# **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. Through its interaction with numerous sequence-specific transcription factors (repressors), 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. 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. 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.

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, 2012 #2966}.

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, 1998 #267} and *in vivo* {Song, 2004 #1161}. However, it does not appear to be a universal requirement for repressive activity in all developmental contexts {Jennings, 2007 #2990}. Evidence from ChIP-PCR experiments suggests Gro spreads over potentially long stretches of chromatin presumably through its ability to self-associate {Winkler, 2010 #2964} {Martinez, 2008 #2287}, 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, 2014 #2204} 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 interactions, 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, 2013 #2377}; 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, 2015 #3045}. 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

1. *Fly strains*

Flies were maintained on standard medium at 25˚C. UAS-*Gro* and UAS-*GroΔGP* transgenic flies were described previously {Turki-Judeh, 2012 #2966}. 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, 2009 #2369}*.* Germ line clones of the *gro* mutant fly allele MB36 (a null allele) were used for Groucho loss-of-function studies {Jennings, 2007 #2990}. These lines were generated using the standard dominant female sterile FLP/FRT protocol {Chou, 1996 #3178}.

1. *Groucho chromatin immunoprecipation (ChIP) and sequencing*

ChIP was carried out as described previously {Bonn, 2012 #3179}. 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).

1. *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, 2012 #3049}. Peak calling was performed using MACS2 (v2.1.0) with default parameters {Zhang, 2008 #2203}. Peak visualizations were generated with Integrated Genome Browser (v8.4.2) {Nicol, 2009 #3050}. 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, 2014 #3187}.

1. *Embryonic RNA isolation and sequencing (RNA-seq)*

Staged embryos were manually homogenized in TRIzol reagent (Life Technologies) according to manufacturer protocols. Purified RNA quality was assessed via Bioanalyzer 2100 (Agilent Technologies). polyA-selected libraries were generated with TruSeq Stranded mRNA Library Prep Kit (Illumina) and sequenced on the Illumina HiSeq 2000 platform.

1. *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, 2013 #1817} against the *Drosophila melanogaster* genome (iGenomes BDGP 5.25 assembly) with iGenomes gene models as a guide. Gene assignment was performed with HTSeq {IAnders, 2015 #3027}.

1. *Gene expression and Groucho target gene identificaiton*

Normalized gene expression values and differential expression analysis generated with DESeq2 (v1.8.0) {Love, 2014 #3031}. Genes exhibiting a log2(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

## 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, 2012 #308}. 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, 2011 #3035} 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, 1995 #3171}. 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). 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 non-overlapping 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, a 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 stages 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, 2010 #759}. 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, 2006 #3119}. Some factors can be recruited to HOT regions independently from their ability to bind and recognize DNA sequence {Li, 2008 #2374}. 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 Groucho regions that overlap a HOT zone is largely invariant between time points, Groucho 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 {Li, 2008 #2374} {Kaplan, 2011 #3172} {Li, 2011 #3173}. DNA accessibility has been mapped across multiple developmental stages {Li, 2011 #3173}, 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, 2007, r08904} {Winkler, 2010, r07182}, widespread recruitment to these sites may indicate that additional undocumented inputs are required to initiate Groucho-mediated chromatin condensation.

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

Choosing the nearest or overlapping gene as a potential Groucho-regulated gene, 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, 2011 #3117}, rather than the broad peaks typical of either highly polymeric factors or histone marks (Fig. 2-7). Interesting, *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, 2007 #1658}, 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 {Dubnicoff, 1997 #2366} {Barolo, 1997 #2365}. 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 therefore predicted to be enriched for regulatory elements {Bradnam, 2008 #3034}.

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 less low-affinity and less specific binding of 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, 1989 #1112}. 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, 1997 #2366} {Kirov, 1994 #3107}. 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, 1998 #3036}. 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, however 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). The downstream peak overlaps the TSS. This hints at 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, 1997 #2366}. Dorsal binding sites necessary for restriction of *dpp* expression to the dorsal portion of the embryo map to a VRR in the gene’s second intron {Huang, 1993 #3037}. 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, 1994 #3107}. 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. A weaker Gro peak is also observed over the transcriptional start site, perhaps indicative of looping, but there is no continuous Gro spreading between the VRR and the start site.

*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, 1998 #3036}. 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, 2006 #3108}. 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 {Thisse, 1987 #3109} {Ip, 1992 #3110}. Dorsal activates both *twist* and *snail* by binding to Ventral Activation Regions (VARs) in the 5’ flanking regions of these genes {Ip, 1992 #3110}. No role for Groucho has been identified in the regulation of either gene. Surprisingly, however Gro binds the VARs in in 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, 2006 #5} {Zeitlinger, 2007 #3025}. 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 form observation of the *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, 2009 #6}. 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 lesser degrees of overlap with one or more additional assayed factor (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 #2280}. 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, 2000 #3120}, 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, two-fold and four-fold higher than endogenous, as well as a line overexpressing a Groucho deletion mutant lacking the central SP domain (Gro∆SP). 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 {Turki-Judeh, 2012 #2966}, a trend that we sought to investigate on a genome-wide scale. Additionally, we analyzed the transcriptome of embryos lacking maternally-contributed functional Groucho. These embryos are derived from maternal germline clones homozygous for *groMB36*, a lethalallele that introduces an ectopic splice site near the 5’ end of *gro* {Jennings, 2007 #2990}. 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 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, while there are enough similarities between overexpression and wild-type embryos that they can cluster together. 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). Gro loss-of-function samples plot independently at all timepoints, exhibiting a strong divergence from all other samples at all timepoints.

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 enhancer 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 the closest feature to a significant Groucho binding region. 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 common between the two full-length Gro overexpression lines when compared 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, incorporating CBP ChIP-seq data and a measurement of a mutant CBP transcriptome {Holmqvist, 2012 #3115}. Similar methodologies have been utilized to integrate transcription factor binding and expression data in other contexts {Wang, 2013 #2256}. 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 *groMB36* 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, 2011 #1659}. 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, 2011 #2284}. 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 activated in either overexpression line compared to repressed genes. This difference is most evident in the first two time windows, where Groucho transcript levels are the highest. 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 in gene activation.

Through this scoring methodology, we identify 351 potential Groucho target genes across all timepoints. Of these, only 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 and are highly enriched over all other groups in the dosage-response group. Other ontologies primarily correspond to different tissue specification processes. Restricting the gene set to the 146 predicted Groucho-activated genes reveals no gene ontologies were significantly enriched (data not shown), leading us to hypothesize that these genes are potentially the result of noise in the gene expression data and do not represent direct Groucho targets.

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, 2007 #3180} 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, 1996 #3181}. Atonal (ato) and Sprouty (sty) are factors with known functions in respiratory and eye development, respectively {Jarman, 1994 #3183} {Hacohen, 1998 #3182}, in which Groucho’s potential roles have not been investigated.

The core regulatory network of targets identified by Groucho occupancy 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, 2001 #3184}. Groucho is recruited to Tinman, a Pannier-interacting protein, to regulate cardiac gene expression {Choi, 1999 #3186}. The association and regulation of multiple Pannier target genes by Groucho may represent a significant contribution by Groucho to cardiac development.

## Discussion

In our current study, we have identified thousands of novel Groucho-recruitment sites throughout the Drosophila genome. The majority of these sites are present during only one of the three timepoints analyzed, supporting the conclusion that the majority of these sites are actively participating in developmental gene regulation. The small carry-over of Gro binding from the 1.5 – 4 hr to 4 – 6.5 hr stages represents a widespread shift in Gro occupancy, consistent with the changing roles of Gro throughout development, as the availability of sequence-specific transcription factors changes across the embryo.

During the 1.5 – 4 hr stage, Gro is essential for correct determination of cell fates along the dorsal-ventral axis through cooperation with Dorsal. Groucho is recruited both within and surrounding two early ventrally-repressed genes, *zen* and *dpp*. The presence of a Groucho-bound peak at the transcription start site often coincides with the presence of Groucho binding upstream or inside of a gene. We hypothesize that this supports a model of repression whereby Groucho is recruited to repressive regulatory regions and precipitates a rearrangement of local chromatin, bringing Gro into contact with promoters or gene bodies. Subsequent repression may be accomplished through multiple mechanisms. Repression may also be initiated by the well-documented interaction of Groucho with HDAC1/Rpd3, leading to deacetylation of histones within and directly upstream of the gene body, resulting in chromatin condensation and repression. The latter mechanism of repression, via alteration of the histone mark landscape, is a potential mechanism for Groucho to act epigenetically, achieving repression that lasts after Groucho is no longer associated with a locus. This is consistent with behavior seen at the *zen* locus, where Groucho occupancy is essentially gone following 4 hours of development, despite Zen remaining repressed throughout later stages of development. Work presented in the next chapter will provide evidence that Gro occupancy positively correlates with stalled PolII in the developing embryo.

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 one strategy common to Groucho regulation. The evolution of regulatory regions within introns is common in *Drosophila*, and in animals more generally. Multiple factors, including Kruppel and Twist have been shown to commonly localize to intronic regions {Matyash, 2004 #3046}{Sandmann, 2007 #3048}{Zeitlinger, 2007 #3025}. 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, 2011 #3051}. Consistent with this hypothesis, developmentally-regulated genes known to exhibit promoter-proximal pausing of RNA PolII have been shown to have a higher frequency of intron loss but not overall shorter introns {Jiang, 2014 #3052}. One potential explanation of the regulatory rational that gives rise to intragenic repressor binding comes from the observation that repression arising from Snail binding in distant CRMs introduces a lag before repression becomes complete, due to the inability of the repressor to affect downstream from the promoter region {Bothma, 2011 #2304}. Due to the relatively slow rate of progression of PolII (~ 1.1 to 1.5 kb per min in *Drosophila)* {Ardehali, 2009 #3053}, this lag time can become significant, especially under developmental contexts in which temporal control of repression can be as important as activation. While microRNAs are known to dampen this effect in some contexts {Biemar, 2005 #3054}, Groucho-mediated repression initiated by binding within genes could potentially be another method to achieve a similar end.

## References