspase-3 marking cell death and if a cell is dead, it is incapable of expressing dILP2. My primary antibodies would include chicken anti-GFP 1:2000, guinea pig anti-Drf 1:500, mouse anti-FasII 1:250, rat anti-dILP2 1:1000, and rabbit cleaved-Caspase-3 1:500. My secondary antibodies would be Alexa Fluor 405, 488, 546, and 633 diluted to 1:500. After staining the brains will be mounted and imaged on a SP8 confocal microscope with slices taken every two µm. The lasers I would use are Diode 405 nm, Argon (Ar) 488 nm, Krypton (Kr) 546 nm, and Neon (Ne) 633 nm.

**Summary**

The creation two new fourth chromosomes will allow for the generation of dCORL MARCM clones. While I was unable to complete this experiment, it will be conducted by a Ph.D. student in the lab. Production of MARCM clones with fourth chromosome genes is enabled by the combination of the two new fourth chromosomes. This method will allow for the identification of dCORL's role in the dActivin pathway in adult Drosophila brains and determine if dCORL acts as a regulator of dILP2 transcription in the adult Drosophila brain.

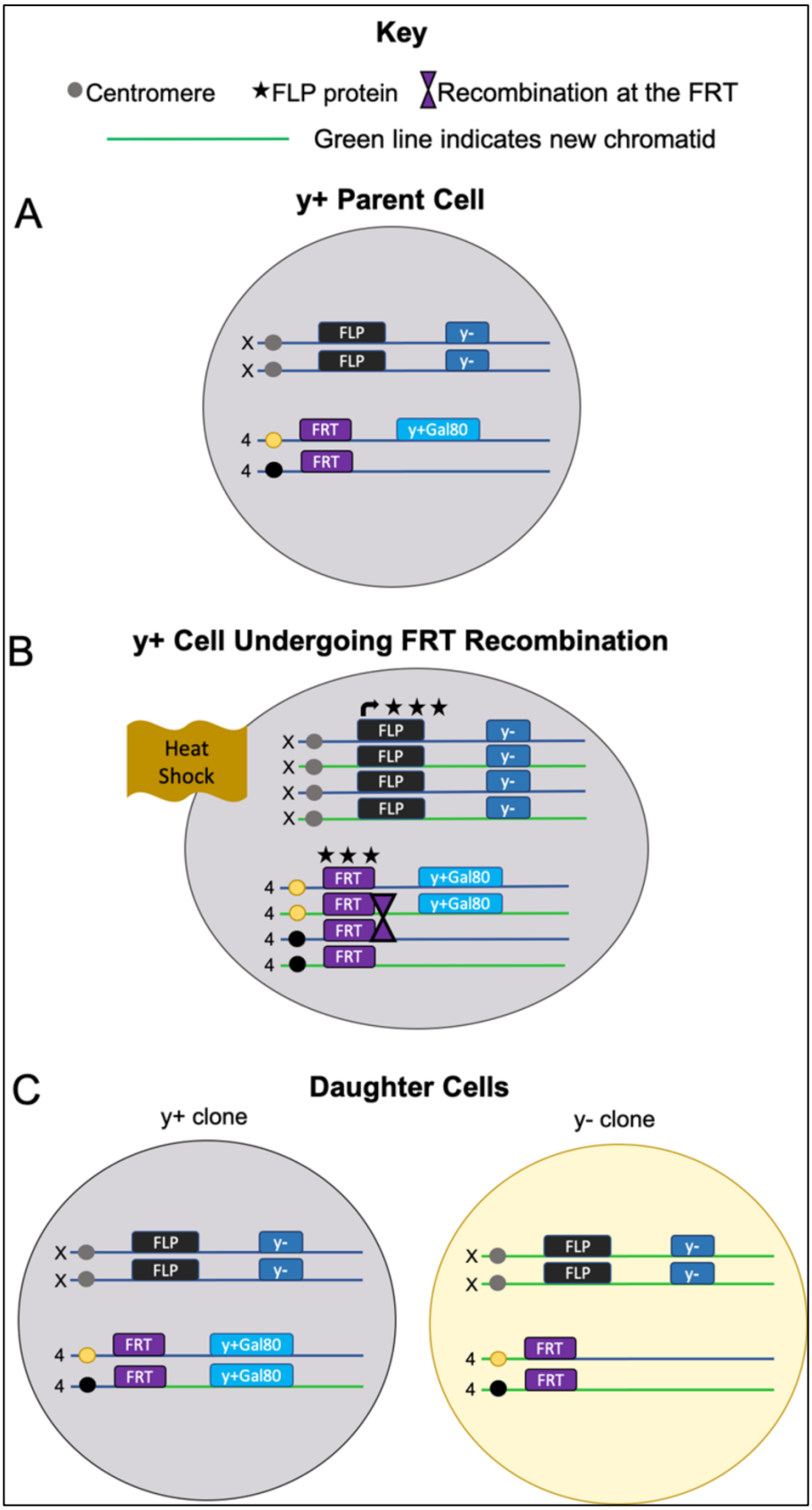
Key
Centromere
FLP protein
Recombination at the FRT
Green line indicates new chromatid
A
y+ Parent Cell
FLP
y-
FRT
y+Gal80
B
y+ Cell Undergoing FRT Recombination
Heat Shock
C
Daughter Cells
y+ clone
y- clone

**Figure 2: How FRT Recombination Creates y- Wing Clones in a y+ Background.** Yellow and black shaded centromeres indicate homologous chromosomes of the fourth chromosome. (A) Displays the parent cell showing only the X chromosome and chromosome four because those two chromosomes contain the four genes of interest for this procedure. FLP recombinase and the y- mutation are located on the X chromosome. PB[w+; FRT-CG2316$^{f00836}$] and P[$y^+$; GAL80-MW1] are on the fourth chromosome. The y+ indicates a dark body phenotype and is dominant over the y- mutation, so all cells with the P[$y^+$; GAL80-MW1] are dark. The parent cell is an example of a cell in the developing wing. (B) Depicts the parent cell undergoing mitotic recombination after DNA replication with the help of a heat shock induced FLP recombinase. FLP in turn initiates recombination at the FRT, visualized by the black stars and purple triangles. After recombination at the FRT has occurred sister chromatids segregate according to their centromeres. (C) Displays two non-identical daughter cells, one that is y+ (dark) and one that is y- (light). The y+ clone has all four genes of interest. The result of the recombination event in (B) is shown in the y+ clone by the black centromere fourth chromosome including portions of the yellow centromere chromosomes newly synthesized (green) chromatid. The y- clone has three of the four genes of interest, without the y+ from P[$y^+$; GAL80-MW1], thus the y- on the X chromosome yields a light-colored cell. The result of the recombination event that occurred in (B) is shown in the y- clone by the yellow centromere fourth chromosome including portions black centromere chromosome original (dark blue) chromatid.

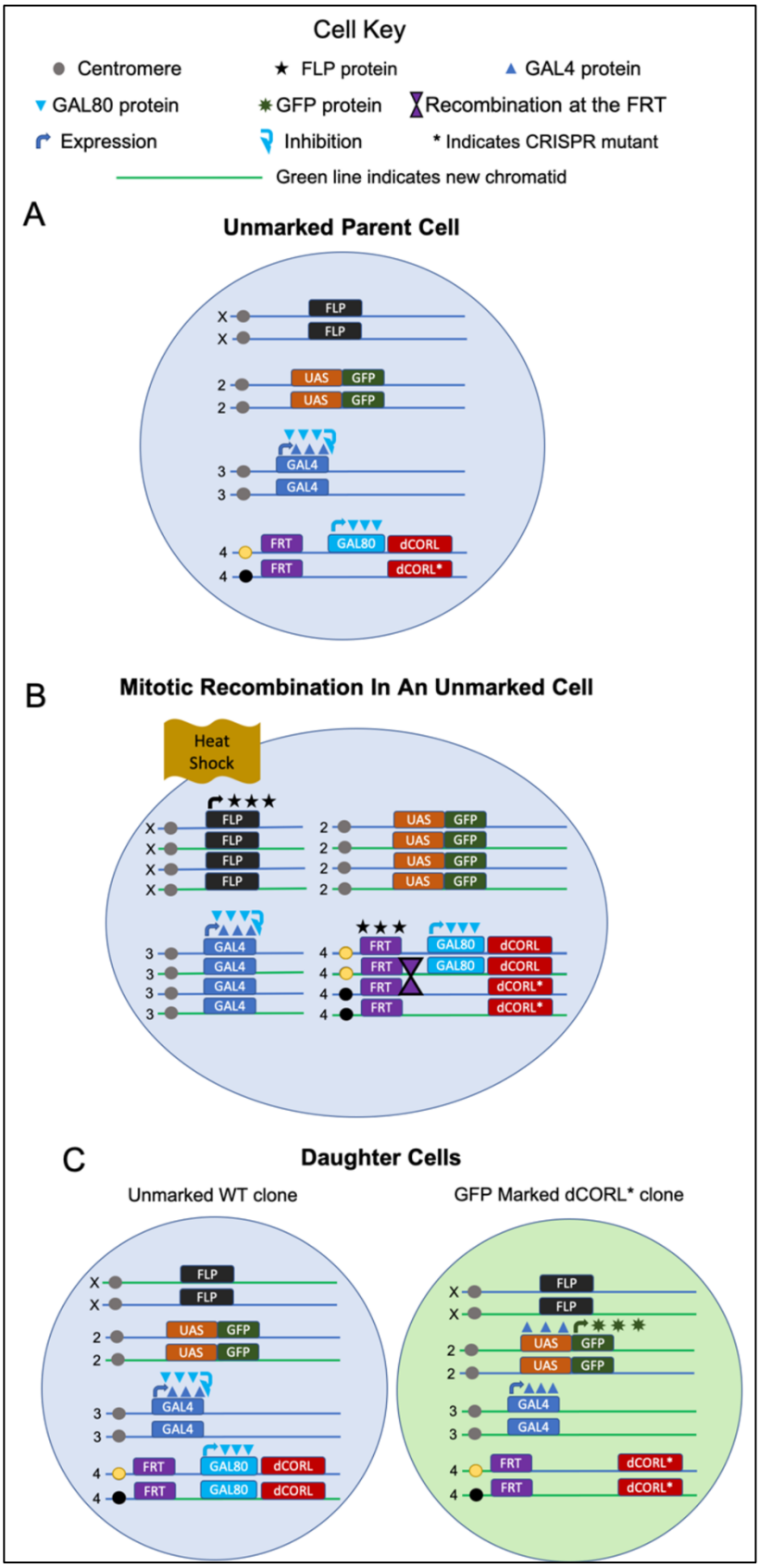
Cell Key
Centromere
FLP protein
GAL4 protein
GAL80 protein
GFP protein
Recombination at the FRT
Expression
Inhibition
* Indicates CRISPR mutant
Green line indicates new chromatid
A
Unmarked Parent Cell
FLP
UAS
GFP
GAL4
FRT
GAL80
dCORL
dCORL*
B
Mitotic Recombination In An Unmarked Cell
Heat Shock
C
Daughter Cells
Unmarked WT clone
GFP Marked dCORL* clone

**Figure 3: How FRT Recombination is Utilized in the MARCM Method of Creating GFP Marked Single Mutant Cell Clones.** Yellow and black shaded centromeres indicate homologous chromosomes of the fourth chromosome. (A) Displays the parent cell containing the X chromosome and chromosomes two, three, and four as genes required for MARCM are present across all four chromosomes. FLP recombinase is located on the X chromosome. UAS.GFP is located on chromosome two. dCORL.GAL4 is located on chromosome three. Chromosome four is heterozygous. The yellow centromere homolog includes PB[w+; FRT-CG2316$^{f00836}$], P[$y^{+}$; GAL80-MW1] and wild type dCORL. The black centromere homolog includes PB[w+; FRT-CG2316$^{f00836}$] and the dCORL CRISPR mutation F7. The parent cell containing all of the listed components results in a cell that is not marked by GFP, as GAL80 present on the fourth chromosome will repress GAL4 on chromosome 3. If the GAL80 was not present the GAL4 on chromosome three would bind to the UAS on chromosome two and GFP would be expressed. The parent cell is an example of an dILP2 expressing neuron. (B) Depicts the parent cell undergoing mitotic recombination after DNA replication with the help of a heat shock induced FLP recombinase. FLP in turn initiates recombination at the FRT, visualized by the black stars and purple triangles. After recombination at the FRT has occurred sister chromatids segregate according to the centromeres. (C) Displays two non-identical daughter cells, one that is unmarked by GFP and one that is marked by GFP. The unmarked clone has all the genes of interest present. The GAL80 continues to repress GAL4 and prevent GFP expression. The recombination that occurred in (B) is shown in the unmarked clone by the black centromere fourth chromosome including portions of the yellow centromere chromosome newly synthesized (green) chromatid. The GFP marked clone lacks the GAL80, allowing for GAL4 to drive GFP expression and has the CRISPR dCORL mutation present. The recombination event that occurred in (B) is shown here by the yellow centromere fourth chromosome including portions of the black centromere chromosome original (dark blue) chromatid.

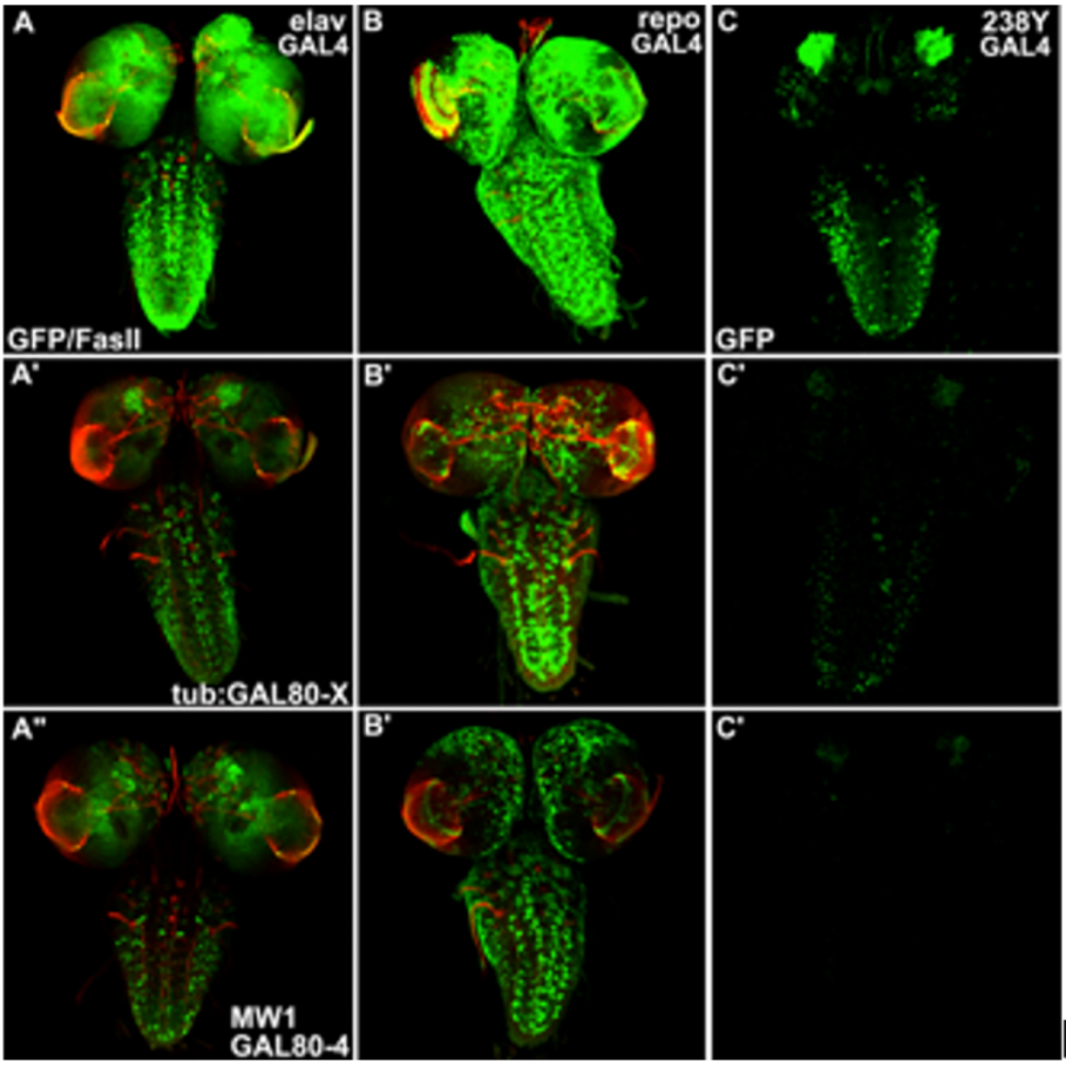


**Figure 4: GAL4 Suppression by *y w*; P[y$^+$; GAL80-MW1].** Confocal stacks of third instar larval brains stained side by side with FASII (red) and GFP. Top Row: UAS.GFP driven by three GAL4 lines. The GAL4 lines Elav, Repo and 238y all express in the brain just like dCORL. Middle Row: Suppression by Tubulin-GAL80 on X. Tubulin-GAL80 is a common GAL80 for MARCM clones on the X chromosome. Bottom Row: In comparison to Tubulin-GAL80 on the X chromosome, the same or better suppression of GAL4 is seen with P[y$^+$; GAL80-MW1] on the fourth chromosome. Staining by Nancy Tran. Note these images do not contain any clones.

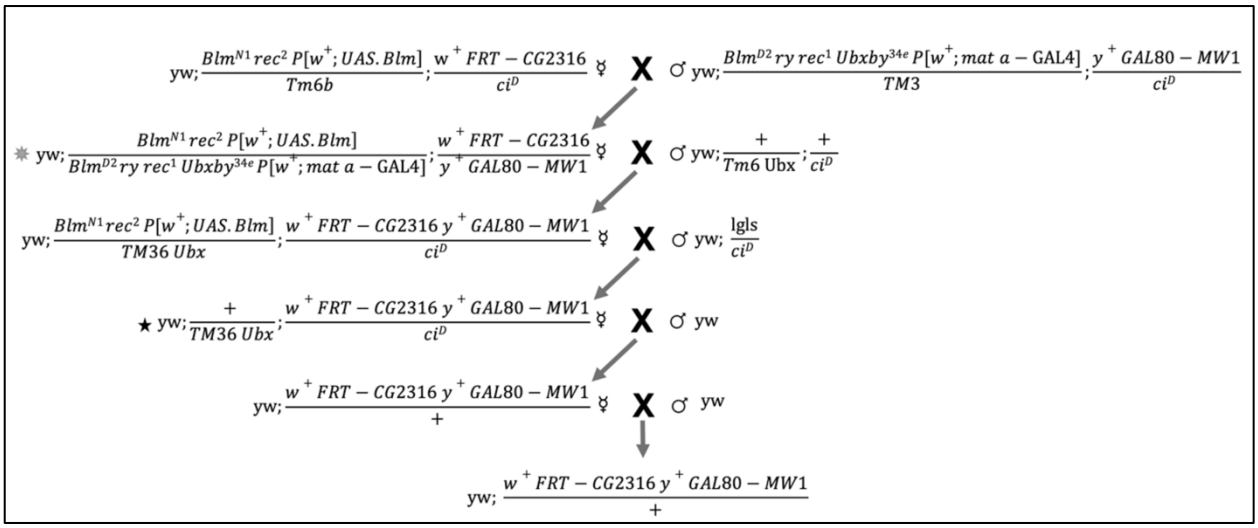

**Figure 5: Crossing Scheme Used to Create a Fourth Chromosome with PB[w+; FRT-CG2316$^{f00836}$] and P[y$^{+}$; GAL80-MW1].** These are the last five crosses in a larger scheme that was used to generate a fourth chromosome with both PB[w+; FRT-CG2316$^{f00836}$] and P[y$^{+}$; GAL80-MW1] (line six). Recombination does not naturally occur on the fourth chromosome, here it was induced through the use of two *Bloom syndrome helicase* mutations and two *recombination defective* mutations (line two). The grey star in line two indicates the female recombination on the fourth chromosome to create a chromosome with both PB[w+; FRT-CG2316$^{f00836}$] and P[y$^{+}$; GAL80-MW1]. The black star (line four) indicates the cross that acted as my test for perfect segregation - identifying if the PB[w+; FRT-CG2316$^{f00836}$] and P[y$^{+}$; GAL80-MW1] segregated away from one another or if the recombination was successful. Three flies produced in line six were then sib-mated to create three stocks: 332, 343 and 360.

| Tested Flies | Candidates | Perfect Segregation | Candidate Success | Success of Perfect Segregation |
|---|---|---|---|---|
| 936 | 9 | 3 | 0.96% | 0.32 % |

**Table 1: The Segregation Test of Nine Candidate Recombinant Lines Yielded Three Perfectly Segregating Chromosomes.** The screening of 936 flies in line five of Figure 5 yielded nine candidates with both PB[w+; FRT-CG2316$^{f00836}$] and P[y$^{+}$; GAL80-MW1]. Three of these eventually displayed perfect segregation together and away from *ci*$^{D}$ in accordance with Mendel's first law.

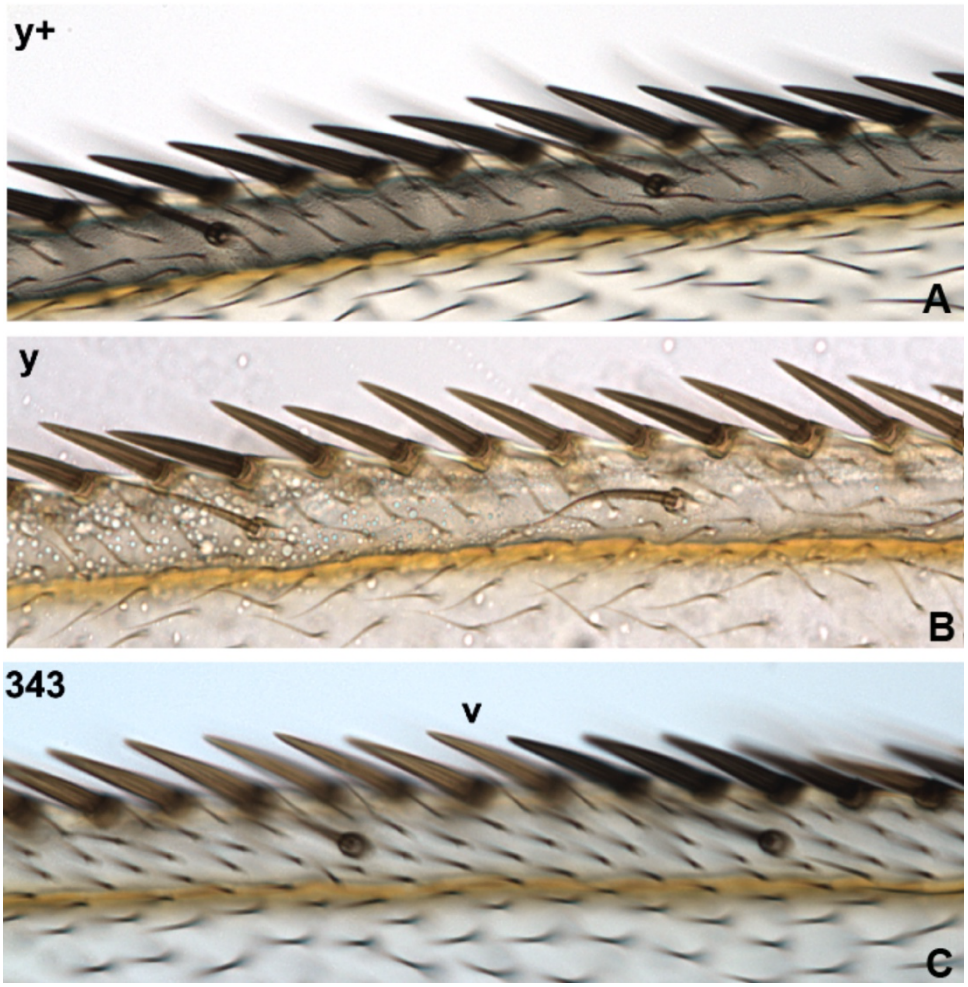


**Figure 6: Heat Shock Clones of Fourth Chromosome Candidates Show FRT is Functional.** (A) is an image of a y+ wing with dark brown bristles along its edge. (B) is an image of a y- (mutant, yellow) wing with blond bristles along its edge. (C) is an image of a y- clone (arrowhead) created in the heterozygous y+ background through recombination on the fourth chromosome. The FRT recombination leading to clones of cells with a y- phenotype is shown in Figure 2. The clone was created in a heterozygous y+ background (PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1] / PB[w+; FRT-CG2316$^{f00836}$]) of line 343.

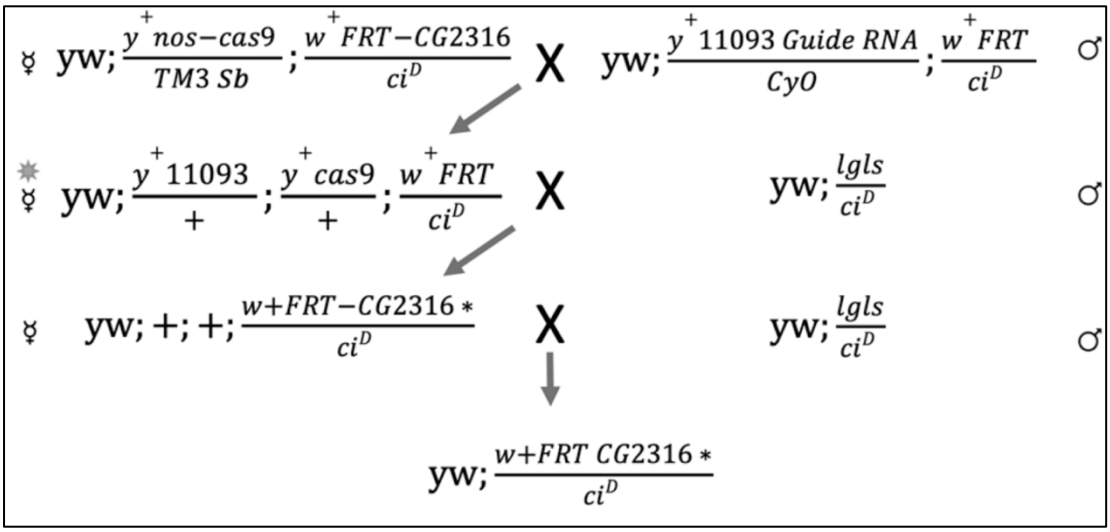


**Figure 7: Scheme to Create dCORL CRISPR Mutations.** The grey star indicates the genotype of the females where CRISPR occurred in oocytes. Progeny of grey star females were crossed to *legless* / $ci^D$ (line three), then their progeny was sib-mated to create stocks (line four). Fifteen stocks were then sent for sequencing to identify lines with mutations in dCORL – four were identified.

| Mutant | Nucleotides Deleted | Result |
|---|---|---|
| A | 9 | 3 amino acid deletion: F-L-P |
| B | 1 | Frameshift: 153 to 183 – Stop |
| J | 5 | Frameshift: 150 to 158 – Stop |
| F | 7 | Frameshift: 151 to 181 – Stop |

**Table 2: Four CRISPR Mutations in dCORL.** *dCORL*$^{A}$ has a three amino acid deletion. *dCORL*$^{B}$ has a single nucleotide deletion resulting in a frameshift starting at amino acid 153 followed by a 30 amino acid nonsense region before a stop. *dCORL*$^{J}$ has a five-nucleotide deletion resulting in a frameshift starting at amino acid 150 followed by an 8 amino acid nonsense region before a stop. *dCORL*$^{F}$ has a seven-nucleotide deletion resulting in a frameshift starting at amino acid 151 followed by a 30 amino acid nonsense region before a stop. Colored arrows correspond to Figure 8, where these mutations are mapped onto the open reading frame of the dCORL.

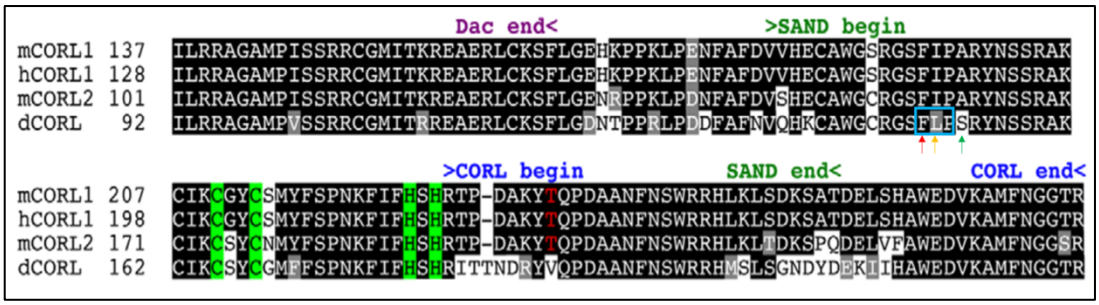


**Figure 8: Four CRISPR Mutations in the dCORL Open Reading Frame**. Image of the CORL Sno homology domain with mouse CORL1, human CORL1, mouse CORL2, and dCORL aligned (Takaesu et al., 2012). Black residues are identical, gray residues are similar and green residues represent the Cys2-His2 zinc finger. Dac, SAND and CORL are conserved motifs found in the Sno homology domain. The blue box represents the three amino acid deletion in *dCORL*$^{A}$. The green arrow indicates the *dCORL*$^{B}$ frameshift at 153. The red arrow indicates the *dCORL*$^{J}$ frameshift at 150. The yellow arrow indicates the *dCORL*$^{F}$ frameshift at 151.

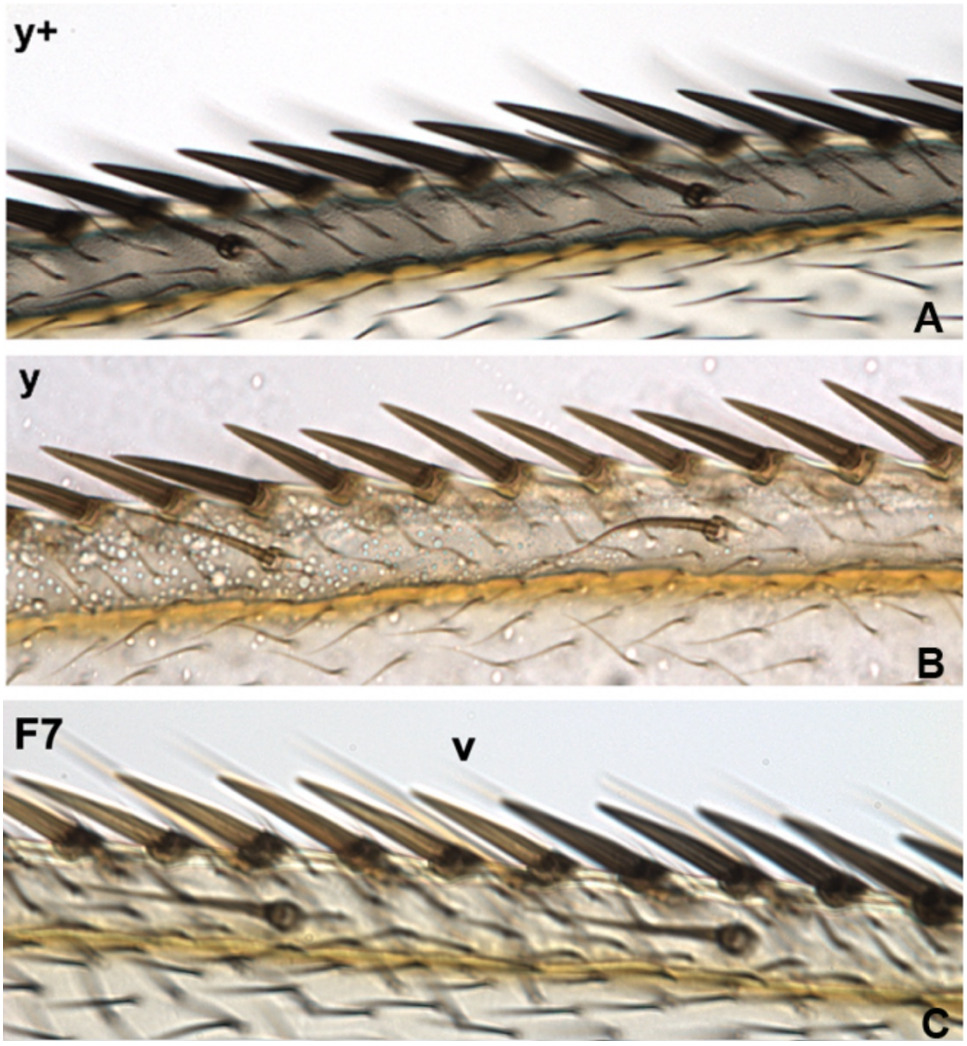


**Figure 9: Heat Shock Clones of dCORL CRISPR Mutant F7.** (A) is an image of a y+ wing with dark brown bristles along its edge. (B) is an image of a y- (mutant, yellow) wing with blond bristles along its edge. (C) is an image of a y- clone (arrowhead) created through recombination on the fourth chromosome. The clone displayed was created in a heterozygous y+ background of *dCORL*$^{F}$, a 7 nucleotide deletion (*y w*; PB[w+; FRT-CG2316$^{f00836}$]-*dCORL*$^{F7}$/ PB[w+; FRT-CG2316$^{f00836}$] P[y$^{+}$; GAL80-MW1]). FRT induced recombination led to clones that did not have a y+ from the P[y$^{+}$; GAL80-MW1] on its fourth chromosome resulting in y- phenotype. Clones and images by Samuel Goldsmith.

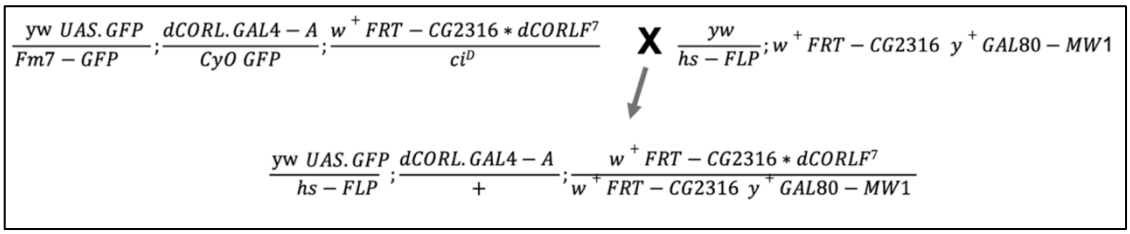


**Figure 10: Scheme to Create a Fly Capable of dCORL Mutant MARCM Clones.** The last cross in a larger scheme that will be conducted to generate the fly containing all components of MARCM for a mutation in dCORL. Progeny shown (line two) will be dissected, and their neuronal clones analyzed. Chromosomes not mentioned in the text are included here for genotype completeness.

# CHAPTER 3

# HUMAN TGF-β FAMILY MEMBERS PRODOMAIN CYSTEINES AND EVIDENCE OF REGULATION FROM CYSTEINE MUTATION PHENOTYPES

## INTRODUCTION

The proteins within the Transforming Growth Factor beta (TGF-β) protein family function within the TGF-β signaling pathway. It is a large family of proteins, identified by their similar structures and the capability to interact with TGF-β receptors. In humans, there are 33 proteins within the TGF-β family and each one is categorized into one of the three subfamilies: TGF-β, Activin, and Bone Morphogenetic Protein (BMP). Proteins are segregated into each of the subfamilies according to structural similarity. Eight of the 33 human TGF-β family proteins belong to the TGF-β subfamily, eight belong to the Activin subfamily, and 17 belong to the BMP subfamily. This project aims to identify new interactions between human TGF-β proteins within the secreting cell, prior to secretion.

TGF-β family proteins have many proposed mechanisms of regulation. One such regulation mechanism is intrinsic to the folding pattern of the three domains found in TGF-β family proteins: the amino terminus, the prodomain (around 250 amino acid residues), and the carboxy terminal ligand domain (around 110 amino acid residues). TGFB1 belongs to the TGF-β subfamily and is one of the most studied TGF-β family proteins.

Described below are the steps involved in the folding mechanism intrinsic to TGF-β family proteins, with TGFB1 as the example protein. The amino terminus contains a signal sequence that is cleaved from the rest of the protein in the endoplasmic reticulum (ER). Then, while in the ER TGFB1 forms a dimer: either a homodimer comprised of two TGFB1 pro-peptide chains or a heterodimer comprised of one TGFB1 pro-peptide chain and one other TGF-β family protein's pro-peptide chain. In the formation of the dimer, the two TFGB ligands are connected through one disulfide bond and the two prodomains are

connected via two disulfide bonds. Once the dimer has formed, it can move to the trans-Golgi where the two prodomains are cleaved from the TGFB1 pro-ligand dimer by furin proteases. After being cleaved from one another the prodomain dimer is now referred to as the Latency-Associated Peptide (LAP) and the TGFB1 pro-ligand dimer is now referred to as the mature TGFB1 dimer. The LAP remains bound non-covalently to the mature TGFB1 dimer, Figure 11A shows this complex, referred to as the small latent complex (SLC) (Finnson et al., 2013). The SLC then moves from the trans-Golgi and is secreted from the cell. TGF-β family proteins can also be secreted in their active form, as only the mature ligand without the LAP. Though, the secretion of active TGF-β ligands is deleterious, as active TGFB1 secretion has only observed in Camurati-Engelman Disease (Saito et al., 2001; Janssens et al., 2003). While the LAP is attached to the mature TGFB1 dimer it prevents TGFB1 from interacting with its receptor. Thus, LAP acts as an intrinsic regulator of TGF-β family signaling (Robertson and Rifkin, 2016).

TGFB1 is secreted from a cell if one of two complexes: the SLC and the large latency complex (LLC). The LLC includes the mature TGFB1 dimer, and the LAP dimer bound to a binding partner protein through two disulfide bonds. Figure 11B shows that one cysteine from each prodomain and two cysteines from the partner protein contribute to the formation of these disulfide bonds (Finnson et al., 2013). This forms the LLC within the ER before the complex is transported to the trans-Golgi. In the trans-Golgi, the two prodomains (still bound to the partner protein) are cleaved from the TGFB1 pro-ligand dimer by furin proteases. The LAP remains bound non-covalently to the mature TGFB1 dimer, keeping the LLC intact and it is secreted by the cell (Robertson and Rifkin, 2016).

There are 10 known binding partner proteins of the 33 TGF-β family members: Latent TGF-β Binding Protein 1-4 (LTBP1-4), Fibrillin 1-3 (FBN1-3), Leucine Rich Repeat Containing families 32 and 33 (LRRC32 and LRRC33), and E-Selectin (SELE). LRRC32 is also known as Glycoprotein-a Repetitions Predominant Protein (GARP). For

this project, binding partners refer only to proteins known to bind to a TGF-β family protein through disulfide bonds. Partner proteins when bound to a TGF-β family protein and are proposed to increase the stability of the LLC and the rate of secretion for the LLC. Partner proteins also have a role in the movement of TGF-β family proteins through the extracellular matrix (ECM) of the secreting cell (Robertson and Rifkin, 2016; Rifkin et al., 2022).

The LLC will then be sequestered to the ECM on the outside of the secreting cell where the activation of the mature TGFB1 ligand will occur. Activation is the process of separating the mature TGFB1 dimer from the LLC so that it can interact with its receptor on another cell. Several methods of activation have been observed in vitro including via proteases, traction-mediated activation, extremes of pH, and reactive oxygen species. Activation via proteases and traction-mediated activation are the most discussed mechanisms. Four proteases are known to activate TGFB1 (calpain, cathepsin D, kallikreins, and metalloproteases) by either degrading the LAP or by cleaving LTBP from the ECM causing the latent TGFB1 ligand to subsequently release from the LLC. For traction-mediated activation, tension originating from the traction between the secreting cell and a surrounding cell deforms the LAP in the LLC, releasing the active mature TGFB1 ligand. This activation method is possible when the two cells are connected, via the LLC. When the LLC is bound to the ECM of the secreting cell and to the integrins in the ECM of the surrounding cell (Robertson and Rifkin, 2016).

Mature TGF-β ligands are secreted from a cell in SLC or LLC, attached to a signaling regulator(s), after passing internal regulation points. TGF-β family ligands undergo the usual points of regulation for proteins within the ER and trans-Golgi. Apoptosis of proteins containing a deleterious mutation or that are misfolded as the base regulation for most proteins. Though, some mutations (neutral and deleterious) make it past this regulation checkpoint and lead to disease. For TGF-β ligands, the LAP acts as the

next point of regulation. The LAP prevents the mature TGF-β ligand from binding to its receptor and is present in both complexes. The SLC undergoes these two points of regulation, while the LLC has three. A partner protein binding to the SLC forms the LLC, adding a third layer of regulation to the signaling of TGF-β family proteins. The partner protein regulates the signaling of TGF-β family proteins by sequestering the LLC in the ECM of the secreting cell, further preventing the mature TGF-β ligand from binding to its receptor. The only known exception to this pattern of secretion was observed in Camurati-Engelman Disease, where active TGF-β ligands are secreted (Saito et al., 2001; Janssens et al., 2003). The importance of regulating the availability of mature TGF-β ligands is demonstrated through the one known case of active TGF-β ligand secretion contributing to a rare disease. Illustrating the importance of understanding the regulatory mechanisms of proteins, especially the TGF-β ligand specific forms of regulation. (Robertson and Rifkin, 2016).

The 10 known TGF-β binding partner proteins are grouped into three categories: secreted, transmembrane, and intracellular. LTBP1-4 and FBs 1-3 are secreted. LTBP1-4 and FBN1-3 also group together because they are the only known proteins to contain TGF-β binding domains. TGF-β binding domains contain eight cysteines and is where LTBP1 binds to LAP in the formation of TGFB1's LLC. LTBP1, LTBP3, and LTBP4 all contain a dipeptide insertion only identified within TGF-β binding domains that bind to the LAP of TGF-β (Saharinen and Keski-Oja, 2000). Of the LTBP proteins, LTBP4 has been shown to inefficiently form the complex with the prodomain and LTBP2 has not been shown to bind to the prodomain. LTBP1 and LTBP3 have the strongest binding affinity to the prodomain. LRRC32 and LRRC33 are transmembrane proteins found in immune cells. SELE acts as an intracellular regulator of latent TGF-β secretion, binding it in the endoplasmic reticulum. LRRC32, LRRC33, and SELE don't contain TGF-β binding domains and where they bind to LAP is currently unknown.

The Association (Assn) region of the prodomain of TGFB1, TGFB2, and TGFB3 have an extrinsic regulation mechanism by binding partner proteins. This regulation is conducted through Cys33 in TGFB1. TGFB1 Cys33 (from each prodomain in LAP dimer) has been experimentally shown to bind biochemically with Cys1359 and Cys1384 in the second TGF-β binding domain of LTBP1 (Figure 11B), and with Cys211 and Cys350 in LRRC32 (Lack et al., 2003; Saharinen and Keski-Oja, 2000; Wang et al., 2012). TGFB2 and TGFB3 are believed to bind in the same manner due to their homology with TGFB1. The specific Assn cysteines equivalent to Cys33 in TGFB1 are Cys24 in TGFB2 and Cys27 in TGFB3.

In humans, 33 TGF-β family proteins have been identified. These 33 proteins are grouped into three subfamilies. Eight proteins belong to the TGF-β subfamily, eight proteins belong to the Activin subfamily, and 17 proteins belong to the BMP subfamily. Of the 33 human TGF-β family proteins only TGFB1, INHBA, and BMP9/GDF2 have the cysteines which facilitate heterodimerization identified. In all three proteins, the identified cysteines were conserved prodomain β8 element cysteines. TGFB1 belongs to the TGF-β subfamily, INHBA belongs to the activin subfamily, and BMP9 is also referred to as GDF2 and belongs to the BMP subfamily. In TGFB1, the β8 element cysteines at positions 223 and 225 facilitate heterodimerization. In INHBA, the β8 element cysteines at positions 244 and 247 facilitate heterodimerization. In GDF2, the β8 element cysteines at positions 156 and 237 facilitate heterodimerization (Shi et al., 2011; Wang et al., 2016; Mi et al., 2015). TGF-β family members form binding partners through disulfide bonds, making other TGF-β family members potential partners for β8 cysteines. All six β8 cysteines from TGFB1, INHBA, and GDF2 are listed with the other identified conserved cysteine residues in Table 5.

Crystal structures determine and clarify the quaternary folding of proteins. Knowing the quaternary folding pattern of proteins aids the understanding of their potential

interactions. Only TGFB1, INHBA, and GDF2 have a known crystal structure, out of the 33 human TGF-β family proteins; Figure 12. INHBA is also known as ActivinA while GDF2 is also known as BMP9. Despite not having the crystal structure of the 30 remaining human TGF-β family proteins, the three proteins with known crystal structures can add understanding for the structures of the three TGF-β protein subfamilies as TGFB1, INHBA, and GDF2 each belong to a different subfamilies. The β8 element near the prodomain carboxy terminus is part of a protruding loop in all three subfamilies. Figure 12 displays this protruding loop is composed of two β-sheets (β8 and β9) in the crystal structure of TGFB1. While the conserved cysteine loop in the other two subfamilies does not contain β-sheets but remains at the surface, Figure 12. The Assn region in all three subfamilies is at the amino terminus and extends beyond their structural cores. The exposed position of the β8 element and Assn region in all three crystal structures show at least one conserved cysteine at a location. All three crystal structures show conserved cysteines within the β8 element and Assn region that are exposed. Meaning that these conserved cysteines could interact with another protein, forming a dimer (β8 element) or binding with a partner protein (Assn region).

A previous study (Wisotzkey and Newfeld, 2020) created an amino acid alignment of the prodomain for 44 TGF-β family proteins and the outgroup GDNF, with a total of 45 sequences across three species. This alignment focused on 12 different structural protein regions and highlighted conservation between the TGF-β family proteins (Mi et al., 2015). Of the 44 TGF-β family proteins included in the alignment, 33 were mouse sequences. Thirteen of the 33 mouse TGF-β family proteins identified conserved cysteines, in either the β8 element or the Assn structural protein region. The β8 element and Assn region of the TGF-β family proteins are highly exposed, Figure 12. The open conformation allows cysteines in the β8 element form dimers between TGF-β proteins via disulfide bonds and cysteines in the Assn region to form disulfide bonds with binding partner proteins.

The studies prodomain alignment does not include one conserved cysteine, at position four in β8 of TGFB2, due to the gap placement in β8. Though this cysteine is present in a supplementary figure in the paper (Wisotzkey and Newfeld, 2020). Once specific cysteines in the prodomain alignment of 44 TGF-β family were identified it was suggested that these residues indicate the widespread use of heterodimers, compared to their current suggested use. It was also suggested that with the identification of these conserved cysteines in the β8 element and Assn region, their potential binding partners and heterodimer partners could be predicted (Wisotzkey and Newfeld, 2020).

Most published papers looking at conserved cysteines focus on the ligand domain. Mice share 85% of their DNA with humans and have a common ancestor 80 million years ago. This relationship is why mouse proteins are commonly compared to human proteins in evolutionary studies to identify homologues, orthologues and estimate potential common functionality of new genes. Thirteen mouse TGF-β family proteins showing conserved cysteine residues in their prodomain β8 element and Assn region, indicating these cysteines may also be present in human versions of these TGF-β proteins. To test this hypothesis, an amino acid alignment of the prodomain for the human versions of the mouse TGF-β family proteins was made, Figure 13 and Figure 14. The original version of this alignment included protein sequences from three different species (*Mus musculus*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) and included 45 total sequences (Wisotzkey and Newfeld, 2020). Focusing on only the included mouse sequences left 33 proteins to identify human orthologues for, including the outgroup (GDNF). Of the 33 human TGF-β family proteins aligned eight are in the TGF-β subfamily, eight are in the Activin subfamily, and 17 are in the BMP subfamily proteins. Again, conserved cysteines were identified in the β8 element and Assn region of the prodomain.

In the prodomain alignment a set of three conserved cysteines were identified in the β8 element (at positions one, three and four; C-terminus) which mediate dimerization, and

a set of two conserved cysteines were identified in the Assn region (at position one and position four; N-terminus) which mediate binding to partner proteins. This identification led to the investigation of five databases for disease phenotypes associated with any of the conserved cysteines. Twelve conserved cysteine mutations were identified with mutant phenotypes across 13 proteins. Common disease phenotypes for conserved cysteine mutations in TGF-β prodomains suggested eleven heterodimer pairs. Two identified heterodimer pairs were previously known, and nine have a common tumor phenotype. Four cancer phenotypes shared by conserved cysteine mutations in TGF-β prodomains and cysteine mutations in eight partner binding proteins suggest 23 new regulatory interactions. Results suggest that specific cysteines in the prodomain of TGF-β family proteins are responsible for establishing disulfide dependent regulation with known partner proteins. This expands previous knowledge of only two known disulfide dependent mechanisms being well understood for three TGF-β family members. The overwhelming frequency of tumor phenotypes in my data (91% of identified interactions have a tumor phenotype) adds support the idea that TGF-β family heterodimers are associated with tumor progression. Identification of the specific cysteines involved allows for the development and testing of new mechanistic hypothesis regarding these two regulatory mechanisms. In this manner, my mutant cysteine data suggests new therapeutic targets in numerous cancers where none currently exist.

## MATERIALS AND METHODS

### Identifying Prodomain Sequences and Cleavage Sites

Using supplemental figures from Wisotzkey and Newfeld, 2020 I gathered the mouse sequences that were used in the creation of their 45 sequence prodomain alignment. The original version of this alignment included protein sequences from three different species: five *C. elegans* proteins sequences, seven *D. melanogaster* protein sequences and 33 *M. musculus* protein sequences. With the accession numbers of each mouse sequence, I used BLASTp to identify the longest versions of their human orthologues. The accession number, gene name description, gene name and the TGF-β subfamily each protein belongs to were all copied over, and a human version of their supplementary table was created (Table 3). Table 3 includes 34 sequences, the human versions of the 33 mouse sequences and the human version of the outgroup (GDNF).

After the 34 human proteins were identified through BLASTp, each identified human protein sequence was copied from NCBI into a word document and aligned by hand with its mouse counterpart. From there I used the supplementary Table S2 from Wisotzkey and Newfeld, 2020 to identify the position of the end of the prodomain and the position of the beginning of the ligand domain, in the mouse sequences. With Table S2 as a guide the end of the prodomain and the beginning of the ligand domain for the human orthologues were identified by hand and later verified though BLASTp. The residue of each cleavage site and 10 amino acids on both sides of the cleavage site where the prodomain and ligand are separated are listed in Table 4 for all 34 Human proteins.

### Prodomain Alignment

I created a complete alignment of the 34 human orthologues. After the Human orthologues were identified using BLASTp, I aligned the 34 sequences in Clustal Omega at EMBL-EBI and annotated them in BoxShade3.21 set to 20% cutoff for shading. After

creating the full alignment all 33 human TGF-β family proteins, it was clear that the N-terminus end where the Assn region is located and the C-terminus where the β8 element is located did not align well because of hypervariability and differences in length. To account for the hypervariability seen in the full alignment two smaller alignments were made focusing on the two regions there the mouse alignment showed conserved cysteines. Figure 13 is the alignment of the last 60 residues of the 34 human proteins and includes the β8 element. While, Figure 14 is the alignment of the first 60 amino acid residues of the 34 human proteins starting with Met and includes the Assn region. Both smaller alignments were this time aligned and annotated for conserved cysteine residues by hand.

Start and end positions for the 12 structural regions were retrieved from Wisotzkey and Newfeld, 2020 along with the positions of conserved cysteines within the β8 element and Assn region. The human homologues of previously identified mouse prodomain cysteines and newly identified human conserved prodomain cysteines are listed in Table 5 and Table 6.

**Prodomain and Binding Protein Cysteine Mutants with Disease Phenotypes**

For all 33 human proteins and 10 known binding partners, four online databases were mined for disease mutations and phenotypes: GeneCards (genecards.org); MalaCards (malacards.org); the NCI Genome Data Commons i.e., GDC (portal.gdc.cancer.gov); and Uniprot (uniprot.org). For the 13 proteins included in Table 5, I also mined Ensembl (uswest.ensembl.org) for benign population variants. All five databases combine genomic and proteomic data from various sources while providing a platform that allows users to search through thousands of data points to find data relevant to their interests. GeneCards provides comprehensive information on all annotated and predicted human genes. Data within GeneCards originates from over 150 online sources including clinical, functional, genetic, genomic, proteomic, and transcriptomic information (Safran et al., 2021).

MalaCards is a database of human maladies and their annotations, with data originating from the GeneCards database and 71 other sources (Rappaport et al., 2017). The NCI CDC Data Portal allows researchers to search and download cancer data for analysis. Date within the NCI CDC Data Portal originates from biospecimen, clinical, and genomic data (Grossman et al., 2016). Uniprot contains protein sequence and functional information. Data in Uniprot is uploaded to one of its sub-databases (the UniProt Reference Clusters, the UniProt Knowledgebase, and the UniProt Archive) from researchers around the world after being released from the International Nucleotide Sequence Database Collaboration (INSDC) and its annotations are checked by Uniprot. Once data is uploaded into Uniprot it can be cross referenced with 180 different online databases (UniProt, 2021). Ensembl includes the annotated genes, computed multiple alignments, predicted regulatory function, and collected disease data for vertebrate genomes. Data within Ensembl originates from researchers around the world after being released from the INSDC (Howe et al., 2021).

Data was collected from each database by hand and included both 'to cysteine' and 'from cysteine' mutations. Mutations labeled 'to cysteine' are gain of function, where another amino acid has mutated into a cysteine. Mutations labeled 'from cysteine' are loss of function, where the cysteine have mutated to another amino acid. Each substitution's location and phenotype were recorded. Most mutations linked to a phenotype originated from NCI, as GeneCards and MalaCards did not contain many phenotypes connected to a specific residue.

Table 7 contains the total number of 'to cysteine' and 'from cysteine' mutations identified in the 13 proteins previously known to contain conserved cysteines. Table 8 and Table 9 contain disease phenotypes associated with mutations in conserved cysteines (match with Table 5 and 6). All 'from cysteine' loss of function mutations in the 10 known binding partners were also collected and examined to identify if any shared phenotypes with those of Assn cysteine mutations. Common disease phenotypes from β8 mutations

suggest potential heterodimer pairs. In Table 10 and Table 11, common disease phenotypes from Assn mutations suggest potential binding partners.

## RESULTS

A previous prodomain study, with a similar alignment, focused on 12 different structural regions and highlighted conservation between the TGF-β family proteins in different species (Wisotzkey and Newfeld, 2020). Table 3 includes 34 sequences, the human versions of the 33 mouse sequences and the outgroup (GDNF) from the previous paper. After my collection of human sequences was completed, the split between the prodomain and ligand domain was identified in each protein sequence. Table 4 contains the residue position of same cleavage site used when analyzing mouse TGF-β proteins.

### Identification of 61 Conserved Cysteines in 33 Human TGF-β Prodomains

Most published papers looking at conserved cysteines focus on the ligand domain. Of the 33 human TGF-β family proteins in the complete amino acid alignment eight sequences are TGF-β subfamily proteins, eight sequences are Activin subfamily proteins, and 17 sequences are BMP subfamily proteins. The full-length alignment is available upon request. In the complete alignment of all 33 human TGF-β family proteins, the central structural regions and elements were consistently and accurately aligned. Though, the variety of different lengths at the C-terminal end and N-terminal end didn't allow for a clean alignment of the β8 element or Assn region. It is easier for software to accurately align the more central elements and regions, as they are surrounded by several amino acids. The ends of proteins tend to be aligned less accurately because they do not have as many surrounding amino acids for the program to use when making comparisons. To account for the length differences seen in the full alignment two smaller alignments were made

focusing on the β8 element and Assn region Figure 13 and Figure 14. The β8 element and Assn region are two structural components where conserved cysteine residues were identified previously in a TGF-β family protein study that contained mouse proteins (Wisotzkey and Newfeld, 2020).

Figure 13 displays up to the last 60 residues of each of the 33 human proteins and includes the β8 element. The red cysteine residues in Figure 13 highlight that 21 of the 33 proteins have at least one conserved cysteine in the β8 element. In red are cysteine residues at positions one, three and four of the β8 element. Figure 14 displays up to the first 60 residues of each of the 33 human proteins, starting with Met, and includes the Assn region. The red cysteine residues highlight that 26 of the 33 proteins and the outgroup GDNF all have at least one conserved cysteine in the Assn region. In red are cysteines at both position one and position four of the Assn region. The green cysteines in Figure 13 and Figure 14 indicate conserved cysteines that do not directly align with the β8 element or Assn region. During this analysis of the human ortholog amino acid sequences additional conserved cysteines were identified in both Figure 13 and Figure 14. Red conserved cysteines in Figures 13 and 14 are listed in Table 5 and Table 6.

Tables 5-11 utilize a naming convention to distinguish individual conserved cysteines within the 33 TGF-β proteins. The positions of conserved cysteines with the β8 structural domain and Assn region each encompass a sequence of four amino acids. Within the two sections of four amino acids, we identified three different positions of conserved cysteines in the β8 element and two positions of conserved cysteines in the Assn domain. Cys@1 denotes the first conserved cysteine in both the β8 element and the Assn domain (e.g., TGFB1 Cys223 and INHBA Cys244 are both in the β8Cys@1 position while MSTN Cys39 and TGFB1 Cys33 are both in the AssnCys@1 position). Cys@3 is the next downstream conserved cysteine within the β8 element and denotes the third residue in the region (e.g., TGFB1 Cys225 is in the β8Cys@3 position). Cys@4 is the next downstream

conserved cysteine in both the β8 element and the Assn region and denotes the fourth residue in the region. In the β8 element, Cys@4 is the third conserved cysteine position while in the Assn domain it is the second conserved cysteine position (e.g., TGFB2 Cys257 and INHBA Cys247 are at the β8Cys@4 position while MSTN Cys42 and INHBA Cys38 are both at the AssnCys@4 position).

There are 13 mouse TGF-β family proteins previously known to contain conserved cysteines, in either their β8 element or Assn region. These 13 proteins have their human orthologues with updated conserved cysteine positions listed in Table 5. Table 5 includes 37 conserved prodomain cysteine positions: the 30 previously and the seven first identified in Figure 13 and Figure 14 (Wisotzkey and Newfeld, 2020).

Table 5 includes six distinct groups of proteins. These groups were created through the identification of shared conserved cysteines at specific positions in the β8 element and Assn region. In Table 5, the four INHB proteins were separated into two distinct groups to highlight that INHBA and INHBB have an additional conserved cysteine at β8Cys@4 that INHBC and INHBE do not share. INHA is placed above the four INHB proteins in Table 5 because it also has a conserved cysteine at AssnCys@1, AssnCys@4 and at β8Cys@4. Being that INHA shares three of four cysteines of the INHBA/B group and only two of three with the INHBC/E group they were grouped closer in the table.

Of the 33 human TGF-β family proteins aligned in Figure 13 and Figure 14, the 13 TGF-β family proteins with identified conserved cysteines in the prodomain of their mice protein homologues were placed in Table 5. Table 6 includes the remaining 17 human TGF-β family ortholog proteins. Not included in Table 6 are GDF10, BMP7, and GDF15 as they did not display any conserved cysteines in their β8 element or Assn region. The outgroup of GDNF is not included in either Table 5 or Table 6.

Table 6 displays all identified conserved cysteines in the β8 element and Assn region shown in Figure 13 and Figure 14 for the 17 human orthologues not included in

Table 5. The 13 proteins in Table 5 were initially separated from Table 6 to highlight the difference in the number of conserved cysteine's present in this group of proteins: between two and four per protein (Table 5) compared to between one and two per protein (Table 6). Most of the conserved cysteines listed in Table 6 are present in the alignments of Figure 13 and Figure 14 with one exception: GDF3. GDF3 does have a conserved cysteine at position 71 believed to be its Assn cysteine, however because the alignment of Figure 13 only went to amino acid 60 it is not in the alignment. GDF1 and GDF2 have their conserved cysteines color coded in green because they do not directly align with the rest of the conserved cysteines in the Assn region.

Unlike the 13 proteins in Table 5, the 17 proteins in Table 6 have fewer conserved cysteines in their β8 element and Assn region. Protein groups in Table 6 are again color coded according to what conserved cysteine positions they have in common. Table 6 includes seven distinct groups of proteins. These groups were created through the identification of shared conserved cysteines at specific positions in the β8 element and Assn region, visualized in Figure 13 and Figure 14.

**Implications of Location and Origin of Conserved Cysteine Substitutions**

Table 7 displays the 13 human TGF-β family proteins previously shown to contain conserved cysteines in their prodomain. All the found cysteines with a substitution in the two groups of data (mutant and variant). Included is the complete number of mutations to a cysteine and from a cysteine. Then this number is broken down into residues in the prodomain and in the ligand domain. Within both the prodomain and the ligand domain it shows the number of mutations to and from a cysteine.

The seven orange numbers in Table 5 correspond with the seven conserved cysteines in the β8 element or Assn region shown to mutate from a cysteine in Table 7. All mutations found originated from the mutant data. Table 7 is restricted to the proteins

included in Table 5. Table 5 includes proteins in all three subfamilies with nearly double the number of conserved cysteines: Table 5 has an average of 2.85 conserved cysteines per protein and Table 6 has an average of 1.53 conserved cysteines per protein. The inclusion on all three subfamilies and the higher number of cysteines means that if there is a trend to be identified in the data it should be identified in the 13 proteins from Table 5. It was deemed the most efficient approach to only focus on the table with the higher number of cysteines (Table 7).

The data in Table 7 originates from four databases. These four databases then created two groups of data: Variant (data collected from Ensembl, with tissue samples from living people to show variants within the living population), Mutation data (collected from gdc.org, a database of sequenced tumors; GeneCards and MalaCards, specific mutations a list of diseases associated with specific proteins).

Here the concept that common phenotypes originating from common mutations within a species indicate a common function by biochemical interaction is applied to the 13 proteins listed in Table 5. Within the data for the 13 proteins four functional criteria were analyzed. Each analysis includes an experimental group, a control group and a prediction based on the hypothesis that each conserved cysteine in the prodomain has an essential function.

The prodomain is known to functionally act as a chaperone for the ligand domain. It is the ligand domain that bonds to TGF-β protein receptors. TGF-β family protein's ability to bind with its receptor is an important function in part regulated by the prodomain. Being that it is the ligand domain which binds to protein receptors, there is a functional prediction that if mutations are to occur in a protein that they would occur in the prodomain at higher rates than in the ligand domain. This is due to the combination of two factors: 1) that the ligand domain is known to be functional, while the prodomain is thought of as acting mainly as a chaperone and 2) that the prodomain is twice as long as the ligand

domain. To account for this, the rate of mutation in prodomain cysteines compared to ligand domain cysteines was analyzed.

Analysis was limited to the proteins in Table 7 displaying at least a two-fold difference in the number of mutations between the prodomain and ligand domain, to account for the two-fold larger size of the prodomain. This prediction was not supported by the data shown in Table 7, as four proteins have more mutations in the ligand domain than in the prodomain compared to three proteins with more mutations in the prodomain than the ligand domain. This lower frequency of mutations in the prodomain than expected indicates that the cysteines in the prodomain have essential functions.

The prediction that benign substitutions (variants) will occur more frequently than deleterious substitutions (mutations), was then investigated. Every protein was examined for this prediction as comparisons were made within the two domains. The prediction is supported by the data in Table 7, though the data supports the prediction less for the ligand domain. In the prodomain 10 proteins have more variants than mutants and three proteins have an equal number of variants and mutants (BMP15, NODAL, MSTN). In the ligand domain nine proteins have more variants than mutants, one protein has an equal number of variants and mutants (TGFB2), and three proteins have more mutants than variants (NODAL, INHBB, TGFB3). The lower frequency of mutations in the prodomain than in the ligand domain indicates that the cysteines in the prodomain have essential functions.

‘To C’ denotes a residue that mutated into a cysteine or a gain of function for a cysteine. “From C’ denotes a mutation in a cysteine that becomes another residue or a loss of function for a cysteine. The functional prediction tested was that cysteines will be created more often than they will be lost in both the prodomain and ligand domain. Every protein was examined for this prediction, as comparisons were made within the two domains. This prediction is supported in the prodomain. In the prodomain all 13 proteins have more mutations that lead to the creation of a cysteine than the loss of one and four

proteins have no ‘From C’ mutations. Again, the prediction is supported less in the ligand domain. In the ligand domain six proteins have more cysteines created than lost, two proteins have more cysteines lost than created, five proteins have mixed results between their variant and mutant data, and only one protein has no ‘From C’ mutations. The pattern of more cysteines being gained than lost further indicates that the cysteines present are functional.

The null hypothesis that the cysteines are conserved because they are functional, was then investigated. Under the null hypothesis the identified conserved cysteines are less likely to undergo substitutions than other cysteines in the protein. The functional prediction that prodomain loss of function mutations will occur more frequently in nonconserved cysteines than in conserved cysteines was tested for all proteins in Table 7. The prediction is supported. Fifteen loss of cysteine mutations in non-conserved cysteines were identified across the 13 prodomain while only seven loss of cysteine mutations were identified in conserved cysteines. Twelve of the 15 cysteine mutations in non-conserved cysteines were found in variant data, indicating that they are not immediately deleterious. The conserved cysteine mutations are deleterious as all seven conserved cysteines mutated lead to disease. The lower frequency of mutations in conserved prodomain cysteines and the mutations in those conserved prodomain cysteines leading to disease phenotypes, adds to the evidence that the cysteines in the prodomain have essential functions.

Data indicating that when conserved cysteines mutate, they lead to disease is further evidence that the identified conserved cysteines are functional. Being that mutations occur more often in non-conserved cysteines and those mutations were mostly found in variant data it indicates that only the conserved cysteines in the prodomain are functional.

Seven mutants occur in conserved cysteines. To test the hypothesis that the identified conserved cysteines are functional, I looked for diseases resulting from mutations in the listed conserved cysteines. Common disease phenotypes for conserved

cysteine mutations would expand on regulation methods in the prodomain of TGF-β family proteins by heterodimerization (for β8 cysteines) or binding partners (for Assn cysteines).

**Disease Phenotypes of Mutated Prodomain Conserved Cysteines**

Six of the seven conserved cysteine loss of function mutations listed in Table 7 occur in the β8 element and one mutation occurs in the Assn region. The eight proteins involved in Table 8 include one BMP family protein (BMP15), three Activin family proteins (INHBA, INHBB and INHBE), and four TGF-β family proteins (INHA, TGFB1, TGFB2 and TGFB3). The orange positions in Table 5 correspond with these mutated cysteines. Table 8 includes specific mutations and their phenotypes for the seven conserved cysteine loss of function mutations identified in Table 7. The identified phenotypes are color coded to indicate potential heterodimer pairs.

In Table 8 three disease phenotypes are highlighted in varying shades of grey, these phenotypes are ungrouped. Ungrouped disease phenotypes indicate no newly identified heterodimer pairs. The three disease types originate from cysteine mutations in BMP15, TGFB1 and TGFB2 respectively. The single mutation in BMP15 β8Cys@4 was identified in bronchus and lung squamous cell neoplasms. The five mutations listed originating in β8Cys@1 and β8Cys@3 of TGFB1 were all identified in Camurati-Engelmann Disease. Camurati-Engelmann Disease is an autosomal dominant disease characterized by skeletal hyperplasia. The mutation in TGFB2 at β8Cys@4 was identified in two different types of disease, depending on the mutation. When the cysteine mutation at β8Cys@4 mutates into a phenylalanine (F) residue it led to Holt-Oram Syndrome. Holt-Oram Syndrome is an autosomal dominant disease with a proximate cause of nonfunctional TBX5 (Boogerd et al., 2010) and is characterized by skeletal abnormalities and heart defects. The second disease that the mutation in TGFB2 at β8Cys@4 was identified in clusters with the pink group discussion below.

The blue group is comprised of a single conserved cysteine loss of function mutation in INHBA β8Cys@1 was identified in lung adenocarcinomas. This mutation clusters with two ligand cysteine loss of function mutations found in TGFB2, were identified in the same type of cancer. Indicating that these two proteins may bind together and form a heterodimer pair. Identifying a common phenotype originating from a ligand cysteine mutation and a conserved prodomain cysteine mutation, provides confidence that the mutated prodomain cysteine is essential for one of the ligand's primary roles. Despite the BMP15 and INHBA cancers both occurring in the lung, their respective disease phenotypes originate in distinct cell types. The presence of different cancers in the lung limits the relevance of these mutations to potential heterodimerization and is why the BMP15 mutation remains ungrouped.

The green group is comprised of a single loss of function mutation in INHBE AssnCys@4 was identified in plasma cell tumors. This mutation clusters with a gain of function mutation to a prodomain cysteine found in INHBB, was identified in the same type of cancer. Indicating that these two proteins may heterodimerize.

The orange group is comprised of a loss of function mutation in TGFB2 at β8Cys@3 was identified in endometrial adenocarcinomas. This mutation clusters with a gain of function mutation in the prodomain of INHBB and a ligand loss of function cysteine mutation in INHA. Both mutations were also identified in endometrial adenocarcinomas. Indicating two new heterodimer pairs, as INHA and INHBB are known to heterodimerize (Walton et al., 2009). Identifying a previously known heterodimer pair bolsters confidence in the identification of potential heterodimer through common disease phenotypes and cysteine mutations.

The pink group is comprised of the loss of function cysteine mutation in TGFB2 at β8Cys@4 which mutates into an early stop codon and was identified in Loeys-Dietz Syndrome (pink in Table 8). Loeys-Dietz Syndrome is an autosomal dominant disease with

systemic effects on connective tissue and blood vessels. This mutation clusters with a ligand loss of function mutation in TGFB3, was also identified in Loeys-Dietz Syndrome. Identification of the common phenotype indicates TGFB2 and TGFB3 heterodimerization. The method of identification of potential heterodimer pairs through common disease phenotypes and cysteine mutations is further validated by this as TGFB2 and TGFB3 are known to heterodimerize (Cheifetz et al., 1988). The β8Cys@4 mutation in TGFB2 suggests a β8 based mechanism for forming a TGFB2 and TGFB3 heterodimer.

Two of the six heterodimer pairs identified were previously known and five were identified in tumors. Of the five identified heterodimer pairs found in tumors, four occur in adenocarcinomas.

After identifying diseases associated with mutations in conserved cysteines of the mostly TGF-β and Activin family proteins in Table 5, the mostly BMP family proteins from Table 6 were then analyzed at for diseases associated with mutations in conserved cysteines for completeness. Of the data listed in Table 6 the BMP subfamily has four prodomain cysteine mutations and the Activin subfamily has one. The five red numbers in Table 6 correspond to the conserved cysteines with disease associations listed in Table 9. In Table 9 the color coding indicates suggested heterodimerization pairs (for β8 cysteines) or binding partners (for Assn cysteines).

The identification of two known heterodimer pairs in Table 8 validated the method of identification of potential heterodimer pairs and binding partners through common disease phenotypes and cysteine mutations. This method was then applied to the proteins listed in Table 6. Five of the 17 proteins listed in Table 6 had mutations in conserved prodomain cysteines leading to four common phenotypes, shown in Table 9.

There is one grey phenotype which is ungrouped. The loss of function cysteine mutation at β8Cys@4 in GDF1 led to a deletion of 145 residues and was identified in a heart defect. The size of the deletion makes this result difficult to interpret. I cannot say if

it is the mutation of the conserved cysteine alone or something else missing in this deletion that contributes to phenotype of Inherited right atrial isomerism.

The yellow group is comprised of the loss of function mutation in BMP3 at AssnCys@1 identified in colon adenocarcinomas. This mutation clusters with a gain of function ligand mutation in GDF5, a loss of function mutation in β8Cys@4 of GDF6 and a loss of function ligand cysteine mutation of GDF6, all of which were also identified in colon adenocarcinomas. Identifying a common phenotype originating from multiple ligand cysteine mutations and multiple conserved prodomain cysteine mutations, provides confidence that the mutated prodomain cysteine is essential for one of the ligand's primary roles. The common phenotype adds evidence to the predicted within subfamily heterodimerization of GDF5 and GDF6 (Wisotzkey and Newfeld, 2020). Having the same phenotype also suggests a cross-subfamily (Activin and BMP) heterodimerization between BMP3 and GDF5, and between BMP3 and GDF6. Colon adenocarcinomas is the one of two phenotypes which indicate potential heterodimer pairs which is caused by conserved cysteine mutations in both the β8 element and Assn region.

The orange group is comprised of a loss of function mutation in BMP10 at AssnCys@1 identified in endometrial adenocarcinomas. This mutation clusters with a loss of function ligand mutation in BMP10, which was also identified in endometrial adenocarcinomas. In addition to the mutations listed in Table 9 the endometrial adenocarcinoma phenotype was also identified in three mutations from Table 8, in INHA, INHBB and TGFB2. The presence of the endometrial adenocarcinoma phenotype in both Table 8 and Table 9 yields five potential cross-subfamily heterodimers and one within subfamily heterodimer pair. The five suggested cross-subfamily heterodimers occur between all three subfamilies. One suggested cross-subfamily heterodimer occurs between the Activin and BMP subfamilies and is comprised of BMP10 and INHBB. Two suggested cross-subfamily heterodimers occur between the Activin and TGF-β subfamilies, and are

comprised of INHA and INHBB, and TGFB2 and INHBB. Two suggested cross-subfamily heterodimers occur between the BMP and TGF-β subfamilies, and are comprised of BMP10 and INHA, and BMP10 and TGFB2. The single one within subfamily heterodimer pair is within the TGF-β subfamily and is comprised of TGFB2 and INHA. BMP10 only has one prodomain conserved cysteine at AssnCys@1, indicating that this cysteine is likely functional in forming heterodimer pairs.

The red group is comprised of a loss of function mutation in AssnCys@1 of GDF5 identified in stomach adenocarcinomas. This mutation clusters with a ligand cysteine gain of function mutation in GDF5, also identified in stomach adenocarcinomas. The shared phenotype indicates the functionality of the cysteine at AssnCys@1 in GDF5.

A total of 12 suggested heterodimer pairs were identified through the comparison of common phenotypes originating from conserved cysteines in Table 8 and Table 9. The number of heterodimer pairs were determined by protein-protein interactions and not individual cysteine mutations. Eleven of the 12 identified heterodimer pairs were found in tumors (adenocarcinomas or plasma cell tumors). Eight of these heterodimer pairs are cross-subfamily and four occur within subfamily. Of the four occurring within subfamilies, two are in the Activin subfamily, one is in the BMP subfamily, and one is in the TGF-β subfamily. Three of the four within subfamily suggested heterodimer pairs are associated with tumors. Two heterodimer pairs identified have previously been demonstrated: INHA with INHBB; and TGFB2 with TGFB3. One of my identified heterodimer pairs was predicted by phylogenetics: Gdf5 and GDF6 (Wisotzkey and Newfeld, 2020).

**Shared Disease Phenotypes of Prodomain Cysteines Mutations and Binding Partner Cysteines**

Twelve potential heterodimer pairs were identified through comparisons of common disease phenotypes caused by mutations in conserved prodomain cysteines. Heterodimer pairs in the TGF-β family form through cysteine-cysteine binding creating disulfide bonds. Common disease phenotypes originating from conserved cysteine mutations indicate the potential function of those cysteines in heterodimerization. There are 10 proteins known to bind to TGF-β family proteins through heterodimerization and binding proteins which regulate these proteins. The 10 known TGF-β binding proteins are LTBP1-4, FBN1-3, LRRC32/33 and SELE. My hypothesis is that a mutated cysteine in a binding protein identified in the same phenotype as a mutated prodomain conserved cysteine in a TGF-β family protein, indicates that the proteins have a shared function mediated by disulfide bonds that is being disrupted, yielding the deleterious phenotype. The following analysis focuses on phenotypes originating from mutations in AssnCys@4, and AssnCys@1. AssnCys@1 in TGFB1 is known to bind to LTBP1 at amino acid 33.

Shared disease phenotypes originating from cysteine mutations in TGF-β proteins and their known binding partners indicate that the cysteine-cysteine bonds used to bind TGF-β proteins and their known binding partners together occur via the conserved cysteines listed in Table 10 and Table 11. This hypothesis is based on the AssnCys@1 at position 33 in the two TGFB1 monomers having previously been identified as the cysteine in the disulfide bond with LTBP1 Cys1359 and Cys1384 in eight-cysteine repeate-3 (Lack et al., 2003); Figure 11B.

In searching for cysteine mutations in the 10 known regulatory binding partner proteins of the TGF-β family which led to plasma cell tumors, 11 cysteine mutations were found across four of the 10 proteins. There is no known binding partner to interact with an Activin domain, Table 10 shows four binding proteins with the same phenotype as a

Activin subfamily member. The phenotype was first identified in a AssnCys@4 mutation in INHBE. The INHBE mutation listed here is the same mutation listed previously in Table 8. In Table 10 this mutation was used to identify common phenotypes originating from cysteine mutations in any of the 10 known TGF-β family binding partners.

The green group is comprised of a loss of function mutation in INHBE at AssnCys@4 identified in plasma cell tumors. This INHBE mutation clusters with a loss of function mutation in LTBP1, two loss of function mutations in FBN1 (Cys1431Tyr and Cys1687Phe), five loss of function mutations in FBN2 and three loss of function mutations in FBN3, all of which were also identified in plasma cell tumors. Indicating that INHBE likely is binding partners with LTPB1, FBN1, FBN2 and FBN3. Evidence for binding partner regulation of INHBE is strongest for FBN2 which has five mutant cysteines each leading to plasma cell tumors followed by FBN3 and FBN1 both of which have three mutant cysteines identified in plasma cell tumors.

The pink group is comprised of two loss of functions mutations in FBN1 (Cys1431Tyr and Cys1431Trp) identified in Loeys-Dietz Syndrome. The FBN1 mutations cluster with a loss of function at β8Cys@4 in TGFB2 and a ligand loss of function mutation in TGFB3, both also identified in Loeys-Dietz Syndrome (Table 8). In FBN1 the Cys1431Tyr mutation is associated with both plasma cell tumors and Loeys-Dietz Syndrome while the Cys1432Trp mutation is only associated with Loeys-Dietz Syndrome (Baetens et al., 2011). The shared phenotypes indicate that FBN1 may act as a binding partner to INHBE (plasma cell tumors) and heterodimerize with TGFB2 and TGFB3 (Loeys-Dietz Syndrome).

Table 11 is comprised of cysteine mutations in partner proteins identified in the three organs where adenocarcinomas were found with mutations at AssnCys@1 in Table 7. The yellow group is comprised of the loss of function mutation in BMP3 at AssnCys@1 identified in colon adenocarcinomas. This mutation is clustered with four loss of function

mutations in FBN1, three loss of function mutations in both FNB2 and FBN3, two loss of function mutations in LTBP1, and one loss of function mutation in SELE, all of which were also identified in colon adenocarcinomas. Evidence for binding partner regulation of BMP3 is strongest for FBN1 which has four mutant cysteines leading to colon adenocarcinomas followed by FBN2 and FBN3 both of which have three mutant cysteines leading to colon adenocarcinomas. the evidence for heterodimerization is weakest with SELE, having only one cysteine mutation identified in the shared colon phenotype.

The orange group is comprised of a loss of function mutation in BMP10 at AssnCys@1 identified in endometrial adenocarcinomas. This mutation is clustered with four loss of function mutations in both LTBP2, four loss of function mutations in FBN1, two loss of function mutations in both FBN2, two loss of function mutations in FBN3, and one loss of function mutation in LTBP1, LTBP3, LRRC32 and SELE, all of which were also identified in endometrial adenocarcinomas. This data indicates potential regulation of BMP10 by eight of the 10 proteins known to bind to TGF-β family members. LTPB2 and FBN1 both have four cysteines associated with the shared phenotype, adding confidence to the phenotypic connection. The evidence for heterodimerization is weakest with LTBP1, LRRC32, and SELE all of which have only one cysteine mutation identified in the shared endometrial adenocarcinoma phenotype.

The red group is comprised of a loss of function mutation in AssnCys@1 of GDF5 identified in stomach adenocarcinomas. This mutation is clustered with four loss of function mutation in FBN1, and one loss of function mutation each in LTBP1, LTBP2, and FBN2, all of which were also identified in stomach adenocarcinomas. The shared phenotype indicates regulation of GDF5 by four of the 10 known binding partners of the TGF-β family. The evidence for regulation by heterodimerization is strongest for FBN1 as it has four cysteine mutations identified in the shared stomach adenocarcinoma phenotype.

Data in Table 10 and Table 11 result from the phenotypic analysis of four conserved

cysteine disease mutations originating from the Assn region. Together this data suggests 22 regulatory interactions with eight of the 10 known binding partners of the TGF-β family, across all three subfamilies. A total of 23 protein-protein interactions were identified across Table 10 and Table 11. Two of these interactions are associated with Loeys-Dietz Syndrome and 21 are associated with tumors. Seventeen of the 21 tumor-specific common phenotypes occur in adenocarcinomas. FBN1 is the only protein associated with Loeys-Dietz Syndrome and is also associated with tumors. LTBP1, FBN1, and FBN2 all displayed the common phenotype found of all four Assn mutations identified which include the Activin and BMP subfamilies. FBN2 the cysteine at 1406 had two mutations, one mutated to a Ser identified in colon adenocarcinomas (Table 11, grouped with BMP3) and one mutated to a phenylalanine identified in plasma cell tumors (Table 10, grouped with INHBE). FBN3 all displayed the common phenotype found in three of the four Assn mutations identified which include the Activin and BMP subfamilies.

LTBP2, LTBP3, LRRC32 and SELE all displayed the common phenotype found only in BMP subfamily proteins. Regulatory mechanisms in the TGF-β family of proteins largely includes heterodimerization with binding partners through disulfide bonds (cysteine-cysteine bonds) to control signaling. TGF-β family members are known to facilitate tumor growth. Having 91.3% of the identified potential interactions were identified in tumors indicated that in these instances partner binding was unable to occur leading to tumors. This suggests that the absence of binding to partner proteins (resulting from prodomain cysteine mutations or partner protein cysteine mutations) could be how TGF-β family members facilitate tumor growth.

Recent advancements in the ability to align TGF-β family proteins has allowed for clean alignments of the C-terminus and N-terminus for all 33 human TGF-β family members. There are three conserved cysteines in the β8 element near the C-terminus, which mediate dimerization. There are also two conserved cysteines in the Association (Assn)

region are near the N-terminus, which mediate partner protein binding. Database mining identified 12 conserved cysteine mutations in 10 proteins and their phenotypes. Common phenotypes for conserved cysteine mutations in TGF-β family proteins suggest eleven heterodimer pairs. Two of the 11 identified heterodimer pairs were previously known and nine have a common tumor phenotype. Conserved cysteine mutations in the Assn region of TGF-β family proteins were connected to eight of 10 partner protein cysteine mutations. Twenty-three regulatory interactions were identified through the identification of four shared cancer phenotypes and Loeys-Dietz Syndrome in eight of the 30 TGF-β family proteins aligned and to eight of 10 known TGF-β family partner proteins. Results suggest that specific cysteines in the prodomain of TGF-β family proteins are responsible for establishing disulfide dependent regulation with known partner proteins. The overwhelming frequency of tumor phenotypes in my data (91% of identified interactions have a tumor phenotype) adds support the idea that TGF-β family heterodimers are associated with tumor progression.

## DISCUSSION

Previously 30 conserved cysteines had been identified across 33 mouse TGF-β family proteins. A total of 61 conserved prodomain cysteines across the 33 human TGF-β family proteins were identified. An additional 31 conserved cysteines in the prodomains of 30 human TGF-β family proteins were identified in the creation of alignments for the C-terminus and N-terminus, Figures 13 and 14. Conserved cysteines were not identified in three human proteins (GDF10, BMP7, and GDF15). The identification of 61 conserved cysteines in the prodomains of 30 human TGF-β family proteins will aid any future research into their folding and function.

In analyzing substitutions, a higher number of substitutions were identified in the variant data than the mutant data in 12 of the 13 proteins listed in Table 7. This pattern is likely because there is a proportion of substitutions that occur and lead to not only a deleterious mutation, but that are fatal. Substitutions should occur at an equal rate at any nucleotide unless a nucleotide is subjected to evolution. The results showed only seven out of 161 substitutions occurred in conserved cysteines – all of which are deleterious mutations with their mutant phenotypes listed.

A well understood process for handling misfolded proteins is to mark them with ubiquitin within the Golgi apparatus for degradation by lysosomes (Foot et al., 2017). Though, this process is well studied it clearly is not perfect. Evidence of this arises from our identification of mutations in many TGF-β proteins from living people. This indicates that either a portion of mutations or a portion of the misfolded proteins are missed by the degradation process.

Tables 5-11 include a naming convention to identify the position of a conserved cysteine within each structural protein region. Two examples are β8Cys@1 and AssnCys@4: indicating the cysteine at the first position of the β8 element and the cysteine in the fourth position of the Assn domain, respectively.

**Disease Phenotypes of Conserved Prodomain Cysteine Mutations Indicate 12 Heterodimer Pairs**

TGFB2 is unique in that it contains two mutated conserved cysteines identified in three different diseases. Thereby TGFB2 is likely involved in two distinct heterodimer pairs. Loeys-Dietz Syndrome likely develops through the previously known heterodimer pair of TGFB2 and TGFB3 being disrupted. My data implies a TGFB2 β8Cys@4 mechanism for heterodimer formation between TGFB2 and TGFB3. Data also implies that TGFB2 β8Cys@3 is capable of heterodimerizing with INHBB, and INHA. TGFB2

potentially capable of forming heterodimer pairs between the TGF-β and Activin subfamilies. Endometrial adenocarcinoma could result from the disruption of the TGFB2-INHBB or the TGFB2-INHA heterodimer. Together the data in Table 8 demonstrates that TGFB2 cysteine mutations have a common phenotype with cysteine mutations in four TGF-β family members: INHA, and TGFB3; and one Activin family member: INHBB.

INHBB has two gain of function mutations identified in two distinct phenotypes. The two gain of function mutations occur at positions 154 and 223, while the conserved cysteines are located at 252 and 255. One explanation for this could be that the creation of a cysteine so close to the two conserved cysteines in the β8 element interfere with the proper function its cysteines. I speculate that the function of one or both β8 cysteines could be disrupted by tertiary protein misfolding or that this new cysteine could be used in place of a β8 cysteine. Either could led to disruption in β8 cysteine function and by extension proper INHBB signaling. As dimerization is required for the proper signaling of TGF-β family proteins, the result of increased INHBB signaling could yield the tumor phenotypes identified.

Patterns of common disease phenotypes listed in Table 8 and Table 9 indicate 12 heterodimer pairs linked to a mutated conserved cysteine. It is reasonable to conclude that the listed mutated conserved cysteine is the cysteine used to create the disulfide bonds between heterodimer pairs. Ten identified heterodimer pairs were in tumors, with nine in adenocarcinomas. Endometrial adenocarcinoma results from five mutations across four proteins in Table 8 and Table 9 suggests one within subfamily heterodimer pair and five potential cross-subfamily heterodimers.

Of the 16 individual mutations in conserved prodomain cysteines found, only four occur in an Assn cysteine. The one AssnC@4 mutation was identified in plasma cell tumors, while the three AssnC@1 mutations were identified in types of adenocarcinomas. Colon adenocarcinomas were found from four mutations in three proteins, suggesting one

within subfamily heterodimer pair and two potential cross-subfamily heterodimers. Both endometrial and colon adenocarcinomas indicate potential heterodimer pairs caused by conserved cysteine mutations in both the β8 element and Assn region. BMP3 and BMP10 both only have one prodomain conserved cysteine at AssnCys@1. Evidence of two mutations in conserved prodomain AssnCys@1 indicates that it functions in the formation of heterodimer pairs. If confirmed by biochemical experiments, this would confirm a unique role for an Assn cysteine - being used to form heterodimers in the Activin and BMP subfamilies.

**Common Disease Phenotypes Indicate 23 Predicted Interactions Between TGF-β Proteins and Known Binding Partners**

There are no previously known binding partners to interact with the Activin subfamily. Shared disease phenotype data indicate that FBN1 and FBN2 may be capable of acting as a binding partner for the Activin subfamily. FBN1 and FBN2 both have at least one mutation identified with all five disease phenotypes. Indicating that one or both proteins are good candidates of regulation by binding partners for nine TGF-β family proteins across all three subfamilies. The following proteins have two connections to FBN1 and FBN2 strengthening the possibility of them being binding partners: GDF5, INHBB, and TGFB2. Previous experiments show that LTBP1 binds with TGFB1 (Yoshinaga et al., 2008), this data is evidence that LTBP1 is likely capable of binding with other TGF-β subfamily members.

Both loss of function mutations in FBN1 were identified in Loeys-Dietz Syndrome, while only Cys1431Tyr mutation was also found in plasma cell tumors (Baetens et al., 2011). FBN1's plasma cell tumor and Loeys-Dietz Syndrome indicate that it may act as a binding partner to INHBE (plasma cell tumors), TGFB2 at β8Cys@4 and TGFB3 (Loeys-

Dietz Syndrome). If true, FBN1 would act as a binding partner to both Activin subfamily and TGF-β subfamily proteins.

The identification of two distinct mutations in Cys1406 of FBN2 indicates that FBN2's EGF-like 22 region as an important binding region for both BMP and Activin proteins. The FBN2 Cys1406Phe mutation was identified in plasma cell tumors and the FBN2 Cys1406Ser mutation was identified in colon adenocarcinomas. This data increases the evidence that FBN2 is capable of being a binding partner to Activin subfamily members.

Across the 47 individual mutations identified in eight known TGF-β family binding partners 38 occur within an EGF-like repeat. EGF-like repeat domains include six cysteines while TGF-β binding domains contain eight cysteines. Only seven mutations occur in TGF-β binding domains, despite their higher volume of cysteines. This is likely due to the difference in the number of EGF-like repeats and TGF-β binding domains in the known binding proteins. Six out of 10 binding partners contain TGF-β binding domains while all 10 known TGF-β family binding partners contain EGF-like repeats. Most EGF-like repeats also act as calcium binding domains. FBN1-3 average 45 EGF-like repeats and only nine TGF-β binding domains. While LTBP1-3 average 17 EGF-like repeats and only four TGF-β binding domains.

**Summary**

Recent advancements in the ability to align TGF-β family proteins have allowed for clean alignments of the C-terminus and N-terminus of 33 family members. Two regulatory mechanisms in the TGF-β family of proteins are heterodimerization via disulfide bonds and binding with binding partners through cysteine-cysteine bonds. There are three conserved cysteines in the β8 element near the C terminal end, which mediate dimerization. Dimerization is required for TGF-β protein secretion and yields the LAP

which keeps the mature ligand from binding to its receptor, regulating TGF-β signaling. There are also two conserved cysteines in the Association (Assn) region that is near the N terminal end, which mediate partner protein binding. When a TGF-β family member is bound to a binding partner it is sequestered to the extra cellular matrix with the LAP in place, additionally regulating TGF-β signaling. Mutant signaling of TGF-β family members are widely associated with facilitate tumor growth.

Database mining identified 12 conserved cysteine mutations in 10 proteins and their phenotypes. Common phenotypes for conserved cysteine mutations in TGF-β family proteins suggest eleven heterodimer pairs. Two of the 11 identified heterodimer pairs were previously known and nine have a common tumor phenotype.

Conserved cysteine mutations in the Assn region of TGF-β family proteins were connected to eight of 10 partner protein cysteine mutations. Twenty-three regulatory interactions were identified through four shared cancer phenotypes and Loeys-Dietz Syndrome in eight of the 33 TGF-β family proteins and eight of 10 known TGF-β family partner proteins. Results suggest that specific cysteines in the prodomain of TGF-β family proteins are responsible for establishing disulfide dependent regulation with known partner proteins.

Of the 10 disease phenotypes identified, tumor phenotypes were overwhelmingly associated with conserved prodomain cysteine mutations, accounting for 91.3% of the identified disease phenotypes. The prevalence of tumor phenotypes adds support to previous findings of mutated TGF-β family proteins being associated with tumor progression. It also suggests that the absence or disruption of disulfide dependent regulation (heterodimerization or binding to partner proteins, resulting from conserved prodomain cysteine mutations or partner protein cysteine mutations) could be how TGF-β family members facilitate tumor growth. This is hypothesized to be the reason why disruption of TGF-β family members binding to their partner proteins via prodomain

cysteine mutations has resulted in tumor phenotypes. As identified disease phenotypes likely result from failed interactions due to cysteine mutations.

One caveat to these findings is that heterodimerization and partner protein binding depend on the predicted proteins being available within the same cell. Unfortunately, exact expression patterns are unknown for all 33 TGF-β family proteins and their 10 known binding partners. An additional complication for these predictions is timing. Some proteins may be expressed in a shared cell type but at different developmental times, which wouldn't allow for the two proteins to act as heterodimers or binding partners. (Guo and Wang, 2009; Dituri at el, 2019).

**Table 3. Human TGF-β Family Sequences, Subfamilies and Accession Numbers.**

| Subfamily | Name | Accession | Description (synonym) |
|---|---|---|---|
| Activin | BMP3 | NP_001192.4 | bone morphogenetic protein 3 preproprotein |
| Activin | GDF10 | NP_004953.1 | growth/differentiation factor 10 preproprotein |
| Activin | GDF11 | NP_005802.1 | growth/differentiation factor 11 preprop. (BMP11) |
| Activin | INHBA | NP_002183.1 | inhibin beta A chain preproprotein |
| Activin | INHBB | NP_002184.2 | inhibin beta B chain preproprotein |
| Activin | INHBC | NP_005529.1 | inhibin beta cysteine chain preproprotein |
| Activin | INHBE | NP_113667.1 | inhibin beta E chain preproprotein |
| Activin | MSTN | NP_005250.1 | growth/differentiation factor 8 prepro. (GDF8) |
| 8 Activin | | | |
| | | | |
| BMP | BMP2 | NP_001191.1 | bone morphogenetic protein 2 preproprotein |
| BMP | BMP4 | NP_001193.2 | bone morphogenetic protein 4 isoform a preprop. |
| BMP | BMP5 | NP_066551.1 | bone morphogenetic protein 5 isoform 1 preprop. |
| BMP | BMP6 | NP_001709.1 | bone morphogenetic protein 6 preproprotein |
| BMP | BMP7 | NP_001710.1 | bone morphogenetic protein 7 preproprotein |
| BMP | BMP8a | NP_861525.2 | bone morphogenetic protein 8A preproprotein |
| BMP | BMP8b | AAP74560.1 | bone morphogenetic protein 8B |
| BMP | BMP10 | NP_055297.1 | bone morphogenetic protein 10 preproprotein |
| BMP | BMP15 | NP_005439.2 | bone morphogenetic protein 15 preprop. (GDF9b) |
| BMP | GDF1 | NP_001483.3 | embryonic growth/differentiation factor 1 precurs. |
| BMP | GDF2 | NP_057288.1 | growth/differentiation factor 2 prepropro. (BMP9) |
| BMP | GDF3 | NP_065685.1 | growth/differentiation factor 3 preproprotein |
| BMP | GDF5 | NP_000548.2 | growth / differentiation factor 5 preproprotein |
| BMP | GDF6 | NP_001001557.1 | growth/differentiation factor 6 preproprotein |
| BMP | GDFf7 | NP_878248.2 | growth/differentiation factor 7 preproprotein |
| BMP | GDF9 | NP_ 005251.1 | growth/differentiation factor 9 preproprotein |
| BMP | NODAL | NP_060525.3 | nodal homolog isoform 1 preproprotein |
| 17 BMP | | | |
| | | | |
| TGF-β | AMH | NP_000470.3 | muellerian inhibiting factor preproprotein (MIS) |
| TGF-β | GDH15 | NP_004855.2 | growth/differentiation factor 15 preproprotein |
| TGF-β | NHA | NP_002182.1 | inhibin alpha chain isoform 1 preproprotein |
| TGF-β | LEFTY1 | NP_066277.1 | left-right determination factor 1 preproprotein |
| TGF-β | LEFTY2 | NP_003231.2 | left-right determination factor 2 isoform 1 preprop. |
| TGF-β | TGFB1 | NP_000651.3 | transforming growth factor beta-1 preproprotein |
| TGF-β | TGFB2 | NP_001129071.1 | tgf beta-2 isoform 1 precursor |
| TGF-β | TGFB3 | NP_001316868.1 | tgf beta-3 isoform 1 preproprotein |
| 8 TGF-β | | | |
| | | | |
| 33 total TGF-β family sequences | | | |
| | | | |
| Outgroup | GDNF | NP_001177397.1 | glial cell line-derived neurotrophic factor isoform 3 |

**Table 4. Cleavage Site Identified Computationally Separating the Prodomain and Ligand Domain**

| Subfamily | Name | Residue | Prodomain | Cleavage Site | Ligand |
|---|---|---|---|---|---|
| Activin | BMP3 | 280 | ALSIER**RKKR** | // | STGVLLPLQN |
| Activin | GDF10 | 331 | LKPRPG**RKDR** | // | RKKGQEVFMA |
| Activin | GDF11 | 289 | VLENTK**RSRR** | // | NLGLDCDEHS |
| Activin | INHBA | 300 | SEDHPH**RRRR** | // | RGLECDGKVN |
| Activin | INHBB | 281 | RLGDSR**HRIR** | // | KRGLECDGRT |
| Activin | INHBC | 226 | RVGGKH**QIHR** | // | RGIDCQGGSR |
| Activin | INHBE | 227 | EPGAGR**ARRR** | // | TPTCEPETPL |
| Activin | MSTN | 257 | VTDTPK**RSRR** | // | DFGLDCDEHS |
| | | | | | |
| BMP | BMP2 | 273 | GHPLHK**REKR** | // | QAKHKQRKRL |
| BMP | BMP4 | 283 | ALTRRR**RAKR** | // | SPKHHSQRAR |
| BMP | BMP5 | 307 | ASEVLL**RSVR** | // | AANKRKNQNR |
| BMP | BMP6 | 365 | VSEVHV**RTTR** | // | SASSRRRQQS |
| BMP | BMP7 | 283 | ATEVHF**RSIR** | // | STGSKQRSQN |
| BMP | BMP8a | 254 | ASPSPI**RTPR** | // | AVRPLRRRQP |
| BMP | BMP8b | 257 | SPIRTP**RAVR** | // | PLRRRQPKKS |
| BMP | BMP10 | 306 | IYDSTA**RIRR** | // | NAKGNYCKRT |
| BMP | BMP15 | 258 | ERESLL**RRTR** | // | QADGISAEVT |
| BMP | GDF1 | 244 | GPGGAC**RARR** | // | DAEPVLGGGP |
| BMP | GDF2 | 310 | AGSTLA**RRKR** | // | SAGAGSHCQK |
| BMP | GDF3 | 241 | DQCHPS**RKRR** | // | AAIPVPKLSC |
| BMP | GDF5 | 371 | EYLFSQ**RRKR** | // | RAPLATRQGK |
| BMP | GDF6 | 344 | KRHGKK**SRLR** | // | CSKKPLHVNF |
| BMP | GDF7 | 337 | AGRGHG**RRGR** | // | SRCSRKPLHV |
| BMP | GDF9 | 310 | GRSSHH**RHRR** | // | GQETVSSELK |
| BMP | NODAL | 228 | SWEWGK**RHRR** | // | HHLPDRSQLC |
| | | | | | |
| TGF-β | AMH | 442 | DPRGPG**RAQR** | // | SAGATAADGP |
| TGF-β | GDF15 | 184 | LRPQAA**RGRR** | // | RARARNGDHC |
| TGF-β | INHA | 223 | PPSGGE**RARR** | // | STPLMSWPWS |
| TGF-β | LEFTY1 | 211 | LASGAH**KLVR** | // | FASQGAPAGL |
| TGF-β | LEFTY2 | 211 | LASGAH**KLVR** | // | FASQGAPAGL |
| TGF-β | TGFB1 | 269 | QHLQSS**RHRR** | // | ALDTNYCFSS |
| TGF-β | TGFB2 | 321 | SQQTNR**RKKR** | // | ALDAAYCFRN |
| TGF-β | TGFB3 | 291 | PGQGGQ**RKKR** | // | ALDTNYCFRN |
| | | | | | |
| Outgroup | GDNF | 99 | QMAVLP**RRER** | // | NRQAAAANPE |

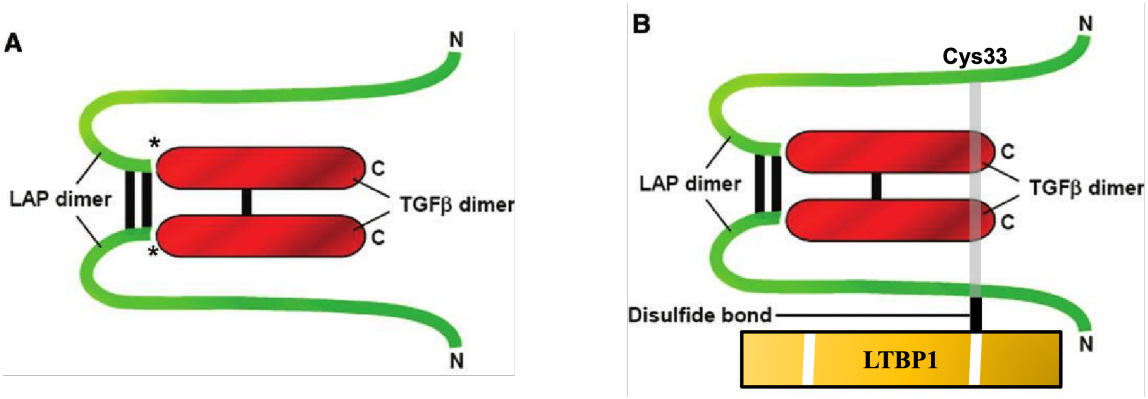


**Figure 11. Latent and Active Forms of the TGFB1 Ligand.** TGFB1 is the example for displaying the regulatory processes TGF-β ligands undergo within the cell prior to secretion. TGFB1 is synthesized as a homodimeric proprotein, called pro-TGFB1. The two monomers within pro-TGFB1 are bound together by two disulfide bonds (black bars) between the two prodomains (green) and one disulfide bond between the two ligand domains (red). [A] Small Latent Complex (SLC): In the trans-Golgi, the two prodomains are cleaved by the furin proteases (* asterisks indicate cleavage site). The prodomain dimer is now called the Latency-Associated Peptide (LAP; green). The cleavage of LAP yields the mature TGFB1 dimer (red). LAP remains bound non-covalently to the mature TGFB1 dimer, forming the SLC. [B] Large Latency Complex (LLC): The SLC bound to a binding partner protein via disulfide bonds forms the LLC. TGF-β family proteins have 10 known binding partner proteins, one of which is Latent TGF-β Binding Protein 1 (LTBP1; yellow). Cys33 in each prodomain of TGFB1's LAP forms a disulfide bond with Cys1359 and Cys1384 in LTBP1's second TGF-β binding domain (white). The disulfide bond between the bottom prodomain of the LAP and LTBP1 is represented by a black bar and the disulfide bond between the top prodomain of the LAP and LTBP1 is represented by the grey bar. Figure was modified from Finnson et al., 2013.

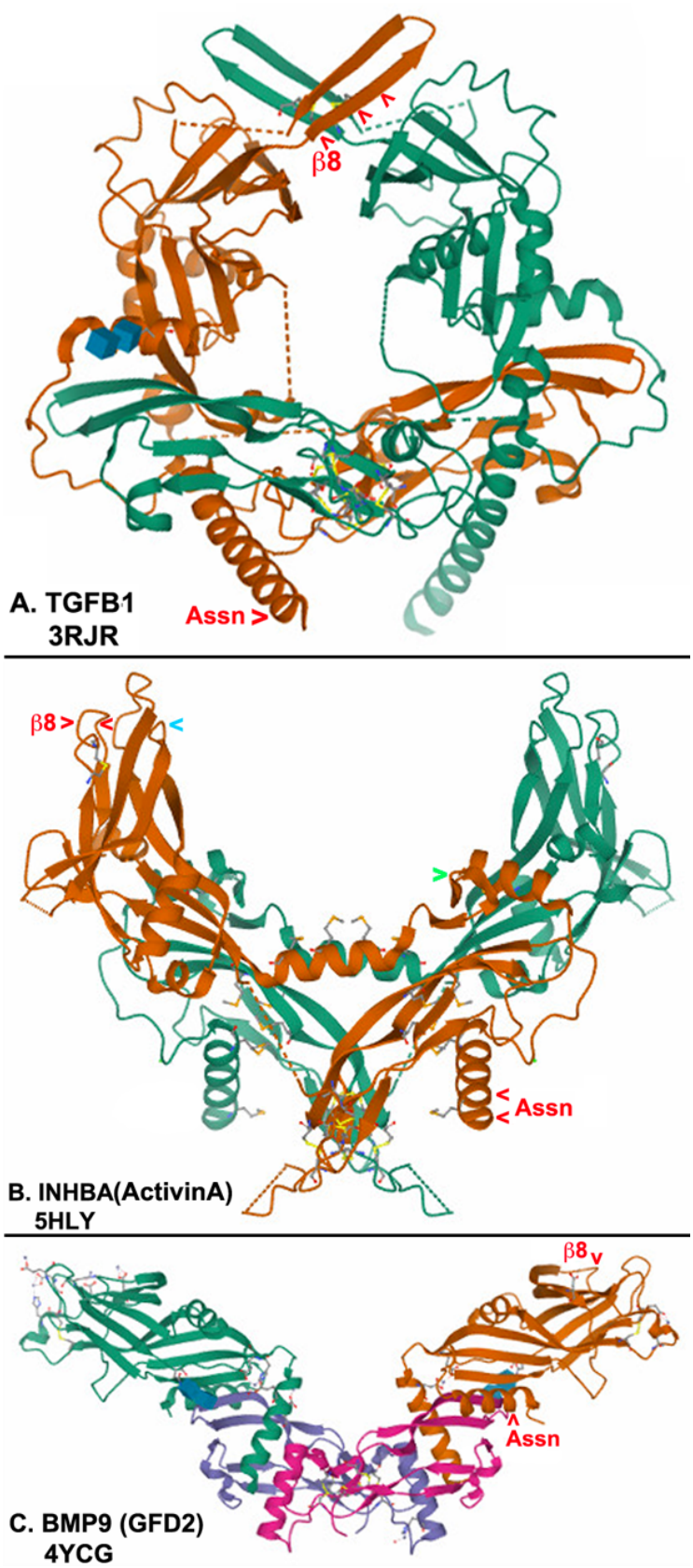
β8
Assn
A. TGFB1
3RJR
β8
Assn
B. INHBA(ActivinA)
5HLY
β8
Assn
C. BMP9 (GFD2)
4YCG

**Figure 12. Crystal Structures of TGFB1, INHBA and GDF2.** The structures of three TGF-β family proteins display conserved prodomain cysteines in two exposed structural regions: β8 element and Association region. [A] TGFB1: Crystal structure of the porcine prodomain dimer with monomers in green and brown. Red arrowheads indicate four conserved cysteines in the brown monomer seen in various combinations in seven of the eight TGF-β subfamily members. The β8 element has three cysteines exposed near the carboxy terminal cleavage site. The Assn region has two cysteines exposed near the amino terminus. The dimer has a closed-ring structure. Image created in Mol* (www.rcsb.org/3d-view/3RJR). [B] INHBA: Crystal structure of the human prodomain dimer. INHBA is also known as ActivinA. Red arrowheads indicate four conserved cysteines in the brown monomer seen in various combinations in seven of the eight Activin subfamily members. The β8 element and Assn region each have two cysteines in exposed locations. The blue arrowhead indicates IHNBB Arg223Cys found in endometrial tumors also containing a β8 element cysteine mutation in TGFB2. The green arrowhead indicates IHNBB Ser154Cys found in plasma cell tumors containing an Assn region cysteine mutation in INHBE. The dimer has an open-arm conformation (www.rcsb.org/3d-view/5HLY). [C] GDF2: Crystal structure of the mouse complete dimer with the prodomains in green and brown. GDF2 is also known as BMP9. Red arrowheads indicate two conserved cysteines in the brown monomer seen in various combinations in 16 of the 17 BMP subfamily members. The β8 element and Assn region each have one cysteine in exposed locations. The dimer has a widely-open conformation (www.rcsb.org/3d-view/4YCG).


**Figure 13: Human Prodomain β8 Element Refined Alignment.** 60 residue alignment begins upstream of β6 and ends downstream of β9 with structural features noted along the top. Red cysteines in the β8 element are conserved in 21 of the 33 human TGF-β family proteins. Green cysteines are additional residues found in INHA and GDF3 that do not appear to be conserved. In 12 TGF-β family proteins there are no cysteines between β6 and the cleavage site. These are: Activin subfamily - BMP3, GDF10, GDF11 and MSTN; BMP subfamily - BMP2, BMP4, BMP6, BMP7 and BMP10; and TGF-β subfamily - GDF15, LEFTY1 and LEFTY2. The blue highlighted R in IHNBB indicates Arg223Cys that causes plasma cell tumors, the same tumors caused by Cys29Tyr, an AssnCys@4 mutation in INHBE. Numbering is accurate and indicates the last residue on each line. The cleavage site is downstream of the C-terminal amino acids included in this alignment. The amino acid sequence and residue for the cleavage site for all 34 proteins is listed in Table 4.


**Figure 14: Human Prodomain Association Region Refined Alignment.** 60 residue alignment begins at Met1 for all sequences except GDF11, INHBB and AMH due to long leaders. It contains the Assn region and ends upstream of α1 with structural features noted along the top. Red cysteines are conserved in 26 of the 33 human TGF-β proteins plus GDNF. Green cysteines are additional cysteines in the vicinity in 10 sequences that do not appear to be conserved. Unalignable single cysteines in GDF1 and GDF2 are also green. The single cysteine in GDF3 is too far downstream to show for a total of 29 TGF-β proteins plus GDNF with an Assn region cysteine. In four proteins there are no cysteines between the initiator methionine and α1: BMP7, GDF6, GDF10 and GDF15. Numbering is accurate and indicates the last residue on each line. The cleavage site is downstream of the C-terminal amino acids included in this alignment. The amino acid sequence and residue for the cleavage site for all 34 proteins is listed in Table 4.

**Table 5. 13 Human TGF-β Proteins with Multiple Conserved Prodomain Cysteines.[a]**

| Protein and Subfamily | Region[b], Position[c] and Residue Number | | | | |
|---|---|---|---|---|---|
| Each group shares 1 or more cysteines | AssnCys@1 | AssnCys@4 | β8Cys@1 | β8Cys@3 | β8Cys@4 |
| BMP15 BMP subfamily | 13 | | | | 209[f] |
| GDF9 BMP subfamily | 13 | | | | **239** |
| NODAL BMP subfamily | 5 | | | | **180** |
| MSTN Act subfamily | **39**[g] | **42** | | | |
| GDF11 Act subfamily | **62** | **65** | | | |
| INHA TGF subfamily | **19** | | 200 | | 203 |
| INHBA Act subfamily | **35** | **38** | 244[d] | | **247**[d] |
| INHBB Act subfamily | **55** | **58** | **252** | | **255** |
| INHBC Act subfamily | **26** | **29** | 190 | | |
| INHBE Act subfamily | **26** | 29 | 192 | | |
| TGFB1 TGF subfamily | **33** | | 223[e] | 225[e] | |
| TGFB2 TGF subfamily | **24** | | **254** | 256 | 257 |
| TGFB3 TGF subfamily | **27** | | **227** | **229** | |

a. Homologs of mouse proteins with conserved cysteines (Wisotzkey and Newfeld, 2020).

b. AssnCys@1 binds LTBP1 and β8Cys@1 plus β8Cys@3 link monomers for TGFB1.

c. Cys@1 is the first in a region with Cys@3 and Cys@4 the 3rd and fourth residue

d. INHBA Cys244/Cys247 form an intrachain bond (Wang et al., 2016).

e. TGFB1 Cys223/Cys225 form an interchain bond (Shi et al., 2011).

f. Disease phenotypes of orange cysteines are shown in Table 8.

g. Bold numbers were previously identified conserved prodomain cysteines (Wisotzkey and Newfeld, 2020).

h. The naming convention of region@position is used in Table 5 through Table 11 to identify the position of a conserved cysteine within each structural protein region. Two examples are β8Cys@1 and AssnCys@4: indicating the cysteine at the first position of the β8 element and the cysteine in the fourth position of the Assn region, respectively.

| Table 6. 17 Additional Human TGF-β Proteins with Conserved Cysteines.[a] | | | | | |
|---|---|---|---|---|---|
| Groups share 1 or 2 cysteines | AssnCys@1 | AssnCys@4 | β8Cys@1 | β8Cys@3 | β8 Cys@4 |
| GDF5 BMP subfamily | 23[d] | | | | 310 |
| GDF7 BMP subfamily | 10 | | | | 231 |
| BMP5 BMP subfamily | 18 | | | | 272 |
| BMP8a BMP subfamily | 16 | | | | 205 |
| BMP8b BMP subfamily | 16 | | | | 205 |
| GDF6 BMP subfamily | | | | | 230 |
| BMP2 BMP subfamily | 7 | | | | |
| BMP4 BMP subfamily | 14 | | | | |
| BMP10 BMP subfamily | 9 | | | | |
| Lefty1 TGF sub family | 7 | | | | |
| Lefty2 TGF subfamily | 7 | | | | |
| BMP3 Act subfamily | 16 | | | | |
| BMP6 BMP subfamily | 12 | | | | |
| AMH TGF subfamily | 55 | | 241 | | |
| GDF1 BMP subfamily | 9[c] | | | | 227 |
| GDF2 BMP subfamily | 2[c] | | | | 237[b] |
| GDF3 BMP subfamily | 71[c] | | | | **222**[f] |
| a. Only BMP7, GDF10 (Act subfamily) and GDF15 (TGF subfamily) have no conserved cysteines. | | | | | |
| b. BMP9 (GDF2) Cys156/Cys237 form an intrachain bond (Mi et al., 2015) | | | | | |
| c. Indicates a cysteine in the vicinity that is unable to be aligned with AssnCys@1. | | | | | |
| d. Disease phenotypes of red cysteines are shown in Table 9. | | | | | |
| f. The bold number was a previously identified conserved prodomain cysteine (Wisotzkey and Newfeld 2020). | | | | | |

| Table 7. Cysteine Variants and Mutations in 13 TGF-β Proteins. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Protein and Subfamily Color groups match Table 1 | Substitution Type | Complete Mutation Totals | Prodomain C substitutions | | | Ligand C substitutions | | |
| | | | Total[a] | From C[b,c] not cons (cons) | To C | Total | From C | To C |
| BMP15 BMP subfamily | Variant | 17 | 8 | 1 | 7 | 9 | 3 | 6 |
| | Mutant | 15 | 8 | 0 (1; β8) | 7 | 7 | 2 | 5 |
| GDF9 BMP subfamily | Variant | 19 | 19 | 3 | 16 | 5 | 2 | 3 |
| | Mutant | 11 | 7 | 0 | 7 | 4 | 2 | 2 |
| NODAL BMP subfamily | Variant | 5 | 2 | 0 | 2 | 3 | 0 | 3 |
| | Mutant | 7 | 2 | 0 | 2 | 5 | 1 | 4 |
| MSTN Act subfamily | Variant | 12 | 7 | 1 | 6 | 5 | 2 | 3 |
| | Mutant | 10 | 7 | 2 | 5 | 3 | 2 | 1 |
| GDF11 Act subfamily | Variant | 8 | 5 | 0 | 5 | 3 | 0 | 3 |
| | Mutant | 4 | 2 | 0 | 2 | 2 | 1 | 1 |
| INHA TGF subfamily | Variant | 14 | 7 | 1 | 6 | 7 | 3 | 4 |
| | Mutant | 8 | 2 | 0 | 2 | 6 | 2 | 4 |
| INHBA Act subfamily | Variant | 14 | 8 | 1 | 7 | 6 | 1 | 5 |
| | Mutant | 8 | 3 | 0 (1; β8) | 2 | 5 | 2 | 3 |
| INHBB Act subfamily | Variant | 11 | 10 | 0 | 10 | 1 | 0 | 1 |
| | Mutant | 7 | 5 | 0 | 5 | 2 | 0 | 2 |
| INHBC Act subfamily | Variant | 16 | 5 | 0 | 5 | 11 | 6 | 5 |
| | Mutant | 4 | 1 | 0 | 1 | 3 | 2 | 1 |
| INHBE Act subfamily | Variant | 13 | 7 | 2 | 5 | 6 | 3 | 3 |
| | Mutant | 5 | 2 | 0 (1; Assn) | 1 | 3 | 1 | 2 |
| TGFB1 TGF subfamily | Variant | 15 | 10 | 0 | 10 | 5 | 2 | 3 |
| | Mutant | 10 | 9 | 0 (2; β8) | 7 | 1 | 1 | 0 |
| TGFB2 TGF subfamily | Variant | 16 | 8 | 2 | 6 | 8 | 6 | 2 |
| | Mutant | 15 | 7 | 0 (2: β8) | 5 | 8 | 5 | 3 |
| TGFB3 TGF subfamily | Variant | 19 | 14 | 1 | 13 | 5 | 3 | 2 |
| | Mutant | 9 | 3 | 1 | 2 | 6 | 2 | 4 |

a. Number of cysteines with a substitution; a single cysteine may have multiple substitutions.

b. Loss of cysteine is abbreviated as "From C" and gain of cysteine is abbreviated as "To C".

c. The specific conserved cysteine mutations shown here are in red in Table 5.

| Table 8. Common Disease Phenotypes of Cysteine Mutations Suggest Seven Heterodimer Pairs in the Activin and TGF-β Subfamilies | | | |
|---|---|---|---|
| Protein | Cysteine | Mutation | Disease Phenotype<br>Colors match potential |
| BMP15 | β8Cys@4 | Cys209Gly | Lung squamous cell neoplasms |
| INHBA | β8Cys@1 | Cys244Tyr | Lung adenocarcinomas |
| INHBB | prodomain | Arg223Cys | Endometrial adenocarcinomas |
| | prodomain | Ser154Cys | Plasma cell tumors |
| INHBE | AssnCys@4 | Cys29Tyr | Plasma cell tumors |
| INHA | ligand | Cys291Trp | Endometrial adenocarcinomas |
| TGFB1 | β8Cys@1 | Cys223Arg<br>Cys233Gly<br>Cys223Ser | Camurati-Engelmann Disease |
| | β8Cys@3 | Cys225Arg<br>Cys225Tyr | |
| TGFB2 | β8Cys@3 | Cys256Stop | Endometrial adenocarcinomas |
| | β8Cys@4 | Cys257Phe | Holt-Oram Syndrome |
| | | Cys257Stop | Loeys-Dietz Syndrome |
| | ligand | Cys246Tyr<br>Cys407Ser | Lung adenocarcinomas |
| TGFB3 | ligand | Cys409Tyr | Loeys-Dietz Syndrome |

| Table 9. Common Disease Phenotypes of Five Additional Cysteine Mutations Suggest Five Cross-Subfamily Heterodimer Pairs and One Within-Subfamily for BMP Proteins | | | |
|---|---|---|---|
| Protein | Cysteine | Mutation | Disease Phenotype<br>Colors match potential |
| BMP3 Act subfamily | AssnCys@1 | Cys16Phe | Colon adenocarcinomas |
| BMP10 BMP subfamily | AssnCys@1 | Cys9Tyr | Endometrial adenocarcinomas |
| | ligand | Cys389Tyr | Endometrial adenocarcinomas |
| GDF1 BMP subfamily | β8Cys@4 | Cys277-delet | Inherited right atrial isomerism |
| GDF5 BMP subfamily | AssnCys@1 | Cys23Arg | Stomach adenocarcinomas |
| | ligand | Arg387Cys | Stomach adenocarcinomas |
| | ligand | Arg438Cys | Colon adenocarcinomas |
| GDF6 BMP subfamily | β8Cys@4 | Cys230Arg | Colon adenocarcinomas |
| | ligand | Cys419Tyr | Colon adenocarcinomas |

| Table 10. Common Disease Phenotype for an INHBE Assn Region Cysteine Mutation Suggest Six New Regulatory Binding Partners for Activin and TGF-β Subfamily Members | | | |
|---|---|---|---|
| Protein | Cysteine | Mutation | Disease Phenotype Colors match Table 4 |
| INHBE | AssnCys@4 | Cys29Tyr | plasma cell tumors |
| LTBP1 | EGF-like repeat 7 | Cys1022Tyr | plasma cell tumors |
| Fibrillin1 | EGF-like repeat 24 | Cys1431Tyr | plasma cell tumors |
| | | ------ | |
| | | Cys1431Trp | Loeys-Dietz Syndrome (match TGFB2 and |
| | EGF-like repeat 28 | Cys1687Phe | plasma cell tumors |
| Fibrillin2 | EGF-like repeat 16 | Cys1131Stop | plasma cell tumors |
| | EGF-like repeat 22 | Cys1378Ser Cys1406Phe[a] | plasma cell tumors |
| | TGF-β binding domain 6 | Cys1579Gly Cys1608Tyr | plasma cell tumors |
| Fibrillin3 | EGF-like repeat 2 | Cys252Tyr | plasma cell tumors |
| | EGF-like repeat 20 | Cys1349Ser | plasma cell tumors |
| | TGF-β binding domain 6 | Cys1519Arg | plasma cell tumors |
| a. FBN2 Cys1406 also found in Table 11 | | | |

| Table 11. Common disease phenotypes for three additional Assn region cysteine mutations suggest 17 regulatory interactions for Activin and BMP subfamily | | | |
|---|---|---|---|
| **Protein** | **Cysteine** | **Mutation** | **Disease Phenotype Colors match potential** |
| BMP3 Act subfamily | AssnCys@1 | Cys16Phe | Colon adenocarcinomas |
| LTBP1 | TGF-β bind 1 | Cys559Tyr<br>Cys594Trp | Colon adenocarcinomas |
| FBN1 | EGF-like 2<br>EGF-like 11<br>EGF-like 21<br>EGF-like 30 | Cys119Gly<br>Cys763Phe<br>Cys1307Tyr<br>Cys1847Tyr | Colon adenocarcinomas |
| FBN2 | EGF-like 22<br>EGF-like 31<br>EGF-like 35 | Cys1406Ser[a]<br>Cys1903Tyr<br>Cys2072Arg | Colon adenocarcinomas |
| FBN3 | EGF-like 6<br>EGF-like 11<br>EGF-like 14 | Cys557Tyr<br>Cys885Tyr<br>Cys1096Phe | Colon adenocarcinomas |
| SELE | Sushi 3 | Cys304Arg | Colon adenocarcinomas |
| BMP10 BMP subfamily | AssnCys@1 | Cys9Tyr | Endometrial adenocarcinomas |
| LTBP1 | EGF-like 2 | Cys413Phe | Endometrial adenocarcinomas |
| LTBP2 | EGF-like 3<br>EGF-like 7<br>TGF-β bind 4<br>EGF-like 20 | Cys648Tyr<br>Cys985Tyr<br>Cys1611Tyr<br>Cys1788Arg | Endometrial adenocarcinomas |
| LTBP3 | EGF-like 13 | Cys1282Tyr | Endometrial adenocarcinomas |
| FBN1 | EGF-like 13<br>EGF-like 19<br>EGF-like 35<br>EGF-like 37 | Cys821Tyr<br>Cys1201Try<br>Cys2053Tyr<br>Cys2190Arg | Endometrial adenocarcinomas |
| FBN2 | EGF-like 4<br>EGF-like 23 | Cys280Trp<br>Cys1412Trp | Endometrial adenocarcinomas |
| FBN3 | EGF-like 35<br>EGF-like 36 | Cys2181Phe<br>Cys2225Arg | Endometrial adenocarcinomas |
| LRRC32 | LRR 12 | Cys342Arg | Endometrial adenocarcinomas |
| SELE | Sushi 6 | Cys506Tyr | Endometrial adenocarcinomas |
| GDF5 BMP subfamily | AssnCys@1 | Cys23Arg | Stomach adenocarcinomas |
| LTBP1 | EGF-like 16 | Cys1506Tyr | Stomach adenocarcinomas |
| LTBP3 | EGF-like 3 | Cys595Met | Stomach adenocarcinomas |
| FBN1 | EGF-like 5<br>EGF-like 10<br>TGF-β bind 3<br>EGF-like 15 | Cys299Arg<br>Cys623Tyr<br>Cys683Arg<br>Cys1044Tyr | Stomach adenocarcinomas |
| FBN2 | EGF-like 46 | Cys2658Tyr | Stomach adenocarcinomas |
| a. FBN2 Cys1406 also found in Table 10 | | | |

# CHAPTER 4

# DISCUSSION

The two projects described here utilize a combination of lab-based techniques and computational methods to study cell signaling from a genetics perspective and expand understanding of regulation in the TGF-β pathway.

dCORL functions in the dActivin side of the TGF-β signaling pathway in Drosophila and acts as a regulator of dSmad2 in the larval brain. However, dCORL's role in TGF-β regulation in the adult brain is currently unknown. To address this question MARCM clones are the best method. Previously this method was unusable for dCORL. In Chapter 2, I created two new fourth chromosomes and tested their viability. The Drosophila lines containing both new fourth chromosomes have functional FRT's and yield healthy progeny. MARCM clones are now possible with fourth chromosome genes. The creation of dCORL MARCM clones will expand our knowledge of TGF-β pathway regulation within the receptor cell. As dCORL MARCM clones will allow the identification of dCORL's role in the dActivin pathway in adult Drosophila brains and determine if dCORL acts as a regulator of dILP2 transcription in the adult Drosophila brain.

Then, in Chapter 3, I produced and examined two prodomain alignments of 33 human TGF-β proteins, to identify conserved prodomain cysteines. After the identification of 61 prodomain conserved cysteine's four online databases were then scoured to document any conserved prodomain cysteines displaying shared mutant phenotypes. The presence of shared mutant phenotypes indicates that there is a disruption in the disulfide bonds formed by conserved prodomain cysteines. Ascertaining shared mutant phenotypes within the TGF-β family members, and between TGF-β proteins and binding partner proteins denotes the potential regulatory roles to which these conserved cysteines participate - by participating in heterodimer formation or partner protein binding. If any of the predicted

interactions are biochemically confirmed it will greatly expand our understanding of TGF-β pathway regulation within the secreting cell.

**Future Directions – dCORL MARCM Clones**

The exact function of dCORL in the dActivin pathway is currently unknown. Evidence suggests that dCORL works either with dSmad2 or after dSmad2 and prior to EcR-B1's transcription. dCORL acts as a model of TGF-β regulation. Producing dCORL MARCM clones will allow for the identification of dCORL's function within the dActivin pathway, in dILP2 transcription. The next step is to create dCORL MARCM clones using the two fourth chromosomes I created in Chapter 2.

Based on previous studies, control cocktail brains stained for GFP, dILP2, Drf and FasII will have dCORL mutant MARCM clones that do not express dILP2 or Drf. My hypothesis predicts that experimental cocktail brains stained for GFP, dILP2, Caspase-3 and FasII will have dCORL mutant MARCM clones that do express dILP2 or Caspase-3. An alternative outcome for the experimental cocktail will be the presence of clones that express Caspase-3. This outcome would support an alternative hypothesis that dCORL has a role in cell death.

If the above experiment shows that dCORL mutant MARCM clone neurons do not express dILP2 and Caspase-3, supporting my hypothesis, a potential next step will be to analyze MARCM clones with dILP3 or dILP5 in place of dILP2 (Nässel et al., 2013). This study will ascertain if dCORL's role in dILP2 expression extends to other dILPs.

If the above experiment shows that dCORL mutant MARCM clone neurons express Caspase-3, a potential next step will be to ascertain if the loss of dCORL leads to cell death in another region of the brain (e.g., the Mushroom Body). Previous experiments found that dCORL has a role in EcR-B1 expression in the Mushroom Body, though cell death has not been investigated (Takaesu et al., 2012).

If MARCM results support the cell death hypothesis, they will add evidence to a tenuous link between the TGF-β signaling pathway and apoptosis (Brown et al., 1999; Zhu et al., 2001). In addition, my results will define the function of dCORL in dILP2 expressing neurons and expand knowledge of dCORL's function.

**Future Directions – Confirm Specific Cysteine Involvement in Heterodimer Pairs**

Multiple hypotheses are provided for the specific cysteine's involved in the formation of heterodimer pairs and partner protein binding. Data predicting potential interactions is relevant for the previously known heterodimer pair of TGFB2 and TGFB3. The heterodimer was identified by co-immunoprecipitation and thus did not specify the cysteines involved (Cheifetz et al., 1988). The hypothesis my data suggests is that heterodimer formation occurs through TGFB2 Cys@256 β8Cys@3) and TGFB3's ligand Cys@409. To test this hypothesis, one could conduct mutagenesis experiments in mice via CRISPR mutations for TGFB2 Cys@256 and TGFB3's Cys@409. The positions of both cysteines involved are at the same location in their human and mouse homologues.

The goal of the CRISPR mutagenesis would be to remove the Cys@256 in TGFB2 and to remove the Cys@409 in TGFB3 in transgenic mice. Four co-immunoprecipitation western blots experiments would be performed on blood samples. One with TGFB2 and TGFB3 mouse proteins (both wild type mice). The second with the CRISPR mutated TGFB2 and unmutated TGFB3 mouse proteins (TGFB2 transgenic mouse). The third with the unmutated TGFB2 and the CRISPR mutated TGFB3 mouse proteins (TGFB3 transgenic mouse). The fourth with CRISPR mutated TGFB2 and TGFB3 (heterozygous transgenic mice). For the first co-immunoprecipitation, TGFB2 and TGFB3 should precipitated indicating that they have formed a heterodimer. In the last three TGFB2 and TGFB3 should fail to bind yielding no precipitation.

If a precipitate is formed in any of the mutated cysteine experiments, it would indicate that despite the shared disease phenotype that the specific cysteine deleted is not involved in the heterodimer between TGFB2 and TGFB3. In this case, alternative conserved prodomain cysteines in TGFB2 and TGFB3 could be tested.

**Summary**

For my two projects in this thesis, computational and lab-based techniques were implemented to study cell signaling from a genetics perspective. Specifically, regulation of the TGF-β pathway at two stages. My first project aims to understand the role of a protein (dCORL) within the TGF-β pathway signal cascade of a receptor cell in Drosophila. While my second project aims to identify new interactions between human TGF-β proteins within a secreting cell prior to their secretion.

Despite being unable to complete my first project due to the COVID pandemic, the work completed and detailed in Chapter 2 will allow a Ph.D. student in the lab to complete the dCORL MARCM experiments. Completing this experiment will allow him to determine if dCORL participates in TGF-β regulation upstream of dILP2 transcription. Though, without the need to pivot to a computational project, the 12 potential heterodimer pairs and 23 regulatory interactions identified via shared disease phenotypes from mutated conserved prodomain cysteines may not have been found. If any of these proposed interactions were to be confirmed it would expand the understanding of intra-subfamily dimer formation and the regulation of TGF-β signaling by binding partners.

The TGF-β pathway is present across vertebrates and invertebrates, together my work expends knowledge of regulatory mechanisms within the TGF-β pathway in both humans and Drosophila while providing experiments to expand this knowledge.

# REFERENCES


Baetens, M., Van Laer, L., De Leeneer, K., Hellemans, J., De Schrijver, J., Van De Voorde, H., Renard, M., Dietz, H., Lacro, R. V., Menten, B., Van Criekinge, W., De Backer, J., De Paepe, A., Loeys, B., and Coucke, P. J. (2011). Applying massive parallel sequencing to molecular diagnosis of Marfan and Loeys-Dietz syndromes. *Human mutation*, *32*(9), 1053-1062.

Boogerd, C. J., Dooijes, D., Ilgun, A., Mathijssen, I. B., Hordijk, R., van de Laar, I. M., Rump, P., Veenstra-Knol, H. E., Moorman, A. F. M., Barnett, P., and Postma, A. V. (2010). Functional analysis of novel TBX5 T-box mutations associated with Holt–Oram syndrome. *Cardiovascular research*, *88*(1), 130-139.

Brown, T. L., Patil, S., Cianci, C. D., Morrow, J. S., and Howe, P. H. (1999). Transforming growth factor β induces caspase 3-independent cleavage of αII-spectrin (α-fodrin) coincident with apoptosis. *J Biol Chem.* 274, 23256-23262.

Cheifetz S, Bassols A, Stanley K, Ohta M, Greenberger J, Massagué J. (1988) Heterodimeric TGF-β. Biological properties and interaction with three types of cell surface receptors. J. Biol. Chem. 263:10783-10789.

Chou, T. B. and Perrimon, N. (1996). The autosomal FLP-DFS technique for generating germline mosaics in Drosophila melanogaster. *Genetics*. 144, 1673-1679.

Dituri, F., Cossu, C., Mancarella, S., and Giannelli, G. (2019). The interactivity between TGFβ and BMP signaling in organogenesis, fibrosis, and cancer. *Cells*, *8*(10), 1130.

Finnson, K. W., McLean, S., Di Guglielmo, G. M., and Philip, A. (2013). Dynamics of transforming growth factor beta signaling in wound healing and scarring. *Advances in wound care*, *2*(5), 195-214.

Foot, N., Henshall, T., and Kumar, S. (2017). Ubiquitination and the regulation of membrane proteins. *Physiological reviews*, *97*(1), 253-281.

Germani, F., Bergantinos, C., and Johnston, L. A. (2018). Mosaic analysis in Drosophila. *Genetics*. *208*, 473-490.

Grönke S, Clarke D-F, Broughton S, Andrews TD, and Partridge L. (2010). Molecular Evolution and Functional Characterization of Drosophila Insulin-Like Peptides. *PLoS Genet.* doi:10.1371/journal.pgen.1000857.

Grossman, Robert L., Heath, Allison P., Ferretti, Vincent, Varmus, Harold E., Lowy, Douglas R., Kibbe, Warren A., Staudt, Louis M. (2016). Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine375:12, 1109-1112

Guo, X., and Wang, X. F. (2009). Signaling cross-talk between TGF-β/BMP and other pathways. *Cell research*, *19*(1), 71-88.

Hatkevich, T., Kohl, K., McMahan, S., Hartmann, M., Williams, A., and Sekelsky, J. (2017). Bloom syndrome helicase promotes meiotic crossover patterning and homolog disjunction. *Curr Biol. 27,* 96-102.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., ... and Flicek, P. (2021). Ensembl 2021. *Nucleic acids research*, *49*(D1), D884-D891.

Janssens, K., ten Dijke, P., Ralston, S. H., Bergmann, C., and Van Hul, W. (2003). Transforming growth factor-β1 mutations in Camurati-Engelmann disease lead to increased signaling by altering either activation or secretion of the mutant protein. *Journal of Biological Chemistry*, *278*(9), 7718-7724.

Kahlem, P., and Newfeld, S. J. (2009). Informatics approaches to understanding TGFβ pathway regulation. *Development*, *136*(22), 3729-3740.

Lack, J., O'Leary, J. M., Knott, V., Yuan, X., Rifkin, D. B., Handford, P. A., and Downing, A. K. (2003). Solution structure of the third TB domain from LTBP1 provides insight into assembly of the large latent complex that sequesters latent TGF-β. *Journal of molecular biology*, *334*(2), 281-291.

Lee, T., and Luo, L. (1999). Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. *Neuron*. *22,* 451-461.

Liu, X., Sun, Y., Weinberg, R. A., and Lodish, H. F. (2001). Ski/Sno and TGF-β signaling. *Cytokine Growth Factor Rev. 12,* 1-8.

Locke, J., and McDermid, H. E. (1993). Analysis of Drosophila chromosome 4 using pulsed field gel electrophoresis. *Chromosoma*, *102*(10), 718-723.

Lue, N. F., Chasman, D. I., Buchman, A. R., and Kornberg, R. D. (1987). Interaction of GAL4 and GAL80 gene regulatory proteins in vitro. *Mol Cell Biol*. 7, 3446-3451.

Mi LZ, Brown CT, Gao Y, Tian Y, Le VQ, Walz T, and Springer, T. A. (2015). Structure of bone morphogenetic protein 9 procomplex. *Proceedings of the National Academy of Sciences*, *112*(12), 3710-3715.

Mizuhara E, Nakatani T, Minaki Y, Sakamoto Y, and Ono Y. (2005). Corl1, a Novel Neuronal Lineage-specific Transcriptional Corepressor for the Homeodomain Transcription Factor Lbx1. *J Biol Chem. 280,* 3645-3655.

Nässel, D. R., Kubrak, O. A., Liu, Y., Luo, J., and Lushchak, O. V. (2013). Factors that regulate insulin producing cells and their output in Drosophila. *Front Physiol*. *4,* 252.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat Protoc. 8,* 2281.

Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., ... and Lancet, D. (2017). MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic acids research*, *45*(D1), D877-D887.

Rifkin, D., Sachan, N., Singh, K., Sauber, E., Tellides, G., and Ramirez, F. (2022). The role of LTBPs in TGF beta signaling. *Developmental Dynamics*, *251*(1), 95-104.

Robertson, I. B., and Rifkin, D. B. (2016). Regulation of the bioavailability of TGF-β and TGF-β-related proteins. *Cold Spring Harbor Perspectives in Biology*, *8*(6), a021907.

Safran, M., Rosen, N., Twik, M., BarShir, R., Stein, T. I., Dahary, D., ... and Lancet, D. (2021). The GeneCards Suite. In *Practical Guide to Life Science Databases* (pp. 27-56). Springer, Singapore.

Saharinen J, Keski-Oja J. (2000) Specific sequence motif of 8-Cys repeats of TGF-β binding proteins, LTBPs, creates a hydrophobic interaction surface for binding of small latent TGF-β. Mol. Biol. Cell 11:2691–2704.

Saito, T., Kinoshita, A., Yoshiura, K. I., Makita, Y., Wakui, K., Honke, K., ... and Taniguchi, N. (2001). Domain-specific mutations of a transforming growth factor (TGF)-β1 latency-associated peptide cause Camurati-Engelmann disease because of the formation of a constitutively active form of TGF-β1. *Journal of Biological Chemistry*, *276*(15), 11469-11472

Shi, M., Zhu, J., Wang, R., Chen, X., Mi, L., Walz, T., and Springer, T. A. (2011). Latent TGF-β structure and activation. *Nature*, *474*(7351), 343-349.

Sousa-Neves, R., and Schinaman, J. M. (2012). A novel genetic tool for clonal analysis of fourth chromosome mutations. *Fly*. 6, 49-56.

Stinchfield, M. J., Miyazawa, K., and Newfeld, S. J. (2019). Transgenic analyses in Drosophila reveal that mCORL1 is functionally distinct from mCORL2 and dCORL. *G3 (Bethesda) 9,* 3781-3789.

Takaesu, N. T., Hyman-Walsh, C., Ye, Y., Wisotzkey, R. G., Stinchfield, M. J., O'connor, M. B., ... and Newfeld, S. J. (2006). dSno facilitates Baboon signaling in the Drosophila brain by switching the affinity of Medea away from Mad and toward dSmad2. *Genetics*, *174*(3), 1299-1313.

Takaesu, N. T., Stinchfield, M. J., Shimizu, K., Arase, M., Quijano, J. C., Watabe, T., ... and Newfeld, S. J. (2012). Drosophila CORL is required for Smad2-mediated activation of Ecdysone Receptor expression in the mushroom body. *Development*, *139*(18), 3392-3401.

Tran, N., Goldsmith, S., Dimitriadou, A., Takaesu, N., Consoulas, C., and Newfeld S. (2018b) CORL expression and function in insulin producing neurons reversibly influences adult longevity in Drosophila. *G3 (Bethesda) 8*, 2979-2990.

Tran, N., Takaesu, N., Cornell, E., and Newfeld S. (2018a) CORL expression in the Drosophila central nervous system is regulated by stage specific interactions of intertwined activators and repressors. *G3 (Bethesda) 8*, 2527-2536.

UniProt. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic acids research*, *49*(D1), D884-D891.

Walton, K. L., Makanji, Y., Wilce, M. C., Chan, K. L., Robertson, D. M., and Harrison, C. A. (2009). A common biosynthetic pathway governs the dimerization and secretion of inhibin and related transforming growth factor β (TGFβ) ligands. *Journal of Biological Chemistry*, *284*(14), 9311-9320.

Wang, R., Zhu, J., Dong, X., Shi, M., Lu, C., and Springer, T. A. (2012). GARP regulates the bioavailability and activation of TGFβ. *Molecular biology of the cell*, *23*(6), 1129-1139.

Wang X, Fischer G, Hyvonen, M. (2016) Structure and activation of pro-ActivinA. Nat. Comm. 7:12052

Wisotzkey, R. G., and Newfeld, S. J. (2020). TGF-β prodomain alignments reveal unexpected cysteine conservation consistent with phylogenetic predictions of cross-subfamily heterodimerization. *Genetics*, *214*(2), 447-465.

Yoshinaga, K., Obata, H., Jurukovski, V., Mazzieri, R., Chen, Y., Zilberberg, L., ... and Rifkin, D. B. (2008). Perturbation of transforming growth factor (TGF)-ß1 association with latent TGF-β binding protein yields inflammation and tumors. *Proceedings of the National Academy of Sciences*, *105*(48), 18758-18763.

Zhu, Y., Ahlemeyer, B., Bauerbach, E., and Krieglstein, J. (2001). TGF-β1 inhibits caspase-3 activation and neuronal apoptosis in rat hippocampal cultures. *Neurochem Int. 38*, 227-235.