UNIVERSITY OF CALIFORNIA

LOS ANGELES

Genome-wide Analysis of Groucho Function in *Drosophila* Embryogenesis

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in Biochemistry and Molecular Biology

by

Michael Douglas Chambers

2015

TABLE OF CONTENTS

ABSTRACT OF THE DISSERTATION

Genome-wide Analysis of Groucho Function in *Drosophila* Embryogenesis

By

Michael Douglas Chambers

University of California, Los Angeles, CA 2009

Professor Albert J. Courey, Chair

Animal developmental patterning is a complex and intricate process, requiring the interpretation of multiple temporally and spatially regulated signals to define the transcriptional profile of each cell. In *Drosophila*, Groucho (Gro), a transcriptional corepressor, participates in these processes through long-range silencing of distant enhancers. Lacking any innate DNA-binding activity, Gro is targeted to these elements through interaction with multiple repressors.

Using a combination of ChIP-seq and RNA-seq techniques, I sought to characterize Groucho activity at multiple stages spanning the initial nine hours of embryonic development, and thereby gain insight into both the mechanisms and extent of Groucho-mediated repression. These data reveal that Groucho is recruited to thousands of sites at each stage of development. Most Gro peaks are < 1kb in width, consistent with recruitment of one or a small number of Gro complexes to a regulatory element, indicating that spreading of Gro along stretches of chromatin is not a common feature

of repression, as previously thought. Gro binding is frequently observed as clusters of discrete peaks, which supports a model in which Gro self-association facilitates crosslinking of non-contiguous regions of chromatin. In some cases, Gro occupancy is observed at transcriptional start sites multiple kilobases from known Gro-binding silencing elements, suggesting that interactions with these sites via crosslinking and looping of chromatin is one method of Gro-mediated repression.

Both Gro gain- and loss-of-function embryos exhibit extensive perturbations in gene expression from the onset of zygotic transcription. Integration of these differentially expressed genes with ChIP-seq-derived Gro occupancy data facilitated the identification of Gro targets within each developmental stage, including known Gro-regulated genes and novel targets. Gro target genes are enriched for transcription factors involved in multiple developmental processes, as well as regulators of multiple signaling pathways. The activity of Gro in regulating both the inputs and outputs of signaling pathways suggests that Gro is utilized to regulate the cellular response to signaling on multiple levels by facilitating crosstalk between pathways.

Gro binds to many genes that it does not repress as exemplified by genes that are regulated by Dorsal, a bifunctional transcription factor that activates some targets and represses others. Gro is essential for repression, but dispensable for activation by Dorsal. However, data presented in this thesis suggest that Dorsal recruits Gro to both the repressed and the activated targets. This shows that Gro recruitment is not sufficient for repression.

Utilizing a technique to isolate and sequence the nascent transcripts of actively transcribed genes provides an accurate profile of the transcriptional activity of the *Drosophila* embryo at multiple stages of development. Nascent transcripts of Gro target genes are significantly enriched for promoter-proximal read density, indicating that Gro targets are enriched for stalled PolII. Coupled with the observation that Gro often binds overlapping or shortly downstream of start sites, we propose that Gro may repress genes by favoring the stalling of PolII, potentially through direct or mediated interaction with PolII or alteration of chromatin structure within transcribed regions.

Finally, I have contributed to a study characterizing the Gro interactome. This study showed that Gro interacts with U1 snRNP. I compared the effect of knocking down a U1 snRNP subunit with that of knocking down Gro on the S2 cell transcription profile. The results suggest that Gro-mediated repression of some targets may require U1 snRNP.

Chapter 1

Introduction:

Groucho – A Multifunctional Regulator of Drosophila Development

Introduction

The Groucho/TLE family of corepressors are ubiquitous regulators of animal development

The Groucho/TLE (Gro) family of corepressors play crucial roles in the interpretation and integration of multiple spatial, temporal, and signaling inputs during development in higher eukaryotes. Groucho, the sole *Drosophila melanogaster* member of this protein family, was first discovered in the context of a slight hypomorphic allele which resulted in the formation of extra supraorbital bristles reminiscent of the bushy eyebrows of Groucho Marx (Lindsley, 1968). Subsequent research on Gro in *Drosophila* has served to characterize this factor's central importance to developmental gene regulation in response to multiple developmental programs and signaling pathways. As a corepressor, Groucho has no documented ability to bind DNA directly in a sequence-specific manner, instead relying on recruitment to genomic loci through interactions with a diverse array of transcriptional repressors (Mannervik, 2014). Through its interactions with these repressors, it is essential to nearly all aspects of embryonic and imaginal *Drosophila* development (Paroush et al., 1994). In humans, Gro/TLE family proteins are involved in such processes as organ development, adipogenesis, neurogenesis, hematopoiesis, and osteogenesis (Bajoghli et al., 2005; Javed et al., 2000; Metzger et al., 2012; Villanueva et al., 2011).

Groucho consists of five domains, two of which are highly conserved throughout higher eukaryotes (Chen and Courey, 2000; Turki-Judeh and Courey, 2012a). A great body of work has arisen documenting the contributions of each domain to the overall function and regulation of Groucho. While much of this work has focused on the N- and C- terminal domains, as they are more conserved and more sensitive to point mutagenesis (Jennings et al., 2006; Jennings et al., 2007), the central domains of Groucho have been investigated for their roles in Groucho activity through interaction with a number of regulatory targets, including protein kinases, histones, and histone modifying enzymes (Turki-Judeh and Courey, 2012a).

Homologs of Groucho with similar roles in developmental decision making have been identified throughout metazoans (Fig. 1.1) (Paroush et al., 1994). Homologs have been identified and characterized in rats (Schmidt and Sladek, 1993), nematodes (Pflugrad et al., 1997), frogs (Choudhury et al., 1997), zebrafish (Wulbeck, 1997), mice (Mallo et al., 1993), and humans (Stifani et al., 1992). While the *Drosophila* and *C. elegans* genomes each encode single Gro family genes, the mouse, chick, and human genomes each encode four members, while zebrafish and medaka each encode six members (Li, 2000). The full-length human Gro orthologs, termed transducin-like Enhancer of Split 1-4 (TLE1-4) (Miyasaka et al., 1993), are expressed combinatorially during cell differentiation and have non-redundant roles during development (Stifani et al., 1992; Yao et al., 1998).

Mammalian genomes additionally encode two truncated Gro homologs, *Amino Enhancer of Splt (AES)*, which is homologous to the two N-terminal domains of Groucho

(Gasperowicz and Otto, 2005), and *Tle6/Grg6,* which possesses a poorly conserved N-terminal region and a C-terminal WD-repeat domain (Dang et al., 2001). Both factors are thought to antagonize the activity of full-length TLE family members. AES may function by directly binding to TLE proteins through Q-domain interactions (Brantjes et al., 2001) or by interacting with a subset of TLE-dependent repressors (Muhr et al., 2001). Similarly, TLE6/Grg6 has been shown to interact with repressors to block recruitment of full-length TLE family proteins and thereby alleviate repression (Marcal et al., 2005). More distantly related Gro homologs have been identified in yeast (Tup1) and plants (TOPLESS) (Courey and Jia, 2001; Lee and Golz, 2012; Smith and Johnson, 2000).

The domain architecture of Groucho/TLE family proteins

The N-terminal Q (glutamine rich) domain is one of the two highly conserved domains and is responsible for the formation of tetramers and potentially higher-order oligomers of Gro (Chen et al., 1998). Additionally, the Q-domain mediates a subset of interactions with transcriptional repressors, including the Tcf/Lef family of proteins (Brantjes et al., 2001). The structure of the Q-domain of TLE1, a human homologue of Gro, was recently solved, revealing that the domain forms a dimer of dimers consisting of two coiled-coils interdigitated in a head-to-head complex (Chodaparambil et al., 2014a). The resulting structure provides an elegant explanation of the mechanics of tetramerization, and corroborates the large frictional coefficient measured in

hydrodynamic studies of the purified Q-domain, as the predicted structure is thin and rod-like (Kuo et al., 2011).

The ability of the Q domain to direct the formation of high-order oligomers has been proposed to mediate the spreading of Gro along chromatin allowing for the establishment of large transcriptionally silent domains. This might explain the documented ability of Gro to direct long-range repression in which entire loci are organized into transcriptionally silent states. In support of this idea, assays involving Grg3, a mouse homolog of Gro, on *in vitro* chromatin arrays showed that oligomerization mediated through the Q-domain is not required for recruitment of Gro to chromatin but is required for subsequent aggregation of chromatinized fragments into a form that was resistant to transcription (Sekiya and Zaret, 2007).

Contrary to the idea that the Q domain could mediate spreading, chromatin immunoprecipitation (ChIP) assays in cell culture revealed that oligomerization-deficient mutants of *Drosophila* Gro exhibited similar median peak widths to wild-type Gro (Kaul et al., 2014). The interpretation of this result is somewhat complicated by the fact that binding data was generated from two *Drosophila* cell lines depleted of endogenous Groucho via RNAi and overexpressing either GFP-tagged wild-type or oligomerization-deficient Groucho. The authors showed a significant reduction of endogenous Gro that nonetheless remained detectable by immunoblot. Thus, it remains a possibility that low levels of endogenous Groucho were contributing to peak formation or spreading in both contexts.

Regardless of the role of oligomerization in the definition of the size of Groucho binding domains, loss of oligomerization does result in significant differences in the recruitment patterns of overexpressed wild-type and oligomerization-deficient mutants. Of the approximately 3000 distinct Groucho binding sites identified in Kc167 cells expressing wild-type or oligomerization-deficient Gro, 48% are unique to a single condition (Kaul et al., 2014). Loss of oligomerization potential therefore, while preserving some aspects of wild-type Gro binding patterns, does disrupt Groucho association with chromatin in some contexts, the nature of which remains unexplained.

(Chodaparambil et al., 2014a)(Kuo et al., 2011)The WD-domain is the second conserved domain of Gro and comprises the C-terminal 329 amino acids of the protein. The WD-domain consists of a seven-bladed β-propeller domain and is responsible for the majority of Groucho interactions with DNA-binding repressors (Table 1-1) (Pickles et al., 2002). The majority of these interactions are mediated through binding of the WD-domain to short peptide motifs (Jennings et al., 2006), which are recognized by the central pore of the propeller domain. Several such peptide motifs have been identified in Groucho-interacting proteins. The majority of these peptide motifs fall into one of two categories. C-terminal WRPW/Y recognition sequences have been found in Hairy/Enhancer of split (HES) and Runt family transcription factors (Aronson et al., 1997; Canon and Banerjee, 2003; Fisher et al., 1996; Jimenez et al., 1997; Paroush et al., 1994). And the engrailed homology domain-1 (eh1) motif is an internal peptide motif with the consensus sequence FxIxxIL that is found in Engrailed, Dorsal, Odd-skipped, and Goosecoid, among others (Copley, 2005; Dubnicoff et al., 1997; Jiménez et al., 1997;

Jimenez et al., 1999; Smith and Jaynes, 1996; Tolkunova et al., 1998). The WD domain binds to these motifs with differing affinities. These differences in affinity are utilized to control the recruitment of Groucho to specific factors. For example, the affinity of Groucho for binding the eh1-like motif of Dorsal is relatively weak (Flores-Saaib and Courey, 2000), necessitating the assistance of additional factors in facilitating a stable interaction between the two proteins. This weak affinity of the Dorsal/Groucho interaction is crucial to allowing Dorsal to function as a bifunctional transcription factor, as mutation of this motif to a higher-affinity sequence abolishes Dorsal's ability to activate genes in the embryo due to constitutive recruitment of Groucho (Ratnaparkhi et al., 2006).

The WD-repeat domain may be involved in additional protein interactions. Studies of Grg3, a mouse Gro/TLE family member, have shown that the WD domain is critical for binding to histone arrays *in vitro* as well as condensation of these arrays (Sekiya and Zaret, 2007). The observation that the Q domain is also capable of strong interaction with K20 methylated H4 tails suggests multiple levels of interaction between Gro/TLE proteins and histones, and may contribute to the protein's ability to associate with histones both locally, at its recruitment site, and distantly, through association with non-contiguous stretches of chromatin (Chodaparambil et al., 2014b).

The central region of Groucho is divided into three domains, the GP, CcN, and SP domains. The GP domain binds to a histone deacetylase (HDAC1/Rpd3), which is involved with some but not all Groucho-repressive activity (Chen et al., 1999). The CcN domain is involved in Groucho regulation, containing multiple Ck2 and Cdc2

phosphorylation sites (Nuthall et al., 2002). The SP domain contains multiple sites phosphorylated in response to MAPK signaling, resulting in down-regulation of Groucho activity (Hasson et al., 2005). There is evidence that the central regions of Groucho are intrinsically disordered (Turki-Judeh and Courey, 2012b), which has emerged as a common strategy among eukaryotic proteins to facilitate participation in diverse protein-protein interactions, expose signaling motifs, and/or accept posttranslational modifications (Dunker et al., 2008).

Groucho integrates multiple signaling pathways to generate specific cellular responses and fates

In *Drosophila,* Groucho's roles in responses to signaling pathways are well documented. The factor participates in Ras/MAPK, Notch, Decapentapletic (Dpp/BMP), and Wingless/Wnt signaling, among others. Groucho activity is down-regulated via the Ras/MAPK pathway in response to signals initiated at multiple receptor tyrosine kinases (RTKs) such as EGFR, FGFR, and Torso (Cinnamon and Paroush, 2008; Hasson et al., 2005). The resulting relief of Groucho-mediated repression is critical to the cellular response to RTK signaling and is thought to precipitate in cellular memory, whereby the attenuation of Groucho activity persists after loss of signaling (Cinnamon and Paroush, 2008; Helman et al., 2011).

In the absence of Notch signaling, Groucho represses *E(spl)* complex genes through interactions with Hairy, which is itself associated with Su(H), a sequence-

specific transcription factor that targets Notch-responsive genes (Delidakis et al., 1991). Recruitment of a Notch ligand to Notch transmembrane receptors activates the pathway, leading to proteolytic cleavage of the receptor and subsequent release of the Notch Intracellular Domain (Notch ICD). The Notch ICD rapidly enters the nucleus, where it displaces Hairy binding at Su(H) sites, relieving Groucho repression and initiating expression of *E(spl)* genes. Groucho then interacts with newly expressed E(spl) family proteins to repress a number of proneural genes (Preiss et al., 1988; Wurmbach et al., 1999). This repressive activity is alleviated by MAPK signaling, which results in the phosphorylation of Gro, negatively affecting its ability to repress these proneural genes in cooperation with E(spl) members (Andersson et al., 2011). The partial or complete negation of Notch signaling through the activation of the MAPK pathway thus represents a Groucho-mediated point of crosstalk between the two pathways (Hasson et al., 2005).

Groucho is also critical to signaling via Decapentaplegic (dpp), a *Drosophila* TGF-β homolog whose diffusion over long distances is essential to patterning during embryogenesis and later during appendage development (Upadhyai and Campbell, 2013). The Dpp morphogen is expressed dorsally in the embryo and is required for the definition of cell-fate along the dorsal-ventral axis (Ferguson and Anderson, 1992). Groucho, through interaction with Dorsal, represses ventral expression of *dpp*, meaning that Gro is involved in both the spatiotemporal definition and interpretation of dpp signaling (Schwyter et al., 1995). In the absence of Dpp signaling, Brinker (Brk) represses a subset of dpp target genes through two independent repressive mechanisms, one

involving dCtBP (a short-range corepressor), and the other involving Gro (Hasson et al., 2001). Upon activation of Dpp signaling, Brinker becomes repressed by Schnurri in dorsal regions of the embryo, while continuing to be expressed in ventrolateral regions (Marty et al., 2000).

Finally, Groucho participates in Wingless/Wnt signaling, through interactions with Tcf/Lef family proteins, to regulate cell-fate choice (Cavallo et al., 1998)(Roose and Clevers, 1999). In unstimulated cells, Groucho assists in repressing Tcf/Lef target genes through interactions with the Q-domain (Clevers, 2006). Upon Wnt activation, nuclear beta-catenin (Armadillo) concentration increases, which binds to Tcf, releasing Groucho and leading to gene activation. In this context, Groucho is essential in guarding against spurious activation of Wnt target genes in unstimulated cells (Daniels and Weis, 2005).

While there are hundreds of cell types in the adult fly, far fewer developmental signaling pathways have been documented (Perrimon et al., 2012). To generate this cellular complexity, informational content from multiple extracellular signals must be interpreted within each cell's specific spatial and temporal context (Hsueh et al., 2009). Even with this ability to simultaneously respond to multiple signals, the high number of discrete transcriptional states required during development necessitates that these signals are integrated non-additively (Housden and Perrimon, 2014). Factors that participate in multiple signaling pathways, such as Groucho, are a necessary component of a non-additive response. Groucho therefore presents a convenient node through which a cell can process limited combinations of inputs to produce a larger number of outcomes.

Groucho is an essential component of the embryonic axial patterning network

It is primarily through the spatially and temporally controlled regulation of gene transcription that Groucho becomes fundamental to embryonic patterning. Many early embryonic patterning proteins can be divided into effectors of the dorsal-ventral and anterior-posterior programs, though these processes are complex and highly interconnected (Jaeger et al., 2012), requiring the coordinated regulation of dozens of transcriptional activators, repressors, and co-regulators (Mannervik, 2014). Definition of the dorsal-ventral axis, which is critical to germ layer development, is carried out by the maternally-contributed gradient of nuclear Dorsal along this axis (Roth et al., 1989). Dorsal is a sequence-specific transcription factor, and the strength, spacing, and grouping of Dorsal binding sites, along with the distribution of adjacent binding sites for other interacting factors modulate Dorsal binding and cofactor recruitment in order to correctly interpret the Dorsal gradient (Zeitlinger et al., 2007).

On the ventral side of the embryo, high concentrations of nuclear Dorsal initiate transcriptional programs that determine the mesoderm (Gonzalez-Crespo and Levine, 1993). In ventrolateral regions, modest Dorsal concentrations help direct a neuroectodermal fate (Ip et al., 1992). Dorsal also acts as a repressor of dorsal ectodermal genes and, by keeping them off in ventral and ventrolateral region, it restricts their expression to the dorsal ectodermal primordium Groucho is required for

this repression and plays a critical role in switching Dorsal from an activator to a repressor (Dubnicoff et al., 1997).

In addition to its roles in dorsal/ventral patterning, Groucho has multiple roles in anterior/posterior pattern formation. For example, it is required for repression by numerous segmentation gene products such as Hairy, Runt, and Engrailed (Levine, 2008). Groucho is also required for the patterning of the anterior and posterior terminal domains by the Torso RTK through its interaction with Capicua (Ajuria et al., 2011), a process regulated by Ras/MAPK signaling (Chen et al., 2009; Paroush et al., 1997). Capicua recruits Gro to *tailless* and *huckebein* throughout the embryo maintaining these genes in an off state. Torso RTK then activates Ras/MAPK signaling at the termini leading to the phosphorylation and consequent inactivation of both Capicua and Gro at the embryonic termini allowing the expression of *tll* and *hkb* as required for specification of terminal fate (Winkler et al., 2010).

Groucho is capable of both short- and long-range repression

Transcriptional repressors in *Drosophila* can be classified as acting as either short- or long-range repressors dependent on their ability to counteract the regulatory potential of local (within ~100 bp) or distal (thousands of bp away or more) activating elements or promoters (Gray and Levine, 1996; Gray et al., 1994). Some repressors are specific for one type of repression, while others can adopt a short- or long-range repressive activity through association with multiple corepressors operating via distinct

mechanisms of repression (Courey and Jia, 2001). Groucho was originally considered a long-range co-repressor recruited exclusively by long-range repressors such as Hairy and Dorsal (Cai et al., 1996; Dubnicoff et al., 1997). CtBP, in contrast, is a well-studied corepressor capable of short-range repression when recruited by such short-range repressors as Kruppel, Giant, and Snail (Nibu and Levine, 2001; Nibu et al., 1998). Evidence that Groucho could oligomerize and potentially crosslink non-contiguous regions of chromatin provides a mechanistic explanation for its ability to quench distant regulatory elements.

More recently, it was found that in some contexts Groucho behaves as a short-range corepressor. Groucho appears to be recruited by Knirps, a short-range repressor capable of interacting with CtBP, to repress the expression of *even-skipped* (Payankaulam and Arnosti, 2009). Sloppy-paired 1 (Slp1), a Groucho-interacting repressor, is involved in the short-range repression of regulatory elements controlling the expression of multiple pair-rule genes (Andrioli et al., 2004). If Groucho is in fact commonly utilized as both a short- and long-range repressor, this sheds light on the observation that Groucho oligomerization is required in a context-dependent manner *in vivo* (Jennings et al., 2007), suggesting a mechanism whereby Groucho oligomerization is necessary for long-range repression but dispensable for short-range. Likely the classification of repressors as short- and long-range actors, while a useful abstraction when classifying repressors, masks much of the complexity of repressive activity that would be provided by a thorough understanding of repressive mechanisms.

The mechanism of Groucho-mediated repression

While a great deal is known about the developmental participation and interactors of Gro, details of the mechanism by which Gro achieves repression have remained elusive. Multiple models have been proposed to explain Groucho's ability to fully and reversibly initiate and maintain both short- and long-range repression, yet a full picture, able to account for all observations of Groucho behavior, has yet to emerge.

Much of the speculation surrounding Groucho activity centers on the possibility that oligomeric structures of Gro form *in vivo*, how these structures interact with chromatin, and what relevance these structures have to repression. Early evidence showed that Groucho tetramerizes *in vitro* via the Q-domain (Chen et al., 1998)(Song et al., 2004). In another experiment, Groucho was found to be associated with chromatin over 2 kb away from its recruitment site, leading to the hypothesis that Groucho spreads from its recruitment site, analogous to the spreading activity of Sir family corepressors (Pirrotta and Gross, 2005). Experiments on a mouse Gro homolog showed that while tetramerization is not required for recruitment to chromatin, it is necessary for the aggregation of nucleosomal arrays in vitro (Sekiya and Zaret, 2007). Monomeric forms of the protein successfully bind to and increase the density of dinucleosomes *in vitro* (Sekiya and Zaret, 2007). *In vivo,* the loss of tetramerization is lethal but does not entirely abolish Gro-mediated repression (Jennings et al., 2007). More recent evidence in cell culture has shown that Gro binds in discrete peaks, though longer stretches of binding do occur (Kaul et al., 2014).

Gro preferentially associates with histone tails and can do so without the involvement of additional DNA-binding interacting factors (Flores-Saaib and Courey, 2000; Sekiya and Zaret, 2007). Additionally, Gro associates with a histone deacetylase, HDAC1/Rpd3 (Chen et al., 1999). This association accounts for some but not all of Groucho's repressive ability *in vivo*, where Groucho binding is associated with decreased acetylation of the tails of histones H3 and H4, as well as increased nucleosome density (Winkler et al., 2010). Colocalization of Gro and Rpd3 is prevalent in Kc167 cells (a cell line derived from *Drosophila* embryos), with over half of Groucho binding sites found to overlap Rpd3 binding (Kaul et al., 2014).

Given the many gaps in our knowledge regarding the mechanisms of Gro-mediated repression, I carried out a genome-wide analysis of Gro function in hopes of filling in some of these gaps. Experiments described in Chapter 2, employing a combination of Gro-ChIP-seq on staged wild-type embryos and RNA-seq on staged embryos expressing different levels of Gro show that Groucho associates with chromatin in discrete < 1 kilobase peaks, often clustered closely upstream or within regulated genes. This data was used to generate a set of high-confidence Groucho targets at multiple developmental stages. Experiments described in Chapter 3, employing nascent-seq on staged wild-type embryos show that Groucho-regulated genes are enriched for promoter-proximal paused polymerase, suggesting a possible role for PolII stalling in Groucho-mediated gene repression. Chapter 4 is a published paper in which we identified the Gro interactome as a way of illuminating mechanisms of Gro-mediated repression.

Figure 1-1. Groucho/TLE family proteins are partially conserved throughout metazoans. The Gro/TLE family of corepressors are typified by five domains defined based on function and sequence. Domain-wise homology to the *D. melanogaster* Groucho is indicated by percentages, when significant. Two domains, the N-terminal Q domain and the C-terminal WD-repeat domain are well conserved while the central region, consisting of the GP, CcN, and SP domains shares little sequence homology between species. The Q domain is involved in association with repressor and the formation of homo-oligomeric Groucho complexes. The WD domain is additionally involved in repressor association. The central region is predicted to be intrinsically disordered and serves as a scaffold for a number of protein interactions, notably with Rpd3, a histone deacetylase involved in some aspects of Groucho-mediated repression. The central regions also serve as a regulatory region of Groucho via being target for multiple post-translational modifications.

Fig. 1-1

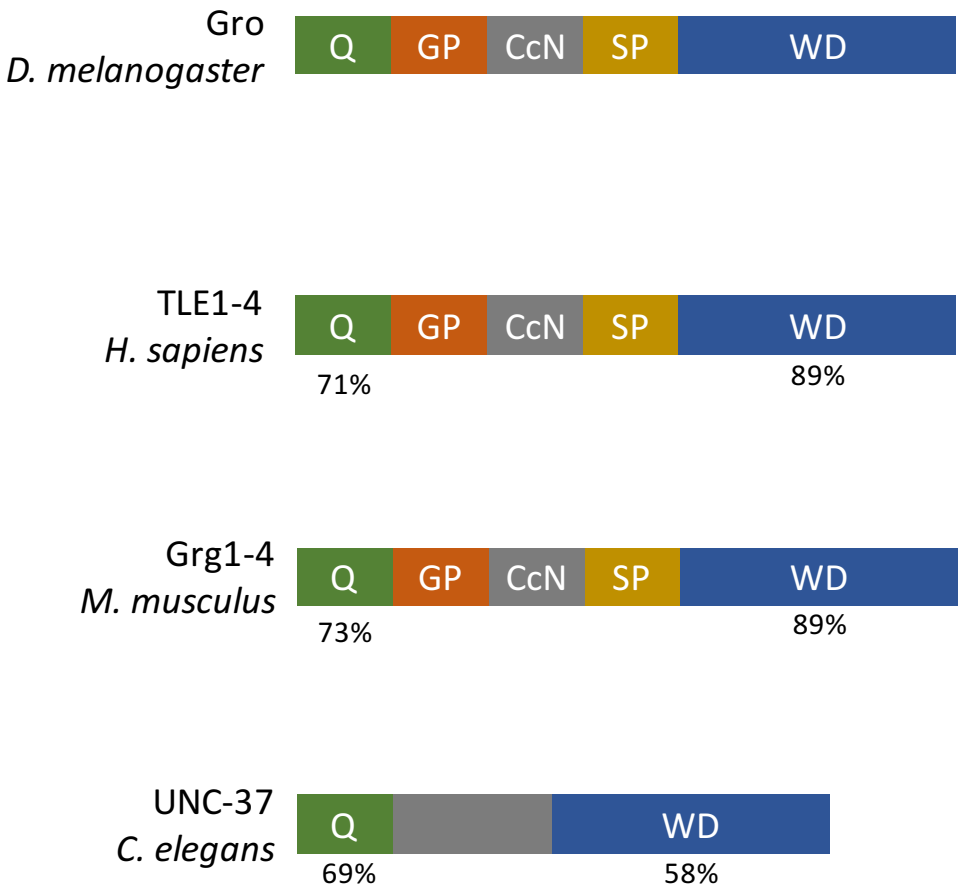

Gro
D. melanogaster
Q
GP
CcN
SP
WD
TLE1-4
H. sapiens
Q
GP
CcN
SP
WD
71%
89%
Grg1-4
M. musculus
Q
GP
CcN
SP
WD
73%
89%
UNC-37
C. elegans
Q
WD
69%
58%

Table 1-1. Groucho-interacting transcription factors

Interacting Protein	Biological Role	Citation
Capicua	RTK signaling; embryonic terminal gene expression	(Jimenez et al., 2000)
Huckebein	Embryonic terminal gene expression	(Goldstein et al., 1999)
Hairy	Segmentation/ Anterior-posterior patterning	(Paroush et al., 1994)
Runt	Segmentation/ Anterior-posterior patterning	(Aronson et al., 1997)
Even-skipped	Segmentation/ Anterior-posterior patterning	(Kobayashi et al., 2001)
Odd-skipped	Segmentation/ Anterior-posterior patterning	(Goldstein et al., 2005)
Sloppy-paired 1	Segmentation/ Anterior-posterior patterning	(Andrioli et al., 2004)
Engrailed	Segmentation/ Anterior-posterior patterning	(Jimenez et al., 1997)
Knirps	Segmentation/ Anterior-posterior patterning	(Payankaulam and Arnosti, 2009)
Goosecoid	Segmentation/ Anterior-posterior patterning	(Jimenez et al., 1999)
Dorsal	Dorsal-ventral patterning	(Dubnicoff et al., 1997)
Brinker	Dorsal-ventral patterning	(Zhang et al., 2001)
Ind	Dorsal-ventral patterning	(Von Ohlen et al., 2007)
Vnd	Dorsal-ventral patterning	(Cowden and Levine, 2003)
Su(H)	Notch signaling	(Barolo et al., 2002)

References

Ajuria, L., Nieva, C., Winkler, C., Kuo, D., Samper, N., Andreu, M.J., Helman, A., Gonzalez-Crespo, S., Paroush, Z., Courey, A.J*., et al.* (2011). Capicua DNA-binding sites are general response elements for RTK signaling in Drosophila. Development (Cambridge, England) *138*, 915-924.

Andersson, E.R., Sandberg, R., and Lendahl, U. (2011). Notch signaling: simplicity in design, versatility in function. Development *138*, 3593-3612.

Andrioli, L.P., Oberstein, A.L., Corado, M.S., Yu, D., and Small, S. (2004). Groucho-dependent repression by sloppy-paired 1 differentially positions anterior pair-rule stripes in the Drosophila embryo. Dev Biol *276*, 541-551.

Aronson, B.D., Fisher, A.L., Blechman, K., Caudy, M., and Gergen, J.P. (1997). Groucho-dependent and -independent repression activities of Runt domain proteins. Mol Cell Biol *17*, 5581-5587.

Bajoghli, B., Aghaallaei, N., and Czerny, T. (2005). Groucho corepressor proteins regulate otic vesicle outgrowth. Dev Dyn *233*, 760-771.

Barolo, S., Stone, T., Bang, A.G., and Posakony, J.W. (2002). Default repression and Notch signaling: Hairless acts as an adaptor to recruit the corepressors Groucho and dCtBP to Suppressor of Hairless. Genes Dev *16*, 1964-1976.

Brantjes, H., Roose, J., van De Wetering, M., and Clevers, H. (2001). All Tcf HMG box transcription factors interact with Groucho-related co-repressors. Nucleic Acids Res *29*, 1410-1419.

Cai, H.N., Arnosti, D.N., and Levine, M. (1996). Long-range repression in the Drosophila embryo. Proc Natl Acad Sci U S A *93*, 9309-9314.

Canon, J., and Banerjee, U. (2003). In vivo analysis of a developmental circuit for direct transcriptional activation and repression in the same cell by a Runx protein. Genes Dev *17*, 838-843.

Cavallo, R.A., Cox, R.T., Moline, M.M., Roose, J., Polevoy, G.A., Clevers, H., Peifer, M., and Bejsovec, A. (1998). Drosophila Tcf and Groucho interact to repress Wingless signalling activity. Nature *395*, 604-608.

Chen, G., and Courey, A.J. (2000). Groucho/TLE family proteins and transcriptional repression. Gene *249*, 1-16.

Chen, G., Fernandez, J., Mische, S., and Courey, A.J. (1999). A functional interaction between the histone deacetylase Rpd3 and the corepressor groucho in Drosophila development. Genes Dev *13*, 2218-2230.

Chen, G., Nguyen, P., and Courey, A. (1998). A role for Groucho tetramerization in transcriptional repression. Molecular and Cellular Biology *18*, 7259.

Chen, Y.C., Lin, S.I., Chen, Y.K., Chiang, C.S., and Liaw, G.J. (2009). The Torso signaling pathway modulates a dual transcriptional switch to regulate tailless expression. Nucleic Acids Res *37*, 1061-1072.

Chodaparambil, J.V., Pate, K.T., Hepler, M.R., Tsai, B.P., Muthurajan, U.M., Luger, K., Waterman, M.L., and Weis, W.I. (2014a). Molecular functions of the TLE tetramerization domain in Wnt target gene repression. EMBO J *33*, 719-731.

Chodaparambil, J.V., Pate, K.T., Hepler, M.R.D., Tsai, B.P., Muthurajan, U.M., Luger, K., Waterman, M.L., and Weis, W.I. (2014b). Molecular functions of the TLE tetramerization domain in Wnt target gene repression. The EMBO Journal *33*, 719-731.

Choudhury, B.K., Kim, J., Kung, H.F., and Li, S.S. (1997). Cloning and developmental expression of Xenopus cDNAs encoding the Enhancer of split groucho and related proteins. Gene *195*, 41-48.

Cinnamon, E., and Paroush, Z. (2008). Context-dependent regulation of Groucho/TLE-mediated repression. Current opinion in genetics & development *18*, 435-440.

Clevers, H. (2006). Wnt/beta-catenin signaling in development and disease. Cell *127*, 469-480.

Copley, R.R. (2005). The EH1 motif in metazoan transcription factors. BMC Genomics *6*, 169.

Courey, A.J., and Jia, S. (2001). Transcriptional repression: the long and the short of it. Genes Dev *15*, 2786-2796.

Cowden, J., and Levine, M. (2003). Ventral dominance governs sequential patterns of gene expression across the dorsal-ventral axis of the neuroectoderm in the Drosophila embryo. Dev Biol *262*, 335-349.

Dang, J., Inukai, T., Kurosawa, H., Goi, K., Inaba, T., Lenny, N.T., Downing, J.R., Stifani, S., and Look, A.T. (2001). The E2A-HLF oncoprotein activates Groucho-related genes and suppresses Runx1. Mol Cell Biol *21*, 5935-5945.

Daniels, D.L., and Weis, W.I. (2005). Beta-catenin directly displaces Groucho/TLE repressors from Tcf/Lef in Wnt-mediated transcription activation. Nat Struct Mol Biol *12*, 364-371.

Delidakis, C., Preiss, A., Hartley, D.A., and Artavanis-Tsakonas, S. (1991). Two genetically and molecularly distinct functions involved in early neurogenesis reside within the Enhancer of split locus of Drosophila melanogaster. Genetics *129*, 803-823.

Dubnicoff, T., Valentine, S.A., Chen, G., Shi, T., Lengyel, J.A., Paroush, Z., and Courey, A.J. (1997). Conversion of dorsal from an activator to a repressor by the global corepressor Groucho. Genes & Development *11*, 2952-2957.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., and Uversky, V.N. (2008). The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics *9 Suppl 2*, S1.

Ferguson, E.L., and Anderson, K.V. (1992). Decapentaplegic acts as a morphogen to organize dorsal-ventral pattern in the Drosophila embryo. Cell *71*, 451-461.

Fisher, A.L., Ohsako, S., and Caudy, M. (1996). The WRPW motif of the hairy-related basic helix-loop-helix repressor proteins acts as a 4-amino-acid transcription repression and protein-protein interaction domain. Mol Cell Biol *16*, 2670-2677.

Flores-Saaib, R., and Courey, A. (2000). Analysis of Groucho–histone interactions suggests mechanistic similarities between Groucho-and Tup1-mediated repression. Nucleic Acids Research *28*, 4189.

Gasperowicz, M., and Otto, F. (2005). Mammalian Groucho homologs: redundancy or specificity? J Cell Biochem *95*, 670-687.

Goldstein, R.E., Cook, O., Dinur, T., Pisante, A., Karandikar, U.C., Bidwai, A., and Paroush, Z. (2005). An eh1-like motif in odd-skipped mediates recruitment of Groucho and repression in vivo. Mol Cell Biol *25*, 10711-10720.

Goldstein, R.E., Jimenez, G., Cook, O., Gur, D., and Paroush, Z. (1999). Huckebein repressor activity in Drosophila terminal patterning is mediated by Groucho. Development *126*, 3747-3755.

Gonzalez-Crespo, S., and Levine, M. (1993). Interactions between dorsal and helix-loop-helix proteins initiate the differentiation of the embryonic mesoderm and neuroectoderm in Drosophila. Genes Dev *7*, 1703-1713.

Gray, S., and Levine, M. (1996). Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila. Genes Dev *10*, 700-710.

Gray, S., Szymanski, P., and Levine, M. (1994). Short-range repression permits multiple enhancers to function autonomously within a complex promoter. Genes Dev *8*, 1829-1838.

Hasson, P., Egoz, N., Winkler, C., Volohonsky, G., Jia, S., Dinur, T., Volk, T., Courey, A.J., and Paroush, Z. (2005). EGFR signaling attenuates Groucho-dependent repression to antagonize Notch transcriptional output. Nat Genet *37*, 101-105.

Hasson, P., Muller, B., Basler, K., and Paroush, Z. (2001). Brinker requires two corepressors for maximal and versatile repression in Dpp signalling. EMBO J *20*, 5725-5736.

Helman, A., Cinnamon, E., Mezuman, S., Hayouka, Z., Von Ohlen, T., Orian, A., Jiménez, G., and Paroush, Z.a.e. (2011). Phosphorylation of Groucho Mediates RTK Feedback Inhibition and Prolonged Pathway Target Gene Expression. Current Biology *21*, 1102-1110.

Housden, B.E., and Perrimon, N. (2014). Spatial and temporal organization of signaling pathways. Trends Biochem Sci *39*, 457-464.

Hsueh, R.C., Natarajan, M., Fraser, I., Pond, B., Liu, J., Mumby, S., Han, H., Jiang, L.I., Simon, M.I., Taussig, R.*, et al.* (2009). Deciphering signaling outcomes from a system of complex networks. Sci Signal *2*, ra22.

Ip, Y.T., Park, R.E., Kosman, D., Bier, E., and Levine, M. (1992). The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo. Genes Dev *6*, 1728-1739.

Jaeger, J., Manu, and Reinitz, J. (2012). Drosophila blastoderm patterning. Curr Opin Genet Dev *22*, 533-541.

Javed, A., Guo, B., Hiebert, S., Choi, J.Y., Green, J., Zhao, S.C., Osborne, M.A., Stifani, S., Stein, J.L., Lian, J.B.*, et al.* (2000). Groucho/TLE/R-esp proteins associate with the nuclear

matrix and repress RUNX (CBF(alpha)/AML/PEBP2(alpha)) dependent activation of tissue-specific gene transcription. J Cell Sci *113 (Pt 12)*, 2221-2231.

Jennings, B.H., Pickles, L.M., Wainwright, S.M., Roe, S.M., Pearl, L.H., and Ish-Horowicz, D. (2006). Molecular recognition of transcriptional repressor motifs by the WD domain of the Groucho/TLE corepressor. Mol Cell *22*, 645-655.

Jennings, B.H., Wainwright, S.M., and Ish-Horowicz, D. (2007). Differential in vivo requirements for oligomerization during Groucho-mediated repression. EMBO reports *9*, 76-83.

Jimenez, G., Guichet, A., Ephrussi, A., and Casanova, J. (2000). Relief of gene repression by torso RTK signaling: role of capicua in Drosophila terminal and dorsoventral patterning. Genes Dev *14*, 224-231.

Jimenez, G., Paroush, Z., and Ish-Horowicz, D. (1997). Groucho acts as a corepressor for a subset of negative regulators, including Hairy and Engrailed. Genes Dev *11*, 3072-3082.

Jiménez, G., Paroush, Z., and Ish-Horowicz, D. (1997). Groucho acts as a corepressor for a subset of negative regulators, including Hairy and Engrailed. Genes & Development *11*, 3072.

Jimenez, G., Verrijzer, C.P., and Ish-Horowicz, D. (1999). A conserved motif in goosecoid mediates groucho-dependent repression in Drosophila embryos. Mol Cell Biol *19*, 2080-2087.

Kaul, A., Schuster, E., and Jennings, B.H. (2014). The Groucho Co-repressor Is Primarily Recruited to Local Target Sites in Active Chromatin to Attenuate Transcription. PLoS Genetics *10*, e1004595.

Kobayashi, M., Goldstein, R.E., Fujioka, M., Paroush, Z., and Jaynes, J.B. (2001). Groucho augments the repression of multiple Even skipped target genes in establishing parasegment boundaries. Development *128*, 1805-1815.

Kuo, D., Nie, M., De Hoff, P., Chambers, M., Phillips, M., Hirsch, A.M., and Courey, A.J. (2011). A SUMO-Groucho Q domain fusion protein: characterization and in vivo Ulp1-mediated cleavage. Protein expression and purification *76*, 65-71.

Lee, J.E., and Golz, J.F. (2012). Diverse roles of Groucho/Tup1 co-repressors in plant growth and development. Plant Signaling & Behavior *7*, 86-92.

Levine, M. (2008). A systems view of Drosophila segmentation. Genome Biol *9*, 207.

Li, S.S. (2000). Structure and function of the Groucho gene family and encoded transcriptional corepressor proteins from human, mouse, rat, Xenopus, Drosophila and nematode. Proc Natl Sci Counc Repub China B *24*, 47-55.

Lindsley, D.L., Zimm, Georgianna G. (1968). Genetic Variations of Drosophila melanogaster, Vol 627 (Carnegie Institution of Washingtion Publication).

Mallo, M., Franco del Amo, F., and Gridley, T. (1993). Cloning and developmental expression of Grg, a mouse gene related to the groucho transcript of the Drosophila Enhancer of split complex. Mech Dev *42*, 67-76.

Mannervik, M. (2014). Control of Drosophila embryo patterning by transcriptional co-regulators. Experimental cell research *321*, 47-57.

Marcal, N., Patel, H., Dong, Z., Belanger-Jasmin, S., Hoffman, B., Helgason, C.D., Dang, J., and Stifani, S. (2005). Antagonistic effects of Grg6 and Groucho/TLE on the transcription repression activity of brain factor 1/FoxG1 and cortical neuron differentiation. Mol Cell Biol *25*, 10916-10929.

Marty, T., Muller, B., Basler, K., and Affolter, M. (2000). Schnurri mediates Dpp-dependent repression of brinker transcription. Nat Cell Biol *2*, 745-749.

Metzger, D.E., Gasperowicz, M., Otto, F., Cross, J.C., Gradwohl, G., and Zaret, K.S. (2012). The transcriptional co-repressor Grg3/Tle3 promotes pancreatic endocrine progenitor delamination and -cell differentiation. Development (Cambridge, England) *139*, 1447-1456.

Miyasaka, H., Choudhury, B.K., Hou, E.W., and Li, S.S. (1993). Molecular cloning and expression of mouse and human cDNA encoding AES and ESG proteins with strong similarity to Drosophila enhancer of split groucho protein. Eur J Biochem *216*, 343-352.

Muhr, J., Andersson, E., Persson, M., Jessell, T.M., and Ericson, J. (2001). Groucho-mediated transcriptional repression establishes progenitor cell pattern and neuronal fate in the ventral neural tube. Cell *104*, 861-873.

Nibu, Y., and Levine, M.S. (2001). CtBP-dependent activities of the short-range Giant repressor in the Drosophila embryo. Proc Natl Acad Sci U S A *98*, 6204-6208.

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

Nuthall, H.N., Joachim, K., Palaparti, A., and Stifani, S. (2002). A role for cell cycle-regulated phosphorylation in Groucho-mediated transcriptional repression. J Biol Chem *277*, 51049-51057.

Paroush, Z., Finley, R.L., Jr., Kidd, T., Wainwright, S.M., Ingham, P.W., Brent, R., and Ish-Horowicz, D. (1994). Groucho is required for Drosophila neurogenesis, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. Cell *79*, 805-815.

Paroush, Z., Wainwright, S.M., and Ish-Horowicz, D. (1997). Torso signalling regulates terminal patterning in Drosophila by antagonising Groucho-mediated repression. Development *124*, 3827-3834.

Payankaulam, S., and Arnosti, D.N. (2009). Groucho corepressor functions as a cofactor for the Knirps short-range transcriptional repressor. Proceedings of the National Academy of Sciences *106*, 17314-17319.

Perrimon, N., Pitsouli, C., and Shilo, B.Z. (2012). Signaling mechanisms controlling cell fate and embryonic patterning. Cold Spring Harb Perspect Biol *4*, a005975.

Pflugrad, A., Meir, J.Y., Barnes, T.M., and Miller, D.M., 3rd (1997). The Groucho-like transcription factor UNC-37 functions with the neural specificity gene unc-4 to govern motor neuron identity in C. elegans. Development *124*, 1699-1709.

Pickles, L.M., Roe, S.M., Hemingway, E.J., Stifani, S., and Pearl, L.H. (2002). Crystal structure of the C-terminal WD40 repeat domain of the human Groucho/TLE1 transcriptional corepressor. Structure *10*, 751-761.

Pirrotta, V., and Gross, D.S. (2005). Epigenetic silencing mechanisms in budding yeast and fruit fly: different paths, same destinations. Mol Cell *18*, 395-398.

Preiss, A., Hartley, D.A., and Artavanis-Tsakonas, S. (1988). The molecular genetics of Enhancer of split, a gene required for embryonic neural development in Drosophila. EMBO J *7*, 3917-3927.

Ratnaparkhi, G.S., Jia, S., and Courey, A.J. (2006). Uncoupling dorsal-mediated activation from dorsal-mediated repression in the Drosophila embryo. Development *133*, 4409-4414.

Roose, J., and Clevers, H. (1999). TCF transcription factors: molecular switches in carcinogenesis. Biochim Biophys Acta *1424*, M23-37.

Roth, S., Stein, D., and Nüsslein-Volhard, C. (1989). A gradient of nuclear localization of the dorsal protein determines dorsoventral pattern in the Drosophila embryo. Cell *59*, 1189-1202.

Schmidt, C.J., and Sladek, T.E. (1993). A rat homolog of the Drosophila enhancer of split (groucho) locus lacking WD-40 repeats. J Biol Chem *268*, 25681-25686.

Schwyter, D.H., Huang, J.D., Dubnicoff, T., and Courey, A.J. (1995). The decapentaplegic core promoter region plays an integral role in the spatial control of transcription. Mol Cell Biol *15*, 3960-3968.

Sekiya, T., and Zaret, K.S. (2007). Repression by Groucho/TLE/Grg Proteins: Genomic Site Recruitment Generates Compacted Chromatin In Vitro and Impairs Activator Binding In Vivo. Molecular Cell *28*, 291-303.

Smith, R.L., and Johnson, A.D. (2000). Turning genes off by Ssn6-Tup1: a conserved system of transcriptional repression in eukaryotes. Trends Biochem Sci *25*, 325-330.

Smith, S.T., and Jaynes, J.B. (1996). A conserved region of engrailed, shared among all en-, gsc-, Nk1-, Nk2- and msh-class homeoproteins, mediates active transcriptional repression in vivo. Development *122*, 3141-3150.

Song, H., Hasson, P., Paroush, Z.a.e., and Courey, A.J. (2004). Groucho oligomerization is required for repression in vivo. Molecular and Cellular Biology *24*, 4341-4350.

Stifani, S., Blaumueller, C.M., Redhead, N.J., Hill, R.E., and Artavanis-Tsakonas, S. (1992). Human homologs of a Drosophila Enhancer of split gene product define a novel family of nuclear proteins. Nat Genet *2*, 119-127.

Tolkunova, E.N., Fujioka, M., Kobayashi, M., Deka, D., and Jaynes, J.B. (1998). Two distinct types of repression domain in engrailed: one interacts with the groucho corepressor and is preferentially active on integrated target genes. Mol Cell Biol *18*, 2804-2814.

Turki-Judeh, W., and Courey, A.J. (2012a). Groucho: A Corepressor with Instructive Roles in Development. In (Elsevier), pp. 65-96.

Turki-Judeh, W., and Courey, A.J. (2012b). The Unconserved Groucho Central Region Is Essential for Viability and Modulates Target Gene Specificity. PLoS ONE *7*, e30610.

Upadhyai, P., and Campbell, G. (2013). Brinker possesses multiple mechanisms for repression because its primary co-repressor, Groucho, may be unavailable in some cell types. Development (Cambridge, England) *140*, 4256-4265.

Villanueva, C.J., Waki, H., Godio, C., Nielsen, R., Chou, W.-L., Vargas, L., Wroblewski, K., Schmedt, C., Chao, L.C., Boyadjian, R.*, et al.* (2011). TLE3 Is a Dual-Function Transcriptional Coregulator of Adipogenesis. Cell Metabolism *13*, 413-427.

Von Ohlen, T., Syu, L.J., and Mellerick, D.M. (2007). Conserved properties of the Drosophila homeodomain protein, Ind. Mech Dev *124*, 925-934.

Winkler, C.J., Ponce, A., and Courey, A.J. (2010). Groucho-Mediated Repression May Result from a Histone Deacetylase-Dependent Increase in Nucleosome Density. PLoS ONE *5*, e10166.

Wulbeck, C.C.-O., J. A. (1997). Two zebrafish homologues of the Drosophila neurogenic gene groucho and their pattern of transcription during early embryogenesis. Dev Genes Evol *207*, 156-166.

Wurmbach, E., Wech, I., and Preiss, A. (1999). The Enhancer of split complex of Drosophila melanogaster harbors three classes of Notch responsive genes. Mech Dev *80*, 171-180.

Yao, J., Liu, Y., Husain, J., Lo, R., Palaparti, A., Henderson, J., and Stifani, S. (1998). Combinatorial expression patterns of individual TLE proteins during cell determination and differentiation suggest non-redundant functions for mammalian homologs of Drosophila Groucho. Dev Growth Differ *40*, 133-146.

Zeitlinger, J., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes & Development *21*, 385-390.

Zhang, H., Levine, M., and Ashe, H.L. (2001). Brinker is a sequence-specific transcriptional repressor in the Drosophila embryo. Genes Dev *15*, 261-266.

Chapter 2

Groucho activity in the developing embryo

Introduction

The corepressor Groucho (Gro) is a crucial regulator of gene expression throughout development and is centrally involved in the establishment of embryonic patterning in the early *Drosophila* embryo (Agarwal et al., 2015). Through interactions with numerous sequence-specific transcription factors, Gro is crucial to the spatial and temporal restriction of gene expression beginning very early in embryonic development and continuing throughout larval and pupal development (Turki-Judeh and Courey, 2012a). As Gro mRNA is maternally deposited in the oocyte, high levels of the protein are present from the onset of development, and as such Gro participates in many of the earliest transcriptional decisions in the embryo (Paroush et al., 1994). Due to the highly-connected position of Gro in the developmental regulatory network, changes in Gro levels or function result in profound developmental abnormalities and disease (Buscarlet and Stifani, 2007).

In this study, we utilize high-throughput sequencing technologies to characterize the dynamics of Groucho genomic binding and to identify Groucho repressive targets. Global analysis of Groucho binding patterns via ChIP-seq allows us to gain insight into the mechanisms of Groucho-mediated repression via characterization of Groucho localization to numerous regulatory regions and analysis of how this localization correlates with binding patterns of additional transcription factors, including those known to interact with Gro. Analysis of the embryonic transcriptome under conditions of perturbed Groucho dosage enables us to dissect Groucho's role in multiple signaling

pathways and, integrated with the ChIP-seq analysis, to identify these targets and Gro's participatory roles with high confidence.

Although Gro is ubiquitously expressed, it is not simply a passive component of the transcriptional machinery. For example, its function can be actively regulated in response to informational signals arising outside of the cell, through, for example, the action of the Ras/MAPK pathway. In addition, although Gro does not bind to DNA directly, it may participate via unknown mechanisms in target gene selection. This is demonstrated by a Gro deletion analysis in which it was shown that deletion of a Gro domain termed the SP domain results in promiscuous repression of genes not normally targeted by Gro (Turki-Judeh and Courey, 2012b).

Despite the extensively documented centrality of Gro in multiple developmental processes, especially in the early embryo, no systematic genome-wide investigation has been undertaken to position Gro in the fly developmental regulatory network. A more thorough understanding of the recruitment patterns of Gro in the early embryo, and the dynamics of such binding, will allow us to address multiple questions about the mechanism of Gro-mediated repression and the position of Gro in the regulatory hierarchy of pattern formation.

Gro tetramerizes and perhaps forms higher order oligomers *in vitro.* This together with the observations that Gro is required for long-range repression and that it binds core histones has led to the suggestion that Gro-mediated repression may involve spreading along chromatin. Indeed, in some contexts Gro oligomerization is necessary

for repression *in vitro* (Chen et al., 1998) and *in vivo* (Song et al., 2004). However, it does not appear to be a universal requirement for repressive activity in all developmental contexts (Jennings et al., 2007). Evidence from ChIP-PCR experiments suggests Gro spreads over potentially long stretches of chromatin presumably through its ability to self-associate (Martinez and Arnosti, 2008; Winkler et al., 2010), although these studies are limited by the resolution of the ChIP-PCR analysis. More recent Gro ChIP-seq data obtained from two Drosophila cell lines (S2 and Kc167)(Kaul et al., 2014) indicate that binding is primarily localized to discrete peaks in those cell lines. However, it is unclear to what degree that binding pattern extends to developing embryos. Genome-wide analysis of binding patterns in embryos presented in this thesis has enabled us to thoroughly investigate the requirement for spreading in Gro-mediated repression. We find that while Groucho is capable of spreading over long regions of chromatin, this spreading appears to be an uncommon feature of repression, with the majority of Groucho binding occurring in discreet peaks characteristic of association with site-specific transcription factors. However, these discrete peaks often cluster over longer stretches of chromatin, potentially indicative of looped or a similar topological rearrangement between distant regions of chromatin.

The accurate assignment of a regulatory region, or even an individual binding region detected by ChIP-seq, to a specific regulatory target (or targets) is a long standing problem in the useful interpretation of ChIP-seq studies (Sikora-Wohlfeld et al., 2013). The inaccuracy of association becomes more significant the further a factor binds from its regulatory target, as genomic complexity often makes assignment of enhancer-gene

interactions uncertain. A common methodology to address this challenge is to incorporate genome-wide binding data with transcriptome measurements in systems perturbed for said factor (Dolinski and Troyanskaya, 2015). To this end, we have employed RNA-seq to examine the effect of Gro-knockdown and Gro-overexpression on the transcriptome measurements at timepoints matching those used in the ChIP-seq analysis. When combined with the ChIP-seq binding profile data, this has allowed the definition of a high-confidence set of Gro target genes across developmental stages, thus enabling a more thorough characterization of the role of Gro during early development and a significant refinement of the factor's influence on the developmentally-regulated gene network. The analysis to be presented here shows that Groucho targets are enriched for numerous transcription factors, confirming its role as a factor near the top of the regulatory hierarchy in the establishment of developmental fate.

Materials & Methods

Fly strains

Flies were maintained on standard medium at 25˚C. UAS-*Gro* and UAS-*GroΔGP* transgenic flies were described previously (Turki-Judeh and Courey, 2012b). The UAS-GroΔGP construct contains a deletion of amino acids 134-194, encompassing the GP domain. Embryos for overexpression studies were obtained from staged embryos collected from crosses of UAS-*Gro* with a maternal driver, *Mat-Gal4* (Nie et al., 2009). Control embryos for RNA-seq were obtained from crossing w^{1118} flies with this *Mat-Gal4* driver. Germ line clones of the *gro* mutant fly allele MB36 (a null allele) were used for Groucho loss-of-function studies (Jennings et al., 2007). These lines were generated using the standard dominant female sterile FLP/FRT protocol (Chou and Perrimon, 1996).

Groucho chromatin immunoprecipation (ChIP) and sequencing

ChIP was carried out as described previously (Bonn et al., 2012). Staged embryos were collected from OregonR population cages and crosslinked with formaldehyde prior to sonication (Diagenode Bioruptor). Immunoprecipitation was carried out using rabbit polyclonal antibodies raised against the Gro-GP domain GST fusion protein that had been affinity purified against the Halo-tagged GP domain. Libraries for multiplex sequencing were prepared using the Nugen Ovatoin Ultralow System V2 kit (catalog # 0344-32).

Groucho ChIP-seq data analysis

Multiplexed libraries were sequenced on Illumina HiSeq 2000 sequencing platforms (High Throughput Sequencing Facility, Broad Stem Cell Research Center, UCLA). Reads were demultiplexed via custom scripts. Demultiplexed libraries were filtered for read quality and PCR duplicates. Alignment was performed against the Drosophila melanogaster genome (iGenomes BDGP 5.25 assembly) with Bowtie2 (v2.2.5) using the following parameters: *-very-sensitive-local* (Langmead and Salzberg, 2012). Peak calling was performed using MACS2 (v2.1.0) with default parameters (Zhang et al., 2008). Peak visualizations were generated with Integrated Genome Browser (v8.4.2) (Nicol et al., 2009). Peaks present in both replicates were used for further analysis, unless otherwise noted. Overlap with HOT regions, chromatin accessibility data, and additional transcription factors was quantified as a minimum of 1bp overlap between a Gro peak and a feature. Motif enrichment analysis was performed with the MEME-chip software suite (Ma et al., 2014).

Embryonic RNA isolation and sequencing (RNA-seq)

Staged embryos were manually homogenized in TRIzol reagent (Invitrogen) and RNA was extracted according to manufacturer protocols. Purified RNA quality was assessed via Bioanalyzer 2100 (Agilent Technologies). Strand-specific polyA-selected libraries

were generated with TruSeq Stranded mRNA Library Prep Kit (Illumina) and sequenced on the Illumina HiSeq 2000 platform.

Transcriptome (RNA-seq) data preparation and genomic alignment

Reads were demultiplexed via custom scripts. Low quality reads were trimmed and remaining reads were aligned with TopHat2 (v2.0.9) (Kim et al., 2013) against the *Drosophila melanogaster* genome (iGenomes BDGP 5.25 assembly) with iGenomes gene models as a guide. Gene assignment was performed with HTSeq (IAnders et al., 2015).

Gene expression and Groucho target gene identification

Normalized gene expression values and differential expression analysis generated with DESeq2 (v1.8.0) (Love et al., 2014). Genes exhibiting a $\log_2$(fold-change) of magnitude 0.5 or later with a multiple-testing corrected p-value of < 0.05 were called as significantly differentially expressed. Genes exhibiting changes in expression in loss- and gain-of-function embryos were identified. For each Gro peak, the nearest or overlapping feature was identified as a potential regulatory target. These two sets were intersected by timepoint to give the high-confidence gene set.

Gro occupancy scores were calculated using a modified scoring algorithm published previously by Sandmann et al., 2007. For each gene, a Gro occupancy score was calculated as the sum of the scores of Gro peaks. Scores for each peak were calculated

on a per-base level and averaged. For each basepair overlapping the gene, a score of 1 was assigned. For each non-overlapping basepair, the score was calculated by

$$\frac{1}{1 + e^{0.0005*(d-15)}}$$

where *d* is the distance between the basepair and the nearest end of the gene.

Results

Groucho is dynamically recruited to thousands of sites throughout embryonic development

The time windows used for the analysis were chosen to overlap significant events in embryonic development that have known Groucho interactions. The first window (timepoint 1: 1.5 – 4 hours post-fertilization) encompasses formation of the syncytial blastoderm and subsequent cellularization. It is during this stage that the expression patterns of the pair-rule and segment polarity genes (including engrailed, a Groucho-interacting TF) are established, a defining step in anterior-posterior patterning. Specification of presumptive germ layers along the dorsal-ventral axis occurs during this stage, primarily guided by the activity of Dorsal in conjunction with Groucho. The second window (timepoint 2: 4 – 6.5 hours post-fertilization) encompasses the growth and segmentation of the germ band, including the formation of neuroblasts, a crucial early step in the onset of neurogenesis. The third window (timepoint 3: 6.5 – 9 hours post-fertilization) encompasses retraction of the germ band and fusion of the anterior and posterior midgut.

ChIP-seq was performed in duplicate on fly embryos representing each time point using an extensively validated affinity purified polyclonal antibody raised against the Gro GP domain. Sequencing libraries were sequenced to a depth that provided at minimum 5 million uniquely mappable reads, far in excess of the minimum recommended by modENCODE ChIP-seq best-practices (Fig. 2-1A) (Landt et al., 2012).

Replicates exhibited high reproducibility in terms of both read density and resulting peak model (Fig. 2-1B, left and right, respectively).

The high degree of correlation between our ChIP-seq data sets and ChIP-chip data sets obtained from 0-12 hour embryos (Negre et al., 2011) using completely independent antibodies also validates our ChIP-seq data (Fig. 2-2A). The modENCODE Groucho peaks were generated from 0 – 12 hour embryos and so should represent a time-averaged superset of our data. Collectively the ChIP-seq peaks from our three data sets identified 79% of the modENCODE ChIP-chip peaks. An additional 81% of our identified Gro binding sites are novel and are not represented in the data generated by the modENCODE consortium. Comparison of our ChIP-seq data with modENCODE Groucho ChIP-chip data generated from white pre-pupae also shows a significant overlap (Fig. 2-2B). A large fraction of embryonic and pre-pupal binding sites are unique to each stage, consistent with the distinct roles of Groucho-mediated repression during pupal development (de Celis and Ruiz-Gomez, 1995). Approximately a third of embryonic peaks are retained to some extent in this later stage, indicating Gro may be utilized in the regulation of a subset of common genes throughout multiple developmental stages.

Peak modeling identified widespread Groucho binding throughout the genome; peaks with overlapping regions between replicates were chosen for further analysis, as they represent a higher confidence subset of all identified peaks (Fig. 2-3A), and peaks overlapping input peaks were removed, as they are assumed to arise from erroneous

read alignment due to abundant or repetitive sequences. Groucho recruitment sites are most numerous during the central timepoint analyzed (5,246 binding sites), compared to the early (1,358) and late (4,232) stages. We detected 5,829 unique binding sites in total, with 535 sites recruiting Groucho across all timepoints, and therefore potentially participating in Grouch-mediated repression in at least one cell type or tissue throughout the developmental timeframe analyzed (Fig 2-3B).

Groucho occupancy is highly dynamic and reversible. Approximately 75% of all Groucho binding sites are unique to a single timepoint. The majority of the sites established during time window 1 that persist into time window 2 continue to persist into timepoint 3, indicating that some Groucho binding sites are utilized throughout early development. Interestingly, few sites are occupied in only the first and third timepoints, indicating that Groucho occupied sites during the first timepoint tend to either be utilized at all timepoints, or are only utilized very early in development and not utilized again in the windows analyzed.

Genome-wide analyses of transcription factor binding in the *Drosophila* embryo has revealed thousands of HOT (Highly Occupied Target) regions to which large numbers of unrelated factors bind concurrently (Consortium et al., 2010). While the cause and regulatory ramifications of these highly-occupied regions remain to be fully explored, they appear to be widespread in eukaryotes, persistent between cell types and developmental stages, and are often located in areas of active transcription (Moorman et al., 2006). Some factors can be recruited to HOT regions independently from their

abilities to bind and recognize DNA sequence (Li et al., 2008). Owing to this and the large number of Groucho-interacting proteins that either bind DNA directly or are otherwise recruited to chromatin, we expected that a significant fraction of Groucho binding sites would localize to these areas (Fig. 2-4). We observe that while the total percentage of Gro regions that overlap a HOT zone is largely invariant between time points, Gro in the 1.5 – 4 hr embryo preferentially localizes to regions with a higher HOTness (i.e. greater numbers of occupying factors), while 6.5 – 9 hr Groucho binding is enriched for overlap with lower HOTness regions.

The clearest theory on the function of the origin of these HOT regions, supported by *in vivo* and computational studies, is that many transcription factors are maintained at sufficiently high nuclear concentrations such that these factors saturate high-affinity binding sites, and as a result also bind to low and intermediate affinity sites in areas of high DNA accessibility (Kaplan et al., 2011; Li et al., 2008). DNA accessibility has been mapped across multiple developmental stages (Li et al., 2011), and Groucho binding is significantly enriched for these regions (Fig. 2-5). As Groucho is known to increase nucleosome density and reduce DNA accessibility (Sekiya and Zaret, 2007; Winkler et al., 2010), widespread recruitment to these sites indicates that association with chromatin may not be in itself sufficient to initiate Gro-mediated chromatin condensation, as this process possibly requires additional undocumented inputs.

Groucho tends to bind in spatially-restricted clusters at promoters and inside genes

Choosing the nearest or overlapping gene as a potential Groucho-regulated target, we see that there are significantly fewer Groucho-associated genes than there are Groucho binding regions (Fig. 2-6A), due to the tendency of Groucho to localize to multiple discrete regions around its potential targets. Half of all Groucho-associated genes predicted in this fashion have two or more Groucho peaks in relative proximity (Fig. 2-6B), with an average of 2.5 binding sites per associated gene (compared to an expected value of 1.5 binding sites per gene, $p < 10^{-10}$ via Monte-Carlo simulation). These peaks have median widths in the 500 – 700 bp range, indicative of point source peaks, as commonly seen for sequence-specific transcription factors (Ho et al., 2011), rather than the broad peaks typical of either highly polymeric factors or histone marks (Fig. 2-7). Interestingly, *in vitro* studies have shown that Grg3/repressor complexes bind to and protect DNA from nuclease activity over the span of 3 to 4 nucleosomes (Sekiya and Zaret, 2007), corresponding to 600 – 800 basepairs of protection, consistent with our observed mean peak width.

At all three timepoints, the distribution of peak widths exhibits a prominent tail of much wider peaks in the 1.5 to 2.5 kb range. This indicates that, consistent with previously proposed models, Groucho may be capable of spreading over relatively large regions of the genome. However, this does not appear to be a widespread mode of chromatin association. Average Groucho peak widths increase slightly at later timepoints, though whether this is indicative of a time-dependent change in the way Groucho interacts with chromatin or slight differences in library composition is unclear.

Groucho binding is enriched close to transcription start sites (Fig. 2-8A). The preference for start sites is somewhat unexpected given extensive evidence that Groucho is a long-range repressor (Barolo and Levine, 1997; Dubnicoff et al., 1997). Groucho sites exhibit a strong preference for binding within genes, with approximately 50% of peaks occurring within gene bodies across all timepoints (Fig. 2-8B).

Within gene bodies, Groucho exhibits a strong preference for binding within introns and UTRs, and is depleted for exon binding when compared to input (Fig. 2-9). Between 60 and 80% of all binding within genes occurs within introns, dependent on timepoint. Of all Groucho intronic binding sites, 40% fall within the first intron. This represents a more than 2-fold enrichment of binding preference for these introns, and is consistent with the observation that the first introns of *Drosophila* genes tend to be longer, more conserved, and more sensitive to mutation than subsequent introns, and are therefore predicted to be enriched for regulatory elements (Bradnam and Korf, 2008).

Motif analysis of Groucho recruitment sites identifies a small number of transcription factor binding motifs enriched at each timepoint, including several factors known to interact with Groucho, including Ventral nervous syndrome defective (vnd), Sloppy paired 1 (slp1), Hairy (h), Huckebein (hkb), and Brinker (brk) (Fig. 2-10). Enrichment of motifs varies by timepoint as well as by the location of the Groucho binding site. The majority of factors analyzed exhibit stronger enrichment for Groucho sites within genes, which can be explained by a smaller group of regulators being

responsible for Groucho recruitment within genes, or fewer low-affinity binding sites recruiting Groucho in these regions.

Groucho is recruited to VRRs in Dorsal-repressed genes, but extensive spreading does not occur

In the early embryo, delineation of the dorsal-ventral axis is accomplished through transcriptional changes arising from a maternally-defined gradient of nuclear Dorsal (DL) along this axis (Roth et al., 1989). In ventral and ventrolateral regions of the embryo, Dorsal facilitates the repression of numerous genes, including *zerknullt* (*zen*), *decapentaplegic* (*dpp*) and *tolloid* (*tld*) through its interaction with Groucho, a critical step in delineating presumptive mesodermal and neuroectodermal regions (Dubnicoff et al., 1997; Kirov et al., 1994) . As a way of assessing the simple model that Gro recruitment by Dorsal leads to ventral repression, I examined the patterns of Gro binding to these three ventrally repressed targets. Since ventral repression is an early event, I focused primarily on my earliest developmental time point (1.5-4 hours).

Ventral repression of *zen* is established through Dorsal recruitment to a well-characterized ventral repression region (VRR) between 1.1 to 1.4 kb upstream of the transcription start site. This region contains four Dorsal binding sites, as well as AT-rich regions responsible for the recruitment of Cut (*ct*) and Dead ringer (*dri,* also known as Retained, *retn*) (Valentine et al., 1998). Through the cooperative action of these factors, Groucho is thought to be recruited to establish repression. ChIP-seq data confirms that

Gro localizes to regions surrounding the VRR. Surprisingly, Gro density is comparatively weak within the VRR region itself and is instead primarily observed both upstream and downstream of the VRR (Fig. 2-11A). This suggests the possibility of limited spreading away from the site of Dorsal-mediated recruitment. At later timepoints, binding to the regions surrounding the VRR is lost, although *zen* remains transcriptionally repressed throughout most of the embryo.

Dorsal is additionally responsible for ventral repression of *decapentaplegic* (*dpp*) in early embryos (1.5 – 2 hours post fertilization) through the recruitment of Gro, and loss of Gro activity at this stage results in complete derepression of *dpp* in ventral regions of the embryo (Dubnicoff et al., 1997). Dorsal binding sites necessary for restriction of *dpp* expression map to a VRR in the gene's second intron (Huang et al., 1993). Our ChIP-seq data confirms extensive Gro recruitment to this site (Fig. 2-11B) in the early embryo. Similarly to what is observed with *zen*, Gro disappears from the VRR at later timepoints.

Three Dorsal binding sites identified upstream of the *tolloid* gene are responsible for the Dorsal-mediated repression of *tolloid* in ventral regions of the early embryo. A region containing two of these sites functions as a VRR (Kirov et al., 1994). Groucho ChIP-seq data indicates that Groucho associates strongly in an asymmetric peak centered on the central Dorsal binding site, approximately 400 bp upstream of the *tolloid* TSS (Fig. 2-11C). While the peak persists through all three time windows, its intensity continuously decreases with time.

Thus, while the details vary, Groucho associates with the VRRs in all three genes during the developmental time frame when the gene is being actively repressed, supporting a model whereby Groucho is recruited specifically to genes by Dorsal to spatially restrict expression. These findings are not, however, consistent with a model involving extensive Gro spreading. This is especially apparent in the case of *dpp*, where I observe binding of Gro in a relatively discrete peak over the intronic VRR and a weaker Gro peak overlapping the transcriptional start site, perhaps indicative of looping.

Groucho localizes extensively to the Dorsal-binding sites of both Dorsal-activated and –repressed genes

In addition to repressing multiple genes in the ventral portion of the embryo, Dorsal can activate genes in both ventral and ventrolateral regions of the embryo in a context-dependent manner. The transition of Dorsal from an activator to a repressor has been ascribed to the presence of adjacent binding sites for additional factors, such as Deadringer and Cut, that could facilitate the association of Groucho with Dorsal, resulting in Groucho-mediated long-range repression (Valentine et al., 1998). The necessity of these factors in generating a stable Dorsal/Groucho interaction is thought to arise from the relatively low binding affinity of Groucho for Dorsal, when compared to factors to which Groucho binds without requiring assistance, such as Engrailed or Brinker (Ratnaparkhi et al., 2006). Due to the inherent weakness of the Dorsal/Groucho interaction, it is not suspected that Groucho would ubiquitously colocalize with Dorsal,

and would instead only associate at those loci at which Dorsal functions as a repressor. Our Groucho ChIP-seq data, however, shows that that is not strictly the case.

In ventral regions of the embryo, Dorsal serves to activate several genes, the two most well-studied being *twist* and *snail,* two transcription factors essential to the specification of the presumptive mesoderm (Ip et al., 1992; Thisse et al., 1987). Dorsal activates both *twist* and *snail* by binding to Ventral Activation Regions (VARs) in the 5' flanking regions of these genes (Ip et al., 1992). No role for Groucho has been identified in the regulation of either gene. Surprisingly, however Gro binds the VARs of both genes in early embryos. We observe extensive Gro binding to both the primary and "shadow" VARs in *snail* (Figure 2-12A), and weaker binding to a VAR in the 5' flanking region of *twist* (Figure 2-12B). Thus, Gro recruitment may not be the critical step in converting Dorsal from an activator to a repressor.

To explore this question further, we looked more broadly at localization of Gro to Dorsal binding sites. These sites can be subdivided into three classes dependent on the resulting expression pattern of the regulated gene (Biemar et al., 2006; Zeitlinger et al., 2007). Class I sites, which are low affinity sites, result in gene expression in the most ventral regions of the embryo (presumptive mesoderm), where Dorsal concentrations are highest. Class II sites are generally of higher affinity than class I sites and are frequently found adjacent to binding sites for other factors (such as bHLH factors) that enable Dorsal to activate transcription at lower concentrations. As a result, these sites are active in in ventrolateral regions (neuroectoderm), an area with intermediate levels

of nuclear Dorsal. Class III sites are associated with genes that are repressed by Dorsal and whose expression is thereby restricted to the dorsal ectoderm. In accord with what we observed at *snail* and *twist* VARs, Groucho is not restricted to the class III sites but is found at all three types of sites (Fig. 2-14A). No single class of Dorsal site is significantly enriched over the others, indicating that Groucho binds to Dorsal more frequently than previously surmised, even at sites where Dorsal is activating transcription.

As Groucho requires additional factors to facilitate interaction with Dorsal, we calculated the combinatorial overlap of each Groucho binding segment with the binding patterns of 25 transcription factors derived from 2 – 4 hr embryos (MacArthur et al., 2009). A factor heatmap of the hierarchically clustered Groucho binding regions reveals two major classes of Groucho binding sites. The first class is characterized by extensive overlap with six factors: Dorsal, Dichaete, Medea, Twist, Daughterless, and Kruppel, and a lesser degrees of overlap with additional assayed factors (Fig. 2-15). While Dorsal is a well-studied Groucho-interacting protein, the degree to which Groucho colocalizes with Dorsal is surprising, given that there are at minimum thirteen other factors capable of recruiting Groucho in processes thought to be Dorsal-independent (Mannervik, 2014). The second major class of Groucho binding site, comprising ~25% of Groucho sites in the early embryo, lacks overlap with any of the assayed transcription factors. This apparent high-level segregation of Groucho recruitment sites has multiple interpretations. Given that overlap was only calculated against 25 of the estimated ~700 transcription factors contained in the *Drosophila* genome (Adams et al., 2000), there could exist factors, or entire classes of factors, to which Groucho is being recruited that have yet to be

identified or assayed in the early embryo. It's also possible that some of these sites represent recruitment of Groucho to chromatin in a manner not dependent on additional factors, for example through interaction with histones, perhaps after delivery to a site by DNA looping.

Identification of Groucho Targets by Developmental Stage

To incorporate our picture of Groucho binding into a framework of Groucho-mediated repression, we analyzed the transcriptomes of staged embryos expressing multiple dosages of Groucho. These included fly lines maternally overexpressing Groucho at two levels, approximately two-fold and four-fold higher than endogenous, as well as a line overexpressing a Groucho deletion mutant lacking the central SP domain (GroΔSP)(Turki-Judeh and Courey, 2012b). Overexpression of a deletion variant of Groucho lacking the SP domain was found to result in faulty targeting and ectopic repression of multiple non-Groucho target genes t(Turki-Judeh and Courey, 2012b), a trend that we sought to investigate on a genome-wide scale (Turki-Judeh and Courey, 2012b). Additionally, we analyzed the transcriptome of embryos lacking maternally-contributed functional Groucho. These embryos are derived from maternal germline clones homozygous for gro^{MB36}, a lethal allele that introduces an ectopic splice site near the 5' end of *gro* (Jennings et al., 2007). The resulting transcript codes the initial 12 amino acids of Groucho followed by ~100 amino acids derived from frameshifted sequence. The allele produces no detectable Groucho protein, and results in severely

decreased levels of transcript, presumably due to nonsense-mediated mRNA decay. Analysis of Gro transcript levels across samples at each timepoint confirms overexpressing lines accumulated increased transcript levels, with the effect being greatest at the first timepoint (Fig. 2-17A). This excess transcript is partially cleared from the embryo by later timepoints, but does not fully return to wild-type levels over the time span analyzed. Groucho loss-of-function embryos failed to accumulate Gro transcripts to any significant degree across all timepoints. Wild-type embryos exhibit the expected pattern of initially high levels of maternally-deposited transcript, which are gradually reduced as development proceeds (Fig. 2-17B).

Clustering of RNA-seq profiles by similarity reveals the transcriptomes cluster first by timepoint, then by Groucho dosage (Fig 2-18). Groucho loss-of-function samples segregate well from wild-type and overexpression samples, while cluster discrimination between wild-type and overexpression is relatively weak, indicating that loss-of-function embryos exhibit a greater degree of transcriptome deviation from all other samples. Groucho loss-of-function samples from the second and third timepoints cluster independently from all other samples at those two timepoints, indicative that accumulated differences in gene expression have put these embryos on a highly divergent and non-viable developmental trajectory (Fig. 2-18, red box).

Principal component analysis (PCA) allows a more detailed dissection of transcriptome profile changes between Groucho dosages, and how those changes evolve over time (Fig. 2-19). PCA is a common technique used to visualize high-

dimensionality data in two dimensions; linear distance between two points is directly proportional to the dissimilarity between those samples. PCA analysis reveals two sources of variance between samples: developmental stage on the x-axis, and Gro dosage on the y-axis, fitting with the major determinants of hierarchical clustering seen in the previous correlation heatmap. Comparison of the overexpression lines with the wild-type embryos shows that while these samples exhibit overall high similarity at early timepoints (upper-left cluster), overexpression samples grow increasingly distinct from wild-type over time, as can be seen by the divergence of these points from the wild-type sample (in red), despite the diminished difference of Gro transcript levels at later time points.

Perturbation of Groucho levels results in the misregulation of a significant proportion of the Drosophila genome over each timespan (Fig 2-20A). The Groucho loss-of-function phenotype was more severe than that obtained from overexpression, with over 10% of expressed genes exhibiting significant changes in expression level at each timepoint, with the greatest effect seen in the second, 4 to 6.5 hour stage (Fig. 2-20B). Overexpression samples exhibit a smaller yet still significant proportion of differentially expressed genes, with between 2 and 16% of the expressed genome undergoing differential expression, with the strongest effect seen at the final, 6.5 to 9 hour stage. Comparison of differentially expressed genes in the three Gro overexpression lines reveals significant correlation between activation or repression of genes regardless of Groucho dosage, with this effect holding across all timepoints (Fig. 2-21).

As Groucho is known to restrict the expression patterns of many developmental regulators including transcription factors, splicing factors, and signaling molecules (e.g., tailless, huckebein, zen, Sxl, dpp, etc.), it is suspected that many of these potential Groucho targets are secondary targets of Groucho and are not regulated by direct Groucho occupancy of their regulatory regions. In order to reduce the inclusion of these secondary effects in our determination of Gro targets, we refined the list of potential Groucho targets using two methodologies.

The first method sought to identify genes both sensitive to multiple levels of Groucho dosage and bound by Gro internally or in adjacent intergenic space. Both sources of data are noisy by nature, as secondary effects could account for the dosage response and Groucho can regulate genes from regulatory regions many kilobases away. First, we focused on genes that exhibit a response of an opposite sign in the loss-of-function and one or both Gro overexpression lines (i.e. up-regulated under conditions of lowered Gro dosage and down-regulated under increased dosage, or vice-versa). This results in a significant restriction of the effected gene list at each timepoint (Fig. 2-22). Secondly, we narrowed this list to only those genes associated with adjacent or overlapping Groucho binding, as determined by ChIP-seq. The resulting gene list is significantly reduced, consisting of 248 genes, of which 151 are identified by comparisons of both full-length Gro overexpression lines to the loss-of-function line (Fig 2-23 & Supplemental Table 1).

The requirement that genes exhibit differential expression under multiple Groucho dosages may be an overly stringent criterion, as it would only capture the set of genes expressed at nominal levels in wild-type embryos and therefore capable of being both up- and down-regulated. Therefore, we utilized an additional method to explore the relationship of Groucho occupancy and regulation. This method involves the use of a scoring algorithm to quantify the predictive power of Groucho binding on changes in expression. A similar procedure has been successfully utilized to predict the targets of CBP, a coactivator that cooperates with Dorsal to activate gene expression in the early embryo, by incorporating CBP ChIP-seq data and a measurement of a mutant CBP transcriptome (Holmqvist et al., 2012). Similar methodologies have been utilized to integrate transcription factor binding and expression data in other contexts (Wang et al., 2013). We modified this method to allow for greater contribution of more distant binding to a gene's score. On a per-gene basis, a "Groucho occupancy score" was calculated taking into account the number, size, and positioning of any Groucho peaks. Operating under a progressively relaxing score cutoff, the number of genes captured with scores above said cutoff that are up- or down-regulated upon Groucho level perturbation were counted (Fig. 2-24). The inflection point of the resulting response curves can than be used as an empirically-derived threshold for classifying Groucho target genes.

We find that the changes in gene expression resulting from Groucho overexpression are significantly more predictive of regulation than changes resulting from loss of Groucho activity (Fig. 2-24B/C). Very few up-regulated genes are captured

by the response curve in overexpressing lines, especially at early timepoints. In *gro*^{MB36} embryos, a slight enrichment of derepressed genes is evident during the first two time spans with clear inflection points (Fig. 2-24A).

Though the Groucho/TLE family of proteins have traditionally been thought of as obligate repressors, TLE3, a human Groucho ortholog, was recently shown to primarily serve as an activator, though the mechanism remains unknown (Villanueva et al., 2011). Additionally, CtBP, a canonical, short-range *Drosophila* corepressor, was shown to serve as a co-activator of certain Wnt-regulated genes, this switch in behavior being controlled by the protein's oligomeric state (Bhambhani et al., 2011). However, the observed asymmetry in the distribution of up- and down-regulated genes between the loss-of-function and overexpression lines can be taken as evidence against Groucho behaving as an activator. Very few high-scoring genes were up-regulated in either overexpression line compared to repressed genes. Additionally, no clear inflection point is present in these up-regulated gene response curves, indicating that high Groucho occupancy is only loosely predictive of gene activation. Though we cannot rule out the possibility that Groucho can serve as an activator under limited and thus far undetected circumstances, we take these two observations as evidence against a widespread role of Groucho as a coactivator.

Through this scoring methodology, we identified 351 potential Groucho target genes across all timepoints. Of these, 90 were also identified by the Groucho dosage-sensitivity analysis. While this overlap is highly significant (*p-value* $<10^{-10}$,

hypergeometric test), the two results do differ substantially. Lacking compelling *a priori* justification to favor one method over the other, we investigated aspects of each data set individually.

Genes in both sets are enriched for transcription factors and factors involved in fly development (Fig. 2-26). In both sets, transcription factors are the most heavily enriched ontology. To identify potentially undocumented processes and regulatory networks in which Groucho may be involved, we annotated each set of potential target genes with genetic and physical interactions curated by FlyMine (Lyne et al., 2007) and integrated these results into a network to search for overrepresented groups of co-regulated genes (Fig 2-27). Both networks exhibit a large core network comprising multiple interconnected hubs corresponding to components of signaling pathways. Both networks contain multiple E(spl)-family proteins, which Groucho is known to repress in the embryo. Delta (Dl) is a transmembrane ligand of the Notch (N) signaling pathway, and complete activation of this pathway requires both Groucho and E(spl)-family proteins (Heitzler et al., 1996). Atonal (ato) and Sprouty (sty) are factors with known functions in respiratory and eye development, respectively (Hacohen et al., 1998; Jarman et al., 1994), in which Groucho's potential roles have not been investigated.

The core regulatory network of targets identified by Groucho occupancy score is somewhat larger and encompasses additional regulatory hubs (Fig. 2-27B). These hubs primarily correspond to components of multiple signaling pathways, including Decapentaplegic (dpp), Wingless (wg), and Ras/MAPK (Egfr and aop). Pannier (pnr) is a

transcription factor activated by Dpp signaling and involved in dorsoventral patterning and cardiogenesis (Herranz and Morata, 2001). Groucho is recruited to Tinman, a Pannier-interacting protein, to regulate cardiac gene expression (Choi et al., 1999). The association and regulation of multiple Pannier target genes by Groucho may represent a significant contribution by Groucho to the regulation of cardiac development.

Discussion

Temporal and spatial patterns of Groucho binding

In our current study, we have identified thousands of novel Groucho-recruitment sites throughout the Drosophila genome. The majority of these sites are detected in a single developmental window, indicating that Gro is often transiently recruited to facilitate repression. This effect is stronger at the earliest stages of development, in which only a small percentage of Gro binding sites are preserved between the 1.5 – 4 and 4 – 6.5 hr stages. These widespread changes in Gro occupancy are indicative of the changing roles of Gro throughout development, as the shifting availability of sequence-specific transcription factors modulates Gro recruitment to chromatin.

We observe that Gro is recruited to the Dorsal-binding regulatory modules of three ventrally-repressed genes, consistent with Dorsal-mediated recruitment and repression. Gro occupies multiple distinct regions within and surrounding two of these genes, as well as at the transcription start sites of all three. We find this trend extends globally to Groucho-associated genes, with the majority of Gro binding in clusters of multiple localized peaks less than 1 kb in width. As Gro tetramers can crosslink chromatin arrays *in vitro* (Sekiya and Zaret, 2007), the presence of these peak clusters may represent the extension of this function to *in vivo* contexts.

Many Gro binding sites correspond to regions of high chromatin accessibility and/or HOT regions occupied by several additional transcription factors. These highly-accessible sites are thought to result from low-affinity interactions between

transcription factors and DNA and generally lack regulatory potential (Fisher et al., 2012). Some factors bound at these are biochemically activate, however, as a recent study found Hairy binds extensively to these presumably inert regions, resulting in widespread but largely non-regulatory H3 and H4 deacetylation, likely through association with Groucho (Kok et al., 2015). Extensive localization of Gro to these sites provides additional evidence that these deacetylation events are Gro-mediated, and that Gro-mediated chromatin modification is not always in itself sufficient to modulate gene repression.

Enrichment of Gro binding within genes

Global analysis of Groucho occupancy additionally reveals that Groucho binding is strongly enriched for binding within genes, specifically within introns, with the highest enrichment exhibited in the 5' intron of genes. Overexpression of Groucho resulted in 10 to 32% of genes bound in this manner by Gro to become repressed, dependent on timepoint, reinforcing that Groucho binding within genes is a common strategy of Groucho regulation. Multiple factors, including Kruppel and Twist have been shown to commonly localize to intronic regions (Matyash et al., 2004; Sandmann et al., 2007; Zeitlinger et al., 2007). The regulatory logic behind intronic cis-regulatory modules is a matter of some debate, as there are significant energetic costs associated with intron maintenance during replication, transcription, and splicing, as well as a regulatory cost in terms of a longer lag-time between transcriptional activation and mature mRNA

formation (Yenerall et al., 2011). Consistent with this hypothesis, genes poised for rapid activation during development have significantly higher frequencies of intron loss (Jiang et al., 2014). One explanation of the regulatory rational behind intronic repressor binding comes from the observation of a significant lag between Snail binding to silencing elements and complete repression of target genes. This is due to the inability of the repressor to affect active polymerases downstream from the promoter region (Bothma et al., 2011). Due to the relatively slow rate of progression of PolII (~ 1.1 to 1.5 kb per min in *Drosophila*), this lag time can become significant, especially under developmental contexts in which precise temporal control of gene expression is required (Ardehali and Lis, 2009; Biemar et al., 2005). Studies have shown that changes in chromatin structure propagate across gene lengths at rates considerably faster than the rate of PolII processivity (Petesch and Lis, 2008). As Gro recruitment has been shown to spread chromatin marks throughout extended regions of target genes (Li and Arnosti, 2011), direct association of Gro with sites within genes may represent a common motif enabling more rapid gene inactivation as opposed to recruitment to distant regulatory elements.

The mechanism of Groucho-mediated repression

Gro interacts with the histone deacetylase HDAC1/Rpd3, leading to localized deacetylation of histones and a consequent increase in nucleosome density and repression (Winkler et al., 2010). Gro recruitment can result in deacetylation of H3 and

H4 histone tails distantly from the site of recruitment, an observation that led to the hypothesis that Gro itself spreads throughout chromatin (Kok et al., 2015; Martinez and Arnosti, 2008). As data presented here suggests Gro does not appear to bind continuous stretches of chromatin in the embryo, we propose that crosslinking could function as a mechanism to transfer these histone marks onto sites distant from Gro recruitment. This crosslinking is potentially mediated by Gro self-association through interactions of the Q-domain. Mutations to this domain which disrupt self-association result in misregulation of a subset of Gro targets (Kaul et al., 2014). This differential requirement for oligomerization can be explained by our observation that Gro frequently localizes within genes and near transcription start sites in the embryo, where the dependence on efficient oligomerization-mediated transfer of histone modification would be reduced in comparison with recruitment to distant silencing regions. Work presented in the next chapter will provide evidence that Gro-mediated repression positively correlates with stalled RNA PolII in the embryo, which may represent another method of transcriptional silencing.

Groucho and Dorsal

Gro is essential for correct determination of cell fates along the dorsal-ventral axis through cooperation with Dorsal and other DNA-binding factors. Dorsal functions as either an activator or repressor through interactions with multiple coregulators (Dubnicoff et al., 1997). Dorsal is thought to recruit Gro only with the cooperation of

additional transcription factors, such as Deadringer (Dri) and Cut (ct), to silence a subset of all Dorsal targets in ventral portions of the embryo (Valentine et al., 1998). We find that Gro is ubiquitously associated with Dorsal regulatory elements, regardless of whether Dorsal is serving as an activator or repressor. Factors thought to assist in strengthening the Dorsal/Gro interaction may instead serve as positive regulators of Gro activity though an alternate, unknown mechanism. Dorsal contains an eh1-like motif that is thought to weakly associate with Gro (Flores-Saaib et al., 2001), and mutation of this sequence to a WRPW motif converts Dorsal to a constitutive repressor (Ratnaparkhi et al., 2006). Recent studies have indicated that these recruitment motifs, in addition to having differing affinities for Gro binding, may cause Gro to adopt different conformations with different regulatory potentials, in some cases possibly resulting in the conversion of Gro from a long-range to a short-range repressor (Payankaulam and Arnosti, 2009). This is supported by crystal structures of the TLE WD domain in complex with WRPW and eh-1 motifs, which bind to the domain in distinct conformations which may result in more significant conformational changes of Gro (Jennings et al., 2006).

The Groucho regulatory network

Perturbation of Groucho activity has severe consequences on the embryonic developmental program. We observe hundreds of misregulated genes at each developmental stage, confirming that Gro is thoroughly integrated into the gene regulatory network. This network is highly sensitive to increased Gro dosage, indicating

that endogenous Gro is not expressed at levels that result in saturated interaction with DNA-bound repressors. These potential Gro targets were filtered using combinations of RNA-seq and ChIP-seq data to obtain lists enriched for direct targets of Gro repression. This list contains 509 genes regulated by Gro at one or more stage in the embryo. Gro targets are enriched for transcription factors controlling multiple aspects of gene expression, explaining how altering Gro levels can generate widespread changes in gene expression. The Gro regulatory targets identified here confirm that Gro regulates both upstream and downstream elements of a highly-interconnected network of signaling pathways. We identified multiple pathways with known Gro involvement, including Dpp, Wingless, and EGFR signaling, as well as novel involvement with downstream effectors of these pathways, such as Pannier, Atonal, and Patched.

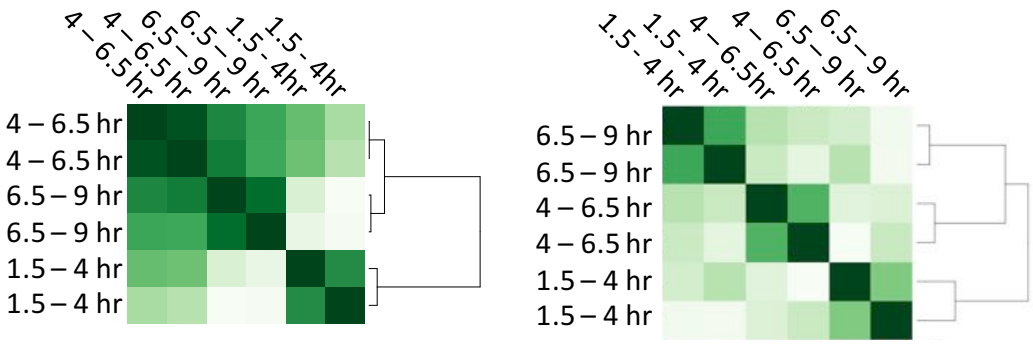
Figure 2-1. Groucho ChIP-seq experiments show high reproducibility in read mapping and peak calling (A) ChIP-seq libraries were sequenced to a depth of ~20M reads, twice the recommended library sizes for ChIP-seq experiments proposed by the modENCODE consortium (Landt et al., 2012). **(B)** *(left)* Overall mapping profiles of ChIP-seq sequenced reads cluster by timepoint. Timepoint 2 and 3 samples cluster more closely together than timepoint 1, which diverges significantly from both other timepoints. Dark green indicates a higher correlation by Spearman's rank correlation coefficient (a value of 1 indicates perfect correspondence). *(right)* Peak calling was performed with MACS2 and called peaks were clustered by similarity. **(C)** Replicate similarity was confirmed using principal component analysis

Fig 2-1

A

Timepoint	Replicate	Total Reads	Uniquely Aligned	Multiply Aligned	Total Aligned	Overall Alignment Rate
Input	A	26,804,524	19,349,084	4,859,241	24,208,325	90%
Input	B	22,770,618	16,367,817	4,175,412	20,543,229	90%
1.5 - 4 hr	A	22,276,144	8,242,625	2,487,432	10,730,057	48%
1.5 - 4 hr	B	23,743,733	5,992,455	1,974,028	7,966,483	34%
4 - 6.5 hr	A	19,973,255	7,925,951	2,463,045	10,388,996	52%
4 - 6.5 hr	B	17,755,470	8,440,270	2,461,968	10,902,238	61%
6.5 - 9 hr	A	18,066,141	7,979,752	2,324,797	10,304,549	57%
6.5 - 9 hr	B	22,367,998	8,179,454	2,740,544	10,919,998	49%

B



C

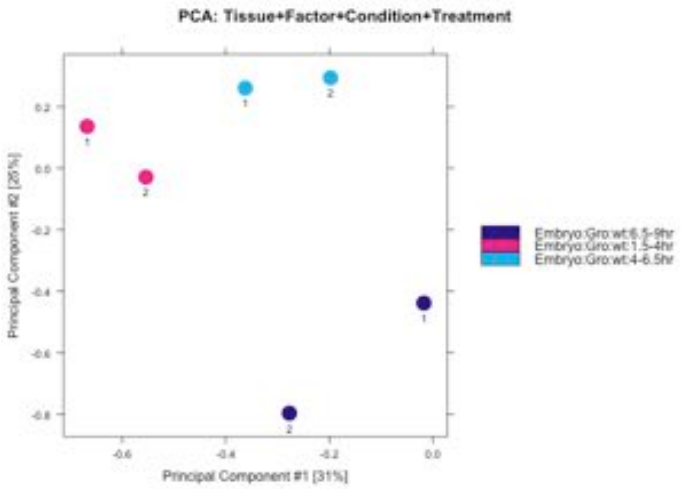


Figure 2-2. The Groucho binding pattern exhibits significant overlap with Groucho ChIP-chip determined peaks utilizing an independently-derived antibody. Significant Groucho peaks were compared to two sets of publicly-available Groucho ChIP-chip data performed on 0 - 12 hour embryos generated using a polyclonal antibody raised against a portion of the Groucho Q domain. The modENCODE data encompasses a timespan beginning 1.5 hours prior to our timepoints, and ending 3 hours afterwards. The degree of overlap is strongest at later timepoints, with the 6.5 - 9 hour data overlapping 68% of all modENCODE binding regions. Comparison of embryonic Groucho binding with modENCODE Groucho ChIP-seq data generated from white pre-pupae reveals that a small subset of embryonic Groucho-bound regions are bound during later development. The majority of Gro bound regulatory regions are unique to each developmental stage. The role of Gro in regulating gene expression during pupal stages, especially in tissue differentiation originating from imaginal discs is well documented, specifically the intepretation of a Brinker gradient arising across the anterior-posterior axis of the wing disc (Hasson et al., 2001).

Fig. 2-2

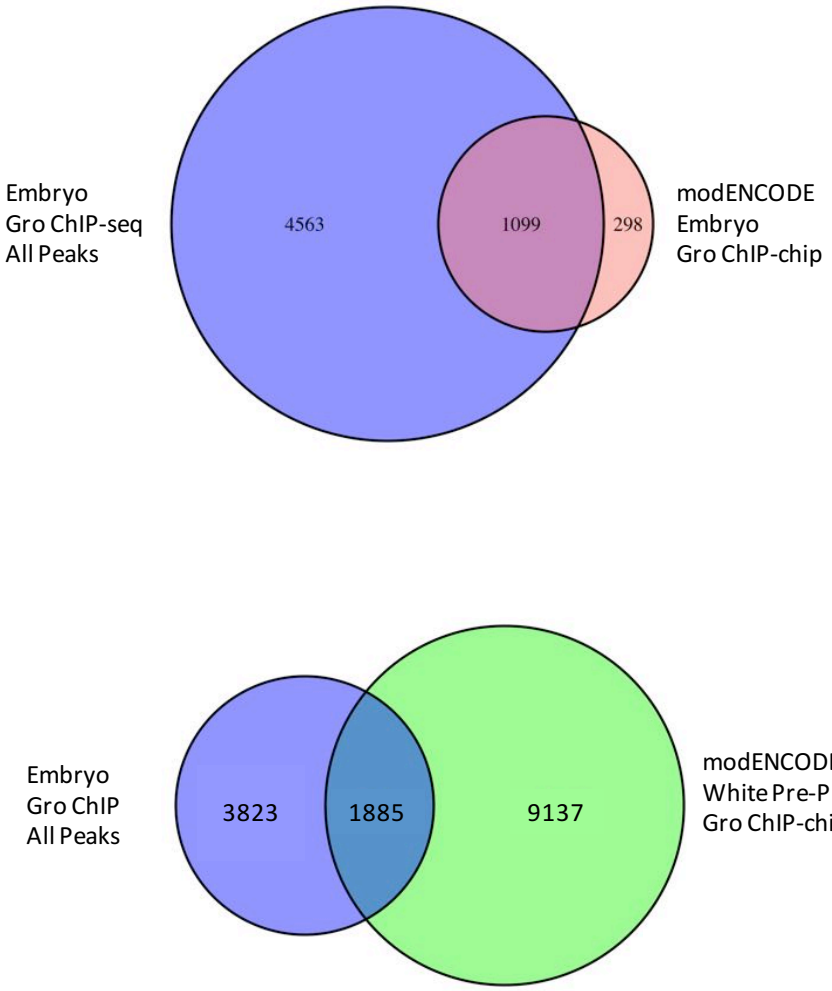

Embryo
Gro ChIP-seq
All Peaks
4563
1099
298
modENCODE
Embryo
Gro ChIP-chip
Embryo
Gro ChIP
All Peaks
3823
1885
9137
modENCODE
White Pre-Pupae
Gro ChIP-chip

Figure 2-3. A large number of Groucho binding regions are preserved throughout all stages analyzed. (A) Peaks present in both replicates were obtained from each timepoint and used for further analysis. Overlapping peaks were identified as those having at least 1 basepair overlap with a peak between replicates. Samples exhibited a replicate overlap rate of approximately 35% of all peaks, for the first and third timepoint, and 61% for the middle timepoint. **(B)** While a large fraction of Groucho binding sites is unique to each of the three timepoints analyzed, are preserved across two or more timepoints. No detected Groucho peak was present in only early and late timepoints, indicated that during the timepoints analyzed removeal of Groucho binding from a locus was a permanent regulatory decision. Additionally, while the middle and late timepoints have a significant fraction of binding sites in common, the early and middle timepoints have very few in common. This is indicative of Groucho genomic localization being relatively dynamic during early timepoints when compared to later times.

Fig. 2-3

A

	Number of Binding Sites
1.5 - 4 hr	1085
4 - 6.5 hr	4764
6.5 - 9 hr	3157

B

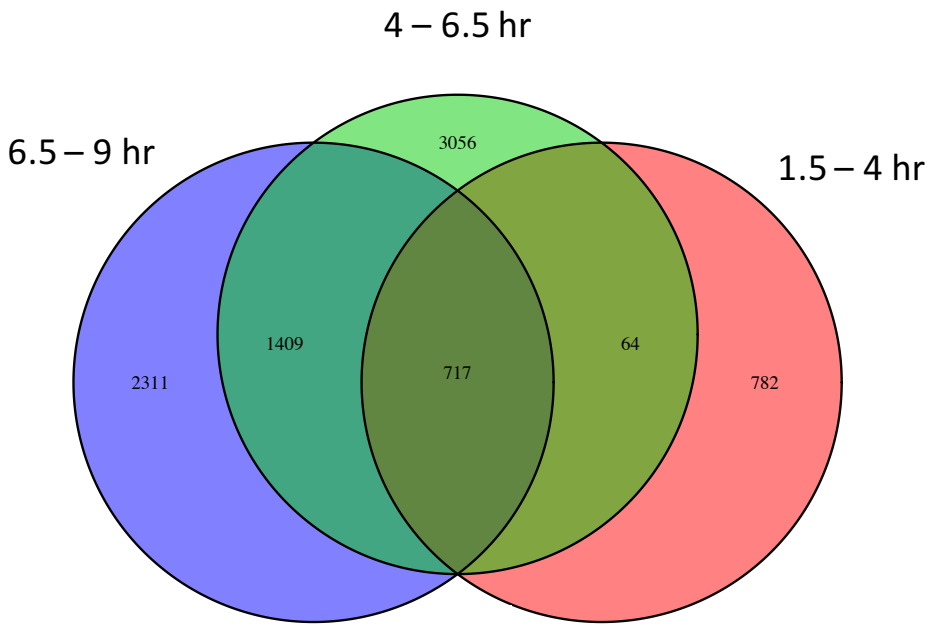


Figure 2-4. Over half of Groucho localizes to highly-occupied target (HOT) regions at all time windows assayed. At earlier timepoints, Groucho peaks prefer to localize to HOT regions with higher average scores, indicative of more colocalizing transcription factors. As development proceeds, Groucho becomes increasingly associated with less-occupied HOT regions. This could represent an expansion of the regulatory program of Groucho as the proliferation of cell and tissue types brings Groucho to more specialized regulatory targets.

Fig. 2-4

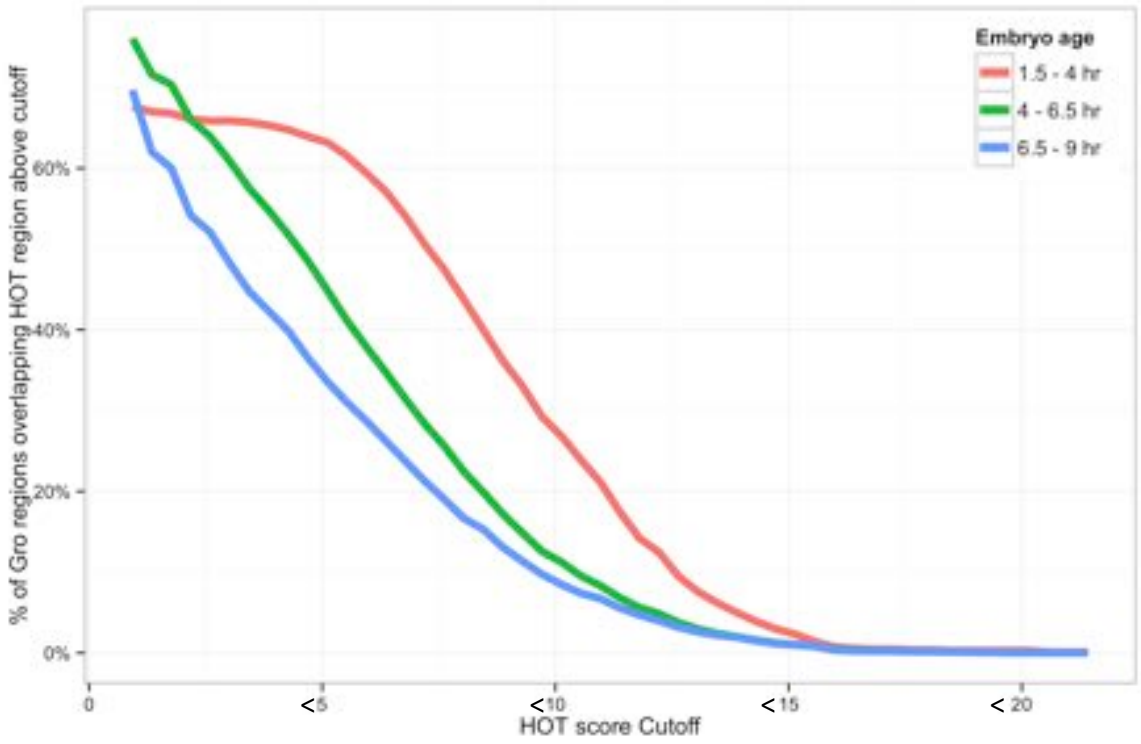

Embryo age
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
% of Gro regions overlapping HOT region above cutoff
60%
40%
20%
0%
0
<5
<10
<15
<20
HOT score Cutoff

Figure 2-5. Groucho frequently localizes to regions of high chromatin accessibility. Previously published developmental stage-specific chromatin accessibility data (from Li. et al., 2011) was intersected with Groucho ChIP-seq data across all timepoints. The percentage of Groucho binding sites that are found within high-accessibility regions was calculated for each pair of data sets (white: low % overlap, purple: high % overlap). The correspondence between Groucho ChIP-seq samples and stages of development is represented by grey boxes (bottom). A pattern of strong enrichment of Groucho binding within these regions is observed in all three assayed developmental stages.

Fig. 2-5

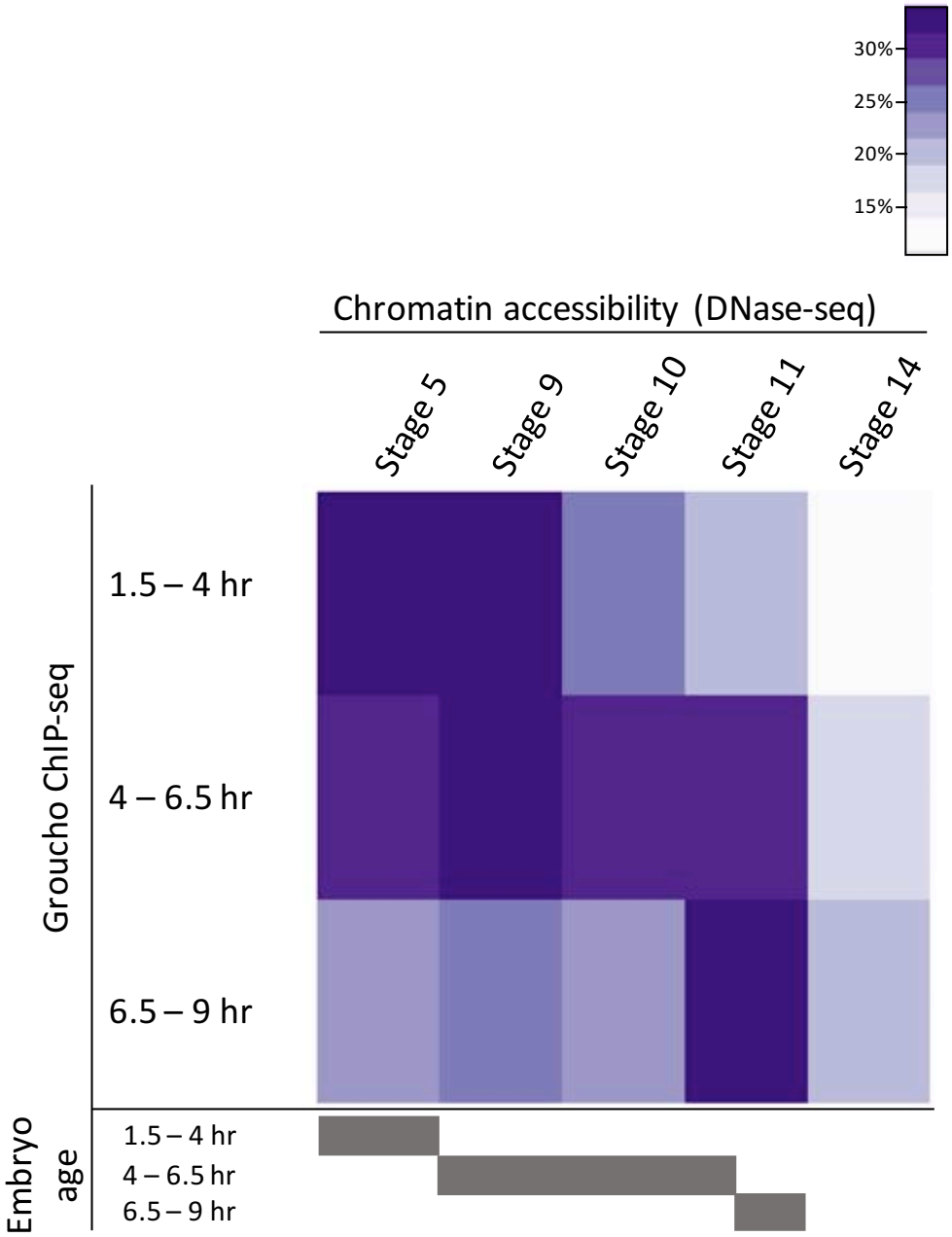

30%
25%
20%
15%
Chromatin accessibility (DNase-seq)
Stage 5
Stage 9
Stage 10
Stage 11
Stage 14
Groucho ChIP-seq
1.5 – 4 hr
4 – 6.5 hr
6.5 – 9 hr
Embryo age
1.5 – 4 hr
4 – 6.5 hr
6.5 – 9 hr

Figure 2-6. Most Groucho bound genes are associated with two or more distinct Gro peaks. (A) Overlap of Groucho-associated genes reveals Groucho binds adjacent to or overlapping hundreds of genes at each timepoint, with a significant number (457) being bound throughout the developmental stages assayed. **(B)** Over half of all Groucho bound genes exhibit two or more distinct Groucho peaks. These situations represent Groucho being recruited to multiple sequence-specific transcription factors or topological rearrangements which bring Gro in contact with multiple genomic loci.

Fig. 2-6

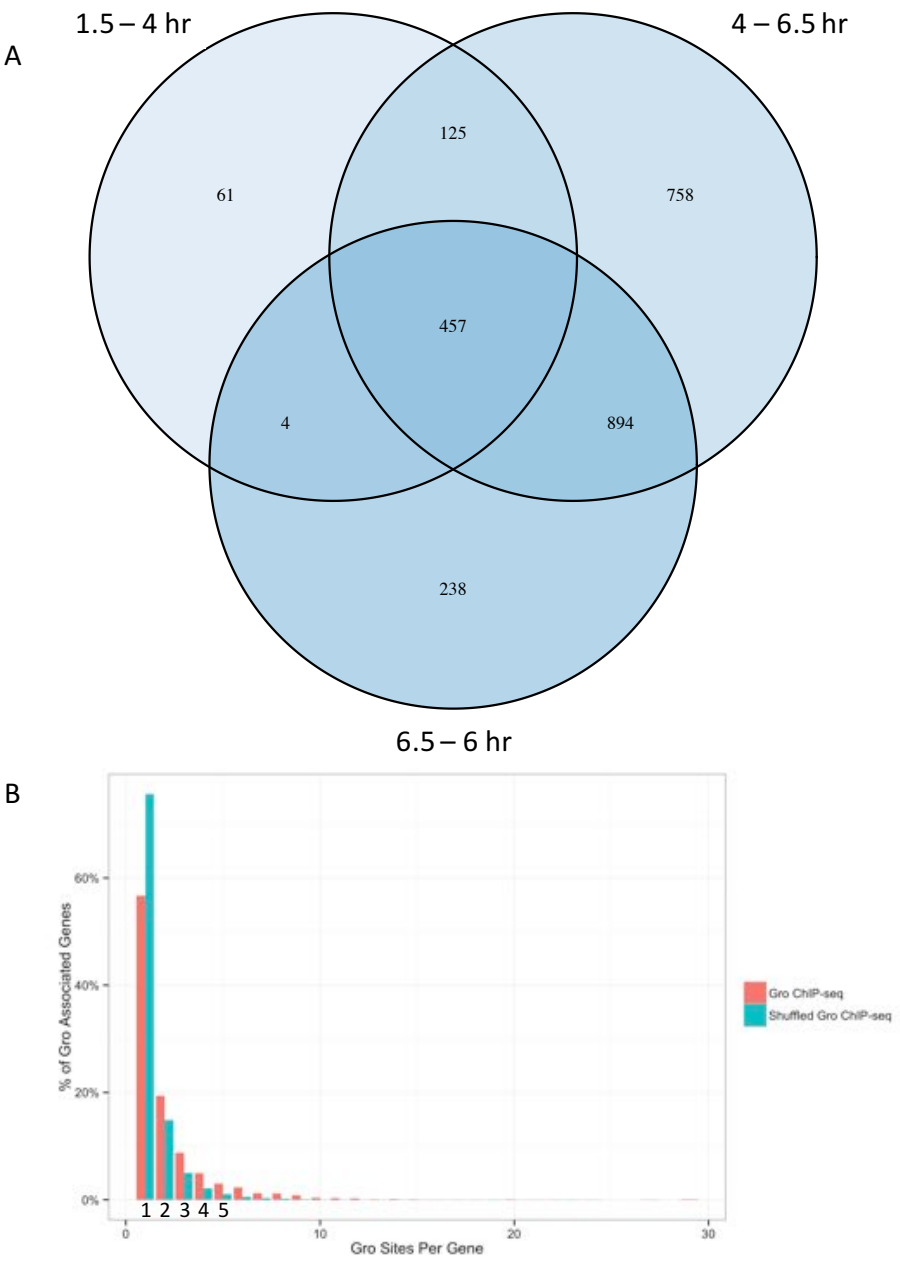

A
1.5 – 4 hr
4 – 6.5 hr
6.5 – 6 hr
61
125
758
457
4
894
238
B
% of Gro Associated Genes
60%
40%
20%
0%
1 2 3 4 5
0
10
20
30
Gro Sites Per Gene
Gro ChIP-seq
Shuffled Gro ChIP-seq

Figure 2-7. Average Groucho peak widths suggest spreading is a limited phenomenon. Groucho binding regions have a median width of between 500 and 700 bp. This binding pattern is more consistent with a transcription factor localizing to a small area of chromatin than with the spreading model that has been theorized to explain the association of Groucho with chromatin. However, at all three timepoints, there are a significant number of outlier Groucho peaks exhibiting wider binding.

Fig. 2-7

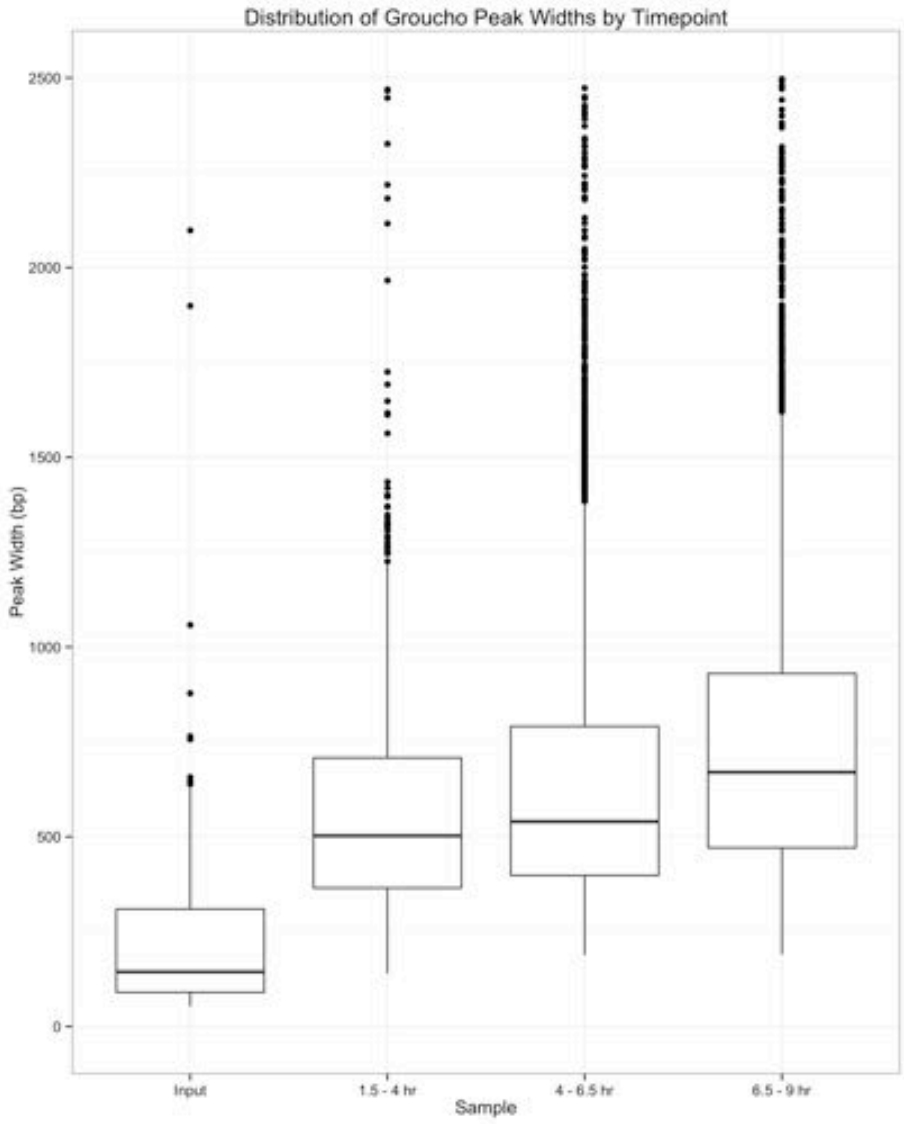
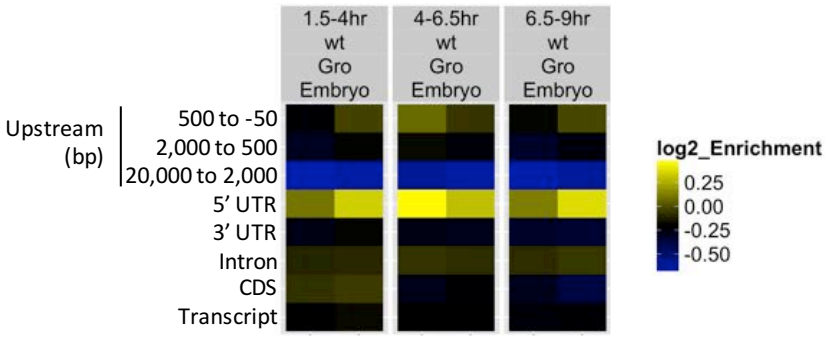

Distribution of Groucho Peak Widths by Timepoint
Peak Width (bp)
2500
2000
1500
1000
500
0
Input
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
Sample

Figure 2-8. Groucho is preferentially recruited to gene bodies at all timepoints. (A) Groucho peaks are enriched within 5’ UTRs, introns, and immediate upstream regions of genes. **(B)** Mapping the location of Groucho binding peaks versus each peak's nearest feature reveals that Groucho preferentially binds within gene bodies, with over half of all Groucho binding at the middle and late timepoints occurring within gene bodies. Groucho binding outside of genes is approximately evenly split between binding upstream and downstream of its nearest feature.

Fig. 2-8

A


1.5-4hr wt Gro Embryo
4-6.5hr wt Gro Embryo
6.5-9hr wt Gro Embryo
Upstream (bp)
500 to -50
2,000 to 500
20,000 to 2,000
5′ UTR
3′ UTR
Intron
CDS
Transcript
log2_Enrichment
0.25
0.00
-0.25
-0.50

B

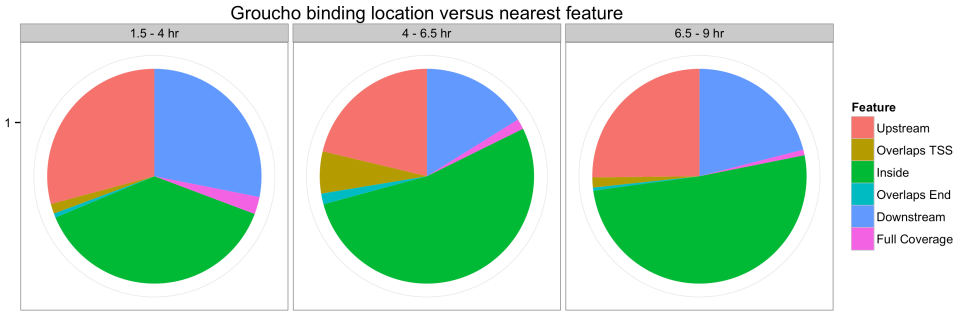
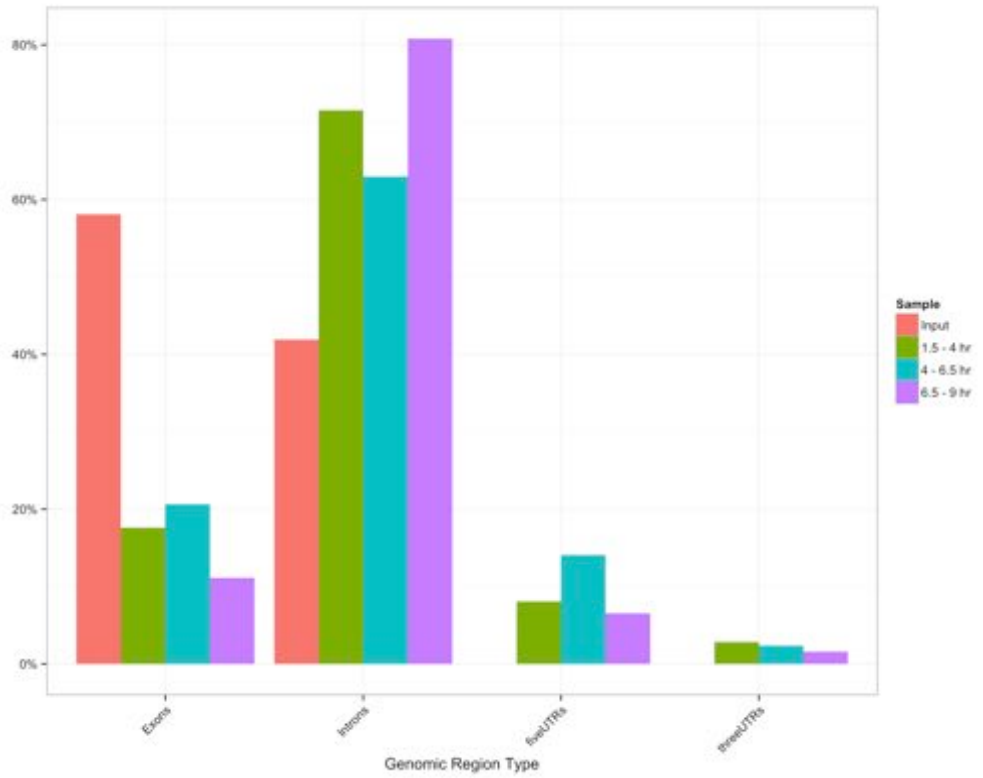

Groucho binding location versus nearest feature
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
1
Feature
Upstream
Overlaps TSS
Inside
Overlaps End
Downstream
Full Coverage

Figure 2-9. The majority of Groucho binding within gene bodies is within introns. (A) The majority of Groucho binding within gene bodies is localized to introns. Binding within exons is depleted in comparison with reads arising from input DNA. Binding is also enriched in 5' and 3' UTR sequences. **(B)** The first intron is particularly enriched for Groucho binding. While initial introns account for 18% of protein-coding gene length in Drosophila, they account for 30% of Groucho binding within gene bodies. Later introns account for 45% of gene sequence and account for 52% of Groucho binding.

Fig. 2-9

A


80%
60%
40%
20%
0%
Exons
Introns
fiveUTRs
threeUTRs
Genomic Region Type
Sample
Input
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr

B

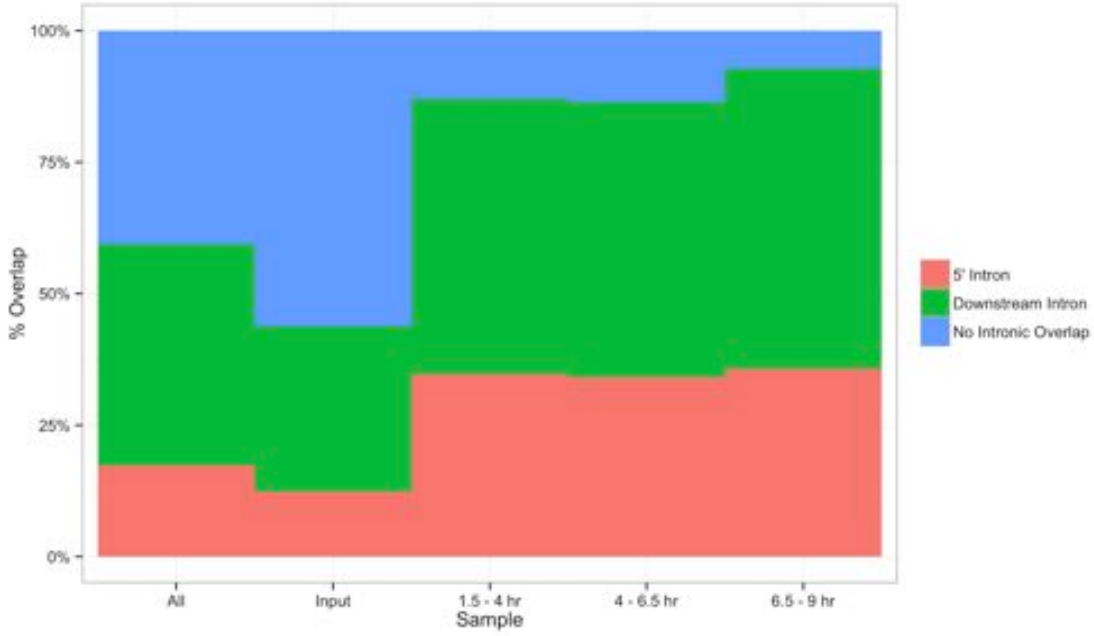

100%
75%
50%
25%
0%
% Overlap
All
Input
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
Sample
5' Intron
Downstream Intron
No Intronic Overlap

Figure 2-10. Motif analysis of Groucho peaks in intergenic and genic regions reveals differential enrichment of coregulators by developmental stage. Binding motifs for several known Groucho-interacting proteins are represented, including hairy (h), huckebein (hkb), sloppy-paired 1 (slp1), brinker (brk), and ventral nervous system defective (vnd). Two factors, serpent (srp) and ultraspiracle (usp), are only enriched in Groucho binding regions arising inside genes.

Fig. 2-10

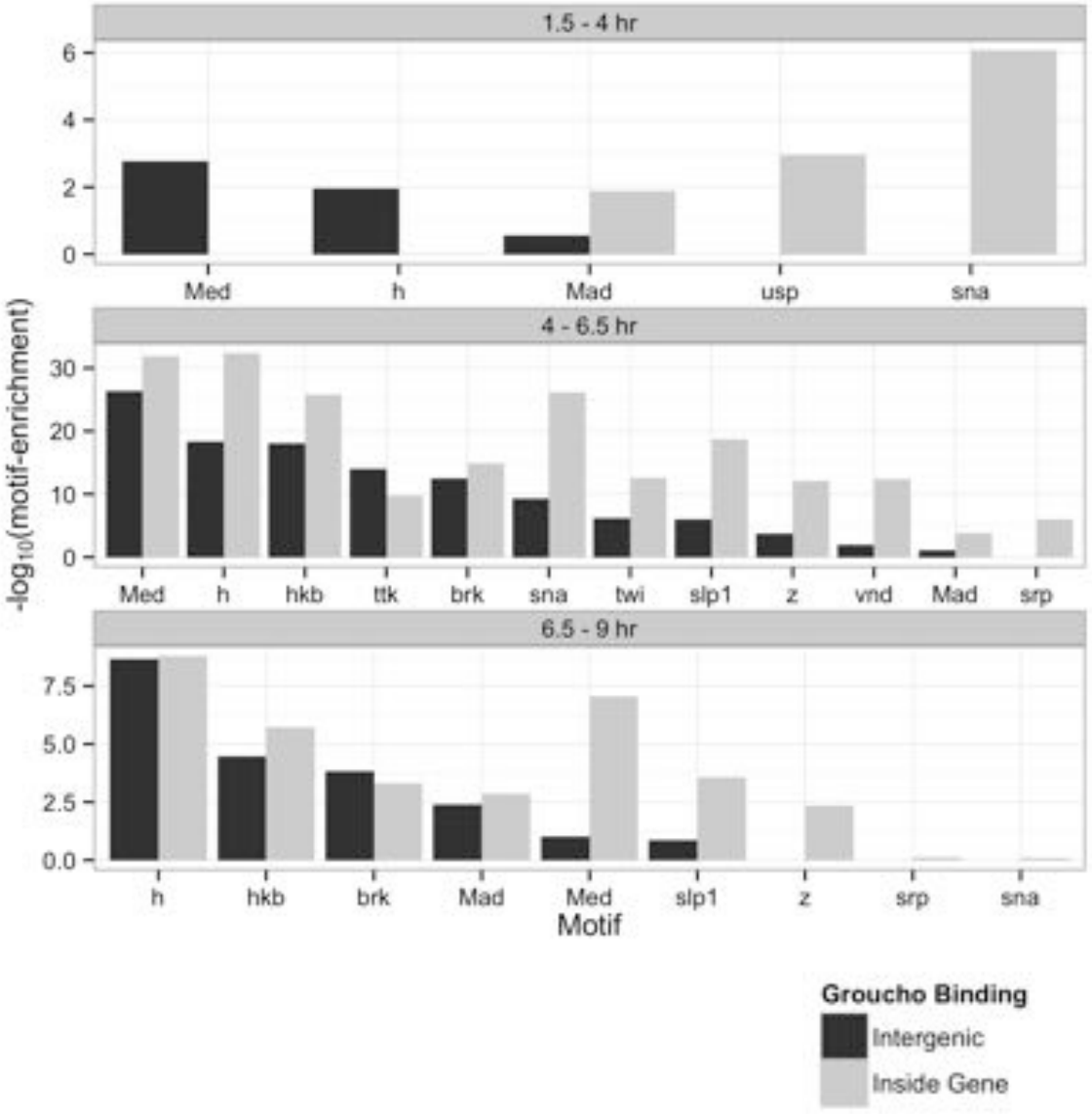

1.5 - 4 hr
6
4
2
0
Med
h
Mad
usp
sna
4 - 6.5 hr
30
20
10
0
Med
h
hkb
ttk
brk
sna
twi
slp1
z
vnd
Mad
srp
6.5 - 9 hr
7.5
5.0
2.5
0.0
h
hkb
brk
Mad
Med
slp1
z
srp
sna
Motif
-log10(motif-enrichment)
Groucho Binding
Intergenic
Inside Gene

Figure 2-11. Groucho binds to early dorsoventral patterning genes with distinct patterns. (A) The region 1.1 to 1.4 kb upstream of *zen* is known as the *zen* ventral repression region (VRR) and contains four Dorsal sites that function, cooperatively with Deadringer/Retained and Cut, to recruit Gro to repress *zen* ventrally in the early embryo (Valentine et al., 1998). Groucho binds within the VRR during the 1.5 - 4 hr timepoint, consistent with Groucho-mediated repression at this stage. However, the majority of binding is outside of the VRR, both immediately upstream of the VRR and downstream. The downstream region overlaps the transcriptional start site of *zen* and continues 700 bp upstream. Groucho binding shifts during the next timepoint, and is lost entirely by the third timepoint analyzed. **(B)** Groucho binds downstream and inside intronic regions of *dpp*, which similarly to *zen* is represssed ventrally by Groucho and Dorsal activity (Dubnicoff et al., 1997) in the early (0 - 2 hr) embryo. At later developmental stages, *dpp* repression is mediated through a 3' cis-regulatory region containing multiple pangolin/TCF and brinker binding sites.

Fig. 2-11

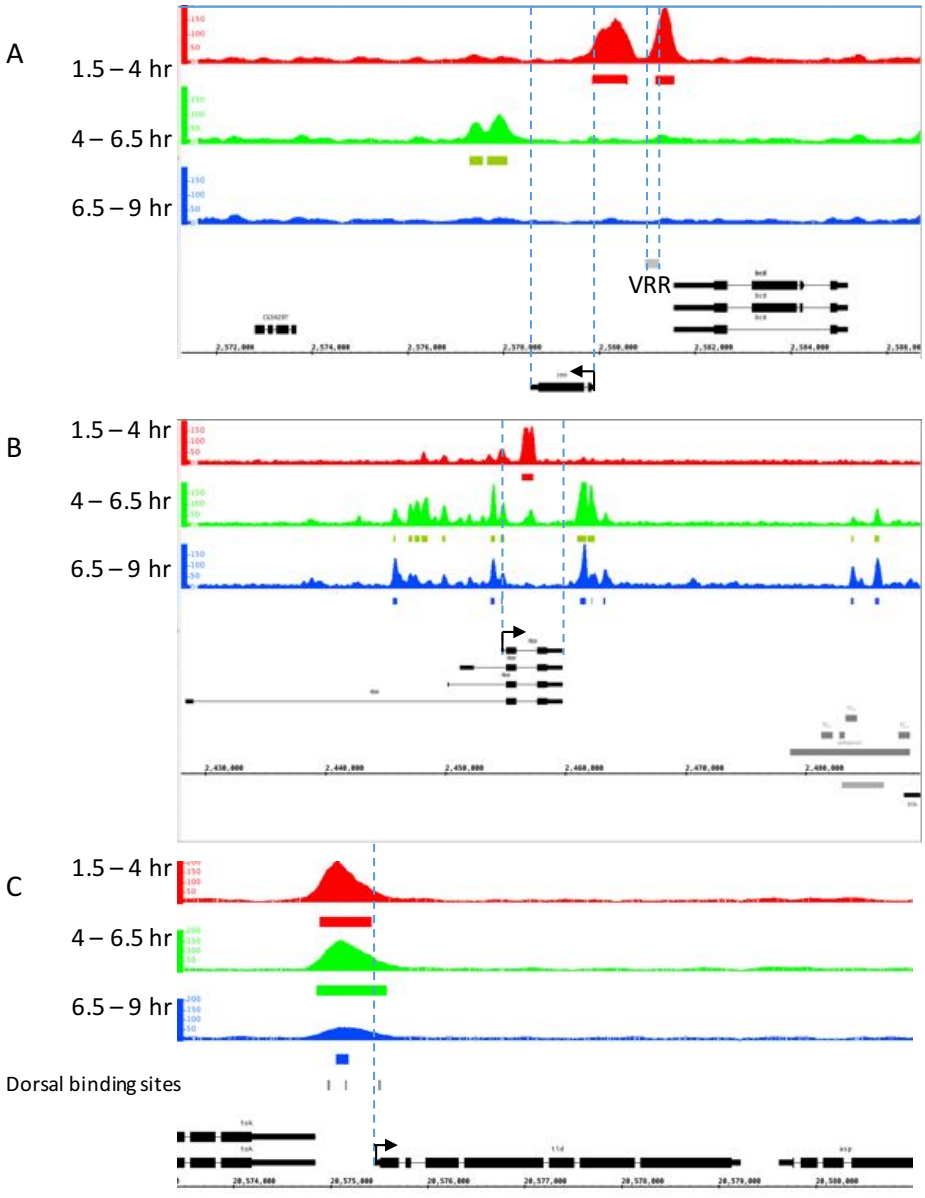

A
1.5 – 4 hr
4 – 6.5 hr
6.5 – 9 hr
VRR
B
1.5 – 4 hr
4 – 6.5 hr
6.5 – 9 hr
C
1.5 – 4 hr
4 – 6.5 hr
6.5 – 9 hr
Dorsal binding sites

Figure 2-12. Groucho associates with a subset of Dorsal-activated genes in the presumptive mesoderm. (A) Two cis-regulatory regions have been identified upstream of *snail*, either of which is sufficient for Dorsal-mediated activation of the gene in ventral regions of the early (2 – 3 hr) embryo, leading to the hypothesis that the shadow enhancer is involved in fine-tuning *snail* expression, or potentially making expression more robust to stochastic fluctuations in transcription factor availability. However, Groucho recruitment patterns are asymmetric over time between these two regions, indicating potentially divergent roles in control of *snail* expression later in development. **(B)** In contrast, recruitment of Groucho to the *twist* locus is relatively weak. Dorsal binds within the ventral activation region (VAR) directly upstream of *twist,* where it serves to activate gene expression via the cooperation of the co-activator dCBP. A small yet significant Gro peak is present within this region during the first time window, but disappears by later stages. While Groucho may be involved in repressing *snail*, in dorsal and dorsolateral regions of the embryo, it appears *twist* repression is initiated or maintained by another, unknown, mechanism.

Fig. 2-12

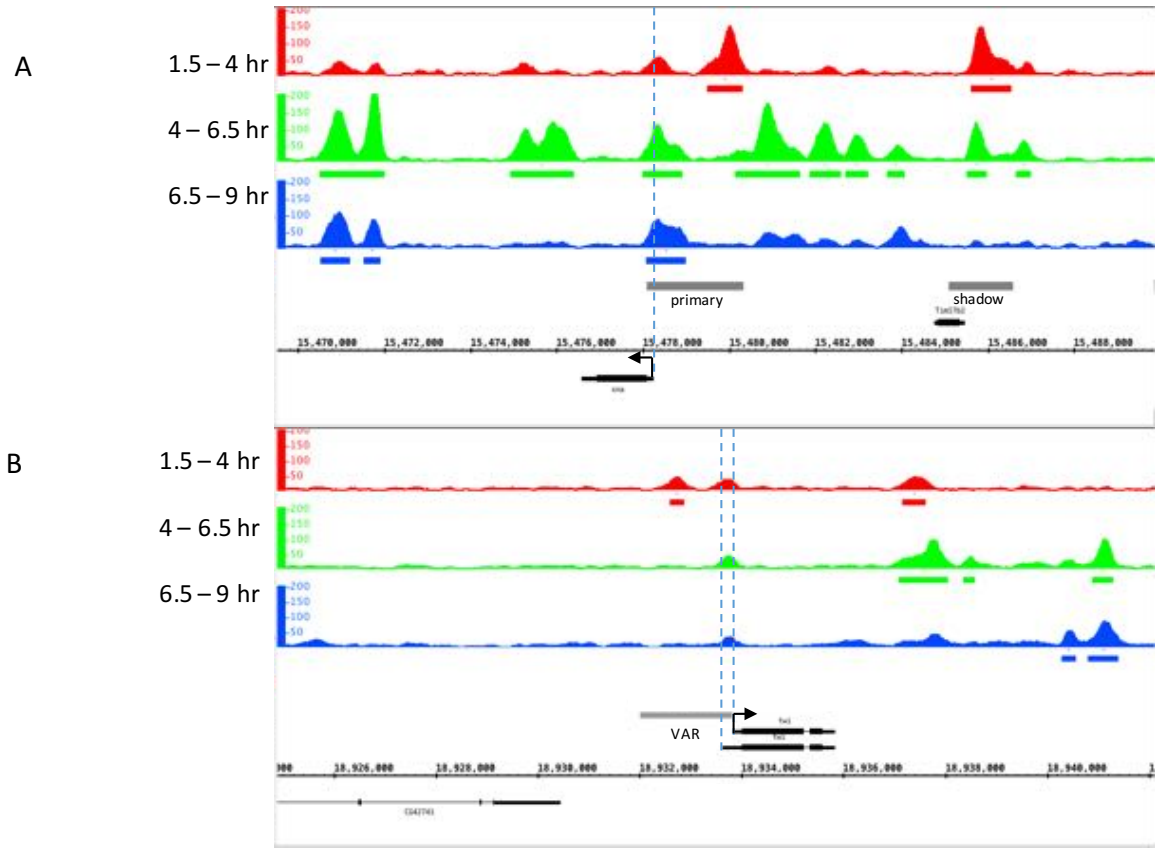


Figure 2-13. Groucho is recruited to Dorsal-activated genes in early embryos.

Rhomboid (rho), Single-minded (sim), and Short gastrulation (sog) are activated through Dorsal activity in ventrolateral regions of the early embryo (1.5 - 2 hours post fertilization). Loss of Gro activity results in decreased expression of these genes, but does not change their expression patterns along the dorsoventral axis, and so Groucho is hypothesized to not play a role in their Dorsal-mediated activation (Dubnicoff et al., 1997). **(A)** However, Groucho is recruited both upstream of *rho* within known two known CRMs at early timepoints and overlapping its TSS, suggesting a previously unidentified role of Gro in regulating *rho* expression. **(B)** and **(C)** Additionally, Gro binds within the intronic regions of *sim* and *sog* at all timepoints.

Fig. 2-13

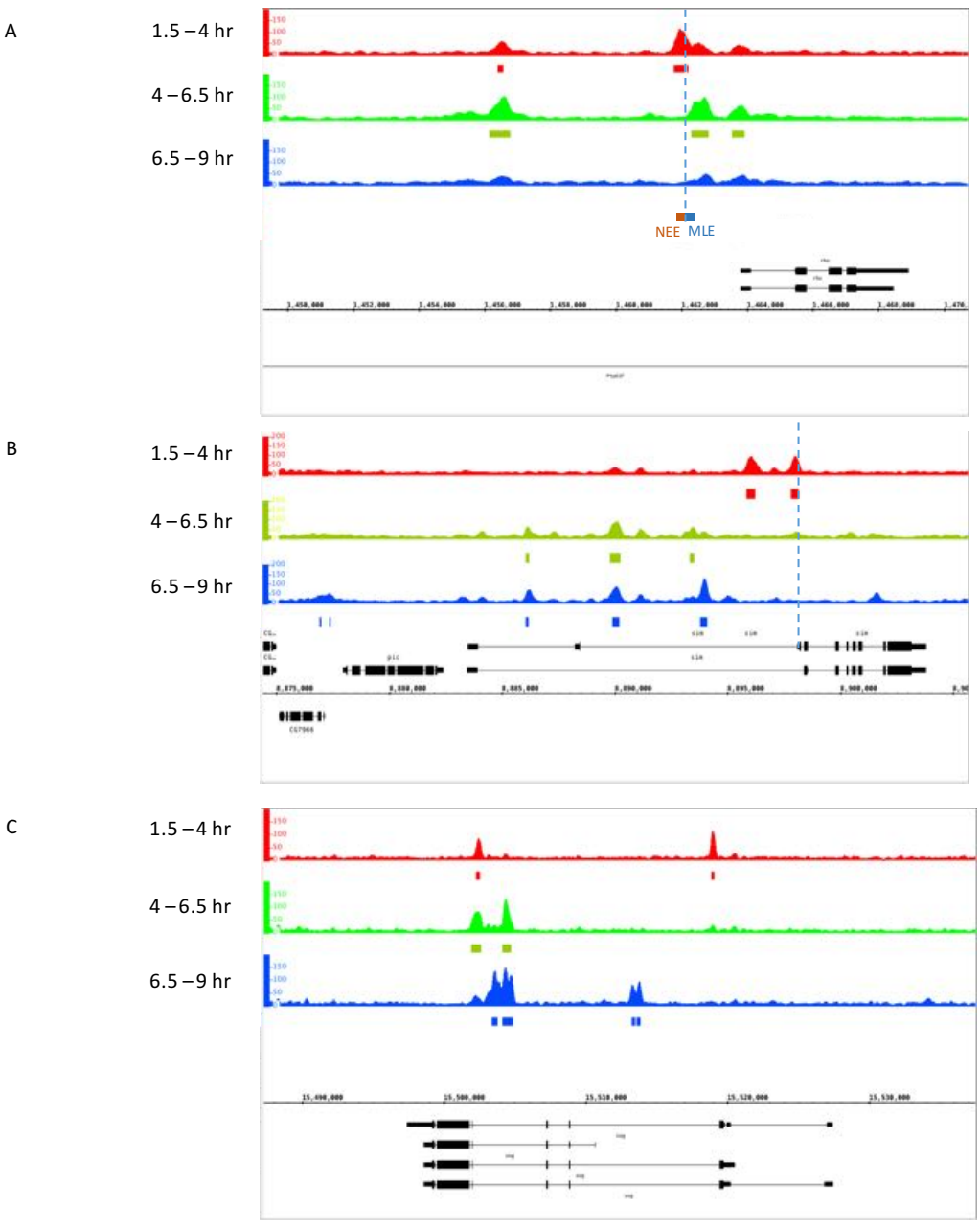

A
1.5 – 4 hr
4 – 6.5 hr
6.5 – 9 hr
NEE
MLE
B
1.5 – 4 hr
4 – 6.5 hr
6.5 – 9 hr
C
1.5 – 4 hr
4 – 6.5 hr
6.5 – 9 hr

Figure 2-14. Groucho is recruited to large subsets of known Dorsal/Twist/Snail-binding regulatory regions. (A) Dorsal binding sites can be subdivided into three classes, dependent on the degree these sites respond to the nuclear Dorsal gradient formed along the embryo's dorsoventral axis. Class I (mesodermal) sites active gene expression in regions of high nuclear Dorsal; Class II (neuroectodermal) sites activate expression in regions of intermediate Dorsal levels; and Class III sites bind Dorsal to repress transcription, resulting in restricted expression in areas of low Dorsal concentration. Groucho overlaps all three types of Dorsal binding site, showing no preference for repressive (Class III) sites. **(B)** Dorsal regulates the dorsoventral patterning of multiple determinants of anteroposterior patterning, here subdivided into determinants of embryonic segmentation (gap, pair-rule, and segment polarity genes) and body plan specification (homeotic genes). Unlike dorsoventral patterning targets, Groucho/Dorsal association with these genes tends to occur earlier in development, potentially indicating a novel regulatory pathway in which Groucho can participate in anteroposterior patterning.

Fig. 2-14

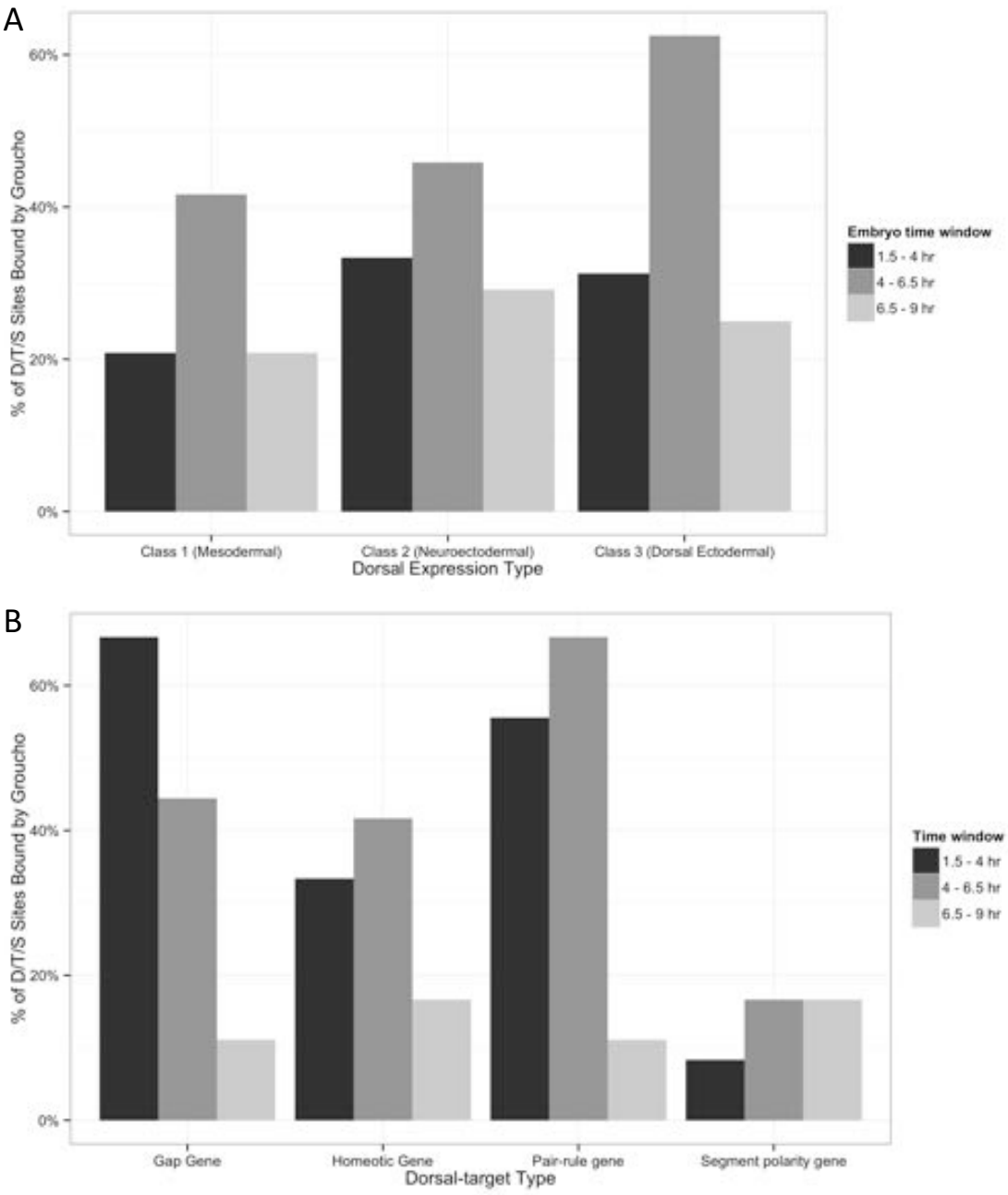
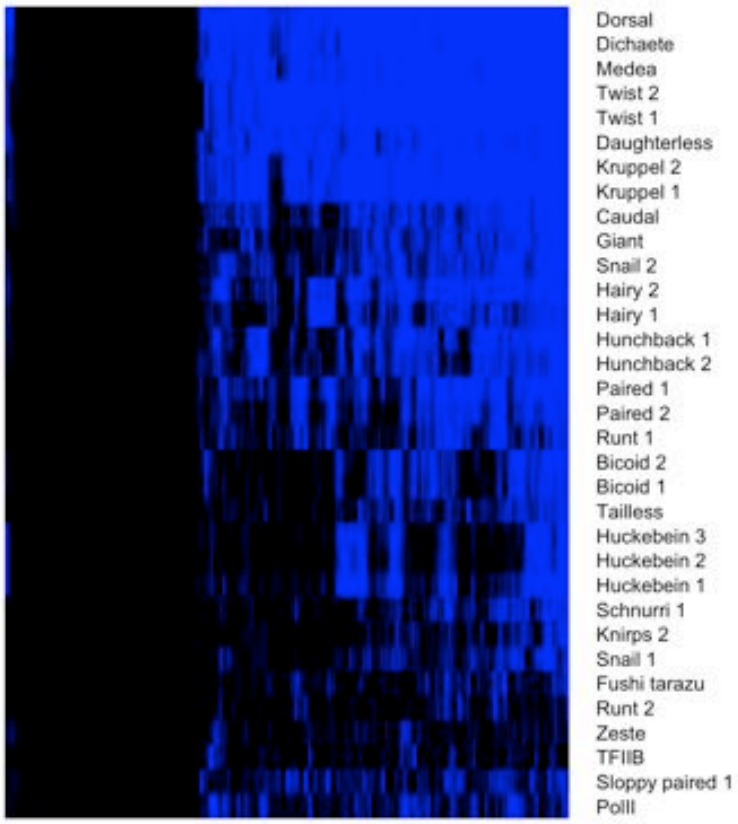

A
60%
40%
20%
0%
% of D/T/S Sites Bound by Groucho
Class 1 (Mesodermal)
Class 2 (Neuroectodermal)
Class 3 (Dorsal Ectodermal)
Dorsal Expression Type
Embryo time window
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
B
60%
40%
20%
0%
% of D/T/S Sites Bound by Groucho
Gap Gene
Homeotic Gene
Pair-rule gene
Segment polarity gene
Dorsal-target Type
Time window
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr

Fig. 2- 15. The majority of Groucho recruitment sites in the early embryo additionally bind Dorsal, Dichaete, and, less frequently, multiple additional factors. A clustered heatmap of the factors that each Groucho binding site overlaps in the developing embryo reveals multiple strategies of Groucho recruitment. Each column represents a single Groucho binding site, with blue representing overlap with the factor given on the y-axis.

Fig. 2-15



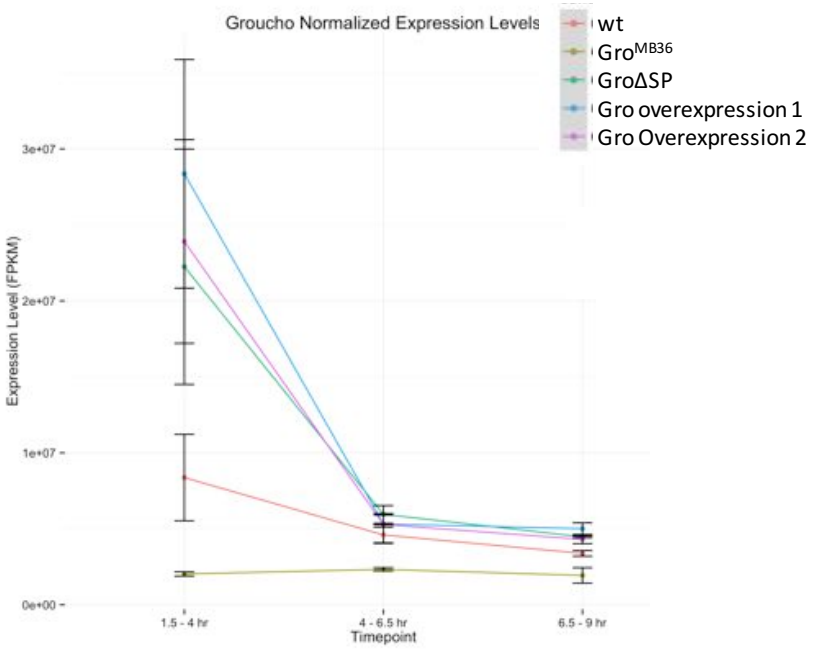
Gro binding sites, 1.5 – 4 hr embryos

Factor occupancy

Figure 2-17. Confirmation of changes in Groucho transcript concentration across timepoints. (A) Analysis of Groucho transcript levels reveals initially high levels of Groucho transcript in early embryos, which steadily declines in Gro wild-type and overexpression embryos. Gro loss-of-function embryos exhibit barely detectable levels of transcript throughout all three developmental stages. **(B)** Our Gro wild-type expresssion pattern is consistent with modENCODE developmental timecourse transcriptome data (Graveley et al., 2011), which shows a peak of Groucho transcript level during 2 to 4 hours post-fertilization, followed by a steady decrease through the remainder of embryonic development.

Fig. 2-17

A


Groucho Normalized Expression Levels
wt
Gro^MB36
GroΔSP
Gro overexpression 1
Gro Overexpression 2
Expression Level (FPKM)
3e+07
2e+07
1e+07
0e+00
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
Timepoint

B

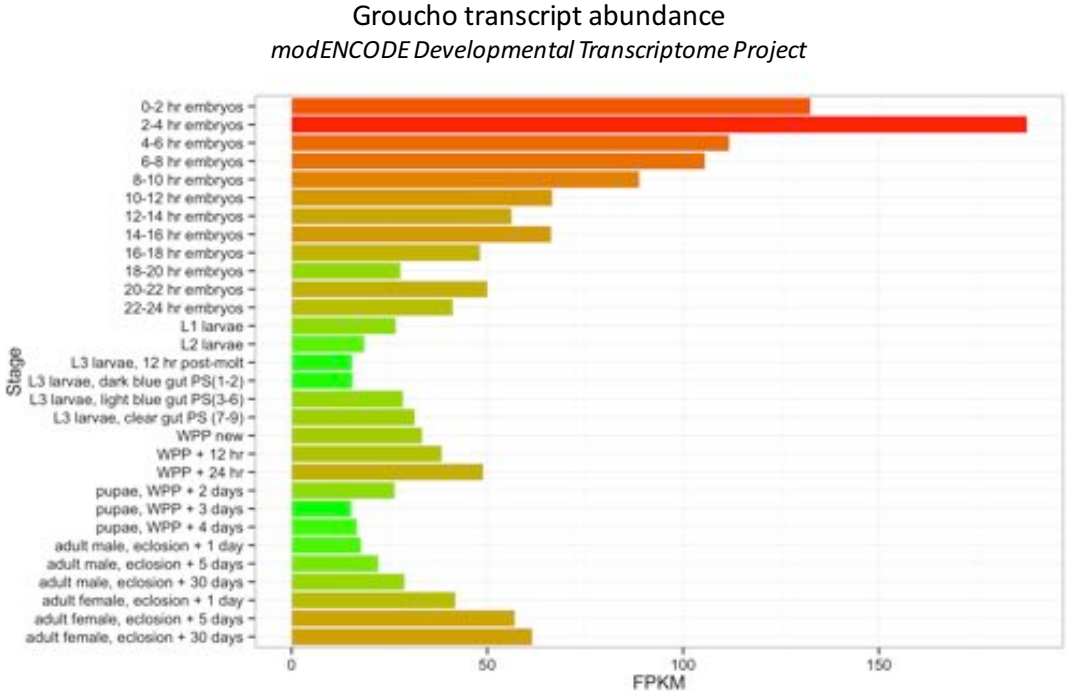

Groucho transcript abundance
modENCODE Developmental Transcriptome Project
Stage
0-2 hr embryos
2-4 hr embryos
4-6 hr embryos
6-8 hr embryos
8-10 hr embryos
10-12 hr embryos
12-14 hr embryos
14-16 hr embryos
16-18 hr embryos
18-20 hr embryos
20-22 hr embryos
22-24 hr embryos
L1 larvae
L2 larvae
L3 larvae, 12 hr post-molt
L3 larvae, dark blue gut PS(1-2)
L3 larvae, light blue gut PS(3-6)
L3 larvae, clear gut PS (7-9)
WPP new
WPP + 12 hr
WPP + 24 hr
pupae, WPP + 2 days
pupae, WPP + 3 days
pupae, WPP + 4 days
adult male, eclosion + 1 day
adult male, eclosion + 5 days
adult male, eclosion + 30 days
adult female, eclosion + 1 day
adult female, eclosion + 5 days
adult female, eclosion + 30 days
0
50
100
150
FPKM

Figure 2-18. Clustering of embryonic transcriptomes across Gro levels and timepoints and between replicates. Pair-wise Spearman correlation coefficients were used to cluster transcriptome profiles by overall similarity. Transcriptomes tend to cluster by timepoint, then by Gro expression level. The notable exception are the later Gro loss-of-function samples, which cluster together (red square), independently from other 4 - 6.5 or 6.5 - 9 hour aged embryos. This is consistent with the significant departure from a viable developmental progression these embryos have taken by this point, which has resulted in significant changes in gene expression.

Fig. 2-18

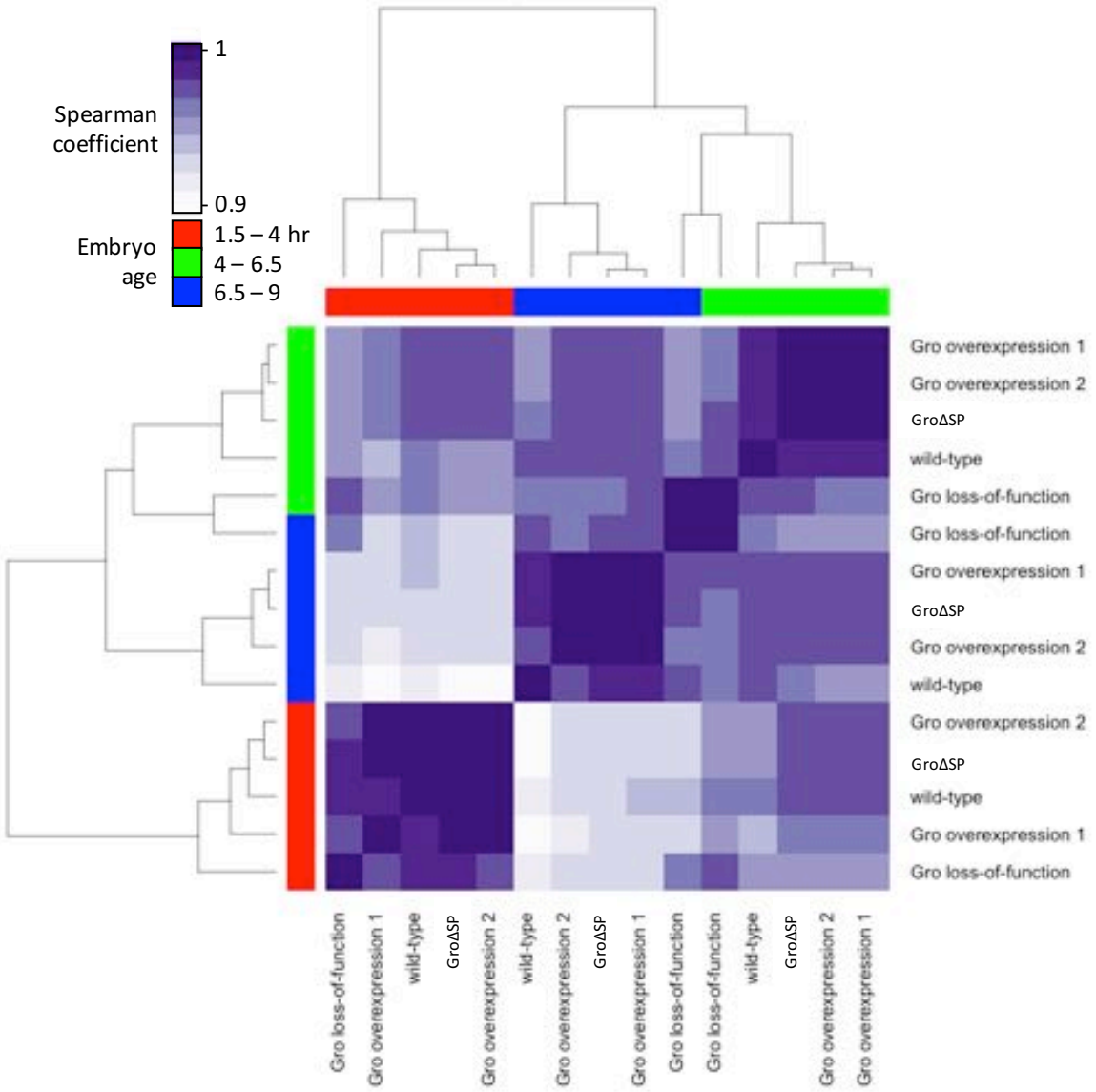

1
Spearman coefficient
0.9
Embryo age
1.5 – 4 hr
4 – 6.5
6.5 – 9
Gro overexpression 1
Gro overexpression 2
GroΔSP
wild-type
Gro loss-of-function
Gro loss-of-function
Gro overexpression 1
GroΔSP
Gro overexpression 2
wild-type
Gro overexpression 2
GroΔSP
wild-type
Gro overexpression 1
Gro loss-of-function
Gro loss-of-function
Gro overexpression 1
wild-type
GroΔSP
Gro overexpression 2
wild-type
GroΔSP
Gro overexpression 2
Gro overexpression 1
Gro loss-of-function

Figure 2-19. Principal component analysis reveals overexpression lines have high inter-group similarity. Principal component analysis was performed on transcriptome profiles from wild-type, Gro loss-of-function, and three Gro overexpression embryos at three timepoints. Principal component analysis is a widely-used technique to visualize relatedness of high-dimensionality data such as transcriptomes, in which the expression level of each gene constitutes a dimension. Relatedness of two transcriptomes then becomes a function of the linear distance between two points (closer distances equate to higher similarity). While the axes have no predetermined physical meaning, they often capture distinct sources of variance between samples. In our case, the x-axis appears to correspond to developmental time point, while the y-axis captures Groucho transcript dosage. Wild-type and Gro loss-of-function samples show significant deviation from overexpression lines. Overexpression lines share a significant degree of overlap across the y-axis, indicative of a high degree of common features. Replicates are joined by lines.

Fig. 2-19

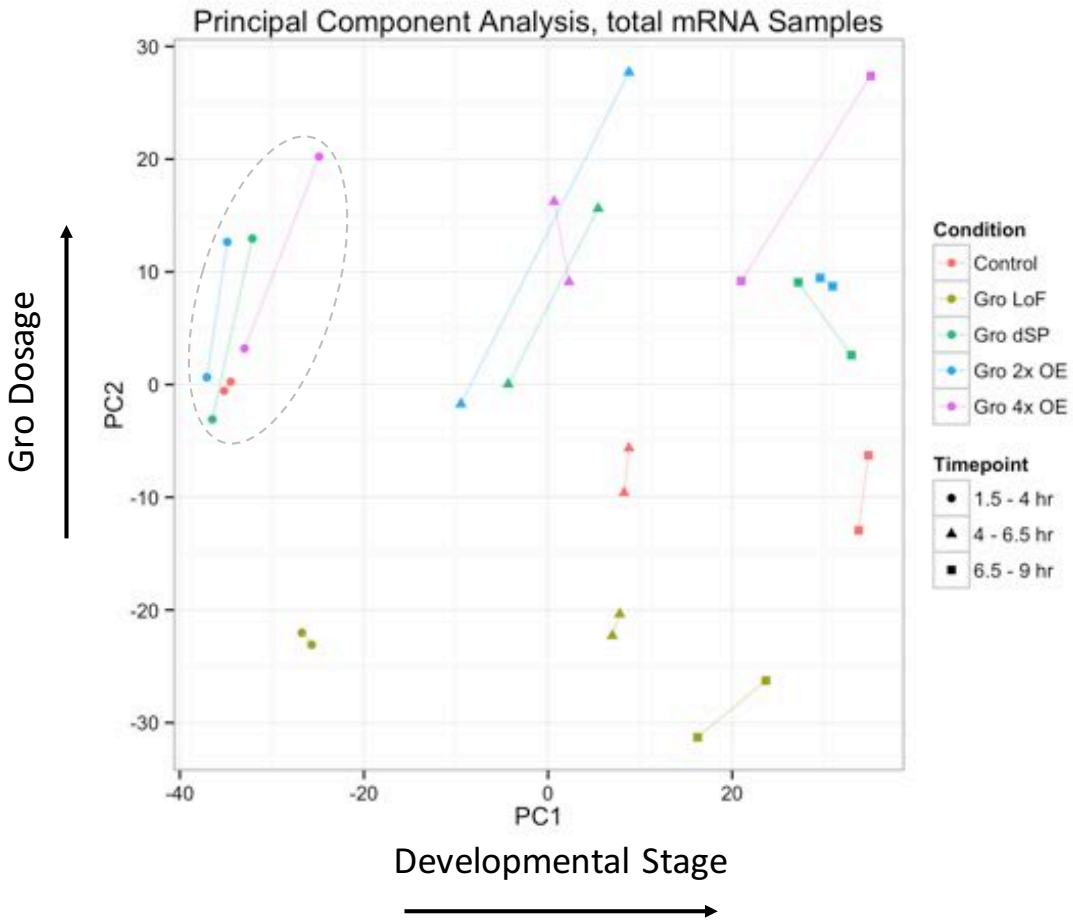

Principal Component Analysis, total mRNA Samples
Gro Dosage
PC2
PC1
Developmental Stage
30
20
10
0
-10
-20
-30
-40
-20
0
20
Condition
Control
Gro LoF
Gro dSP
Gro 2x OE
Gro 4x OE
Timepoint
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr

Figure 2-20. Perturbation of Groucho expression levels results in the mis-regulation of thousands of genes. Maternal deficiency of Gro activity results in a large proportion (>10%) of expressed genes to become misregulated in the Drosophila embryo across all timepoints. The fraction of misregulated genes is approximately evenly split between up- and down-regulation. Overexpression of wild-type Gro at two levels (approx. 2x and 4x endogenous), or a Gro mutant lacking the SP domain (GrodSP), results in a smaller, but still significant alteration of the embryonic transcription profile.

Fig. 2-20

A

		1.5 - 4 hr	4 - 6.5 hr	6.5 - 9 hr
Gro LoF	Down	1459	2253	2043
Gro LoF	Up	1437	1837	1632
GroΔSP	Down	241	284	500
GroΔSP	Up	100	230	616
Overexpression (2x)	Down	698	566	463
Overexpression (2x)	Up	244	599	496
Overexpression (4x)	Down	674	655	1171
Overexpression (4x)	Up	219	490	1204

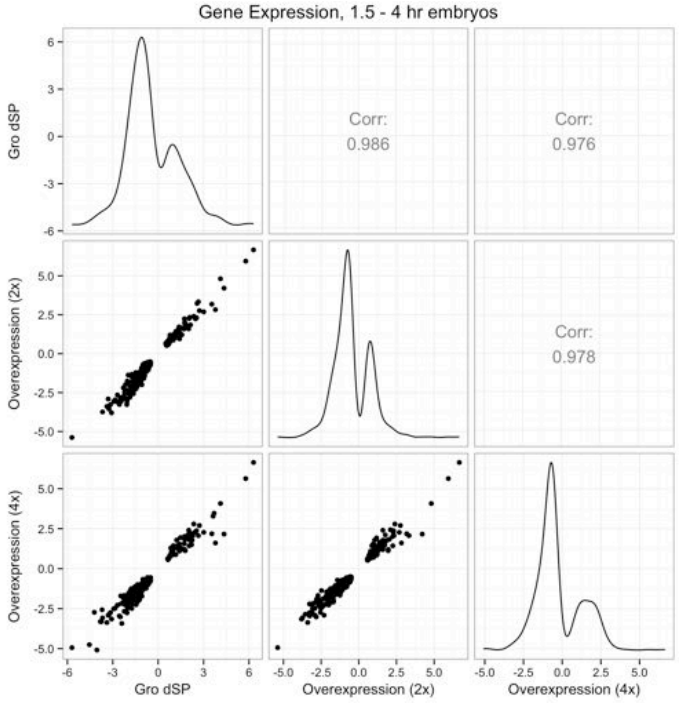
B

% of Protein-Coding Genome Differentially Expressed vs wild-type			
Sample	1.5 - 4 hr	4 - 6.5 hr	6.5 - 9 hr
Gro LoF	19%	27%	25%
GroΔSP	2%	3%	7%
Gro 2x Overexpression	6%	8%	6%
Gro 4x Overexpression	6%	8%	16%

Figure 2-21. The three Groucho overexpression lines show similar patterns of altered gene expression, though significant differences in the magnitude of gene expression changes are evident. Paired scatterplots of $\log_2$(fold-changes) in expression level of each differentially expressed gene (in comparison to wild-type embryos) across all timepoints reveals that overexpression of Groucho results in similar changes in expression of the majority of genes, indicating that not all genes exhibit a strong-dosage response, though some significant differences are evident. **(A)** At the earliest timepoint, the majority of effected genes exhibit decreased expression, consistent with increased repression via Gro. This effect becomes less pronounced at later timepoints. **(B & C)**.

Fig. 2-21

A


Gene Expression, 1.5 - 4 hr embryos
Gro dSP
Overexpression (2x)
Overexpression (4x)
Corr:
0.986
Corr:
0.976
Corr:
0.978

B

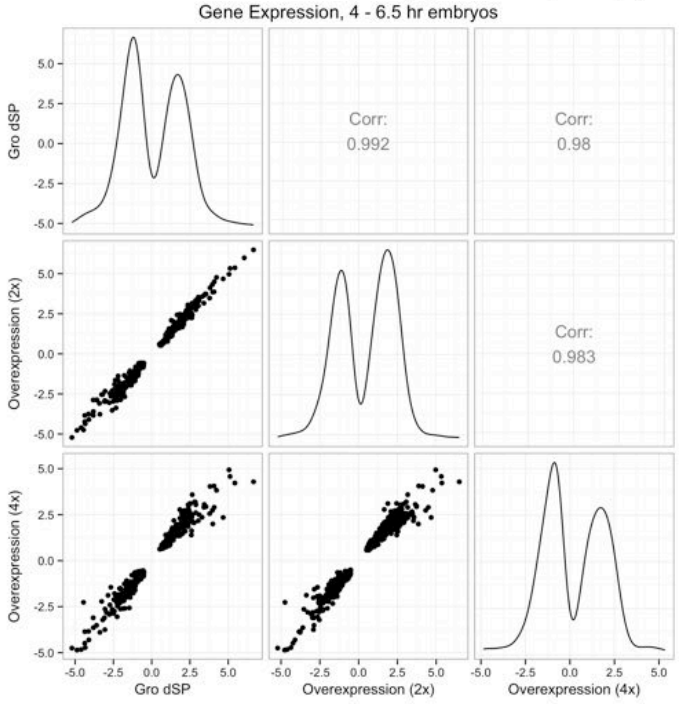

Gene Expression, 4 - 6.5 hr embryos
Gro dSP
Overexpression (2x)
Overexpression (4x)
Corr:
0.992
Corr:
0.98
Corr:
0.983

Fig. 2-21 (cont)

C

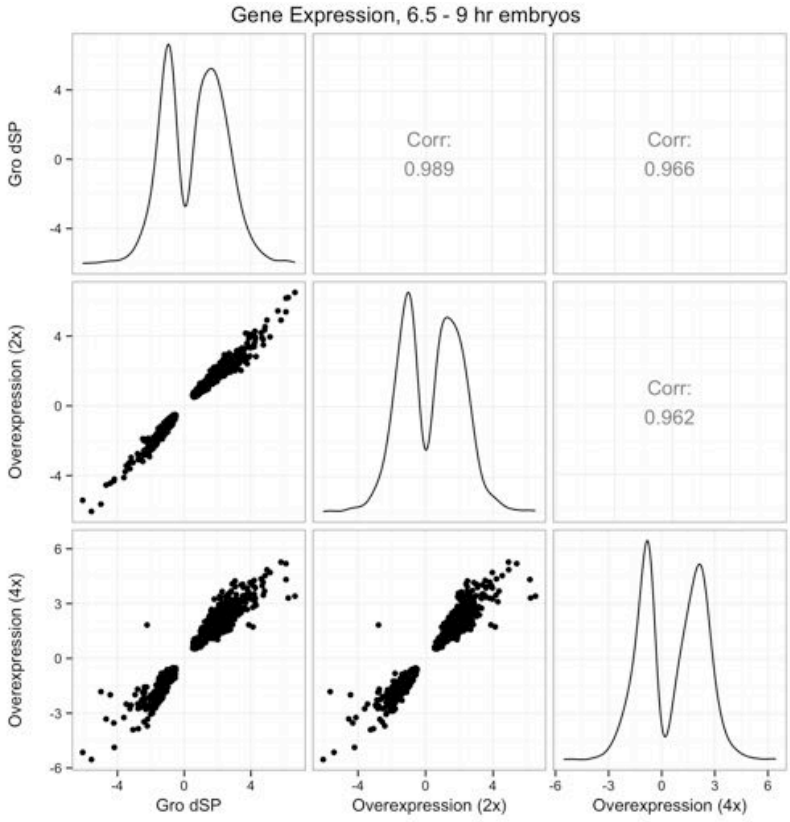


Figure 2-22. A subset of genes is differentially expressed under both loss and gain of Groucho dosage. The $\log_2$ fold-change of gene expression of each gene was calculated by comparing expression levels in Groucho mutant embryos versus wild-type. A subset of the total Groucho-effected genes was identified as being misregulated under both loss and gain of Groucho function. A portion of these genes show changes in expression of the opposite sign under both conditions (i.e. increased expression under Gro loss-of-function and decreased expression under Gro overexpression, or vice-versa). We hypothesize that this set of genes is further enriched for direct targets of Groucho-mediated repression.

Fig. 2-22

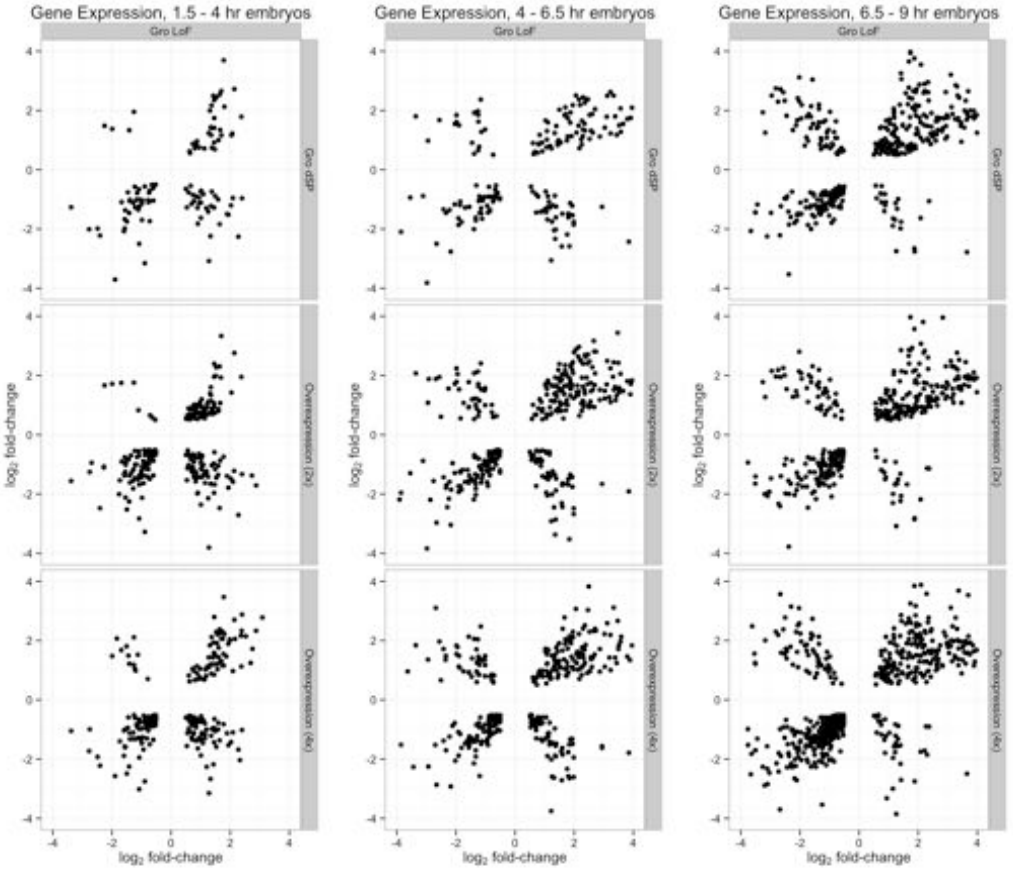

Gene Expression, 1.5 - 4 hr embryos
Gene Expression, 4 - 6.5 hr embryos
Gene Expression, 6.5 - 9 hr embryos
Gro LoF
Gro dSP
Overexpression (2x)
Overexpression (4x)
log2 fold-change
log2 fold-change

Figure 2-23. The majority of high-confidence Groucho targets are differentially expressed in both Gro overexpression lines. Limiting the list of Groucho regulated genes by ChIP-seq and RNA-seq data resulted in a total of 248 target genes. Of these genes, 61% are repressed in both the 2x and 4x Gro overexpression embryos.

Fig. 2-23

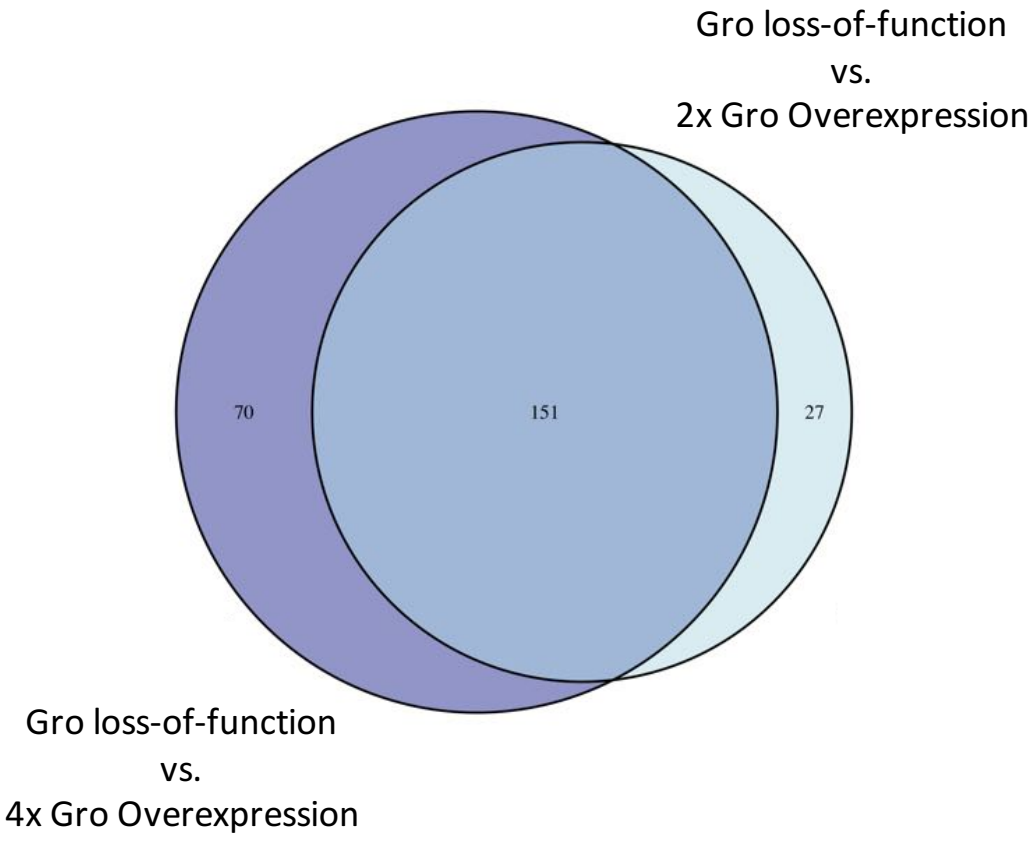
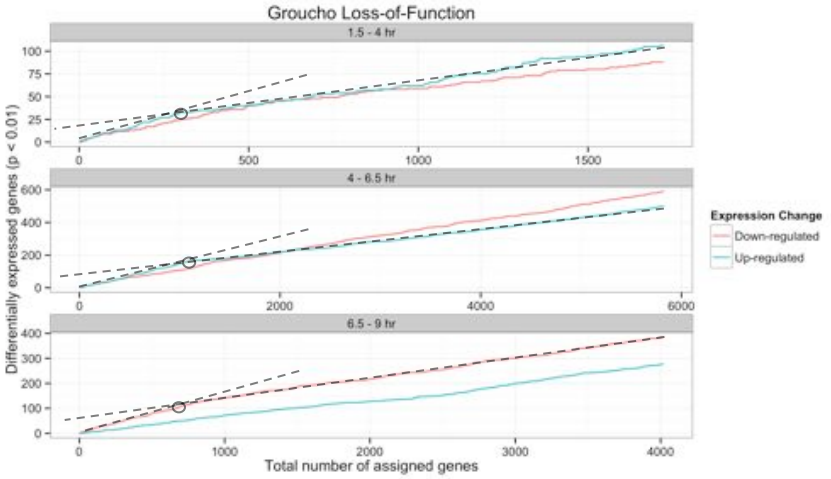

Gro loss-of-function
vs.
2x Gro Overexpression
70
151
27
Gro loss-of-function
vs.
4x Gro Overexpression

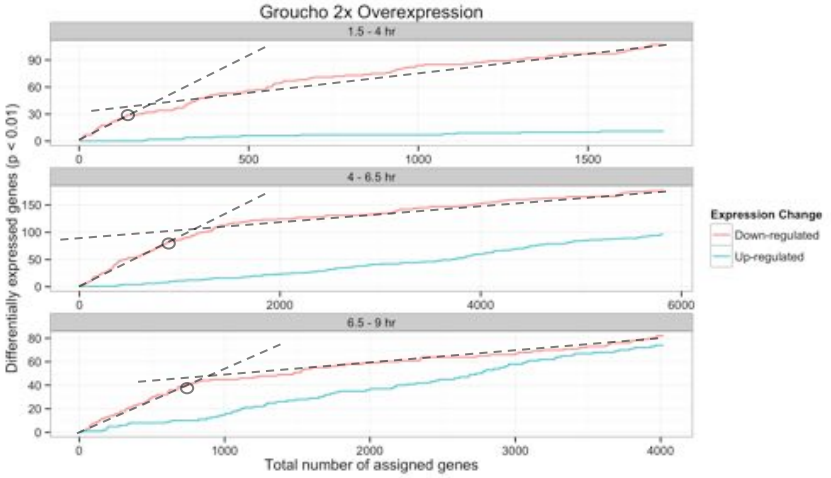
Figure 2-24. High-confidence Groucho targets were identified through a scoring algorithm integrating binding data (ChIP-seq) with expression data measured under multiple Groucho dosages (RNA-seq). A score corresponding to the distribution of Groucho occupancy within and in adjacent areas of a gene was calculated using a previously published algorithm (Sandmann et al., 2007). The algorithm was adjusted to allow for increased score contribution from regions binding more distantly from the target gene. Plotted are the number of genes differentially expressed under the indicated Groucho dosage out of the total number of genes meeting a score cutoff of decreasing stringency. Where a change in slope is clearly evident, the score cutoff selected for the high-confidence set of Groucho targets is indicated.

Fig. 2-24

A


Groucho Loss-of-Function
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
Differentially expressed genes (p < 0.01)
Total number of assigned genes
Expression Change
Down-regulated
Up-regulated

B


Groucho 2x Overexpression
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
Differentially expressed genes (p < 0.01)
Total number of assigned genes
Expression Change
Down-regulated
Up-regulated

C

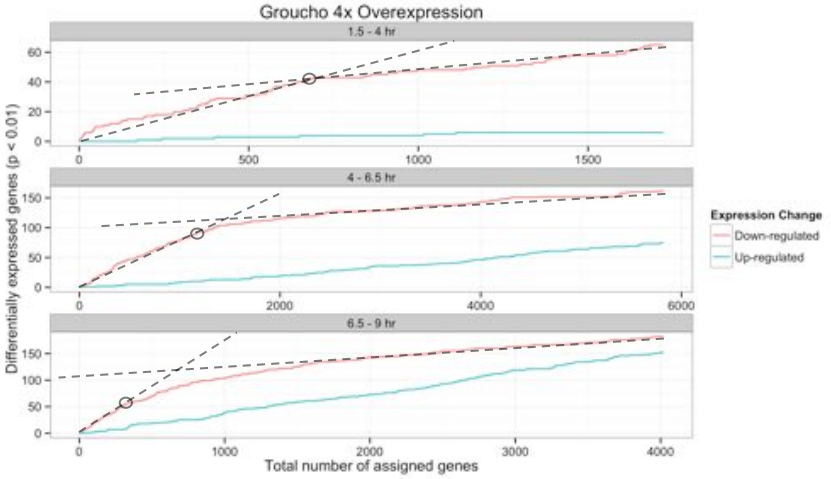

Groucho 4x Overexpression
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
Differentially expressed genes (p < 0.01)
Total number of assigned genes
Expression Change
Down-regulated
Up-regulated

Figure 2-25. Potential Groucho targets identified by the Gro-dosage responsive and occupancy scoring methods. A total of 248 genes were identified as being responsive to multiple levels of Groucho dosage and associated with Groucho peaks (dark-blue). The alternative method, of choosing genes sensitive to Groucho level under a single condition and exhibiting a level of Groucho occupancy above an empirically-derived score threshold, identified 351 genes. Ninety genes were identified by both methods. The resulting target gene sets therefore differ significantly, though the overlap is statistically significant ($p < 10^{-10}$, hypergeometric test).

Fig. 2-25

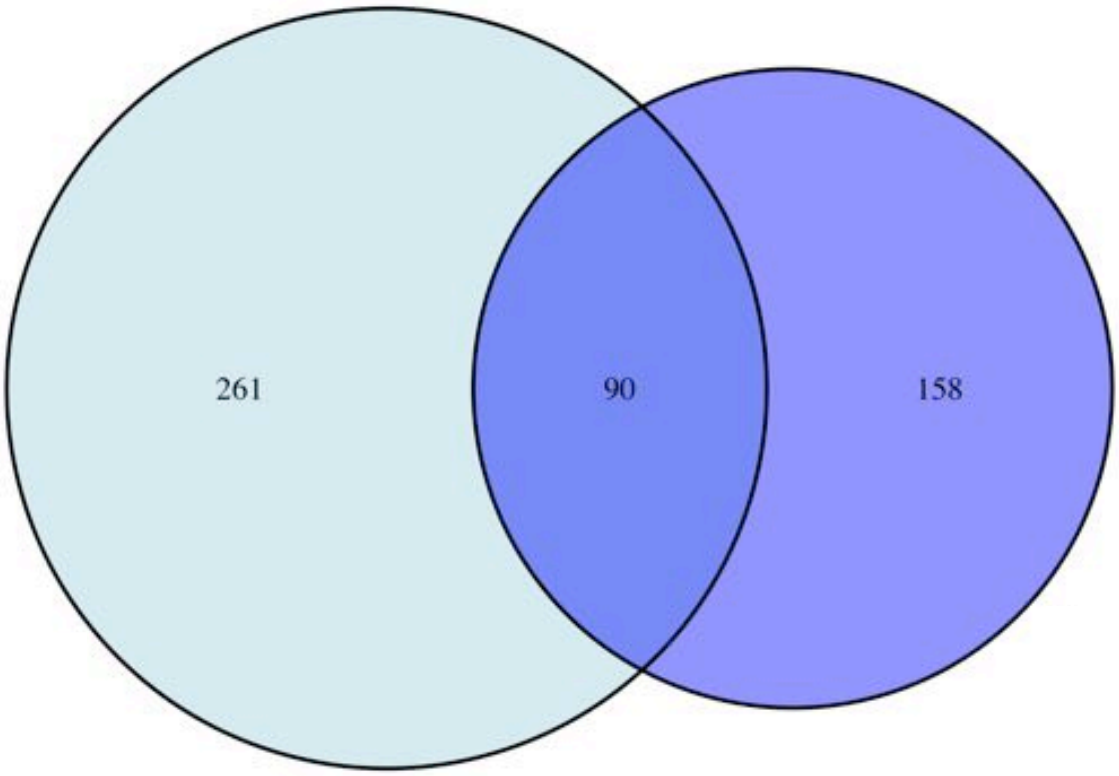
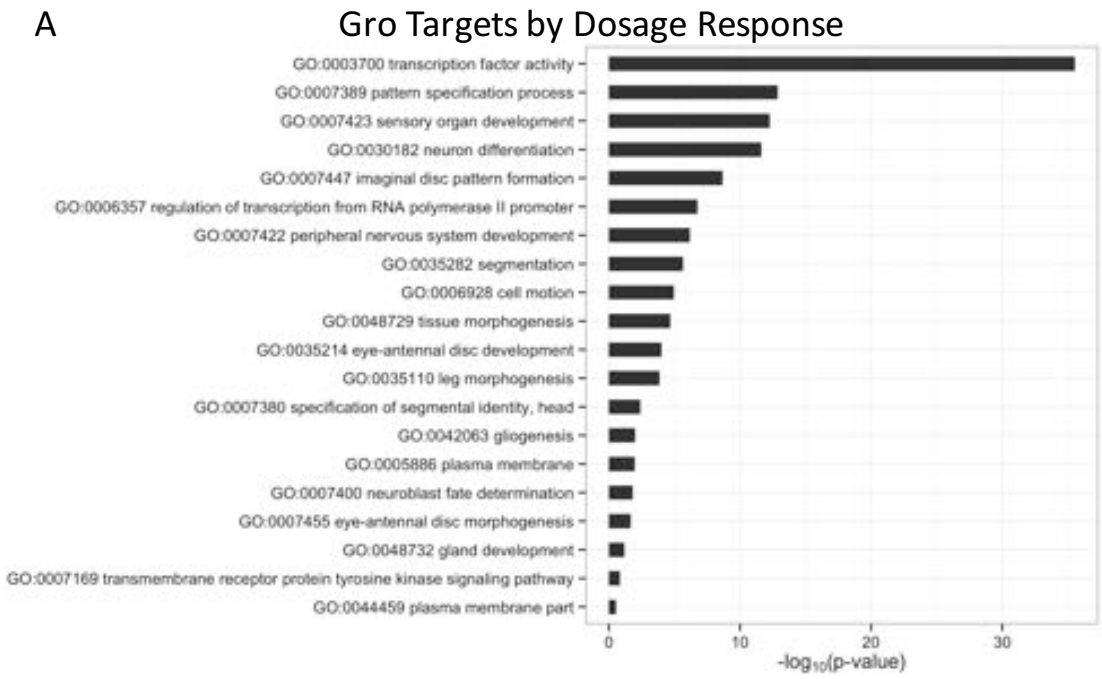

261
90
158
Genes above Groucho occupancy score cutoff
Groucho Dosage Sensitive And Groucho Bound Genes

Figure 2-26. Groucho-regulated genes are enriched for developmentally-regulated transcription factors. The most significantly enriched gene ontology groups of high-confidence Groucho target genes are uniformly related to developmental regulation, confirming Groucho's role as a high-level regulatory node in the establishment of tissue fate during development.

Fig. 2-26


A
Gro Targets by Dosage Response
GO:0003700 transcription factor activity
GO:0007389 pattern specification process
GO:0007423 sensory organ development
GO:0030182 neuron differentiation
GO:0007447 imaginal disc pattern formation
GO:0006357 regulation of transcription from RNA polymerase II promoter
GO:0007422 peripheral nervous system development
GO:0035282 segmentation
GO:0006928 cell motion
GO:0048729 tissue morphogenesis
GO:0035214 eye-antennal disc development
GO:0035110 leg morphogenesis
GO:0007380 specification of segmental identity, head
GO:0042063 gliogenesis
GO:0005886 plasma membrane
GO:0007400 neuroblast fate determination
GO:0007455 eye-antennal disc morphogenesis
GO:0048732 gland development
GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway
GO:0044459 plasma membrane part
0
10
20
30
-log10(p-value)

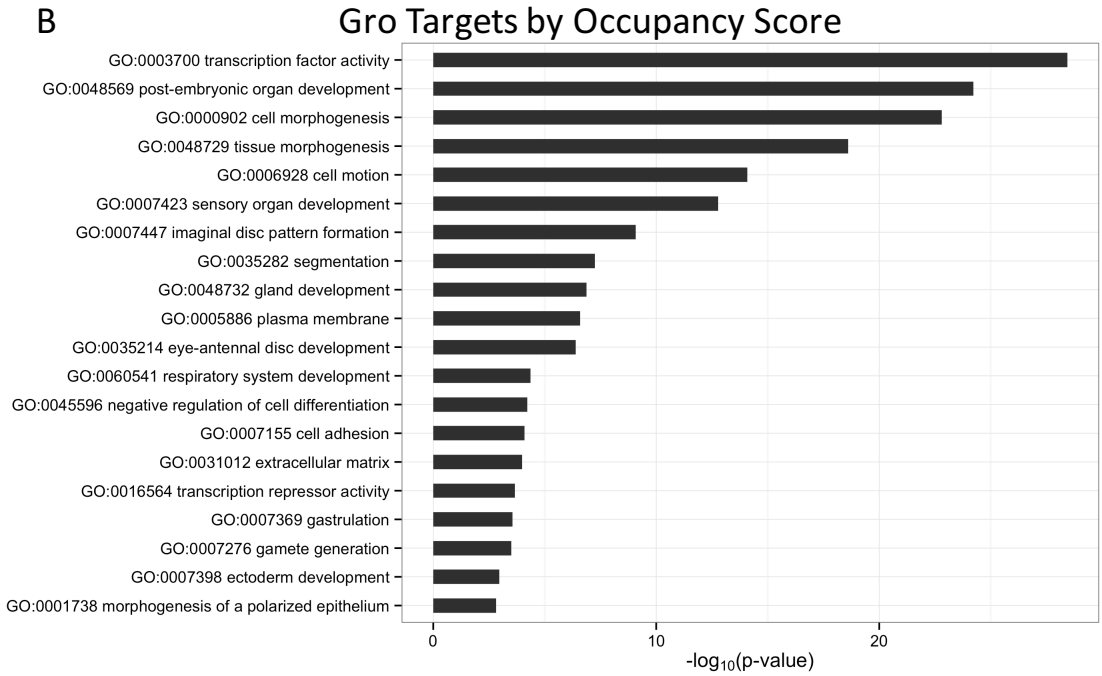

B
Gro Targets by Occupancy Score
GO:0003700 transcription factor activity
GO:0048569 post-embryonic organ development
GO:0000902 cell morphogenesis
GO:0048729 tissue morphogenesis
GO:0006928 cell motion
GO:0007423 sensory organ development
GO:0007447 imaginal disc pattern formation
GO:0035282 segmentation
GO:0048732 gland development
GO:0005886 plasma membrane
GO:0035214 eye-antennal disc development
GO:0060541 respiratory system development
GO:0045596 negative regulation of cell differentiation
GO:0007155 cell adhesion
GO:0031012 extracellular matrix
GO:0016564 transcription repressor activity
GO:0007369 gastrulation
GO:0007276 gamete generation
GO:0007398 ectoderm development
GO:0001738 morphogenesis of a polarized epithelium
0
10
20
-log10(p-value)

Figure 2.27. Groucho target genes form a highly-interconnected network with multiple hubs. Potential Groucho-target genes identified by the two methods outlined above were integrated into a network analysis to visualize genetic and physical interactions of these target genes. Genetic (blue edges) and physical (orange edges) interactions were obtained from a curated set maintained by FlyMine (Lyne et al., 2007). Both gene sets result in highly-connected networks with multiple hubs (8 or more edges, yellow nodes) interconnected by multiple genetic interactions. **(A)** The Groucho dosage responsive gene list identifies a large network containing multiple E(spl)-family proteins, as well as Delta (Dl), Sprouty (sty), Atonal (ato), and Patched (ptc) hubs. **(B)** The Gro-targets identified by Groucho occupancy score reveals a similar, but larger, network. Hubs representing genes regulated by the Decapentapletic (Dpp), Wingless (wg), and Ras/MAPK (EGFR and anterior-open;aop) pathways. Additional regulatory hubs include Thickveins (tkv), Pannier (pnr), Patched (ptc), and Cyclin G (CycG).

Fig. 2-27

A

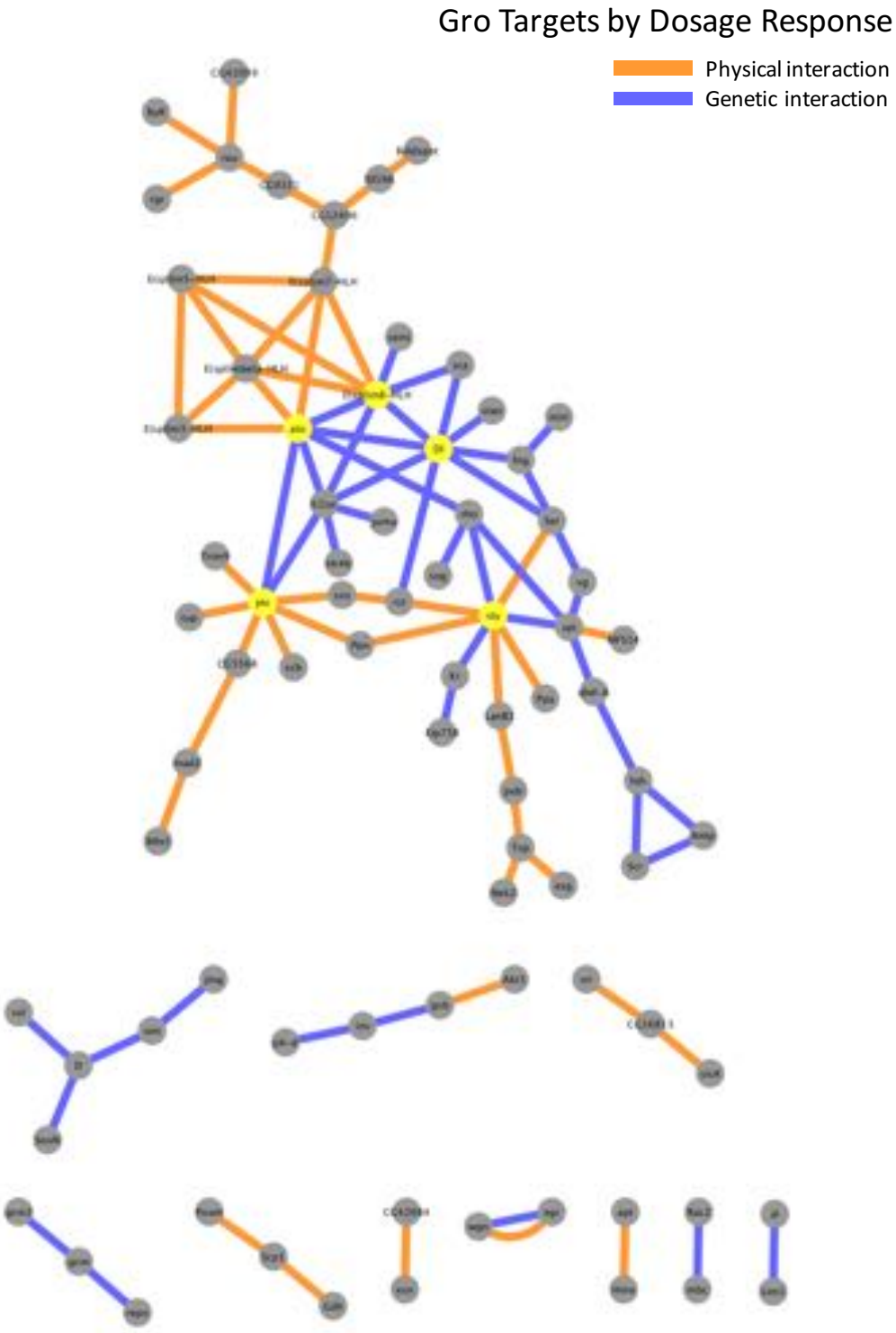
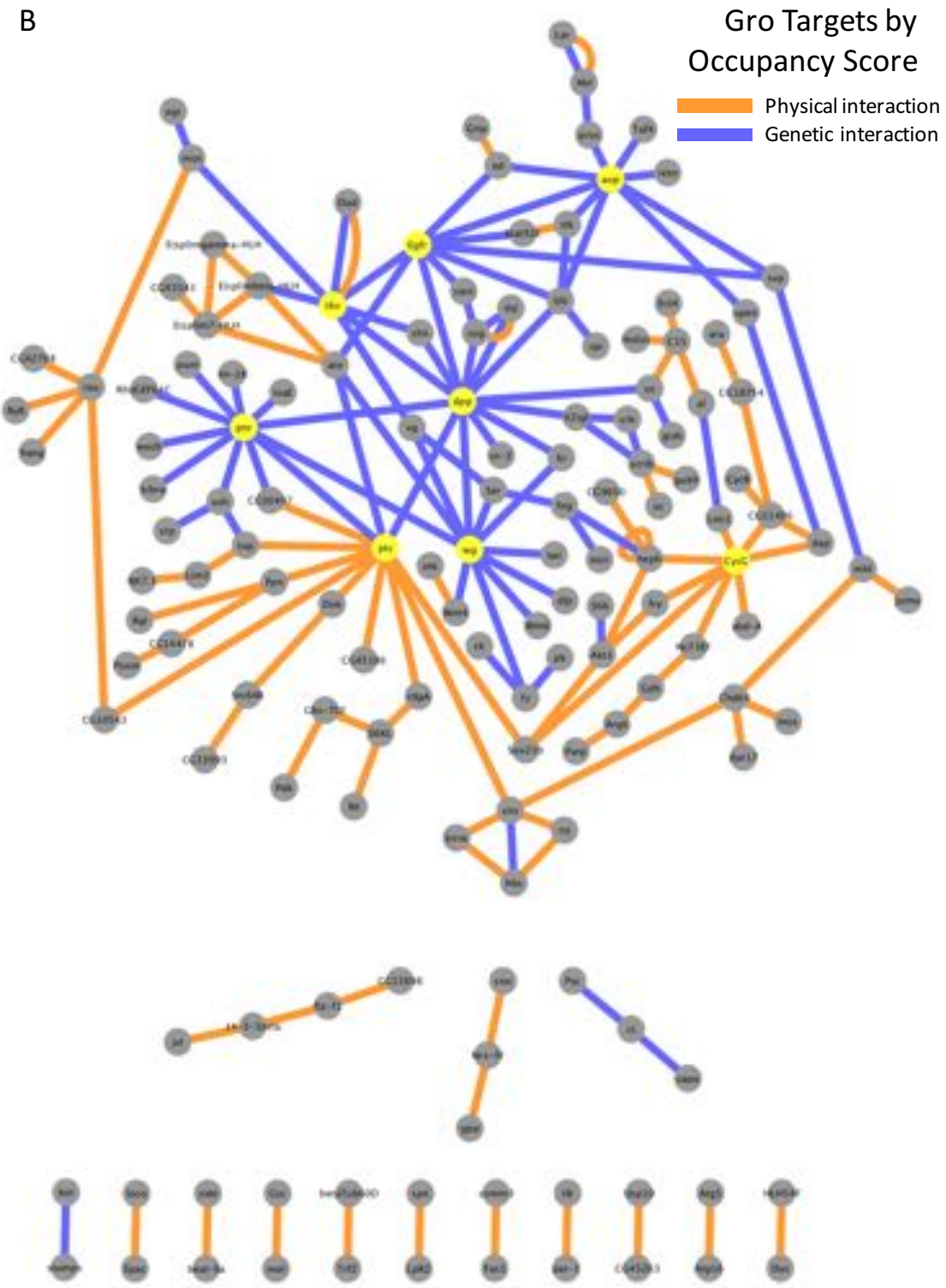

Gro Targets by Dosage Response
Physical interaction
Genetic interaction

Fig. 2-27 (cont'd)



References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F*., et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Agarwal, M., Kumar, P., and Mathew, S.J. (2015). The Groucho/Transducin-like enhancer of split protein family in animal development. IUBMB Life *67*, 472-481.

Ardehali, M.B., and Lis, J.T. (2009). Tracking rates of transcription and splicing in vivo. Nat Struct Mol Biol *16*, 1123-1124.

Barolo, S., and Levine, M. (1997). hairy mediates dominant repression in the Drosophila embryo. The EMBO Journal *16*, 2883-2891.

Bhambhani, C., Chang, J.L., Akey, D.L., and Cadigan, K.M. (2011). The oligomeric state of CtBP determines its role as a transcriptional co-activator and co-repressor of Wingless targets. The EMBO Journal *30*, 2031-2043.

Biemar, F., Nix, D.A., Piel, J., Peterson, B., Ronshaugen, M., Sementchenko, V., Bell, I., Manak, J.R., and Levine, M.S. (2006). Comprehensive identification of Drosophila dorsal-ventral patterning genes using a whole-genome tiling array. Proc Natl Acad Sci U S A *103*, 12763-12768.

Biemar, F., Zinzen, R., Ronshaugen, M., Sementchenko, V., Manak, J.R., and Levine, M.S. (2005). Spatial regulation of microRNA gene expression in the Drosophila embryo. Proc Natl Acad Sci U S A *102*, 15907-15911.

Bonn, S., Zinzen, R.P., Perez-Gonzalez, A., Riddell, A., Gavin, A.C., and Furlong, E.E. (2012). Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. Nat Protoc *7*, 978-994.

Bothma, J.P., Magliocco, J., and Levine, M. (2011). The Snail Repressor Inhibits Release, Not Elongation, of Paused Pol II in the Drosophila Embryo. Current Biology *21*, 1571-1577.

Bradnam, K.R., and Korf, I. (2008). Longer first introns are a general property of eukaryotic gene structure. PLoS One *3*, e3093.

Buscarlet, M., and Stifani, S. (2007). The 'Marx' of Groucho on development and disease. Trends in Cell Biology *17*, 353-361.

Chen, G., Nguyen, P., and Courey, A. (1998). A role for Groucho tetramerization in transcriptional repression. Molecular and Cellular Biology *18*, 7259.

Choi, C.Y., Lee, Y.M., Kim, Y.H., Park, T., Jeon, B.H., Schulz, R.A., and Kim, Y. (1999). The homeodomain transcription factor NK-4 acts as either a transcriptional activator or repressor and interacts with the p300 coactivator and the Groucho corepressor. J Biol Chem *274*, 31543-31552.

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

Consortium, T.m., Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Nègre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L.*, et al.* (2010). Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science *330*, 1787-1797.

de Celis, J.F., and Ruiz-Gomez, M. (1995). groucho and hedgehog regulate engrailed expression in the anterior compartment of the Drosophila wing. Development *121*, 3467-3476.

Dolinski, K., and Troyanskaya, O.G. (2015). Implications of Big Data for cell biology. Mol Biol Cell *26*, 2575-2578.

Dubnicoff, T., Valentine, S.A., Chen, G., Shi, T., Lengyel, J.A., Paroush, Z., and Courey, A.J. (1997). Conversion of dorsal from an activator to a repressor by the global corepressor Groucho. Genes & Development *11*, 2952-2957.

Fisher, W.W., Li, J.J., Hammonds, A.S., Brown, J.B., Pfeiffer, B.D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J.A., Eisen, M.B.*, et al.* (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. Proc Natl Acad Sci U S A *109*, 21330-21335.

Flores-Saaib, R.D., Jia, S., and Courey, A.J. (2001). Activation and repression by the C-terminal domain of Dorsal. Development *128*, 1869-1879.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W.*, et al.* (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473-479.

Hacohen, N., Kramer, S., Sutherland, D., Hiromi, Y., and Krasnow, M.A. (1998). sprouty encodes a novel antagonist of FGF signaling that patterns apical branching of the Drosophila airways. Cell *92*, 253-263.

Hasson, P., Muller, B., Basler, K., and Paroush, Z. (2001). Brinker requires two corepressors for maximal and versatile repression in Dpp signalling. EMBO J *20*, 5725-5736.

Heitzler, P., Bourouis, M., Ruel, L., Carteret, C., and Simpson, P. (1996). Genes of the Enhancer of split and achaete-scute complexes are required for a regulatory loop between Notch and Delta during lateral signalling in Drosophila. Development *122*, 161-171.

Herranz, H., and Morata, G. (2001). The functions of pannier during Drosophila embryogenesis. Development *128*, 4837-4846.

Ho, J.W., Bishop, E., Karchenko, P.V., Negre, N., White, K.P., and Park, P.J. (2011). ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. BMC Genomics *12*, 134.

Holmqvist, P.H., Boija, A., Philip, P., Crona, F., Stenberg, P., and Mannervik, M. (2012). Preferential genome targeting of the CBP co-activator by Rel and Smad proteins in early Drosophila melanogaster embryos. PLoS Genet *8*, e1002769.

Huang, J.D., Schwyter, D.H., Shirokawa, J.M., and Courey, A.J. (1993). The interplay between multiple enhancer and silencer elements defines the pattern of decapentaplegic expression. Genes Dev *7*, 694-704.

IAnders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Ip, Y.T., Park, R.E., Kosman, D., Yazdanbakhsh, K., and Levine, M. (1992). dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. Genes Dev *6*, 1518-1530.

Jarman, A.P., Grell, E.H., Ackerman, L., Jan, L.Y., and Jan, Y.N. (1994). Atonal is the proneural gene for Drosophila photoreceptors. Nature *369*, 398-400.

Jennings, B.H., Pickles, L.M., Wainwright, S.M., Roe, S.M., Pearl, L.H., and Ish-Horowicz, D. (2006). Molecular recognition of transcriptional repressor motifs by the WD domain of the Groucho/TLE corepressor. Mol Cell *22*, 645-655.

Jennings, B.H., Wainwright, S.M., and Ish-Horowicz, D. (2007). Differential in vivo requirements for oligomerization during Groucho-mediated repression. EMBO reports *9*, 76-83.

Jiang, L., Li, X.N., and Niu, D.K. (2014). Higher frequency of intron loss from the promoter proximally paused genes of Drosophila melanogaster. Fly (Austin) *8*, 120-125.

Kaplan, T., Li, X.Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D., and Eisen, M.B. (2011). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet *7*, e1001290.

Kaul, A., Schuster, E., and Jennings, B.H. (2014). The Groucho Co-repressor Is Primarily Recruited to Local Target Sites in Active Chromatin to Attenuate Transcription. PLoS Genetics *10*, e1004595.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

Kirov, N., Childs, S., O'Connor, M., and Rushlow, C. (1994). The Drosophila dorsal morphogen represses the tolloid gene by interacting with a silencer element. Mol Cell Biol *14*, 713-722.

Kok, K., Ay, A., Li, L.M., and Arnosti, D.N. (2015). Genome-wide errant targeting by Hairy. Elife *4*.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P.*, et al.* (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research *22*, 1813-1831.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Li, L.M., and Arnosti, D.N. (2011). Long- and Short-Range Transcriptional Repressors Induce Distinct Chromatin States on Repressed Genes. Current Biology *21*, 406-412.

Li, X.-Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C.L.L., *et al.* (2008). Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. PLoS Biology *6*, e27.

Li, X.Y., Thomas, S., Sabo, P.J., Eisen, M.B., Stamatoyannopoulos, J.A., and Biggin, M.D. (2011). The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol *12*, R34.

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

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., McLaren, P., North, P., *et al.* (2007). FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome Biol *8*, R129.

Ma, W., Noble, W.S., and Bailey, T.L. (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nat Protoc *9*, 1428-1450.

MacArthur, S., Li, X.-Y., Li, J., Brown, J.B., Chu, H.C., Zeng, L., Grondona, B.P., Hechmer, A., Simirenko, L., Keränen, S.V., *et al.* (2009). Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome biology *10*, R80.

Mannervik, M. (2014). Control of Drosophila embryo patterning by transcriptional co-regulators. Experimental cell research *321*, 47-57.

Martinez, C.A., and Arnosti, D.N. (2008). Spreading of a Corepressor Linked to Action of Long-Range Repressor Hairy. Molecular and Cellular Biology *28*, 2792-2802.

Matyash, A., Chung, H.R., and Jackle, H. (2004). Genome-wide mapping of in vivo targets of the Drosophila transcription factor Kruppel. J Biol Chem *279*, 30689-30696.

Moorman, C., Sun, L.V., Wang, J., de Wit, E., Talhout, W., Ward, L.D., Greil, F., Lu, X.J., White, K.P., Bussemaker, H.J., *et al.* (2006). Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proc Natl Acad Sci U S A *103*, 12027-12032.

Negre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R., *et al.* (2011). A cis-regulatory map of the Drosophila genome. Nature *471*, 527-531.

Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A., and Loraine, A.E. (2009). The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics *25*, 2730-2731.

Nie, M., Xie, Y., Loo, J.A., and Courey, A.J. (2009). Genetic and Proteomic Evidence for Roles of Drosophila SUMO in Cell Cycle Control, Ras Signaling, and Early Pattern Formation. PLoS ONE *4*, e5905.

Paroush, Z., Finley, R.L., Jr., Kidd, T., Wainwright, S.M., Ingham, P.W., Brent, R., and Ish-Horowicz, D. (1994). Groucho is required for Drosophila neurogenesis, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. Cell *79*, 805-815.

Payankaulam, S., and Arnosti, D.N. (2009). Groucho corepressor functions as a cofactor for the Knirps short-range transcriptional repressor. Proceedings of the National Academy of Sciences *106*, 17314-17319.

Petesch, S.J., and Lis, J.T. (2008). Rapid, transcription-independent loss of nucleosomes over a large chromatin domain at Hsp70 loci. Cell *134*, 74-84.

Ratnaparkhi, G.S., Jia, S., and Courey, A.J. (2006). Uncoupling dorsal-mediated activation from dorsal-mediated repression in the Drosophila embryo. Development *133*, 4409-4414.

Roth, S., Stein, D., and Nüsslein-Volhard, C. (1989). A gradient of nuclear localization of the dorsal protein determines dorsoventral pattern in the Drosophila embryo. Cell *59*, 1189-1202.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E.E. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev *21*, 436-449.

Sekiya, T., and Zaret, K.S. (2007). Repression by Groucho/TLE/Grg Proteins: Genomic Site Recruitment Generates Compacted Chromatin In Vitro and Impairs Activator Binding In Vivo. Molecular Cell *28*, 291-303.

Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E.G., Singaravelu, K., and Beyer, A. (2013). Assessing Computational Methods for Transcription Factor Target Gene Identification Based on ChIP-seq Data. PLoS Computational Biology *9*, e1003342.

Song, H., Hasson, P., Paroush, Z.a.e., and Courey, A.J. (2004). Groucho oligomerization is required for repression in vivo. Molecular and Cellular Biology *24*, 4341-4350.

Thisse, B., el Messal, M., and Perrin-Schmitt, F. (1987). The twist gene: isolation of a Drosophila zygotic gene necessary for the establishment of dorsoventral pattern. Nucleic Acids Res *15*, 3439-3453.

Turki-Judeh, W., and Courey, A.J. (2012a). Groucho: A Corepressor with Instructive Roles in Development. In (Elsevier), pp. 65-96.

Turki-Judeh, W., and Courey, A.J. (2012b). The Unconserved Groucho Central Region Is Essential for Viability and Modulates Target Gene Specificity. PLoS ONE *7*, e30610.

Valentine, S.A., Chen, G., Shandala, T., Fernandez, J., Mische, S., Saint, R., and Courey, A.J. (1998). Dorsal-mediated repression requires the formation of a multiprotein repression complex at the ventral silencer. Mol Cell Biol *18*, 6584-6594.

Villanueva, C.J., Waki, H., Godio, C., Nielsen, R., Chou, W.-L., Vargas, L., Wroblewski, K., Schmedt, C., Chao, L.C., Boyadjian, R.*, et al.* (2011). TLE3 Is a Dual-Function Transcriptional Coregulator of Adipogenesis. Cell Metabolism *13*, 413-427.

Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C.A., Zhang, Y., and Liu, X.S. (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nature Protocols *8*, 2502-2515.

Winkler, C.J., Ponce, A., and Courey, A.J. (2010). Groucho-Mediated Repression May Result from a Histone Deacetylase-Dependent Increase in Nucleosome Density. PLoS ONE *5*, e10166.

Yenerall, P., Krupa, B., and Zhou, L. (2011). Mechanisms of intron gain and loss in Drosophila. BMC Evol Biol *11*, 364.

Zeitlinger, J., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes & Development *21*, 385-390.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based Analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Chapter 3

Investigating the dynamics of the embryonic transcriptome

Abstract

In dynamic systems such as the *Drosophila* embryo, the transcription rates of genes can rapidly fluctuate in response to regulatory events. The levels of processed mRNA, as measured by transcriptome assays such as RNA-seq, therefore are less accurate as measurements of such transcription rates, as mRNA levels are not at a steady-state. Nascent-seq is one method developed to assay the transcription rate of genes directly in these dynamic systems, bypassing the effect of differential rates of transcript synthesis, processing, and degradation on mRNA accumulation. It accomplishes this by isolating and sequencing only those transcripts bound within ternary elongation complex and therefore in the process of synthesis.

In this study, we apply Nascent-seq to embryos at multiple stages of development to measure these transcription rates. From this data, we obtain information about the developmental stage-specific expression of each gene, as well as the distribution of nascent transcript lengths for each gene. The latter serves as a readout for patterns of accumulated positioning of RNA Polymerase II within gene bodies indicative of promoter-proximal pausing. We observe that Groucho-regulated genes at each stage of development are enriched for promoter-proximal paused polymerase. This holds true of both genes with internal or adjacent Gro binding, and genes up-regulated, but not down-regulated in Gro loss-of-function embryos.

Introduction

Drosophila development involves the coordinated expression of a vast number of genes under strict temporal and spatial control (Brown et al., 2014). Transcript levels in the embryo are therefore highly dynamic, undergoing rapid shifts in level dependent on altered rates of accumulation, processing, and degradation. In excess of ~7,000 unique transcripts, arising from at least 3,000 distinct genes are expressed from the earliest onset of zygotic transcription (De Renzis et al., 2007; Graveley et al., 2011) . Of these, at least 1,000 are expressed in a spatially restricted manner (Tomancak et al., 2002). Regulatory systems go to great lengths to minimize even relatively slight stochastic changes in expression, which can nevertheless have a negative effect on viability (Perry et al., 2010). Additional processes influencing mRNA abundance are also tightly controlled, with regulated degradation pathways resulting in significant and transcript-specific differences in the kinetics of mRNA decay in the embryo (Thomsen et al., 2010). Ideally, accurate and quantitative measurements of protein abundance could be utilized to investigate the mechanics of development, but lacking such a technique, mRNA abundance is often substituted as a measurement providing a strong correlation with protein abundance (Fu et al., 2009).

A complicating factor in the interpretation of transcriptome data from the early *Drosophila* embryo arises from the significant effects of maternally-contributed mRNAs and the timeline of activation of the zygotic genome. This latter process, the maternal-to-zygotic transition (MZT) is a common feature of animal development and

encompasses a cascade of processes whereby widespread alterations to the chromatin landscape are engendered by the activity of pioneering transcription factors (Li et al., 2014; Tadros and Lipshitz, 2009) . In *Drosophila,* the zinc-finger transcription factor Zelda is a well-studied example of such a pioneering factor (Harrison et al., 2011; Liang et al., 2008) (Xu et al., 2014). In most animals, the MZT process consists of two distinct stages. An initial "minor wave" of activation becomes significant at ~1.5 hours post-fertilization in *Drosophila*, followed by a more rapid and synchronous "major wave" at ~2.5 hours (Pritchard and Schubiger, 1996). During this time the majority of the maternally-contributed transcriptome is destabilized and undergoes coordinated degradation (Benoit et al., 2009; Tadros et al., 2007).

As mRNA abundance is a complex process, determined by the integrative inputs of the rates of transcription, processing, and degradation, measuring the rates of transcription in a temporally-discriminate manner becomes challenging in an evolving system such as the embryo. A number of techniques have been proposed to address this question, one popular technique being the direct sequencing of RNA populations enriched for nascent and chromatin-associated RNAs (nascent-seq). Nascent RNA-seq, or nascent-seq, has been shown to be an effective strategy for identifying the actively transcribed genes in a cell or tissue, as well as quantitating the relative transcriptional rate of these genes. This is accomplished through the use of existing deep-sequencing platforms to specifically sequence the fraction of RNA that is chromatin-associated, and therefore enriched for transcripts undergoing active elongation. By using this method in *Drosophila melanogaster* embryos, we seek to obtain a timeline of transcriptional

activation and repression to a high degree of temporal accuracy, which will aid us in identifying genes regulated by Groucho as well as the timeframes over which this regulatory ability is exercised.

Nascent-seq has been successfully applied to track the transcriptional changes in a number of biological contexts, including macrophages (Bhatt et al., 2012), where it was utilized to obtain a timeline of transcriptional changes following induction of an immune response, adult *D. melanogaster* tissues to analyze the prevalence of cotranscriptional splicing (Khodor et al., 2011), and circadian transcript cycling (Rodriguez et al., 2013), in which the authors saw significant differences in total mRNA and nascent mRNA levels between consecutive ninety minute embryo collections. We have adopted the method to developing embryos, using an established protocol for embryo nuclei isolation (Nechaev et al., 2010) followed by isolation of a chromatin-associated fraction from these nuclei. Purification of RNA from the chromatin fraction yields a RNA pool significantly enriched for nascent RNA.

Integrating this data with whole RNA-seq data will additionally aid in eliminating false-positives from our derived list of Groucho-regulated genes. As Groucho's ability to repress transcription is regulated both spatially and temporally throughout development, discreet measurements of transcription over time will allow us to more accurately describe and understand Groucho's roles in fly development.

Materials & Methods

Chromatin-associated RNA isolation in embryos

Wild-type (OregonR) fly embryos were collected in three 2.5 hour cohorts beginning 1.5 hours post-deposition. Between 3 to 5 grams of embryos were utilized for each fractionation. The chromatin-associated RNA isolation protocol was adapted from Nechaev et al. (2010) and Khodor et al. (2011). Embryos were dechorionated in 50% bleach for 90 sec and transferred to a chilled Dounce homogenizer. Embryos were then rinsed three times with 25 ml of homogenization buffer (15 mM HEPES-KOH pH 7.6; 10 mM KCl; 3 mM $CaCl_2$; 2 mM $MgCl_2$; 0.1% Triton X-100; 1 mM DTT; 0.1 mM PMSF; 0.1x RNAase inhibitor). Embryos were then suspended in homogenization buffer containing 0.3 M (15 ml) sucrose and dounced five times each with loose and tight pestles. Embryo lysate was filtered through 50-micron nylon cell strainer. Clarified lysate was layered over a sucrose cushion consisting of a layer of 1.7 M sucrose (15 ml) underneath a layer of 3 M sucrose (15 ml) in homogenization buffer.. The samples were centrifuged at 15,000 RCF for 10 min at 4˚C. Pelleted nuclei were resuspended in 250 µl of nuclear lysis buffer (10 mM HEPES-KOH pH 7.6; 100 mM KCl; 0.1 mM EDTA; 10% glycerol; 0.15 mM spermine; 0.5 mM spermidine; 0.1 mM NaF; 0.1 mM Na_3VO_4; 0.1 mM $ZnCl_2$; 1 mM DTT; 0.1 mM PMSF; 1x RNAase inhibitor). While gently vortexing, an equal volume of NUN buffer (25 mM HEPES-KOH pH 7.6; 300 mM NaCl; 1M urea; 1% NP-40; 1 mM DTT; 0.1 mM PMSF) was added drop-by-drop over a period 5 minutes. Condensed chromatin became visible as a fluffy white precipitate. The solution was then incubated for 20 min

on ice and centrifuged at 14,000 rpm for 30 min at 4˚C. The supernatant (primarily nucleoplasm) was discarded and the pellet was resuspended in Trizol reagent (Qiagen). RNA was then purified following the manufacturer's protocol.

rRNA removal

RNA samples were depleted of ribosomal, poly(A)+, and additional RNA contaminants through an affinity depletion procedure adopted from Khodor et a. (2011). An equimolar mixture of biotinylated affinity oligomers (Table 3-1; Eurofins MWG Operon) was added to 6 µg of purified RNA in annealing buffer (10 mM EDTA; 0.5x SSC) in a volume of 100 µl. RNA was denatured at 75˚C for 5 min and annealed at 37˚C for 30 min. Annealed mixture was added to 1ml streptavidin paramagnetic beads (Promega) and incubated at 25˚C for 15 min, followed by 2 hours at 4˚C with gentle rocking, and the supernatant retained for library preparation. This procedure was performed twice per sample.

RNA-seq library construction and sequencing

rRNA-depleted RNA was concentrated via ethanol precipitation. Size distribution of samples was determined via Agilent 2100 Bioanalyzer (Agilent Technologies). Indexed RNA-seq libraries were generated with the ScriptSeq v2 RNA-seq Library Preparation Kit (Epicentre). Sequencing was performed on Illumia HiSeq 2000 sequencing platform

(High Throughput Sequencing Core Facility, Broad Stem Cell Research Center, UCLA). Reads were demultiplex via custom scripts and mapped to the BDGP5/dm3 *D. melanogaster* genome with Tophat2 (v2.1.0) (Kim et al., 2013) using the following parameters: -g 1 –solexa1.3-quals. A gene model annotation (iGenomes UCSC dm3) was provided as a mapping guide. Assignment of mapped reads to transcripts was performed with HTSeq (IAnders et al., 2015).

Data analysis

Mean normalized transcript expression levels (FPKM) were generated with DESeq2 (v1.10.0) (Love et al., 2014). Significant changes in transcript abundance were quantified with the same software by comparison with poly(A)+ RNA-seq from wild-type embryo data described in Chapter 2 of this thesis. RNA-seq read mapping density analysis was performed using PicardTools (http://broadinstitute.github.io/picard/). Additional metagene analysis was performed using the 'metagene' package of R/Bioconductor (Beauparlant, 2014).

Results

RNA from fractionated embryos exhibits multiple characteristics of nascent pre-mRNA

Total RNA was extracted from chromatin isolated from *D. melanogaster* embryos collected over three time spans in early development and subjected to high-throughput sequencing. This chromatin-associated RNA is expected to be enriched for nascent transcripts, as well as additional RNA species associated with chromatin in structural, catalytic, or regulatory capacities (Cernilogar et al., 2011). Isolated RNA was affinity-depleted for polyadenylated RNA in order to further minimize the contribution of mature mRNA from analysis. The level of enrichment for nascent transcript was validated and quantified through various measures. The efficiency of chromatin isolation was confirmed through analysis of protein compartmental markers (Fig. 3-1A/B), confirming that the sequenced RNA was derived from a embryonic fraction enriched for histones and deficient for a cytoplasmic marker. Sequencing reads obtained from mature transcripts ideally map uniformly across genes, though this is dependent on the quality of the RNA utilized for library generation. Non-uniformity in poly(A)+ libraries generally manifests as a 3' bias in mappable reads as a result of partially fragmented mRNA being purified by affinity selection to polyA sequence (Roberts et al., 2011). Chromatin-associated RNA exhibits a significant 5' bias at each developmental stage, and is partially depleted at the 3' end (Fig. 3-2). The large enrichment of reads arising from the initial 15% of gene bodies may be indicative of promoter-proximal paused polymerase. The sharp decrease in read occupancy near the 3' terminus may

result from frequent polymerase pausing in terminal exons (Carrillo Oesterreich et al., 2010). Pausing in terminal exons is thought to promote recognition of polyadenylation sites and transcriptional termination (Gromak et al., 2006).

Chromatin-associated RNA is enriched for intronic sequence when compared to poly(A)+ libraries prepared from the same developmental stages (Fig. 3-3). Our data indicates that on average, 13% of poly(A)+ RNA-seq reads map to constitutive introns compared to 35% of chromatin-associated RNA reads. While 60-70% of gene sequence in *D. melanogaster* is annotated as intronic, the large majority of introns are believed to be cotranscriptionally spliced, with only 16% of introns exhibiting little or no splicing (Khodor et al., 2011; Wuarin and Schibler, 1994) . Therefore, an intronic content of between 13 and 60% should be expected for a library enriched for pre-mRNA.

The levels of many nascent transcripts differ significantly from levels of mature mRNA

Analysis of nascent pre-mRNA levels in multiple contexts has shown that the rate of accumulation of a particular transcript can be strongly uncoupled from the rate of transcript synthesis, owing to differential rates of accumulation, processing, and degradation (Khodor et al., 2011; Rodriguez et al., 2013) . In developmental contexts, a significant proportion of the transcriptome is far from steady-state. Comparison of chromatin-associated RNA transcript profiles with mRNA profiles obtained from the same timepoints by principal component analysis indicates significant differences, with the majority of expressed genes exhibiting some deviance in expression rate and

accumulation level (Fig. 3-4). Samples continue to cluster by developmental time stage, but segregate first by degree of "nascentness." Comparison of the normalized expression levels of each gene indicates that many genes exhibit comparable levels of expression in poly(A)+ and nascent samples (Fig 3-5). A small number of genes are significantly enriched in the nascent population, however, while showing very little accumulation in the poly(A)+ RNA-seq data. A large fraction of these correspond to non-polyadenylated RNAs, including histones, snRNAs, and snoRNAs. These non-polyadenylated RNA species were removed from further analysis.

Of the remaining transcripts, a significant number were found to be under- or over-represented in the nascent mRNA pool in comparison to mature mRNA, comprising between 40 and 50% of all expressed genes across timepoints (Fig. 3.6A). Analysis of these genes in the 1.5 – 4 hr developmental window reveals differences in the expression patterns enriched in the over- and under-represented gene sets. Genes with lower abundance in nascent RNA are enriched for maternally deposited genes, consistent with these genes being transcribed before nascent RNA was isolated and less frequently zygotically transcribed. Over-represented nascent transcripts are enriched for spatially constricted expression within portions of the embryo, and are therefore enriched for genes being actively transcribed.

Groucho-regulated genes are enriched for stalled RNA polymerase

Promoter-proximal pausing of RNA Polymerase II has been identified as a crucial step in gene regulation. Pausing was originally characterized in *Drosophila* at multiple heat-shock genes, presumably to facilitate rapid induction of gene expression upon receipt of an appropriate regulatory signal (Lis and Wu, 1993). Since this discovery, polymerase stalling has been found to be a ubiquitous regulatory mechanism in higher eukaryotes (Conaway et al., 2000), with strong peaks of PolII present in the promoter regions of a diverse array of genes throughout the *Drosophila* genome. Expression of the majority of protein-coding genes in humans is regulated to some degree after the initiation of transcription (Guenther et al., 2007), as is a large fraction of the *Drosophila* developmental genome (Zeitlinger et al., 2007).

To explore whether Groucho regulation potentially promotes the stalling of polymerase *in vivo*, we undertook to compare Groucho-regulated genes with publically available genome-wide PolII localization data (Zeitlinger et al., 2007). In this data set, the authors classified each gene into one of several states including the lack of detected PolII, active (elongation phase) PolII, or stalled PolII. Genes found to bind Gro internally or in adjacent intergenic regions were found to be significantly enriched for stalled PolII at each timepoint (Fig. 3-7A). We observe a significant correlation between PolII pausing within genes and those genes becoming up-regulated in Gro loss-of-function embryos ($p < 10^{-10}$), while no significant correlation is observed between PolII pausing and down-regulation in Gro loss-of-function embryos ($p > 0.05$) (Fig. 2-7B). Conversely, down-regulated genes are enriched for active PolII ($p < 10^{-10}$), while Gro repressed genes are not. Together, this provides evidence that, at least at early developmental timepoints, a

significant fraction of Groucho-associated genes exhibit characteristics of PolII pausing, suggesting that retention or prevention of PolII from transitioning to an active elongation complex is a potential mechanism of Groucho-dependent repression.

Groucho target genes are enriched for promoter-proximal read density indicative of polymerase pausing

Evidence presented in the previous section hypothesized that many Groucho-repressed genes possess significant levels of promoter-proximal stalled polymerase. Much of this was established using previously-published data from the 2 -4 hour *Toll*^{10B} mutant embryos (Zeitlinger et al., 2007). These embryos generate a more homogenous population of cells, as all portions of the embryo adopt a cell type representative of the presumptive mesoderm (Schneider et al., 1991), and so simplified the embryo-wide classification of PolII pausing state. Our data allows us to quantify the accumulation of promoter-proximal nascent transcript at later stages of development, albeit in a more heterogeneous population of cell types. This heterogeneity limits the interpretation of Groucho's involvement with promoter-proximal stalled polymerase, as we can determine whether a gene is regulated by Groucho and possesses stalled PolII at each developmental time span, but we cannot make definite conclusions as to whether those events are occurring in identical populations of cells. A correlation is still informative, as association of the two states potentially represents a program of regulation whereby

Groucho either promotes stalling itself, or is recruited to repress genes that undergo stalling at the same developmental stage but in different tissues.

Focusing on genes that are responsive to increasing levels of maternal Groucho overexpression, we see that at all three timepoints genes negatively regulated by increased Groucho dosage are enriched for promoter-proximal accumulation of transcript when compared both to genes up-regulated in this genetic background as well as unresponsive genes (Fig. 3-8).

Discussion

Quantification of chromatin-associated pre-mRNA is a useful metric for the exploration of dynamic transcriptional systems such as the *Drosophila* embryo. The relatively high stability of the RNA Polymerase II ternary elongation complex facilitates the purification of nascent transcripts in a highly specific manner, thereby enabling us to more thoroughly characterize the dynamics of this transcriptional system and relate aspects of gene expression to the activity of Groucho. We find that chromatin-associated RNA is enriched for nascent transcripts, as evidenced by the increase in unspliced intronic content, a 5' bias in read density, and enrichment for actively transcribed genes in early stages of development. In addition to a modest 5' bias throughout the gene body, nascent RNA exhibits a significant spike in transcript density at the 5' transcription start site, likely corresponding to nascent transcripts locked in stalled or slowed ternary PolII complexes. Investigations of stalled PolII in the embryo have previously shown that in 2-4 hour embryos, 12% of all protein-coding genes have stalled promoter-proximal PolII (Zeitlinger et al., 2007). Additionally, purification of chromatin-associated RNA from *Drosophila* S2 cells predicted that 30% of protein-coding genes experienced some degree of PolII pausing, characterized by an enrichment of 5' transcripts (Nechaev et al., 2010).

The manner in which PolII pausing is utilized to regulate transcription remains poorly understood, though multiple non-exclusive mechanisms have been proposed, (Adelman and Lis, 2012). One of these mechanisms posits that sustained or transient pausing facilitates the participation of additional regulatory elements in the

determination of transcriptional activity (Nechaev and Adelman, 2008). This allows the expression level of a gene to be regulated through multiple, independent pathways, potentially at the behest of independent signaling pathways (Blau et al., 1996). Combinatorial control of gene expression is a common regulatory motif in eukaryotes, so it is feasible that the capability to exert influence over expression both before the assembly of the PolII complex as well as after transcription has began would be useful in such scenarios. Members of the Rel family of transcription factors, of which the Groucho-interactor Dorsal is a member, have been found to promote both PolII pausing and release in mammals (Barboric et al., 2001).

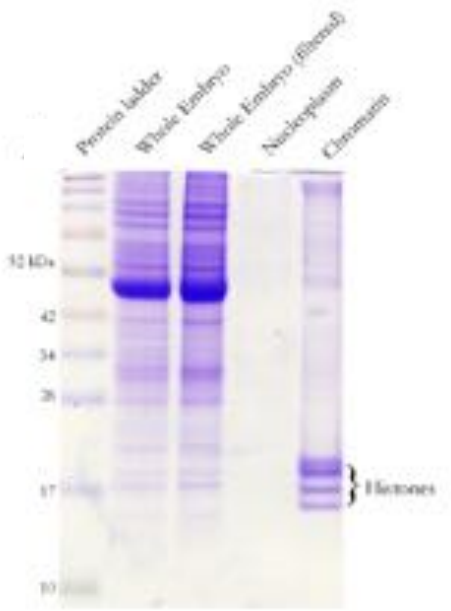
We find that Groucho-regulated genes are enriched for paused PolII in the early embryo, and that this correlation applies to both genes bound by Gro at each time window, as well as genes differentially expressed in Gro loss-of-function and gain-of-function embryos. In Gro loss-of-function embryos, up-regulated genes are enriched for stalled PolII, while down-regulated genes are enriched for active PolII. The converse is true in embryos overexpressing Gro. PolII stalling in embryos has been hypothesized to enable the rapid, synchronous activation of genes across the embryo, as opposed to stochastic activation observed from genes lacking poised PolII (Boettiger and Levine, 2009). Given that genes possessing stalled PolII often continue to be expressed at high levels (Nechaev and Adelman, 2008; Rougvie and Lis, 1990), PolII stalling in Gro-regulated genes may not be a primary mechanism of repression, but instead indicate that these genes are primed for rapid activation once Gro-mediated repression is

relieved. The ability of Gro-mediated repression to be rapidly reversible may be an important aspect of its activity in the embryo.

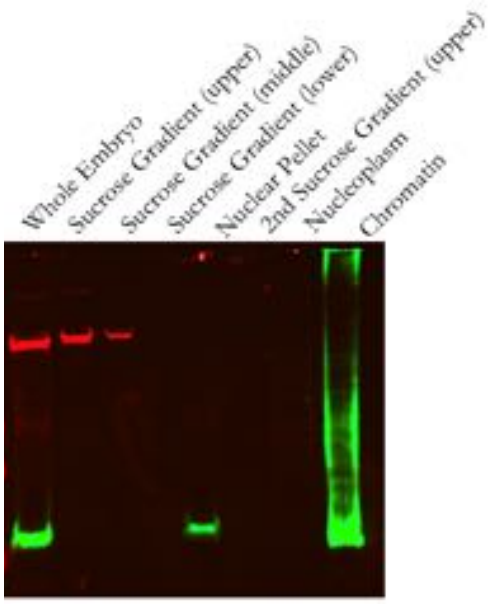
Figure 3-1. RNA enriched for nascent transcripts was isolated from the chromatin fraction of fractionated embryos. (A) A typical fractionation resulted in a chromatin fraction enriched for proteins consistent with *Drosophila* histone cores. **(B)** Fractionation was confirmed via immunoblot. Chromatin fractions were enriched for the histone H3 (green) and depleted for tubulin, which is predominately cytoplasmic (red).

Fig. 3-1

A



B



α-tubulin
α-H3

Figure 3-2. Nascent RNA is enriched for reads originating from the 5' end of transcripts and depleted for 3' transcript reads. The distribution of mappable sequencing reads generated via RNA-seq of chromatin-associated RNA was calculated for each expressed gene. Gene distributions were then normalized for total gene length and expression level. The resulting metagene distribution shows a strong increase in read density arising from 5' portions of expressed genes, consistent with the isolation of incompletely transcribed transcripts.

Fig. 3-2

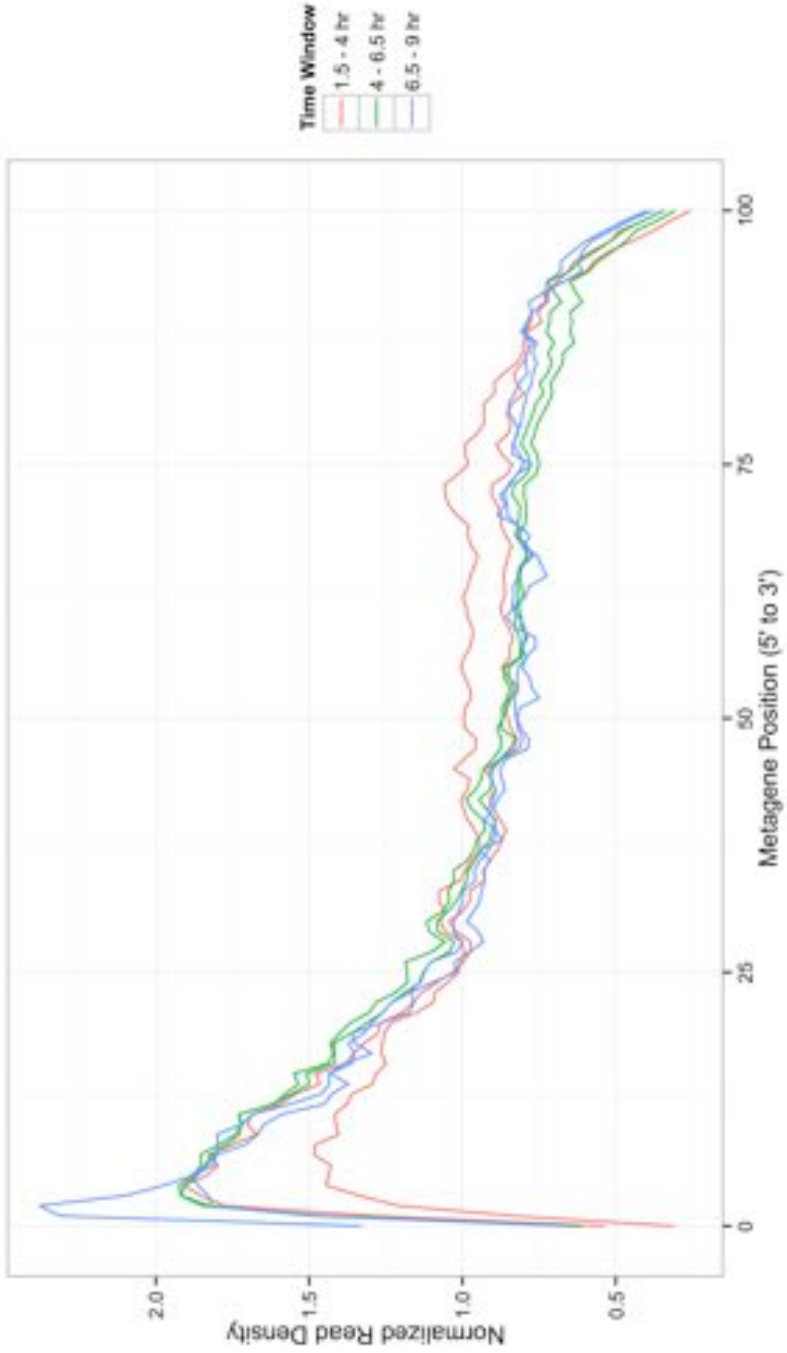

Time Window
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
Normalized Read Density
0.5
1.0
1.5
2.0
0
25
50
75
100
Metagene Position (5' to 3')

Figure 3-3. Chromatin-associated RNAs are enriched for unspliced introns.

Cotranscriptional splicing is widespread in *Drosophila,* with the majority of introns being cotranscriptionally processed; some introns are known to be spliced post-transcriptionally (Khodor et al., 2011). Chromatin-associated RNA remains enriched for intronic content compared to poly(A)-selected mRNA obtained at the same embryonic stages. Over twice as much chromatin-associated RNA maps to intronic sequences as compared to poly(A)-selected (35% vs. 13%), indicating that these samples are enriched for nascent transcripts.

Fig. 3-3

Condition	Time window	Total Reads	Transcript Reads	%	Exon Reads	%	Intron Reads	%	5'UTR Reads	%	3'UTR Reads	%
Nascent	1.5 - 4 hr	13,396,998	12,869,419	96%	9,630,349	72%	3,991,330	30%	827,902	6%	717,773	5%
Nascent	1.5 - 4 hr	14,402,702	14,074,143	98%	13,143,526	91%	1,256,036	9%	384,863	3%	405,031	3%
Nascent	4 - 6.5 hr	19,102,357	18,102,323	95%	10,249,092	54%	9,274,889	49%	1,495,488	8%	1,038,051	5%
Nascent	4 - 6.5 hr	15,913,811	15,082,718	95%	8,935,177	56%	7,285,111	46%	1,198,370	8%	883,736	6%
Nascent	6.5 - 9 hr	15,747,237	15,201,712	97%	10,200,282	65%	5,976,815	38%	1,035,707	7%	853,559	5%
Nascent	6.5 - 9 hr	13,952,494	13,493,631	97%	8,784,412	63%	5,694,037	41%	1,117,133	8%	897,345	6%
poly(A) RNA	1.5 - 4 hr	23,796,734	20,604,743	87%	19,752,186	83%	2,168,260	9%	975,884	4%	3,986,880	17%
poly(A) RNA	1.5 - 4 hr	38,083,224	30,341,487	80%	27,976,746	73%	4,214,078	11%	1,392,513	4%	5,761,724	15%
poly(A) RNA	4 - 6.5 hr	37,870,358	32,424,416	86%	28,651,441	76%	5,631,405	15%	1,385,828	4%	6,037,661	16%
poly(A) RNA	4 - 6.5 hr	36,076,928	29,887,172	83%	26,704,160	74%	4,787,606	13%	1,198,679	3%	5,352,807	15%
poly(A) RNA	6.5 - 9 hr	30,597,716	26,509,005	87%	24,452,990	80%	3,777,231	12%	1,258,219	4%	5,190,917	17%
poly(A) RNA	6.5 - 9 hr	26,551,585	23,282,466	88%	20,952,590	79%	3,886,388	15%	1,196,059	5%	4,597,938	17%

Figure 3-4. Chromatin-associated RNA samples present a significantly different transcriptional profile in comparison to poly(A)-selected RNA. Principal component analysis was performed to compare nascent-seq transcriptome profiles to the wild-type and Gro mutant embryos presented in Chapter II. This technique allows the visualization of global correlation of largely multidimensional data in two dimensions. Each point represents a transcriptome profile (normalized expression level across all expressed genes). The distance between two points is inversely proportional to the overall similarity of those two points (closer = more similar). The procedure defines the two axes to encompass the largest variance between samples. This analysis indicates that nascent-seq samples differ significantly from poly(A)+ RNA-seq samples, with the x-axis discriminating the two types of RNA. The y-axis encompasses the second largest contributor of variance between samples, in this case the developmental stage of the transcriptome being profiled.

Fig. 3-4

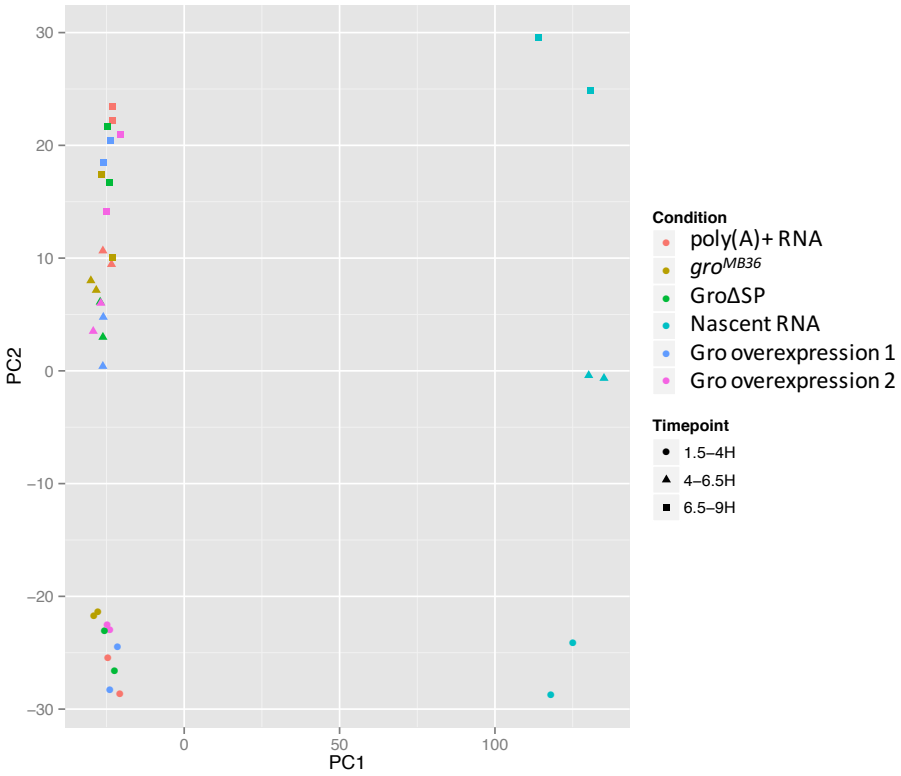

30
20
10
0
−10
−20
−30
PC2
0
50
100
PC1
Condition
poly(A)+ RNA
groMB36
GroΔSP
Nascent RNA
Gro overexpression 1
Gro overexpression 2
Timepoint
1.5–4H
4–6.5H
6.5–9H

Figure 3-5. A small number of transcripts are significantly over-represented in chromatin-associated RNA. The nascent-seq and poly(A)+ RNA-seq expression levels of all annotated genes in wild-type embryos were normalized via the FPKM method. This normalization method accounts both for differences in library size, as well as the length of each gene, such that the expression levels of different genes in different samples are meaningfully comparable. Each point corresponds to a gene expressed in both samples, with the color gradient representing the squared ratio of nascent-seq FPKM to poly(A)+ FPKM; blue indicating a larger disparity in expression value. Many genes exhibit significant changes in transcript level, with a small number of genes (dark blue) corresponding to several RNA species highly overrepresented in the nascent-seq transcriptomes.

Fig. 3-5

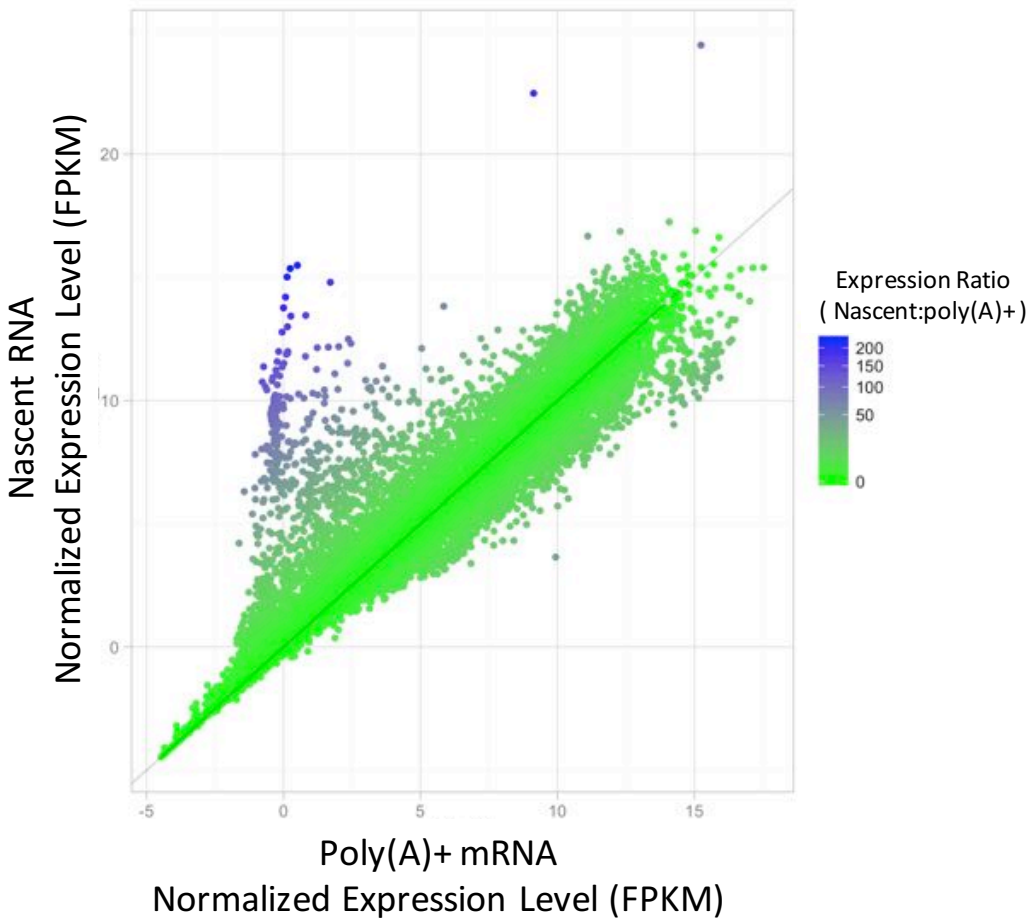

Nascent RNA
Normalized Expression Level (FPKM)
20
10
0
-5
0
5
10
15
Poly(A)+ mRNA
Normalized Expression Level (FPKM)
Expression Ratio
(Nascent:poly(A)+)
200
150
100
50
0

Figure 3.6. Nascent transcript abundance differs broadly from mature polyadenylated transcripts. (A) Comparison of normalized transcript abundance between nascent and poly(A)+ RNA reveals 40-50% of transcripts are either over- or under-represented in nascent samples compared to mature mRNA across all timepoints. **(B)** Over- and under-represented transcripts from the first timepoint were analyzed for enrichment of spatial expression categories (ImaGo Database, Berkeley Drosophila Genome Project). Over half of under-represented transcripts are classified as being primarily maternally deposited. These transcripts are already present in the embryo and are often not significantly zygotically transcribed, and so should be under-represented in nascent RNA. Transcripts over-represented in nascent RNA are enriched for categories of spatially-restricted expression within the embryo, many of which should be actively transcribed during early embryogenesis.

Fig. 3-6

A

	Relave nascent transcript abundance	
	Lower	**Higher**
1.5 - 4 hr	2,831	3,097
4 - 6.5 hr	3,200	3,735
6.5 - 9 hr	2,910	3,187

B

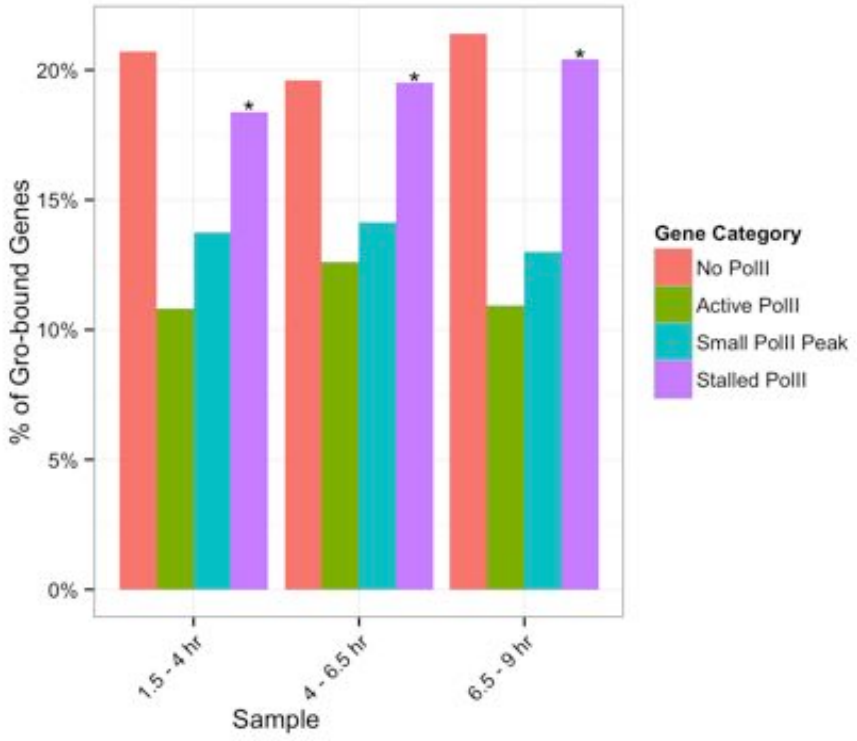
Lower Relative Abundance		
BDGP Term Enrichment	**p-value**	**# genes**
maternal	3.09E-86	1541
ubiquitous	7.16E-45	847
anterior midgut primordium	6.21E-29	505
posterior midgut primordium	2.40E-26	514
trunk mesoderm primordium	1.80E-20	411
anterior endoderm primordium	3.51E-15	292
embryonic midgut	5.09E-15	587
posterior endoderm primordium	5.20E-14	299
dorsal prothoracic pharyngeal muscle	1.13E-12	220
head mesoderm primordium P2	1.76E-09	253
embryonic/larval muscle system	3.69E-09	277
head mesoderm primordium	8.15E-09	153
posterior endoderm primordium P2	2.19E-08	218
anterior endoderm anlage	6.20E-08	205
faint ubiquitous	1.26E-07	437

Higher Relative Abundance		
BDGP Term Enrichment	**p-value**	**# genes**
ventral nerve cord	1.19E-24	452
ventral epidermis primordium	5.92E-24	159
dorsal ectoderm primordium	9.17E-23	152
dorsal ectoderm anlage in statu nascendi	3.07E-22	137
ventral ectoderm anlage in statu nascendi	1.25E-20	125
embryonic brain	1.14E-19	431
dorsal epidermis primordium	2.17E-19	167
procephalic ectoderm anlage in statu nascendi	5.27E-19	123
ventral ectoderm primordium P2	2.40E-18	169
ventral ectoderm primordium	7.18E-17	139
gap	6.36E-13	56
embryonic ventral epidermis	1.43E-12	245
procephalic ectoderm anlage	1.44E-12	140
embryonic dorsal epidermis	2.56E-12	266
tracheal primordium	7.27E-12	90

Figure 3-7. Groucho regulated genes are enriched for stalled PolII. Published data classifying all Drosophila genes into four categories of PolII enrichment or depletion in 2 - 4 hour embryos was used to classify all Groucho-regulated genes at each timepoint (Zeitlinger et al., 2007). **(A)** Between 18 and 22% of all genes with internal or adjacent Gro binding were found to contain stalled PolII. Stalled PolII genes were the only category to exhibit significant enrichment among Groucho-associated genes (*p-value* < 10^{-20}, Fisher's Exact Test). **(B)** Genes differentially expressed in Gro loss-of-function embryos are enriched for classes of PolII stalling dependent on their response to loss of Gro. Genes up-regulated in loss-of-function embryos (potential Gro-repressed genes) are enriched for stalled PolII, while down-regulated genes are enriched for active PolII. This latter result likely arises from the fact that these genes are at least moderately expressed in wild-type embryos and so enriched for active PolII. These genes become repressed by ectopic expression of a secondary repressor that becomes derepressed in Gro loss-of-function embryos.

Fig. 3-7

A


20%
15%
10%
5%
0%
% of Gro-bound Genes
1.5 - 4 hr
4 - 6.5 hr
6.5 - 9 hr
Sample
Gene Category
No PolII
Active PolII
Small PolII Peak
Stalled PolII

B

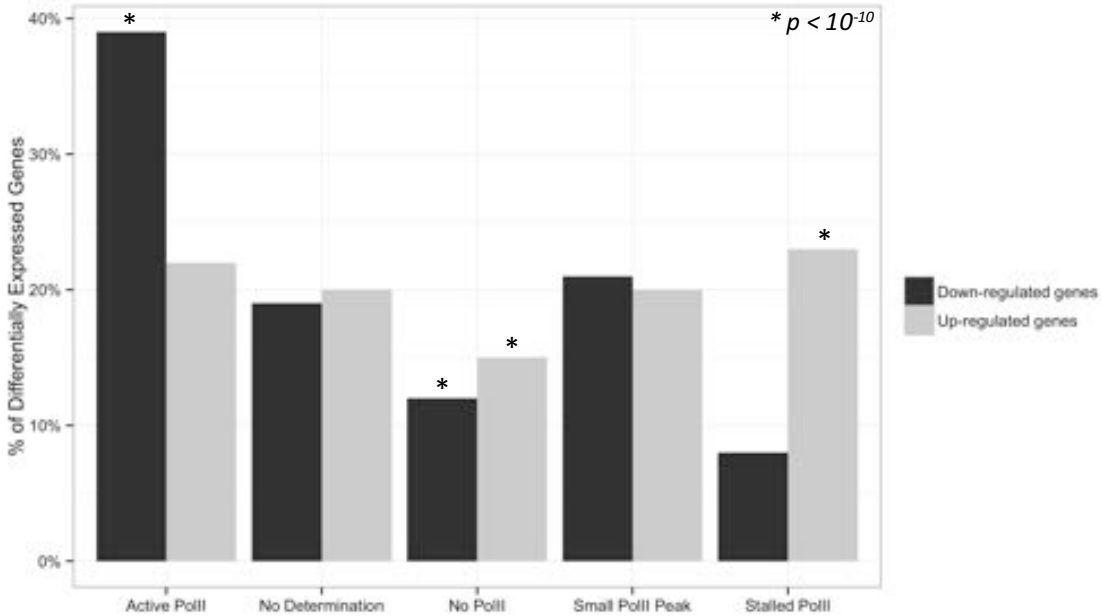

* p < 10-10
40%
30%
20%
10%
0%
% of Differentially Expressed Genes
Active PolII
No Determination
No PolII
Small PolII Peak
Stalled PolII
Down-regulated genes
Up-regulated genes

Figure 3-8. Genes responsive to changes in Groucho level exhibit increased accumulation of promoter-proximal transcript. Transcript density across all expressed genes was calculated independently for genes exhibiting diferent responses to Groucho overexpression in three time stages. At each time window, genes that decrease in expression under the influence of increased Groucho dosage are enriched for 5' proximal transcript density (green), indicating these genes are potentially enriched for stalled polymerase when compared to both unresponsive (red) and up-regulated (blue) genes. Transparent ribbons represent a 95% confidence interval for each position.

Fig. 3-8

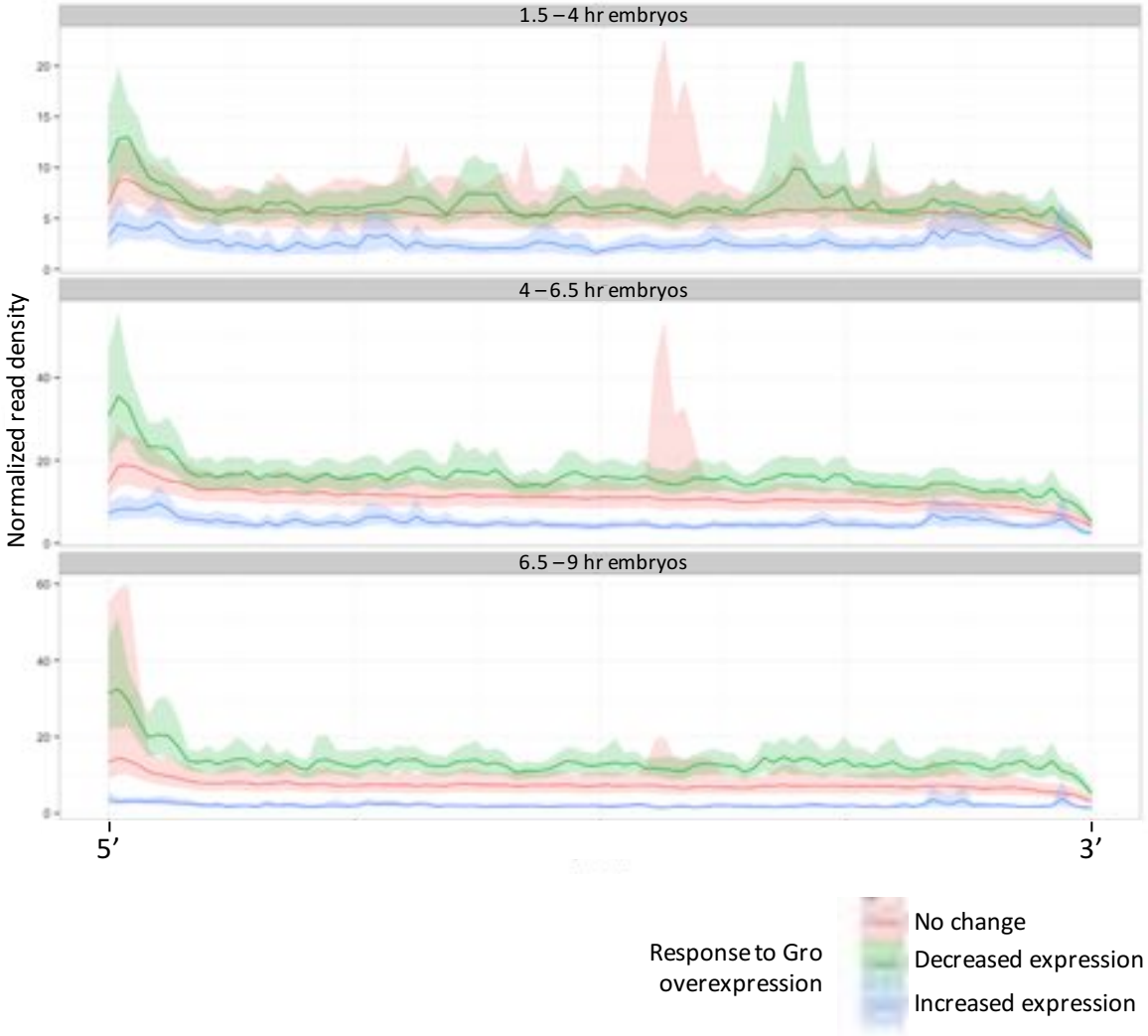

1.5 – 4 hr embryos
4 – 6.5 hr embryos
6.5 – 9 hr embryos
Normalized read density
5’
3’
Response to Gro overexpression
No change
Decreased expression
Increased expression

Table 3-1. Primers for rRNA depletion of embryonic total-RNA.

Table 3-1

Target	Sequence
2s	CTTACAACCCTCAACCATATGTAGTCCAAGCAGC
18s	CAATAATGATCCTTCCGCAGGTTC
5.8s	CAGCATGGACTGCGATATGCGTTC
28s alpha	ATTTTCGCTTTCGCCTTGAAC
28s Beta	TCGAATCATCAAGCAAAGGATAAGC
28s	GTGTTAATTAGCTATAAATAGCTAAAAAACTAATC
28s	CAGGTTACGGAATTGGAACCGTATTCCCTTTCGTT
28s	CAATCTTCAGAGCCAATCCTTATCCCGAAGTTACG
28s	GCCCGTTCCCTTGGCTGTGGTTTCGCTAG
18s	GAACAGAGGTCTTATTTCATTATCCCATGCACAGA
18s	CGGTACAAGACCATACGATCTGCATGTTATCTAGA
18s	TTTAATTGCATGTATTAGCTCTAGAATTACCACAG
5s	AAGTTGTGGACGAGGCCAACAACACGCGGTGTTCCC
5'_end_of_rRNA	TATTCCTATTATCCGCGGAG
5'_end_of_rRNA	CCATTCGAATACGGCCATTT
nodavirus RNA1	ACCTCCGCCCTTTCGGGCTAGAAC
nodavirus RNA2	ACCTTAGTCGGCTGACTTAAACTGTC
totivirus SW-2009a	CGACTATATCTTCTGCGTTATCCAGC
oligo dT	TTTTTTTTTTTTTT

References

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet *13*, 720-731.

Barboric, M., Nissen, R.M., Kanazawa, S., Jabrane-Ferrat, N., and Peterlin, B.M. (2001). NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. Mol Cell *8*, 327-337.

Beauparlant, C.J.L., F.C.; Samb, R.; Deschenes, A.L. Droid, A. (2014). metagene: A package to produce metagene plots. R package version 220.

Benoit, B., He, C.H., Zhang, F., Votruba, S.M., Tadros, W., Westwood, J.T., Smibert, C.A., Lipshitz, H.D., and Theurkauf, W.E. (2009). An essential role for the RNA-binding protein Smaug during the Drosophila maternal-to-zygotic transition. Development *136*, 923-932.

Blau, J., Xiao, H., McCracken, S., O'Hare, P., Greenblatt, J., and Bentley, D. (1996). Three functional classes of transcriptional activation domain. Mol Cell Biol *16*, 2044-2055.

Boettiger, A.N., and Levine, M. (2009). Synchronous and stochastic patterns of gene activation in the Drosophila embryo. Science *325*, 471-473.

Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J., Park, S., Suzuki, A.M.*, et al.* (2014). Diversity and dynamics of the Drosophila transcriptome. Nature *512*, 393-399.

Carrillo Oesterreich, F., Preibisch, S., and Neugebauer, K.M. (2010). Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol Cell *40*, 571-581.

Cernilogar, F.M., Onorati, M.C., Kothe, G.O., Burroughs, A.M., Parsi, K.M., Breiling, A., Lo Sardo, F., Saxena, A., Miyoshi, K., Siomi, H.*, et al.* (2011). Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. Nature *480*, 391-395.

Conaway, J.W., Shilatifard, A., Dvir, A., and Conaway, R.C. (2000). Control of elongation by RNA polymerase II. Trends in biochemical sciences *25*, 375-380.

De Renzis, S., Elemento, O., Tavazoie, S., and Wieschaus, E.F. (2007). Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. PLoS Biol *5*, e117.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R.*, et al.* (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics *10*, 161.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W.*, et al.* (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473-479.

Gromak, N., West, S., and Proudfoot, N.J. (2006). Pause sites promote transcriptional termination of mammalian RNA polymerase II. Mol Cell Biol *26*, 3986-3996.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Harrison, M.M., Li, X.Y., Kaplan, T., Botchan, M.R., and Eisen, M.B. (2011). Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet *7*, e1002266.

IAnders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.H.A., Marr, M.T., and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes & Development *25*, 2502-2512.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

Li, X.Y., Harrison, M.M., Villalta, J.E., Kaplan, T., and Eisen, M.B. (2014). Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition. Elife *3*.

Liang, H.L., Nien, C.Y., Liu, H.Y., Metzstein, M.M., Kirov, N., and Rushlow, C. (2008). The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature *456*, 400-403.

Lis, J., and Wu, C. (1993). Protein traffic on the heat shock promoter: parking, stalling, and trucking along. Cell *74*, 1-4.

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

Nechaev, S., and Adelman, K. (2008). Promoter-proximal Pol II: when stalling speeds things up. Cell Cycle *7*, 1539-1544.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global Analysis of Short RNAs Reveals Widespread Promoter-Proximal Stalling and Arrest of Pol II in Drosophila. Science *327*, 335-338.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow Enhancers Foster Robustness of Drosophila Gastrulation. Current Biology *20*, 1562-1567.

Pritchard, D.K., and Schubiger, G. (1996). Activation of transcription in Drosophila embryos is a gradual process mediated by the nucleocytoplasmic ratio. Genes Dev *10*, 1131-1142.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol *12*, R22.

Rodriguez, J., Tang, C.-H.A., Khodor, Y.L., Vodala, S., Menet, J.S., and Rosbash, M. (2013). Nascent-Seq analysis of Drosophila cycling gene expression. Proceedings of the National Academy of Sciences *110*, E275-284.

Rougvie, A.E., and Lis, J.T. (1990). Postinitiation transcriptional control in Drosophila melanogaster. Mol Cell Biol *10*, 6041-6045.

Schneider, D.S., Hudson, K.L., Lin, T.Y., and Anderson, K.V. (1991). Dominant and recessive mutations define functional domains of Toll, a transmembrane protein required for dorsal-ventral polarity in the Drosophila embryo. Genes Dev *5*, 797-807.

Tadros, W., Goldman, A.L., Babak, T., Menzies, F., Vardy, L., Orr-Weaver, T., Hughes, T.R., Westwood, J.T., Smibert, C.A., and Lipshitz, H.D. (2007). SMAUG is a major regulator of maternal mRNA destabilization in Drosophila and its translation is activated by the PAN GU kinase. Dev Cell *12*, 143-155.

Tadros, W., and Lipshitz, H.D. (2009). The maternal-to-zygotic transition: a play in two acts. Development *136*, 3033-3042.

Thomsen, S., Anders, S., Janga, S.C., Huber, W., and Alonso, C.R. (2010). Genome-wide analysis of mRNA decay patterns during early Drosophila development. Genome biology *11*, R93.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., and Celniker, S.E. (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol *3*, 0081-0088.

Wuarin, J., and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Molecular and Cellular Biology *14*, 7219-7225.

Xu, Z., Chen, H., Ling, J., Yu, D., Struffi, P., and Small, S. (2014). Impacts of the ubiquitous factor Zelda on Bicoid-dependent DNA binding and transcription in Drosophila. Genes & Development *28*, 608-621.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nature genetics *39*, 1512-1516.

Chapter 4

The Central Region of the Drosophila Co-Repressor Groucho as a Regulatory Hub

In this chapter I present data from a publication that identified multiple Groucho-interacting proteins. The influence of these proteins on Gro-mediated repression was investigated through a reporter assay and transcriptome analyses of RNAi-treated cells. I carried out the bioinformatic analyses of the RNA-seq data presented in figures 4 and 6.

JBC Papers in Press. Published on October 19, 2015 as Manuscript M115.681171
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.M115.681171

The Central Region of the Drosophila Co-Repressor Groucho as a Regulatory Hub

Pak N. Kwong[1], Michael Chambers[1], Ajay A. Vashisht[2], Wiam Turki-Judeh[1,3], Tak Yu Yau,[1] James A. Wohlschlegel[2,3], and Albert J. Courey[1,3]

[1]Department of Chemistry and Biochemistry
[2]Department of Biological Chemistry
[3]Molecular Biology Institute
University of California, Los Angeles 90095

Running title: *Groucho/Spliceosome Interactions*

To whom correspondence should be addressed: Albert J. Courey, Department of Chemistry and Biochemistry, 607 Charles E. Young Drive East, Los Angeles, CA 90095-1569; Telephone: (310) 825-2530; Email: courey@chem.ucla.edu

Keywords: Groucho, transcription regulation, transcription co-repressor, protein complex, spliceosome, proteomics, intrinsically disordered domain

Background: The co-repressor Groucho has an essential, but disordered, central region.
Results: We identified over 160 central region-binding proteins, many of which, including components of the spliceosome, modulate Groucho-mediated repression.
Conclusion: Groucho regulates transcription by multiple mechanisms and may link the transcriptional and splicing machineries.
Significance: Its central region may serve as the hub of a regulatory network.

ABSTRACT

Groucho (Gro) is a Drosophila co-repressor that regulates the expression of a large number of genes, many of which are involved in developmental control. Previous studies have shown that its central region is essential for function, even though its three domains are poorly conserved and intrinsically disordered. Using these disordered domains as affinity reagents, we have now identified multiple embryonic Gro-interacting proteins. The interactors include protein complexes involved in chromosome organization, mRNA processing, and signaling. Further investigation of the interacting proteins using a reporter assay showed that many of them modulate Gro-mediated repression either positively or negatively. The positive regulators include components of the spliceosomal subcomplex U1 small nuclear ribonucleoprotein (U1 snRNP). A co-immunoprecipitation experiment confirms this finding and suggests that a sizable fraction of nuclear U1 snRNP is associated with Gro. The use of RNA-seq to analyze the gene expression profile of cells subjected to knockdown of Gro or snRNP-U1-C (a component of U1 snRNP) showed a significant overlap between genes regulated by these two factors. Furthermore, comparison of our RNA-seq data to Gro and Pol II ChIP data led to a number of insights including the finding that Gro-repressed genes are enriched for promoter proximal Pol II. We conclude that the Gro central domains mediate multiple interactions required for repression thus functioning as a regulatory hub. Furthermore, interactions with the spliceosome may contribute to repression by Gro.

Groucho (Gro) is a conserved metazoan co-repressor that may be particularly critical for

long-range repression, whereby repressors are able to establish large transcriptionally silent domains that can spread over many thousands of basepairs (1-3). Gro is essential in many developmental processes, including sex determination, neurogenesis, and pattern formation in Drosophila, as well as myogenesis and hematopoiesis in vertebrates (2,4,5). Gro also has roles in multiple signal transduction pathways, including the Ras and Notch pathways (6-8). Furthermore, increased Gro activity correlates with the appearance of certain forms of cancer, such as lung cancer (9,10). Thus, understanding the mechanism of Gro-mediated repression should contribute to our understanding of long-range repression and its role in development, signaling, and disease.

Sequence comparison of Gro family proteins reveals five domains (2,10). The C-terminal WD-repeat domain forms a β-propeller that interacts with the WRPW and eh1 motifs found in many Gro-dependent DNA-binding repressors (11). The N-terminal Q domain folds into a coiled-coil structure that forms tetramers and perhaps higher order oligomers, and this self-association is required for robust repression (12-15). The central GP, CcN, and SP domains are believed to have essential functions even though their primary sequences are not well conserved. The GP domain interacts with the histone deacetylase Rpd3/HDAC1 (16,17). Histone deacetylation is broadly associated with gene silencing, and treatment of flies with histone deacetylase inhibitors attenuates Gro-mediated repression (18). In addition, the GP domain is essential for nuclear localization, since deletion of this domain prevents Gro nuclear uptake (19). The SP domain regulates Gro function negatively, as its deletion leads to promiscuous repression and developmental defects (19). Phosphorylation of the SP domain by Ras/MAPK signaling was shown to attenuate repression, providing a mechanism for regulating repression in response to environmental cues (20). Finally, the CcN domain is also targeted for phosphorylation by protein kinases and is required for repression by Gro (19,21).

Sequence analysis of the Gro central domains strongly suggests that they are intrinsically disordered (19). Intrinsically disordered regions in proteins lack rigid three-dimensional structures under native conditions and can serve as hubs of large regulatory networks by mediating a wide array of highly specific protein interactions (22,23). Increasing evidence suggests that intrinsically disordered domains have critical functions in transcriptional regulation (24,25).

In this study, we set out to illuminate the mechanisms of Gro mediated repression by identifying proteins that interact with the N-terminal Q domain and the three central domains. A proteomic screen revealed over 160 interacting proteins, many of which are components of protein complexes in a variety of functional categories such as chromatin remodeling and RNA processing. Perhaps most notably, the interactors included multiple components of the spliceosome, and a co-immunoprecipitation experiment suggests that a sizable fraction of U1 snRNP (a subcomplex of the spliceosome) is associated with Gro in embryonic nuclei.

As a means of systematically validating the functional significance of these interactions, we carried out a novel reporter assay employing three different luciferase reporters that could be monitored simultaneously. These assays showed that many of the interacting proteins, including the protein components of U1 snRNP, are required for optimal Gro mediated repression. Lastly, we compared the effects on gene expression profile of Gro and U1 snRNP knockdown, finding a significant overlap in the regulated genes. Our results indicate that the central domains of Gro mediate multiple interactions required for repression, and reveal a possible mechanism of Gro mediated repression through an interaction with the spliceosome complex or subcomplexes. This reinforces previous studies suggesting that the spliceosome has roles in transcriptional

regulation in addition to its roles in RNA processing (26-30).

EXPERIMENTAL PROCEDURES

Plasmids–To generate plasmids for expression of GST fusion proteins, sequences encoding the Gro domains were amplified by PCR and inserted between the BamHI and XhoI sites of pGEX4T (GE Healthcare Life Sciences). The Q domain included Gro amino acids 1-133; the GP domain included amino acids 134-194, the CcN domain included amino acids 195-257, and the SP domain included amino acids 258-390. Sequences of PCR primers are provided in Table 1.

Plasmids used in the reporter assay were generated as follows. The red luciferase plasmid, G5DE5-pCBR, was generated by inserting the G5 DE5 enhancer region (14) into pCBR-basic vector (Promega Cat.# E1411) between the KpnI and XhoI sites. The green luciferase plasmid, DE5G5-pCBG68, was generated by inserting the luciferase gene using NcoI and SalI from pCBG68-basic vector (Promega Cat.# E1431) into the DE5 G5 vector, which has UAS elements downstream of the reporter (unpublished data). Actin promoter driven Dorsal (pPac Dl), Twist (pPac Twi) and Gal4-Gro (pAct Gal4-Gro) plasmids have been previously described (14). The RpIII128 promoter driven Renilla luciferase plasmid, RpIII128-Rluc, was obtained from Addgene (ID #37380) (31).

Affinity purification and identification of Gro interacting proteins–Plasmids encoding the recombinant domains fused to GST or GST alone were transformed into BL21 cells. 250 ml of mid-log cells were induced with 0.25 mM IPTG for an hour. Cells were pelleted at 4000 x g, resuspended in 25 mL of Salty TE (0.15 M NaCl, 10 mM Tris pH8, 1 mM EDTA) with protease inhibitor (Life Technologies, Cat.# 88266), and incubated on ice for 30 min. Samples were incubated at 4˚ for 15 min after DTT and Triton X-100 were added to final concentrations of 5 mM and 1%, respectively. Cells were then disrupted through a microfluidizer (Microfluidics M110L) using standard conditions. The lysate was collected and centrifuged at 14000 x g for 10 min at 4˚. Supernatant was collected, and 1 ml of glutathione agarose resin (50% slurry) was added. After overnight incubation, the resin was washed with cold PBS three times and stored at 4˚.

Drosophila embryo nuclear extracts were prepared as previously described (32). To isolate Gro-interacting proteins, 20 µg of glutathione bead-immobilized recombinant domains were mixed with nuclear extract containing 30 mg of protein (20 mg/ml) in 8 ml of HEMNK buffer (40 mM HEPES pH 7.5, 5 mM MgCl2, 0.2 mM EDTA, 1 mM DTT, 0.5% NP-40, 0.1 M KCl) at 4˚ overnight. Samples were washed six times for 15 minutes with 5 ml of HEMNK buffer. Proteins were first eluted with 5 ml of 2 M NaCl in HEMNK buffer and then with 2.5 ml of 2 M NaCl in HEMNK buffer for 20 min each. Eluted proteins were subjected to TCA precipitation prior to MudPIT analysis. MudPIT analysis was performed as previously described (33). Peptide identifications were filtered using a false discovery rate (FDR) cutoff of 0.05 as determined by the decoy database approach. Protein-level false positive rates were less than 0.03 for all individual runs.

Table S1C includes all the mass spectroscopy data for the two independent replicate screens carried out with each GST fusion protein and GST alone, while Tables S1A and B include selective data for 159 proteins that were detected in both replicates, as well as three proteins (Histone H3, Caf1, and bic) that were only detected in one replicate, but for which other data confirm the significance of the interaction (see notes 2 and 3 to Table S1B). Ribosomal proteins were excluded from the lists in Tables S1A and B.

Gro immunoprecipitation and reverse transcriptase qPCR (RT-qPCR) analysis of U1 snRNA–500 µg of nuclear extract was incubated with 1.875 µg of affinity purified rabbit antibody against the Gro GP domain or rabbit IgG in a final

volume of 250 µl HEMNK buffer overnight at 4˚. 225 µg of Protein A Dynabeads (Invitrogen Cat.# 10001D) were incubated with the samples at 4˚for 1 hour. Samples were then washed with HEMNK buffer three times for 10 minutes each. For RT-qPCR, RNA was eluted in 10 µl water by heating to 80˚ for 2 minutes. Samples were treated with DNase I according to the manufacturer's protocol (Promega Cat.# M6101). Reverse transcription was performed with 300 ng of random primer (Invitrogen Cat.# 48190-011), and qPCR was performed using primers amplifying U1 snRNA (Table 2). Threshold cycle values were converted to percent input values by comparison to a standard curve generated from multiple serial dilutions of RNA isolated by Trizol extraction (Life Technologies, Cat.# 10296010) from the input nuclear extract. Primer specificity was validated by melting curve analysis of the amplification products (data now shown).

For immunoblotting, samples were eluted in SDS-PAGE loading buffer. Proteins were detected with a mixture of mouse anti-Gro (Developmental Studies Hybridoma Bank, 1:650 dilution) and affinity purified rabbit anti-GP domain (1:100 dilution) antibodies. Immunoblots were subsequently probed with goat anti-mouse 680 and goat anti-rabbit 800 IR-dye coupled secondary antibodies (Li-Cor) and imaged with a Li-Cor Odyssey imager.

Three-reporter luciferase assay–To guard against off-target effects, each candidate gene was knocked down with three non-overlapping dsRNAs when possible (the complete list of dsRNAs used is available upon request). Each dsRNA was tested in triplicate. dsRNA was synthesized by the Drosophila RNAi Screening Center and re-aliquoted into white flat bottom 96 well plates (USA scientific Cat.# CC7682-7968) with 150 ng/well in 10 µl of water using a Beckman Coulter BioMek FX Workstation.

Transfections were carried out with Effectene reagent (Qiagen Cat.# 301425). 6 µg each of G5DE5-pCBR and DE5G5-pCBG68, 0.6 µg of RpIII128-Rluc, 1 µg of pPac Dl, 0.3 µg of pPac Twi, and 1.2 µg of the pAct Gal4-Gro were suspended in 600 µl buffer EC. 33 µl of this mixture was added to 25 µl of enhancer. After 2-3 minutes, 7.5 µl Effectene was added and mixed by pipetting up and down. 6 µl of this mixture was immediately added into each well of a 96-well plate containing 150 ng of dsRNA. 4-8 minutes later, 100 µl of S2 cells (diluted to 1 x 10^6 cell/ml) was added to each well. Cells were incubated at 24˚ for 2 days before assaying.

The luminescence signal was measured with a Molecular Devices LJL Analyst HT microplate reader using emission filters ET510/80m and E610LP (Chroma Cat.#S-022658 and #138951). 50 µl of D-luciferin (Chroma-Glo system, Promega Cat.# E2980) was added to each well. Five minutes later the reaction was stopped by the addition of 50 µl of stop buffer containing coelenterazine (Dual-Luciferase system, Promega Cat.# E1960). The luminescence signal was measured immediately without applying a filter.

To address the issue of signal overlap, raw signals were subjected to filter correction. The corrected red luminescence signal R' and green luminescence signal G' were calculated according to the following equations:

$$R' = \frac{Lrf - Lgf \times \left(\frac{Grf}{Ggf}\right)}{\left(\frac{Rrf}{R}\right) - \left(\frac{Rgf}{R}\right) \times \left(\frac{Grf}{Ggf}\right)}$$

$$G' = \frac{Lgf - R' \times \left(\frac{Rgf}{R}\right)}{\left(\frac{Ggf}{G}\right)}$$

Parameters were determined by expressing the individual luciferases and recording the luminescence signals with red and green filters, and without filter (data not shown). The ratio of green signal passed through the red filter, Grf/Ggf, was determined to be 0.0975; the ratio of red signal passed through the red filter, Rrf/R, was determined to be 0.42; the ratio of red signal passed through the green filter, Rgf/R, was determined to be 0; the ratio of green signal passed through the green filter, Ggf/G, was determined to be 0.47. Lrf and Lgf are luminescence signals in

which cells are co-transfected with both red and green luciferases. Lrf is the signal recorded with red filter, and Lgf is the signal recorded with green filter.

The signal from untransfected cells was then subtracted from the corrected data to eliminate background. Processed data were then normalized to the internal control Renilla luciferase. Finally, data were compared to the signal from cells in the same plate that were treated with control GFP dsRNA. A change in long or short-range repression was considered significant if $p < 0.1$. If multiple dsRNAs were tested for a given gene (as was true in most cases, Table S2), then a change is only listed if $p<0.1$ for at least two separate dsRNAs.

RNA-seq library preparation–Gro dsRNA was generated by PCR amplification of the first 800 nucleotides of the coding sequence using primers containing T7 promoters followed by in vitro transcription with T7 RNA polymerase. snRNP-U1-C dsRNA was generated by PCR and in vitro transcription of the snRNP-U1-C coding sequencing with primers 5'-taatacgactcactatagggtactCAAAGTACTATTGCG ACTACTGC and 5'-taatacgactcactatagggtactCTTGGGTCCGTTCATG ATTCC. Transfection was carried out as previously described (34). RT-qPCR was used to determine the knockdown efficiency prior to RNA-seq library preparation. RT-qPCR primers targeted the 3' UTRs of Gro and snRNP-U1-C. Rpl32 was used as a reference gene. The specificity of all primers was validated by melting curve analysis of the amplification products (data not shown). Sequences of the qPCR primers are listed in Table 2.

Total RNA was extracted with Trizol according to the manufacturer's protocol (Life Technologies, Cat.# 10296010). RNA integrity was determined with an Agilent 2100 Bioanalyzer using the RNA 6000 Nano Kit (Agilent Cat.# 5067-1511). Isolation of mRNA was carried out as follows. Streptavidin magnetic beads (Promega Cat.# Z5481) were prepared in aliquots of 120 µl and 60 µl in 0.5X SSC with 10 mM EDTA. 15 µg of total RNA was mixed with 1.5 µM of biotinylated 15-mer poly(T) oligonucleotide in 0.5X SSC with 10 mM EDTA. Samples were first incubated at 75˚ for 5 minutes, followed by 15˚ for 10 minutes and 10˚ for 10 minutes. Samples were then incubated with 120 µl of magnetic beads at 4˚ for 2 hours, followed by 60 µl of magnetic beads at 4˚ for 30 min. The two aliquots of beads were combined and washed four times with 300 µl of ice cold 0.1X SSC containing 10 mM EDTA. mRNA was first eluted with 100 µl of water followed by 150 µl of water at 37˚ for 10 min each. Samples were precipitated with ethanol and stored at -80˚. Pulldown efficiency of mRNA and depletion efficiency of 18S rRNA were determined by RT-qPCR (data not shown).

The RNA-seq library was prepared according to the manufacturer's protocol (Epicentre, Cat.# SSV21124 and Cat.# RSBC10948). The concentration of the library was determined with Pico Green (Life Technologies, Cat.# Q32851) according to the manufacturer's directions. Fluorescence signal was measured using a TECAN M1000 fluorescent plate reader.

Bioinformatics–Alignment of paired-end reads to the D. melanogaster genome (assembly BDGP 5/dm3) was performed with Tophat2 (v2.0.9) (35) using default parameters. DESeq2 (v1.6.3) (36) was used for gene expression-level normalization and differential expression significance testing. Histone modification and motif enrichment analysis was carried with i-cisTarget (37) using default parameters. Enriched gene ontology analysis was done with Flymine (v31.0) (38) using default parameters.

RESULTS

Identification of Gro interacting proteins–A previous study showed that deletion of the GP or CcN domains in the Gro central region led to a loss of Gro-mediated repression and to lethality, while deletion of the SP domain led to reduced

specificity of Gro-mediated repression and to reduced viability (19). To identify possible regulatory partners of these domains, we used them as affinity reagents to purify interacting proteins, which were then identified by mass spectrometry. The three central domains of Gro were expressed as glutathione-S-transferase (GST)-tagged proteins and purified from E. coli lysates (Figure 1A, B). We also constructed a similarly tagged form of the N-terminal Q domain since previous studies suggested that, in addition to mediating Gro oligomerization, the Q domain engages in interactions with regulatory targets (39,40).

The glutathione bead-immoblized GST-fused domains (or, as a negative control, immobilized unfused GST) were incubated with a Drosophila embryo nuclear extract. After extensive washing, interacting proteins were eluted with 2 M salt and analyzed by multi-dimensional protein identification technology (MudPIT) (33) (Table S1C). Duplicate extract preparations and affinity purifications were carried out and analyzed on separate dates and there was a high degree of overlap between the sets of proteins identified in these duplicate experiments (Figure 1C). With three exceptions (see Experimental Procedures), only proteins that appeared in both replicates were included in our list of Gro interacting proteins (Figure 1C, Table S1A, B.) Gene ontology analysis of this list of 162 proteins revealed a variety of functions including regulation of gene expression, RNA processing, and developmental processes (Table 3).

89 the 162 Gro-interacting proteins associated uniquely with one domain (in all but one case, the SP domain), while 32 interacted with two domains. In the case of 23 of the 32 proteins that interacted with two domains, one of these domains was the Q domain (Table S1A). This is consistent with the known role of the Q domain in homo-oligomerization (12-15). In accord with this role, chromatography using GST-Q as the affinity reagent resulted in the purification of some full-length endogenous Gro (Table S1C and data not shown). This could lead to the co-purification of Gro-interacting proteins that bind to regions outside the Q domain. Thus, 112 (89 plus 23) of the 162 detected interacting proteins can, in principal, be accounted for by the binding of Gro to a single central domain. However, at least 50 proteins (162 minus 112) are able to bind independently to two or three central domains. The ability to interact with multiple Gro domains could allow tighter binding or more versatile control of binding.

The list of interacting proteins (Table 4, Table S1A, B) contains multiple components of known multisubunit protein complexes. For example, we identified the α and β subunits of casein kinase II (CKII), a previously identified regulator of Gro activity (21). We also detected protein complexes involved in chromosome organization, including both components of the ATP-dependent chromatin remodeling and assembly factor (ACF), Acf1 and Iswi (41). Our proteomic screens also identified all the core protein components of the nucleosome (the core histones) as well as histone variant H2Av, consistent with previous studies demonstrating functional interactions between Gro and nucleosomes (42-44).

Perhaps most surprisingly, we discovered a number of components of the spliceosome among the group of Gro-interacting proteins, including all three proteins unique to U1 snRNP, components of U4/U6 snRNP, U2 snRNP, and the Sm complex (45,46). To validate the interaction between Gro and U1 snRNP, Drosophila embryo nuclear extracts were subjected to immunoprecipitation using an affinity purified antibody against the Gro GP domain or, as a negative control, rabbit IgG. An anti-Gro immunoblot of the immunopreciptated material demonstrates the efficiency of the immunoprecipitation (Figure 2A). RNA was extracted from the immunoprecipitates and analyzed by RT-qPCR with primers specific for

U1 snRNA (a component of U1 snRNP). The results show that ~13% of the U1 snRNA in the nuclei of 0-12 hour embryos is associated with Gro (Figure 2B).

Functional analysis of Gro interacting proteins–We next carried out functional assays to determine if the interacting proteins are required for regulation of a Gro-responsive reporter gene. Previous studies established a reliable reporter assay for Gro function employing a luciferase reporter containing Gal4 binding sites (UAS elements), as well as an artificial enhancer containing binding sites for the Dorsal and Twist activators (14,16,18,47). Dorsal/Twist activated transcription of this reporter is strongly repressed upon introduction of a Gal4-Gro fusion protein. By altering the position of UAS elements relative to the artificial enhancer, we were able to examine both short-range and long-range Gro-mediated repression simultaneously (Figure 3A, B). The reporter system relied on two variants of click beetle luciferase that use D-luciferin as a substrate and emit either red or green light (48). In addition, a plasmid encoding Renilla luciferase, which uses coelenterazine as a substrate, was used as an internal control for transfection efficiency, cell viability, and general effects on transcription and translation. We validated the three-reporter system using dsRNA against Dorsal, Gro, and Rpd3 (which is partially required for Gro-mediated repression (18)) (Figure 3C). As predicted, Dorsal knockdown resulted in a complete loss of activation, Gro knockdown resulted in a complete loss of repression, and Rpd3 knockdown resulted in a partial loss of repression.

Each of the candidates from the screen for Gro-interacting proteins was knocked down by RNAi using up to three dsRNAs per gene to guard against off-target effects. We excluded the histones from this analysis under the assumption that knockdown of these essential chromatin components would have pleiotropic deleterious effects on cell metabolism, and because each histone is encoded by multiple genes making efficient knockdown problematic. We therefore tested 157 genes in this S2 cell luciferase assay, in most cases with multiple dsRNAs per gene (three if available), and each dsRNA was tested in triplicate. In total, we carried out approximately 1300 assays (including controls) in a 96 well plate format using a partially automated approach (see Experimental Procedures).

A candidate was scored as a regulator of Gro-mediated repression if knockdown reproducibly resulted in either an increase or a decrease in the level of repression (see Experimental Procedures for explanation of the statistical test of significance). Forty-four candidates met these criteria, of which 28 interfered with optimal repression (i.e., repression increased upon knockdown; these were termed "negative regulators of Gro") and 16 were required for optimal repression (i.e., repression decreased upon knockdown; these were termed "positive regulators of Gro"). We provide representative data for one negative regulator (vir), one positive regulator (snRNP-U1-C, and one protein that is neither a positive nor a negative regulator (Figure 3D); a list of all the positive and negative regulators (Table 5); and a separate list showing the quantitative effect of RNAi knockdown of each of the 44 regulators on repression by Gal4-Gro (Table S2). Of particular interest, four spliceosomal proteins, including two components of U1 snRNP, act as positive regulators of Gro, confirming the functional significance of the interaction between Gro and U1 snRNP. A few other noteworthy examples among the Gro regulators (Tables 5 and S2) include both components of the CKII complex (CKIIα, CKIIβ), which act as negative regulators, and the chromatin remodeling factor Acf1, which acts as a positive regulator (see discussion)

Expression profiling of Gro and snRNP-U1-C knockdown cells–snRNP-U1-C is one of the components of the U1 snRNP complex, which is responsible for 5' splice site recognition (46). In addition to its role in RNA processing, it has been

shown to repress transcription of EWS/FLI-transactivated genes (30). Since our data indicated that snRNP-U1-C may also modulate Gro function, we examined the genome-wide role of snRNP-U1-C in Gro mediated repression. Using RNA-seq, we compared the effects of snRNP-U1-C knockdown to that of Gro knockdown on the gene expression profile in S2 cells. Cells were treated with Gro or snRNP-U1-C dsRNA for four days, leading to four-fold or greater knockdown of the Gro and snRNP-U1-C mRNA (Figure 4A). The transcriptomes in wild-type and Gro knockdown S2 cells were quantitatively similar to those published previously (49,50) (Figure 4B, C). We note that the genes differentially expressed in the snRNP-U1-C knockdown are enriched for genes containing introns as would be expected given the role of U1 snRNP in splicing. However, this set of genes also contains a number of intron-less genes consistent with the idea that snRNP-U1-C has roles in gene regulation apart from its role in splicing (Figure 4D). We note that changes in the expression of an intron-less gene could also reflect a requirement for the product of an intron-containing gene in the expression of the intron-less gene.

98 genes were differentially expressed in both Gro and snRNP-U1-C knockdown cells (Figure 4E), of which 36 were upregulated in either case. These coordinately upregulated targets included genes in various signaling pathways, such as the Wnt, Notch, and Toll pathways (Table 6). Comparison with publically available ChIP-seq data on histone modification and transcription factor binding revealed that these coordinately regulated genes were most enriched for histone H3K36 methylation and the H3K36 methyltransferase ASH1 (Figure 4F).

To determine if the regulatory effects of knocking down Gro are likely to be direct, we compared our RNA-seq data from Gro knockdown S2 cells to available S2 cell Gro ChIP data (49). Gro appears to bind many genes that it does not repress (Figure 5A). This is consistent with observations made with numerous regulatory factors (51,52) and suggests that binding, while required, is not sufficient for regulation. We observe an enrichment of Suppressor of Hairless (Su(H)) and Brinker (Brk) binding motifs within Gro ChIP-seq peaks in the differentially expressed genes but not in the non-differentially expressed genes (Figure 5B). Comparison of our RNA-seq data from Gro knockdown cells to available Pol II ChIP-chip data (53) also reveals an enrichment in Pol II pausing near the transcriptional start site in genes that are up-regulated upon Gro knockdown (i.e., genes that are repressed by Gro; Figure 6).

DISCUSSION

Previous studies showed that the disordered Gro central domains are essential for properly regulated transcriptional repression (2,19). To shed light on the mechanism by which these domains function, we used them as affinity reagents to purify interacting proteins in Drosophila embryo nuclear extracts, which were then identified by MuDPIT. We identified over 160 interacting polypeptides, many of which associate with one another in a variety of multi-protein complexes. Several of these interacting proteins (e.g., the core histones, CKII) were previously characterized as Gro interactors thus partially validating the screen. In addition, we validated the interaction between Gro and U1 snRNP by demonstrating the presence of U1 snRNA in an anti-Gro immunoprecipitate of embryonic nuclear extracts.

As a means of systematically validating interactions, we employed a functional assay in Drosophila cells, in which 157 of the interactors were each knocked down by RNAi to determine their requirement for Gal4-Gro-mediated repression of a luciferase reporter. In this way, we obtained evidence that 44 of the interactors have functional roles in Gro mediated repression. 28 of these are required for repression while 16 of them antagonize repression. The number 44 is probably an underestimate of the true number of functional

interactors due to the artificiality of the reporter assay. For example, because we artificially recruit Gro to the reporter by tethering it to the Gal4 DNA binding domain, any interactions that work to help recruit Gro to the template will not be required. In addition, the reporters are introduced by transient transfection, and certain chromatin structures or modifications that contribute to Gro-mediated repression may not be reproduced in this context.

Gro-interactors include chromatin remodelers, protein kinases, and protein complexes involved in RNA processing–Gro-mediated repression may be associated with changes in chromatin structure including histone deacetylation and possibly increased nucleosome density (3,18,54). Consistent with this possibility, our proteomic screen identified a number of histone modifiers and ATP-dependent chromatin remodelers, including subunits of the ACF chromatin remodeling complexes (Acf1 and Iswi), the histone chaperone NAP1, and the histone kinases JIL-1 and Ball. Consistent with the idea that chromatin remodelers may be required for Gro-mediated repression by catalyzing changes in nucleosome density or higher order chromatin structure, our reporter assay showed that Acf1 is required for optimal repression by Gro.

CKII is a heterotetrameric complex consisting of two copies of a catalytic subunit (CKIIα) and two copies of a regulatory subunit (CKIIβ) (55,56). A previous study showed that CKII phosphorylates Gro at multiple sites including serines 239 and 253 to promote repression (21). We identified both the α and β subunits of CKII and the CKII negative regulator Nopp140 in our proteomic screen; but our findings are inconsistent with the view that CKII is a positive regulator of Gro and that Nopp140 acts by inhibiting CKII. This is because our reporter assays show that CKIIα, CKIIβ, and Nopp140 are all negative regulators of Gro. However, our results are consistent with other findings showing that Gro phosphorylation can block repression (2). Furthermore, the effect we observe due to Nopp140 knockdown could reflect the role of this factor in processes other than CKII regulation (57).

In addition to several expected protein complexes, we have also isolated many novel Gro interacting proteins, one of which is the RNA helicase Rm62 (also known as p68). Rm62 is a DEAD box RNA helicase that has multiple functions including roles in RNA processing, RNAi, and transcriptional regulation (58). Previous studies have shown a dual role for Rm62 in transcriptional regulation – its interaction with coactivator CBP/p300 may lead to gene activation (59), while its interaction with HDAC1 may lead to repression (60,61). Our reporter assay confirms its function as a positive regulator of Gro-mediated repression, as knocking down Rm62 resulted in attenuated Gro activity. Interestingly, Rm62 was also shown to be an essential splicing component through its action on the U1 snRNP (62,63). The possible significance of the spliceosome in Gro mediated repression is discussed below.

An unanticipated role for the spliceosome in Gro mediated repression–One of the most surprising findings from our proteomic screen was the purification of a significant portion of the spliceosome complex, which suggests a potential role for the spliceosome in transcriptional regulation.

Pre-mRNA processing frequently occurs co-transcriptionally (64-66). Splicing factors are often recruited to nascent transcripts by the C-terminal domain (CTD) of the RNA Pol II large subunit and elongation factors (67,68). In addition, there is evidence that co-activators are able to interact with splicing factors (27). The interaction between the transcriptional and splicing machinery may be functionally relevant since different promoters can yield transcripts that are subject to differential alternative splicing (69,70). While many studies have focused on the effect of transcription factors in splicing, there is also increasing evidence that promoter proximal

splicing elements can influence transcription (26,28,71).

U1 snRNP, a part of the spliceosome, consists of U1 snRNA, three U1 snRNP specific proteins, and the seven subunit Sm complex (46). Our list of 162 Gro-interacting proteins (Tables S1B and C) includes all three U1 snRNP specific proteins (snRNP-U1C, snRNP-U1-70K, and snRNP-U1-A), as well as two subunits of the Sm complex (Sm-D2 and Sm-D3). We note that we also detected at least four other Sm complex subunits in one of the two replicate screens (Sm-B, Sm-F, Sm-D1, and Sm-G) (Table S1C). Additionally, we showed by co-immunoprecipitation that approximately 13% of U1 snRNA, the RNA component of the U1 snRNP, is associated with Gro in embryonic nuclei. Thus, we have detected essentially the entire U1 snRNP in our proteomic screens for Gro-interacting proteins.

Data from our reporter assay suggests that the U1 snRNP complex is required for optimal Gro mediated repression, as snRNP-U1-C and snRNP-U1-70K knockdown attenuated repression. Consistent with our finding, it has been shown that snRNP-U1-C overexpression can decrease EWS/FLI-activated transcription (30). It is worth noting that the U1 snRNA is known to associate with TFIIH and promote transcriptional initiation *in vitro* (29). Thus, the effect of the U1 snRNP complex in transcription regulation may be context dependent.

Gro recruitment is insufficient for repression–The available S2 cell Gro ChIP-seq data (49) reveals 1242 Gro binding sites in the S2 cell genome associated with 748 genes, while our RNA-seq analysis revealed that only 46 of these 748 genes are differentially expressed in Gro knockdown S2 cells implying that Gro binds to many genes that it does not regulate. The apparent contradiction could be explained by the absence of a required transcriptional activator in S2 cells to activate these genes upon Gro depletion. Regardless of the reason for the finding that Gro binds to many more genes than it regulates, this is a phenomenon that is common to many (perhaps most) eukaryotic gene-specific transcriptional regulators (51,52). Gro ChIP-seq peaks associated with genes differentially expressed upon Gro knockdown are enriched for Su(H) and Brk binding motifs. This is in agreement with the known roles of Su(H) and Brk in the recruit of Gro to target genes in the Notch and Dpp signaling pathways, respectively (72-74).

Genes that are up-regulated in Gro knockdown cells (and which are therefore candidate Gro repression targets) exhibit enrichment in Pol II pausing near the transcriptional start site. This finding is in agreement with the hypothesis that Pol II pausing is one mechanism to repress gene expression (75,76). We note that our proteomic screen revealed the Pol II C-terminal domain (CTD) kinase Cdk12 as a Gro-interacting protein (Table S1). By phosphorylating the CTD on Ser 2, Cdk12 may function to allow release of paused Pol II (77). Consistent with this idea, our reporter assay shows that Cdk12 functions to alleviate Gro-mediated repression (Tables 5 and S2).

Genes that are differentially expressed in Gro and snRNP-U1-C knockdown cells are enriched for H3K36me1 as well as the H3K36 methyltransferase ASH1. While H3K36me is involved in multiple functions including transcriptional regulation, splicing, and DNA repair (78,79), these findings suggest a previously unknown role for this histone mark in Gro mediated repression.

The Gro central region as a regulatory hub of repression activity–In conclusion, our findings reinforce the idea of that the Gro central domains, which are intrinsically disordered, are indispensable for repression (19). Previous studies from our lab and other labs show that the GP domain interacts with the histone deacetylase Rpd3/HDAC1, which may promote local histone deacetylation and alter nucleosome density (16,18). The identification of the ACF chromatin

remodeling complexes as a central region interacting protein complex, and our demonstration that knockdown of this protein attenuates Gro-mediated repression, provides further support for the idea that regulation of chromatin structure is a critical aspect of Gro mediated repression. On the other hand, modulation of chromatin structure is likely not the only mechanism of Gro mediated repression as histone deacetylase inhibitors and Rpd3 knockdown reduce, but do not abolish Gro-mediated repression (16,18) (Figure 3C). Through a combination of proteomic screening, reporter assays, and genome-wide expression profiling, our results suggest a possible new mechanism of Gro mediated repression involving the action of the spliceosome. Future experiments will focus on elucidating the underlying mechanisms by which these interacting partners act in Gro-mediated repression.

Acknowledgements: This work was supported by NIH grant GM44522 to AJC and NIH grant GM089778 to JAW. We thank Robert Damoiseaux, Scientific Director of the UCLA Molecular Screening Shared Resource Center, for assistance in developing the mechanized multiple luciferase reporter screen. We are grateful to all the members of the Courey lab (past and present) for their insight and advice.

Conflict of interest: The authors declare that they have no conflict of interest with the contents of this article.

Author contributions: AJC, PNK, and WT-J conceived and planned the study, which was coordinated by AJC. PNK and AJC wrote the manuscript. PNK and WT-J conducted most of the experiments. TYY assisted with the co-immunoprecipitation study. MC carried out the bioinformatics analysis. AAV and JAW carried out the MuDPIT analysis. All authors reviewed and approved the manuscript.

REFERENCES

1. Courey, A. J., and Jia, S. (2001) Transcriptional repression: the long and the short of it. *Genes Dev* **15**, 2786-2796
2. Turki-Judeh, W., and Courey, A. J. (2012) Groucho: a corepressor with instructive roles in development. *Current topics in developmental biology* **98**, 65-96
3. Martinez, C. A., and Arnosti, D. N. (2008) Spreading of a corepressor linked to action of long-range repressor hairy. *Mol Cell Biol* **28**, 2792-2802
4. Chen, G., and Courey, A. J. (2000) Groucho/TLE family proteins and transcriptional repression. *Gene* **249**, 1-16
5. Gasperowicz, M., and Otto, F. (2005) Mammalian Groucho homologs: redundancy or specificity? *J Cell Biochem* **95**, 670-687
6. Nagel, A. C., Krejci, A., Tenin, G., Bravo-Patino, A., Bray, S., Maier, D., and Preiss, A. (2005) Hairless-mediated repression of notch target genes requires the combined activity of Groucho and CtBP corepressors. *Mol Cell Biol* **25**, 10433-10441
7. Blair, S. S. (2007) Wing vein patterning in Drosophila and the analysis of intercellular signaling. *Annual review of cell and developmental biology* **23**, 293-319
8. Hasson, P., and Paroush, Z. (2006) Crosstalk between the EGFR and other signalling pathways at the level of the global transcriptional corepressor Groucho/TLE. *British journal of cancer* **94**, 771-775
9. Allen, T., van Tuyl, M., Iyengar, P., Jothy, S., Post, M., Tsao, M. S., and Lobe, C. G. (2006) Grg1 acts as a lung-specific oncogene in a transgenic mouse model. *Cancer Res* **66**, 1294-1301

10. Buscarlet, M., and Stifani, S. (2007) The 'Marx' of Groucho on development and disease. *Trends Cell Biol* **17**, 353-361
11. Jennings, B. H., Pickles, L. M., Wainwright, S. M., Roe, S. M., Pearl, L. H., and Ish-Horowicz, D. (2006) Molecular recognition of transcriptional repressor motifs by the WD domain of the Groucho/TLE corepressor. *Mol Cell* **22**, 645-655
12. Song, H., Hasson, P., Paroush, Z., and Courey, A. J. (2004) Groucho oligomerization is required for repression in vivo. *Mol Cell Biol* **24**, 4341-4350
13. Chodaparambil, J. V., Pate, K. T., Hepler, M. R., Tsai, B. P., Muthurajan, U. M., Luger, K., Waterman, M. L., and Weis, W. I. (2014) Molecular functions of the TLE tetramerization domain in Wnt target gene repression. *EMBO J* **33**, 719-731
14. Chen, G., Nguyen, P. H., and Courey, A. J. (1998) A role for Groucho tetramerization in transcriptional repression. *Mol Cell Biol* **18**, 7259-7268
15. Pinto, M., and Lobe, C. G. (1996) Products of the grg (Groucho-related gene) family can dimerize through the amino-terminal Q domain. *J Biol Chem* **271**, 33026-33031
16. Chen, G., Fernandez, J., Mische, S., and Courey, A. J. (1999) A functional interaction between the histone deacetylase Rpd3 and the corepressor groucho in Drosophila development. *Genes Dev* **13**, 2218-2230
17. Choi, C. Y., Kim, Y. H., Kwon, H. J., and Kim, Y. (1999) The homeodomain protein NK-3 recruits Groucho and a histone deacetylase complex to repress transcription. *J Biol Chem* **274**, 33194-33197
18. Winkler, C. J., Ponce, A., and Courey, A. J. (2010) Groucho-mediated repression may result from a histone deacetylase-dependent increase in nucleosome density. *PLoS One* **5**, e10166
19. Turki-Judeh, W., and Courey, A. J. (2012) The unconserved groucho central region is essential for viability and modulates target gene specificity. *PLoS One* **7**, e30610
20. Hasson, P., Egoz, N., Winkler, C., Volohonsky, G., Jia, S., Dinur, T., Volk, T., Courey, A. J., and Paroush, Z. (2005) EGFR signaling attenuates Groucho-dependent repression to antagonize Notch transcriptional output. *Nat Genet* **37**, 101-105
21. Nuthall, H. N., Joachim, K., and Stifani, S. (2004) Phosphorylation of serine 239 of Groucho/TLE1 by protein kinase CK2 is important for inhibition of neuronal differentiation. *Mol Cell Biol* **24**, 8395-8407
22. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J* **272**, 5129-5148
23. Uversky, V. N., and Dunker, A. K. (2010) Understanding protein non-folding. *Biochim Biophys Acta* **1804**, 1231-1264
24. Tantos, A., Han, K. H., and Tompa, P. (2012) Intrinsic disorder in cell signaling and gene transcription. *Mol Cell Endocrinol* **348**, 457-465
25. Bondos, S. E., and Hsiao, H. C. (2012) Roles for intrinsic disorder and fuzziness in generating context-specific function in Ultrabithorax, a Hox transcription factor. *Adv Exp Med Biol* **725**, 86-105
26. Damgaard, C. K., Kahns, S., Lykke-Andersen, S., Nielsen, A. L., Jensen, T. H., and Kjems, J. (2008) A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo. *Mol Cell* **29**, 271-278
27. Monsalve, M., Wu, Z., Adelmant, G., Puigserver, P., Fan, M., and Spiegelman, B. M. (2000) Direct coupling of transcription and mRNA processing through the thermogenic coactivator PGC-1. *Mol Cell* **6**, 307-316
28. Furger, A., O'Sullivan, J. M., Binnie, A., Lee, B. A., and Proudfoot, N. J. (2002) Promoter proximal splice sites enhance transcription. *Genes Dev* **16**, 2792-2799
29. Kwek, K. Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N. J., and Akoulitchev, A. (2002) U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat Struct Biol* **9**, 800-805

30. Knoop, L. L., and Baker, S. J. (2000) The splicing factor U1C represses EWS/FLI-mediated transactivation. *J Biol Chem* **275**, 24865-24871
31. Armknecht, S., Boutros, M., Kiger, A., Nybakken, K., Mathey-Prevot, B., and Perrimon, N. (2005) High-throughput RNA interference screens in Drosophila tissue culture cells. *Methods in enzymology* **392**, 55-73
32. Soeller, W. C., Poole, S. J., and Kornberg, T. (1988) In vitro transcription of the Drosophila engrailed gene. *Genes Dev* **2**, 68-81
33. Wohlschlegel, J. A. (2009) Identification of SUMO-conjugated proteins and their SUMO attachment sites using proteomic mass spectrometry. *Methods Mol Biol* **497**, 33-49
34. Zhou, R., Mohr, S., Hannon, G. J., and Perrimon, N. (2013) Inducing RNAi in Drosophila cells by transfection with dsRNA. *Cold Spring Harbor protocols* **2013**, 461-463
35. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36
36. Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550
37. Herrmann, C., Van de Sande, B., Potier, D., and Aerts, S. (2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res* **40**, e114
38. Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., McLaren, P., North, P., Rana, D., Riley, T., Sullivan, J., Watkins, X., Woodbridge, M., Lilley, K., Russell, S., Ashburner, M., Mizuguchi, K., and Micklem, G. (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome biology* **8**, R129
39. Brantjes, H., Roose, J., van De Wetering, M., and Clevers, H. (2001) All Tcf HMG box transcription factors interact with Groucho-related co-repressors. *Nucleic Acids Res* **29**, 1410-1419
40. Daniels, D. L., and Weis, W. I. (2005) Beta-catenin directly displaces Groucho/TLE repressors from Tcf/Lef in Wnt-mediated transcription activation. *Nature structural & molecular biology* **12**, 364-371
41. Bouazoune, K., and Brehm, A. (2006) ATP-dependent chromatin remodeling complexes in Drosophila. *Chromosome Res* **14**, 433-449
42. Flores-Saaib, R. D., and Courey, A. J. (2000) Analysis of Groucho-histone interactions suggests mechanistic similarities between Groucho- and Tup1-mediated repression. *Nucleic Acids Res* **28**, 4189-4196
43. Sekiya, T., and Zaret, K. S. (2007) Repression by Groucho/TLE/Grg proteins: genomic site recruitment generates compacted chromatin in vitro and impairs activator binding in vivo. *Mol Cell* **28**, 291-303
44. Edmondson, D. G., Smith, M. M., and Roth, S. Y. (1996) Repression domain of the yeast global repressor Tup1 interacts directly with histones H3 and H4. *Genes Dev* **10**, 1247-1259
45. Patel, S. B., and Bellini, M. (2008) The assembly of a spliceosomal small nuclear ribonucleoprotein particle. *Nucleic Acids Res* **36**, 6482-6493
46. Will, C. L., and Luhrmann, R. (2011) Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3**
47. Fisher, A. L., Ohsako, S., and Caudy, M. (1996) The WRPW motif of the hairy-related basic helix-loop-helix repressor proteins acts as a 4-amino-acid transcription repression and protein-protein interaction domain. *Mol Cell Biol* **16**, 2670-2677
48. Thorne, N., Inglese, J., and Auld, D. S. (2010) Illuminating insights into firefly luciferase and other bioluminescent reporters used in chemical biology. *Chemistry & biology* **17**, 646-657
49. Kaul, A., Schuster, E., and Jennings, B. H. (2014) The Groucho co-repressor is primarily recruited to local target sites in active chromatin to attenuate transcription. *PLoS genetics* **10**, e1004595

50. Celniker, S. E., Dillon, L. A., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K. P., Waterston, R. H., and mod, E. C. (2009) Unlocking the secrets of the genome. *Nature* **459**, 927-930
51. Fisher, W. W., Li, J. J., Hammonds, A. S., Brown, J. B., Pfeiffer, B. D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J. A., Eisen, M. B., Bickel, P. J., Biggin, M. D., and Celniker, S. E. (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 21330-21335
52. Walhout, A. J. (2011) What does biologically meaningful mean? A perspective on gene regulatory network validation. *Genome biology* **12**, 109
53. Muse, G. W., Gilchrist, D. A., Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., Zeitlinger, J., and Adelman, K. (2007) RNA polymerase is poised for activation across the genome. *Nat Genet* **39**, 1507-1511
54. Li, L. M., and Arnosti, D. N. (2011) Long- and short-range transcriptional repressors induce distinct chromatin states on repressed genes. *Current biology : CB* **21**, 406-412
55. Niefind, K., Guerra, B., Ermakowa, I., and Issinger, O. G. (2001) Crystal structure of human protein kinase CK2: insights into basic properties of the CK2 holoenzyme. *EMBO J* **20**, 5320-5331
56. Niefind, K., Raaf, J., and Issinger, O. G. (2009) Protein kinase CK2 in health and disease: Protein kinase CK2: from structures to insights. *Cell Mol Life Sci* **66**, 1800-1816
57. Meier, U. T., and Blobel, G. (1992) Nopp140 shuttles on tracks between nucleolus and cytoplasm. *Cell* **70**, 127-138
58. Fuller-Pace, F. V. (2006) DExD/H box RNA helicases: multifunctional proteins with important roles in transcriptional regulation. *Nucleic Acids Res* **34**, 4206-4215
59. Rossow, K. L., and Janknecht, R. (2003) Synergism between p68 RNA helicase and the transcriptional coactivators CBP and p300. *Oncogene* **22**, 151-156
60. Buszczak, M., and Spradling, A. C. (2006) The Drosophila P68 RNA helicase regulates transcriptional deactivation by promoting RNA release from chromatin. *Genes Dev* **20**, 977-989
61. Wilson, B. J., Bates, G. J., Nicol, S. M., Gregory, D. J., Perkins, N. D., and Fuller-Pace, F. V. (2004) The p68 and p72 DEAD box RNA helicases interact with HDAC1 and repress transcription in a promoter-specific manner. *BMC molecular biology* **5**, 11
62. Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A., and Mann, M. (1998) Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat Genet* **20**, 46-50
63. Liu, Z. R. (2002) p68 RNA helicase is an essential human splicing factor that acts at the U1 snRNA-5' splice site duplex. *Mol Cell Biol* **22**, 5443-5450
64. Reed, R. (2003) Coupling transcription, splicing and mRNA export. *Curr Opin Cell Biol* **15**, 326-331
65. Han, J., Xiong, J., Wang, D., and Fu, X. D. (2011) Pre-mRNA splicing: where and when in the nucleus. *Trends Cell Biol* **21**, 336-343
66. Bentley, D. L. (2005) Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol* **17**, 251-256
67. Maniatis, T., and Reed, R. (2002) An extensive network of coupling among gene expression machines. *Nature* **416**, 499-506
68. Bentley, D. (2002) The mRNA assembly line: transcription and processing machines in the same factory. *Curr Opin Cell Biol* **14**, 336-342
69. Cramer, P., Caceres, J. F., Cazalla, D., Kadener, S., Muro, A. F., Baralle, F. E., and Kornblihtt, A. R. (1999) Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol Cell* **4**, 251-258

70. Cramer, P., Pesce, C. G., Baralle, F. E., and Kornblihtt, A. R. (1997) Functional association between promoter structure and transcript alternative splicing. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 11456-11460
71. Merkhofer, E. C., Hu, P., and Johnson, T. L. (2014) Introduction to cotranscriptional RNA splicing. *Methods Mol Biol* **1126**, 83-96
72. Hasson, P., Muller, B., Basler, K., and Paroush, Z. (2001) Brinker requires two corepressors for maximal and versatile repression in Dpp signalling. *EMBO J* **20**, 5725-5736
73. Paroush, Z., Finley, R. L., Jr., Kidd, T., Wainwright, S. M., Ingham, P. W., Brent, R., and Ish-Horowicz, D. (1994) Groucho is required for Drosophila neurogenesis, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. *Cell* **79**, 805-815
74. Grbavec, D., and Stifani, S. (1996) Molecular interaction between TLE1 and the carboxyl-terminal domain of HES-1 containing the WRPW motif. *Biochem Biophys Res Commun* **223**, 701-705
75. Adelman, K., and Lis, J. T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-731
76. Gaertner, B., and Zeitlinger, J. (2014) RNA polymerase II pausing during development. *Development* **141**, 1179-1183
77. Bartkowiak, B., Liu, P., Phatnani, H. P., Fuda, N. J., Cooper, J. J., Price, D. H., Adelman, K., Lis, J. T., and Greenleaf, A. L. (2010) CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev* **24**, 2303-2316
78. Lee, J. S., and Shilatifard, A. (2007) A site to remember: H3K36 methylation a mark for histone deacetylation. *Mutation research* **618**, 130-134
79. Wagner, E. J., and Carpenter, P. B. (2012) Understanding the language of Lys36 methylation at histone H3. *Nature reviews. Molecular cell biology* **13**, 115-126

TABLES

Table 1: PCR primers used in construction of plasmids encoding GST fusion proteins

Domain	Sequence
Q	ATTATTAGGATCCATGGATTACAAGGACGATGACGATAATATCCCTCACCGGTGCGCCACCCC
	ATTATTACTCGAGTCACTGCTGGGCGTGGATCTGTTGCCCA
GP	ATTATTAGGATCCATGGATTACAAGGACGATGACGATAAGTGCCAGGTGGACCACCTCAGCCGA
	ATTATTACTCGAGTCACGAATTGAGCAATCGCTCCTCGGC
CcN	ATTATTAGGATCCATGGATTACAAGGACGATGACGATAAGTTTCGCCGGCCGATCGTGAGAAGT
	ATTATTACTCGAGTCACATAGACACGTGCTCGCCGTTGGGA
SP	ATTATTAGGATCCATGGATTACAAGGACGATGACGATAAGAGGTGCGCGATCGGGAAAGCTTGA
	ATTATTACTCGAGTCAACCCGTTAGGGCCGAGGGATGTGGA

Table 2: RT-qPCR primers

Gene	Sequence
Gro	TTTATTACAACATGTTCGAAATCATGC
	TTCGCTTTTTGATGCGTTGCTAC
snRNP-U1-C	CTCAGGAACGGCATCAACGTT
	TATAATTAATTGTTTTCGCTATCGGG
Rpl32	CCCAAGGGTATCGACAACAGA
	CGATCTCGCCGCAGTAAAC
U1 snRNA	ATACTTACCTGGCGTAGAGGTTAACC
	AACGCCATTCCCGGCTA

Table 3. Enriched gene ontology groups of Gro-interacting proteins[1]

Enriched gene ontology	# of genes
Gene expression	83
Chromosome organization	21
Chromatin modification	11
mRNA processing	53
Cell cycle	30
Cell differentiation	66
Developmental process	76
Neurogenesis	57
Anatomical structure development	74

[1] $p<0.05$

Table 4. Representative Gro-interacting proteins[1]

Chromosome organization	
Acf1	ACF chromatin remodeling complex
Iswi	ACF chromatin remodeling complex
Caf1	dNuRD chromatin remodeling complex
Nap1	Histone chaperone
Ball	H2A Thr 119 kinase
JIL-1	H3 Ser 10 kinase
Top1	Topoisomerase
Top2	Topoisomerase
Developmental process	
CKIIα	CKII complex
CKIIβ	CKII complex
Nopp140	Negative regulator of CKII
Nito	Positive regulator of Wnt signaling pathway
Rm62	DEAD box helicase
Fmr1	Fragile X protein; Interacting partner of Rm62
vir	Involvement in sex determination
Snama	Involvement in eye morphogenesis
nonA	Involvement in visual perception
mRNA processing	
snRNP-U1-A	U1 snRNP complex
snRNP-U1-C	U1 snRNP complex
snRNP-U1-70K	U1 snRNP complex
U4-U6-60K	U4/U6 snRNP complex
CG7028	U4/U6 snRNP complex
Prp31	U4/U6 snRNP complex
Prp8	U5 snRNP complex
U2af38	U2 snRNP complex
U2af50	U2 snRNP complex
SF2	U2 snRNP complex
SmD2	Sm complex
SmD3	Sm complex
Nop60B	H/ACA ribonucleoprotein complex subunit
NHP2	H/ACA ribonucleoprotein complex subunit

[1]For the complete list, full gene/protein names, and UniProt identifiers, see Table S1.

Table 5. Positive and negative Gro-regulators[1]

Potential negative regulators of Gro[2]	Potential positive regulators of Gro[3]
CKIIα, CKIIβ, Nopp140, fl(2)d, l(2)35Df, l(3)72Ab, vir, nonA, nito, x16, Nap1, JIL-1, nop5, NHP2, FK506-bp1, CG3605, Prp31, Fmr1, Cdk12, CG6418, CG7372, CG7946, Srp68, Srp72, Ssrp, Pitslre, Pep, Nab2	snRNP-U1-C, snRNP-U1-70K, U2af50, U4-U6-60K, Rm62, Orc2, smid, Acn, Acf1, snama, CG1622, ZCHC8, CG4709, CG4806, lat, Srp19

[1]See Table S2 for quantitative information on positive and negative regulation by these factors. See Experimental Procedures for an explanation of the test of statistical significance that genes had to pass to be included in this list.

[2]Negative regulators are defined as the products of those genes the knock down of which led to increased repression by Gal4-Gro in the reporter assay.

[3]Positive regulators are defined as the products of those genes the knock down of which led to decreased repression by Gal4-Gro in the reporter assay.

Table 6. Genes up regulated upon knockdown of either Gro or snRNP-U1-C

Name	Function
Secreted Wg-interacting molecule	Wnt signaling pathway
Wnt oncogene analog 5	Wnt signaling pathway
E(spl)m2-BFM	Notch signaling
spatzle	Toll signaling pathway
SH2 ankyrin repeat kinase	JNK cascade
Dawdle	SMAD protein signal transduction
CG33275	Rho protein signal transduction
Epac	Rap protein signal transduction
Boundary element-associated factor of 32kD	H3K9 methylation
Syncrip	Dorsal/ventral axis specification
Fasciclin 1	Neuron recognition
axotactin	Transmission of nerve impulse
Muscle-specific protein 300 kDa	Skeletal muscle tissue development
cheerio	Lamellocyte differentiation

Table S1: Gro interacting proteins

See Excel spreadsheet

Table S2: Gro-interacting proteins that have significant effects on repression by Gal4-Gro

See Excel spreadsheet

FIGURE LEGENDS

Figure 1. Purification of Gro-interacting proteins. (A) Schematic representation of Gro. The Q, GP, CcN, and SP domains were tagged with GST. (B) The GST-tagged domains were expressed in E. coli and purified with glutathione agarose beads. They were then resolved by 10% SDS PAGE and visualized by Coomassie Blue staining. These proteins were then used as affinity reagents in the purification of Gro-interacting proteins from embryonic nuclear extracts, which were subsequently identified by MuDPIT (see Tables 4 and S1). (C) Venn diagram showing overlap between the non-ribosomal proteins identified in two replicate sets of affinity purification experiments. Fisher's exact test indicates that the overlap between the two sets is highly significant ($p < 2.2x10^{-16}$).

Figure 2. Validation of the interaction between Gro and U1 snRNP. (A) 0-12 hour Drosophila embryo nuclear extracts were subjected to immunoprecipitation using an affinity purified polyclonal antibody directed against the Gro GP domain, or, as a control, rabbit IgG. To assess immunoprecipitation efficiency and specificity, immunoprecipitates were subjected to SDS-PAGE and immunoblotting. The blot was probed with a mixture of the rabbit anti-GP domain antibody and a mouse monoclonal anti-Gro antibody, and IR-dye labeled secondary antibodies. The signal from the rabbit antibody was detected in the green channel of the IR imager, while the signal from the mouse antibody was detected in the red channel. Rabbit IgG heavy chain (IgG) and Gro bands are indicated with arrows on the right. The orange-yellow color of the Gro band is indicative of the overlap between the red and green signals. Lane 1) Markers labeled in kD; Lane 2) 10% input; Lane 3) Anti-Gro immunoprecipitate; Lane 4) Rabbit IgG immunoprecipitate, Lane 5) Mock anti-Gro immunoprecipitate from which input nuclear extract was omitted. (B) RNA was extracted from immunoprecipitates prepared as described in A. The RNA from the immunoprecipitates as well as the RNA extracted from the input nuclear extracts was analyzed by RT-qPCR as described in Experimental Procedures to determine U1 snRNA levels. Error bars based on two independent biological replicates indicate standard deviation. A two-tailed t-test gives $p = 0.016$.

Figure 3: The three-reporter high throughput luciferase assay. (A) Schematic representation of the three reporters. Constructs are not drawn to scale. In the red luciferase reporter, the Gal4 binding sites (UAS elements) are immediately upstream of the enhancer, while in the green luciferase reporter, the UAS elements are about 2 kb downstream of the transcriptional start site. Expression is induced by the Dorsal and Twist activators and repressed by Gal4-Gro. The Renilla luciferase reporter was used as an internal control for transfection efficiency and cell viability. (B) Flow chart of the reporter assay. (C) Validation of the reporter assay. Co-transfection with Dorsal and Twist (Dl/Twi) encoding plasmids activated both the red and green reporters, while addition of a plasmid encoding the Gal4-Gro fusion resulted in repression of the reporters. Dorsal, Gro (including Gal4-Gro), and the histone deactylase Rpd3, which is partially required for Gro-mediated repression (18), were knocked down by RNAi.. Data are normalized to the red and green signals from the Gro dsRNA sample. Error bars based on triplicate transfection assays represent standard deviation. (D) Representative results of the reporter assay. The luciferase reporter assay was carried out using three non-overlapping dsRNAs from the genes encoding vir, snRNP-U1-C, and SRPK. The result of transfection with each dsRNA was compared to that of transfection with GFP dsRNA. Error bars based on triplicate transfection assays represent standard deviation.

Figure 4: Genome-wide expression profiling reveals co-regulation of genes by Gro and snRNP-U1-C. (A) Expression of Gro and snRNP-U1-C mRNA after dsRNA treatment. RT-qPCR was performed after extraction of total RNA. Data was normalized to reference gene Rpl32. (B) Comparison of transcriptomes from our wild-type S2 cell RNA-seq data and the modENCODE S2 cell RNA-seq data. (C) Comparison of transcriptomes from our Gro knockdown RNA-seq data and previously published Gro knockdown RNA-seq data (49). The transcripts that were detected at significant levels in only the previously published Gro knockdown study (represented by the points in contact with the vertical axis) correspond primarily to non-polyadenylated transcripts. In B and C, the scale on both axes is $\log_2$(CPM) where CPM is counts per million sequence reads. (D) Based on RNA-seq analysis of wild-type and snRNP-U1-C knockdown cells, genes were categorized as non-differentially expressed upon knockdown (non-DE, 12,028 genes), up-regulated upon knockdown (1,431 genes), and down-regulated upon knockdown (1,691genes). Percent of genes in each category with no introns is shown. Some Drosophila genes lack annotated transcripts and thus it was not possible to determine their intron count. This results in a small numerical discrepancy between the number of differentially expressed genes included in this analysis and the number of snRNP-U1C differentially expressed genes shown in part E of this figure. (E) Venn diagram showing numbers of differentially expressed genes in Gro and snRNP-U1-C knockdown cells and the overlap between these sets. Fisher's exact test indicates that the overlap is highly significant ($p < 2.2 \times 10^{-16}$). (F) Enrichment of Gro/snRNP co-regulated genes for various features. Normalized enrichment scores are calculated using cumulative recovery curves (37). Scores above 2.5 are considered significant.

Figure 5. Gro binding regions in differentially expressed genes. (A) S2 cell ChIP-seq data (49) identified 1242 Gro binding sites, which map to 748 genes, 46 of which were differentially expressed when we knocked-down Gro. Of the 46 differentially expressed genes, 39 were up regulated and 7 were down regulated in response to Gro knock-down. (B) Gro binding regions in the 46 differentially expressed genes are significantly enriched for Su(H) and Brk binding sites.

Figure 6. Gro-repressed genes are enriched for promoter proximal Pol II. Percent of non-differentially expressed genes, and genes that are either up-regulated or down regulated in Gro knockdown cells containing no Pol II bound, Pol II bound, or enriched for promoter proximal Pol II as ascertained by Pol II ChIP-chip analysis (53).

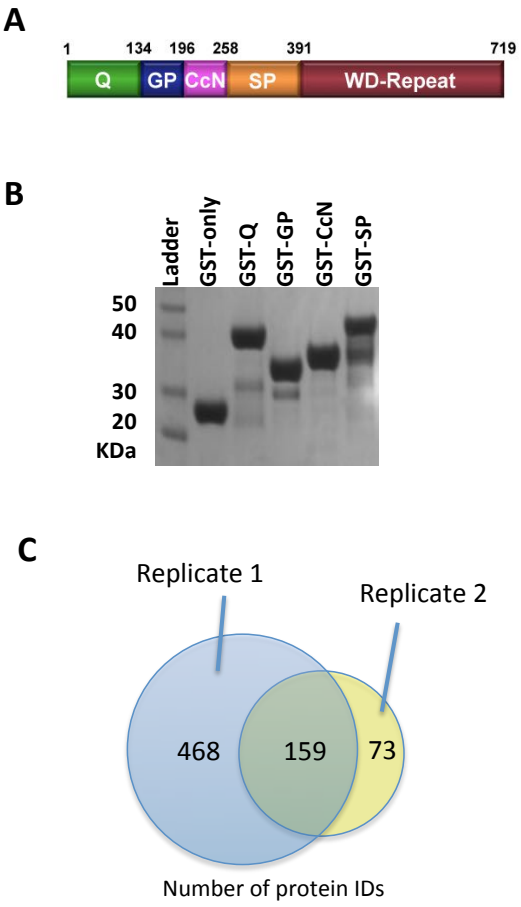


Fig 1

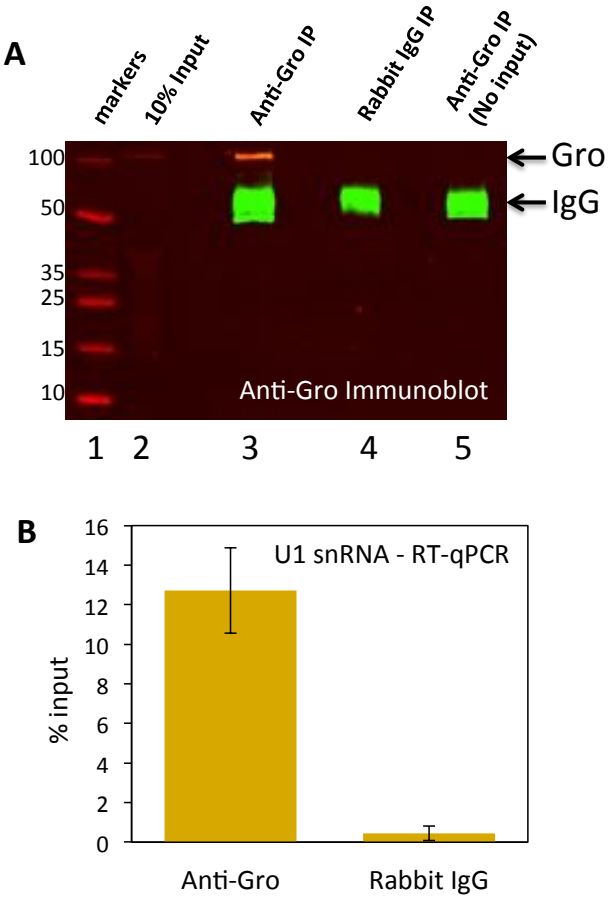


Fig 2

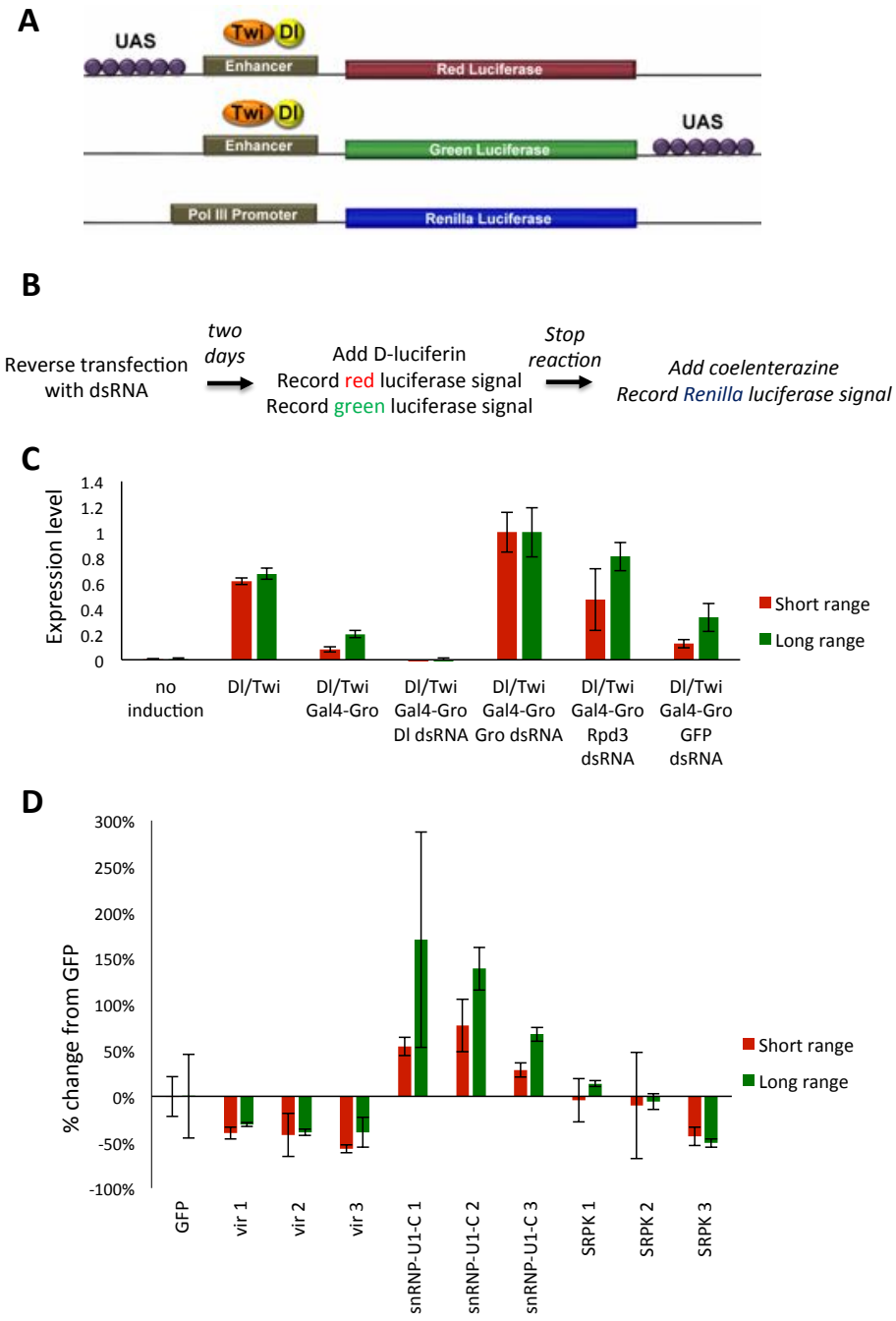


Fig 3

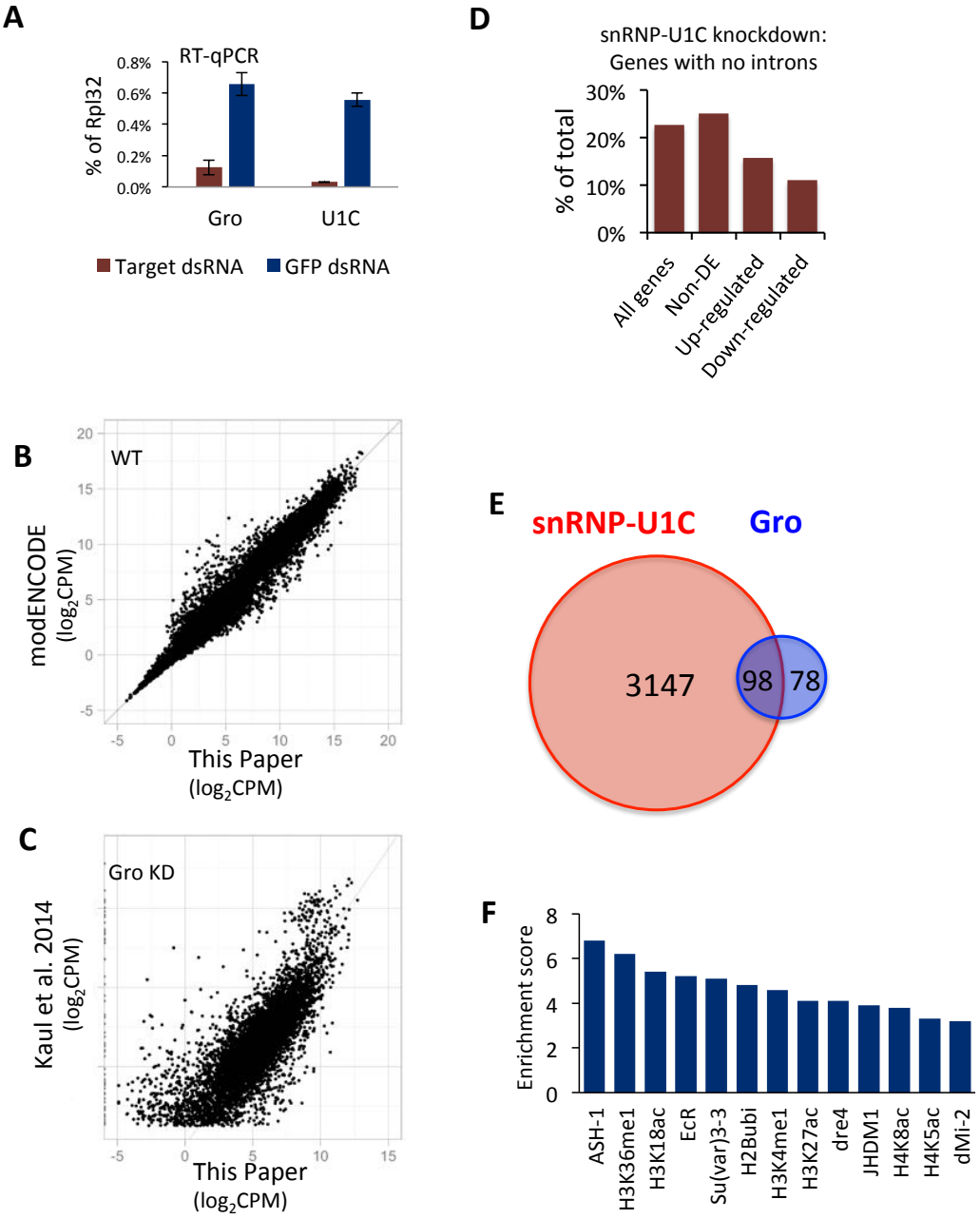


Fig 4

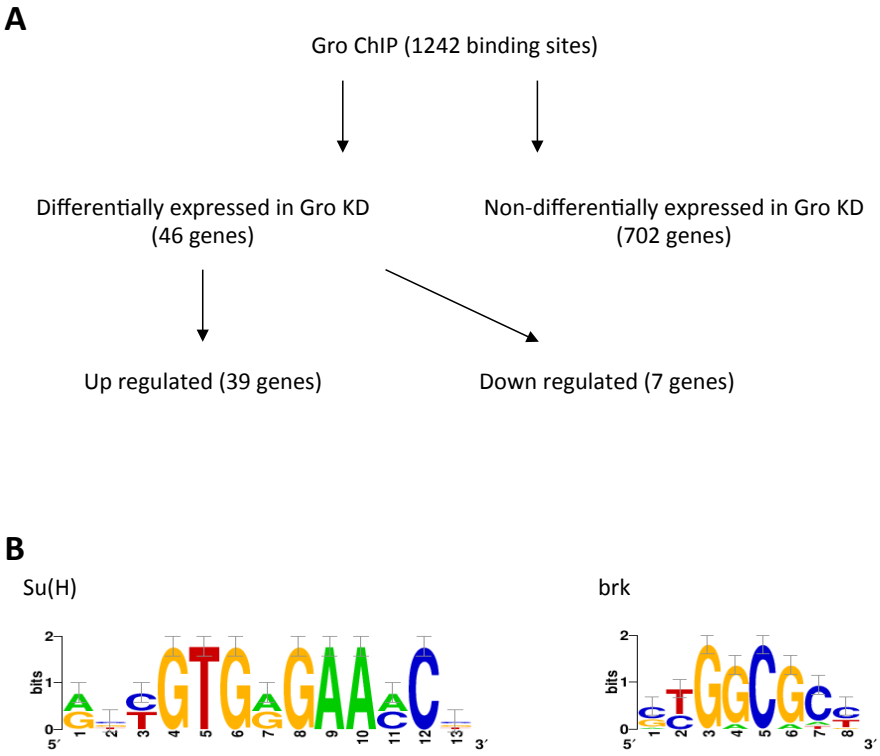


Fig 5

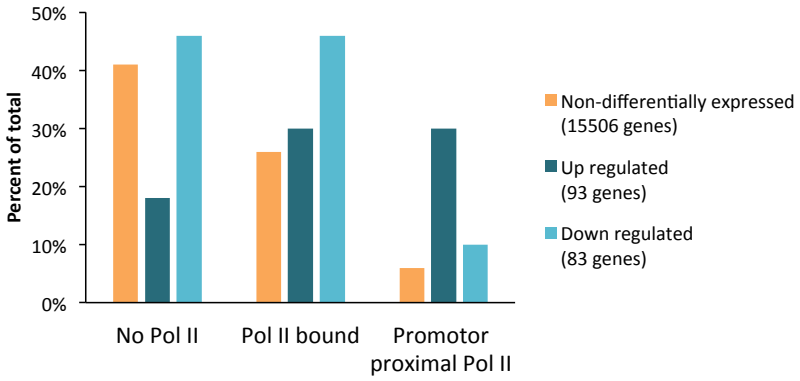


Fig 6