# Chapter X: Groucho activity in the developing embryo

## Abstract

Animal developmental patterning is an vastly complex and intricate process, requiring the integration of multiple temporally and spatially variant signals to define the transcription profile of each cell. Groucho, a transcriptional corepressor, plays a central role in this process in *Drosophila*. Groucho is thought to repress gene expression primarily through its recruitment to the genome through the action of multiple DNA binding factors. While Groucho is ubiquitously expressed in the early fly embryo, these factors are not, so Groucho mediated expression becomes dependent on the presence and concentration of these recruiting elements. Despite the broad importance of Groucho in fly development, a full picture of its regulatory network in the developing embryo has yet to be established. To this end, we have undertaken a multiomics approach to identify Groucho targets during three discrete stages of embryonic development. At each stage, we have analyzed the embryonic transcriptome of wild-type and Groucho mutant embryos. Additionally, we have utilized high-throughput sequencing of chromatin-associated RNAs (Nascent-seq) to confirm transcriptional rates at each timepoint. Groucho ChIP-seq provides information about the dynamics of the localization of Groucho to the chromatin in wild-type embryos. By combining these data sources, we established a temporally discrete high-confidence set of Groucho regulated genes, illuminating Groucho's multiple roles in developmental processes. Groucho appears to be involved in the regulation of hundreds of genes at each stage, a significant proportion of which are regulatory genes themselves, reinforcing the idea that Groucho is a highly-connected node or hub in the developmental regulatory network. Groucho's recruitment is highly dynamic, with a widespread transition in genomic localization occurring in 4 to 6.5 hours post embryo deposition. Human homologs of Groucho have been shown to act as both coactivators and corepressors, Drosophila Gro appears to be a dedicated repressor. While Groucho has long been considered to be a long-range corepressor, often binding thousands of base pairs away from its target genes, we find that this is not an obligate condition for repression, as actively repressing Groucho is often bound directly adjacent to transcription start sites. Furthermore, while Groucho is known to oligomerize *in vivo*, the role and necessity of this oligomerization in repression remains unknown. We find that, while Groucho can bind over large portions of the genome, possibly through self-association, most binding events are more spatially constricted, though clustering of peaks is apparent at multiple genomic loci. Around some classes of regulated genes, Groucho exhibits binding to both distal and proximal regions, perhaps indicative of Groucho serving a role in bringing distant portions of the genome in close proximity.

## Introduction

Groucho 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 a number of sequence-specific transcription factors, Groucho is crucial to the spatial and temporal restriction of gene expression beginning very early in development, and remains involved in tissue differentiation and specification throughout larval and pupal stages. As Groucho is maternally deposited, high levels of the transcript are present from the onset of development, allowing Groucho to play a role as one of the earliest available transcription factors in the embryo. Groucho is one of a vital group of factors that enables a cell to alter its transcriptional profile, and therefore developmental fate, in response to informational signals arising outside of the cell. This regulatory role of Groucho homologues is conserved throughout metazoans. Due to Groucho’s highly-connected position in the developmental regulatory network, mutations in the function or regulation of Groucho can result in profound developmental abnormalities and disease.

Despite the extensively documented centrality of Groucho to multiple developmental processes, especially in early embryonic development, no systematic genome wide investigation has been undertaken to position Groucho in the fly developmental regulatory network. A more thorough understanding of the recruitment patterns of Groucho in the early embryo, and the dynamics of such binding, will allow us to address multiple facets of Groucho-mediated repression that have eluded thorough investigation. While Groucho has been shown to tetramerize *in vitro,* and that in some contexts Groucho oligomerization is necessary for repression *in vitro* {Chen, 1998 #267} and *in vivo* {Song, 2004 #1161}, this does not appear to be a universal requirement for repressive activity in embryos {Jennings, 2007 #2990}. . Some evidence from ChIP experiments suggests Groucho spreads over potentially long stretches of chromatin presumably through its ability to self-associate {Winkler, 2010 #2964}, though these studies are limited by the resolution of the ChIP regions analyzed. More recent Groucho 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. It is unclear to what degree that binding pattern extends to embryos.

Accordingly, multiple, non-mutually exclusive mechanisms of Groucho-mediated repression have been proposed, both independent and dependent on Groucho oligomerization {Turki-Judeh, 2012 #2385}. Genome-wide analysis of binding patterns in embryos will enable us to analyze the recruitment strategies of Groucho *in vivo* during the developmental stages when Groucho is active and crucial to generating viable patterns of gene expression, specifically the 1.5 to 9 hours of development post-fertilization. Analysis of the size and distribution of Groucho peaks, and how that distribution evolves over time, will enable a thorough investigation of whether Groucho-mediated repression arises from spreading over long stretches of chromatin, and to what extent it do so. Analysis of Groucho recruitment to several known Groucho targets will enable us to explore how Groucho activity at those targets evolves over time, as the factor is bound and released from regulatory elements both within and adjacent to such targets. On a genome-wide scale, we can make determinations as to whether Groucho functions primarily as a long-range repressor, as is generally thought, or whether it is associated with promotors, transcription start sites, and gene bodies. Analysis of motifs different classes of Groucho peaks will allow us to predict preference for binding partners under specific situations, and potentially predict factors Groucho preferentially colocalizes with.

Additionally, integration of Gro ChIP-seq data with transcriptome analyses of embryos in which Gro activity has been perturbed will enable us to more accurately ascribe Groucho repression arising from binding regions to individual gene targets. While ChIP-seq provides a genome-wide picture of how Groucho associates with genes and intergenic regions, as well as how these dynamics change over time, it does not provide evidence that Groucho is actively involved in regulating the expression of those genes it is binding within or adjacent to. The accurate assignment of a 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 (a factor’s cistrome) with transcriptome measurements in systems perturbed for said factor. To this end, we have generated transcriptome measurements (via RNA-seq) of Groucho deficient embryos collected at paired timepoints to the ChIP-seq data as well as embryos overexpressing Groucho at various levels. Comparison of gene expression levels under these conditions to expression under endogenous levels of Gro activity will enable us to predict Groucho regulatory targets. When combined with ChIP-seq binding profile data, we can produce a high-confidence set of Groucho target genes by timepoint. This will enable us to more thoroughly explore the role of Groucho within these three stages of development through a broader investigation of which components of the developmentally-regulated gene network it plays a role.

## Materials & Methods

1. *Groucho chromatin immunoprecipation (ChIP) and sequencing*
2. *Groucho ChIP-seq data analysis*
3. *Embryonic RNA isolation and sequencing (RNA-seq)*
4. *RNA-seq data analysis*

## Results

*Groucho is recruited ubiquitously and dynamically throughout the genome*

The timepoints used for the analysis were chosen to overlap significant events in embryonic development that have known Groucho interactions. The first timepoint (1.5 – 4 hours post deposition) 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. The second timepoint (4 – 6.5 hours) encompasses the growth and segmentation of the germ band, including the formation of neuroblasts. Timepoint 3 (6.5 – 9 hours) 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. We used an affinity purified antibody raised against the Gro GP domain, which we validated extensively in immunoprecipitation and immunoblot assays. Sequencing libraries were sequenced to a depth that provided at minimum 5M uniquely mappable reads, far in excess of the minimum recommended by modENCODE ChIP-seq best-practices (Fig. 2-a//A) {Landt, 2012 #308}. Replicates exhibited high reproducibility in terms of both read density and resulting peak model (Fig. 2-a//B, left and right, respectively). The high degree of correlation between our ChIP-seq data set and a ChIP-chip data set obtained from 0-12 hour embryos{Negre, 2011 #3035} using completely independent antibodies also validates our ChIP-seq data (Fig. 2-o). The modENCODE Groucho peaks were obtained from 0 – 12 hour embryos and so represent a time-averaged superset of our data. Despite each of our datasets comprising only 21% of the development time represented by the modENCODE data, peak overlap is significant. The greatest degree of similiatiry is seen when comparing the 6.5 – 9 hour data set, which captures 68% of all modENCODE-identified peaks. Additional comparison with modENCODE Groucho ChIP-chip data generated from white pre-pupae indicates that a significant fraction of the Groucho-regulated sites in embryos continue to be bound in this later developmental stage (Fig. 2-p). Though the utilization of a large fraction of Groucho binding sites appears to be restricted to either embryonic or pupal stages.

Peak modeling identified widespread Groucho binding throughout the genome; peaks with overlapping regions between replicates were chosen for further analysis (Fig. 2-b). Groucho recruitment sites appear 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 at 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-c).

Groucho occupancy is highly dynamic and reversible. Approximately 50% of all Groucho binding sites are unique to a single timepoint. The majority of the sites established during timepoint 1 that persist into timepoint 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.

*Groucho associates with multiple ventrally-repressed genes in patterns unique to early developmental stages*

In the early embryo, delineation of the dorsal-ventral axis is accomplished through establishment of a gradient of nuclear Dorsal (dl) along the axis {Roth, 1989 #1112}. In ventral and ventrolateral regions of the embryo, Dorsal facilitates the repression of *zen* and *dpp* through its interaction with Groucho {Dubnicoff, 1997 #2366}. Ventral repression of *zen* is established through the Dorsal recruitment to a so-called 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 recruited to establish repression. ChIP-seq data confirms that Gro localizes the the VRR, however Gro density is comparatively weak within the VRR region (Fig. 2-q//A). Stronger Gro signal is seen both directly upstream of the VRR (compared to *zen*), as well as downstream. This downstream region overlaps the TSS of *zen* and extends an additional 700 bp upstream of the gene. Following the 1.5 – 4 hour timespan, Groucho recruitment to the *zen* locus shifts to a pair of peaks centered approximately 500 bp downstream of *zen*, and all upstream binding is lost. As no *zen* regulatory regions have been characterized 3’ of the gene, it is now known if these regions represent actively repressive Groucho. Though the mechanism has not been extensively documented, *zen* expression contracts in stage 5 (2 – 3 hr) embryos, and becomes permanently repressed afterwards.

Additionally, Dorsal is responsible for ventral repression of decapentaplegic (dpp) in early embryos (1.5 – 2 hours post fertilization) through the recruitment of Gro. 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 were mapped to the second intron {Huang, 1993 #3037}. Our ChIP-seq data confirms extensive Gro recruitment to this site (Fig. 2-q//B), indicating also that Groucho is lost from this locus at later timepoints, consistent with the finding that in later embryonic development multiple elements upstream of *dpp* confine its expression into multiple lateral stripes along the extending germ band {Schwyter, 1995 #3038}. Extensive binding of Gro within these regions suggests that it also plays a role in *dpp* expression resulting from these elements. Additionally, *dpp* plays a crucial role in correct anterior-posterio patterning of limb imaginal discs post-embryonic development {Spencer, 1982 #3039}. Expression of *dpp* in imaginal discs is controlled by an extensive 3’ cis-regulatory region containing multiple Pangolin/TCF and Brinker binding sites {Blackman, 1991 #3040}{Theisen, 2007 #3041}. While this enhancer region is not known to play a role in *dpp* expression during embryogenesis, Groucho does bind both immediately downstream of *dpp* and within the second of three Pangolin binding sites between 4 to 9 hours of development, indicating that these sites may potentially play a role earlier in development than previously hypothesized.

Dorsal is also necessary for the ventral activation of a number of genes, a process that is thought to be Groucho-independent. These ventrally-activated genes include rhomboid (rho), single-minded (sim), and short gastrulation (sog) {Ip, 1992 #3042}{Gonzalez-Crespo, 1993 #3043}. Loss of Gro activity was shown to result in restricted expression of these genes in 1.5 – 2 hour embryos, but did not result in significant change in expression pattern along the dorsoventral axis, so it is hypothesized that Gro is not involved in Dorsal-mediated activation of these genes.

ChIP-seq data reveals, however, that Groucho potentially plays a direct role in regulating expression of these genes in some portions of the embryo. A significant Gro peak overlaps a regulatory region termed the neuroectoderm element (NEE) in 1.5 – 4 hour embryos. The area contains multiple Dorsal, Twist, and Snail binding sites, all of which are restriction of rhomboid expression to the presumptive neuroectoderm {Ip, 1992 #3042}. The Gro peak shifts towards an adjacent CRM termed the midline element (MLE) during 4 – 6.5 hours post fertilization. It is unknown which factors bind to the MLE and would be responsible for recruitment. At both timepoints, recruitment of Groucho to the regulatory region is associated with additional binding at the TSS of *rho*. This may represent a looping of the enhancer region over the 1.5 – 2 kb intervening sequence, which is depleted for Gro, potentially indicative of a repressive mechanism whereby Gro interacts with or blocks assembly of the primary transcriptional machinery.

*Groucho-associated genes tend to be associated with multiple, small clustered peaks*

Groucho binding regions are associated with fewer genes than there are binding sites (Fig. 2-d//A), due to the tendency of Groucho to localize to multiple regions around its potential targets. Half of all Groucho-associated genes have two or more Groucho peaks in relative proximity (Fig. 2-d//B), with an average of 2.5 binding sites per associated gene. These peaks have median widths in the 500 – 700 bp range, indicative of point source peaks, as commonly seen for sequence-specific transcription factors, rather than the broad peaks typical of polymeric factors or histone marks. However, 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 of Groucho binding, it may be capable of spreading over relatively large regions of the genome, but this does not appear to be a common mode of chromatin association. Average Groucho peak widths slightly increase 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 in promoter regions within 500bp of transcription start sites (Fig. 2-e//A). Regions further upstream (20 kb to 2 kb) are depleted for Gro occupancy, and intermediate range regions (2 kb to 500 bp) show neither depletion nor enrichment. Binding within introns and 5’ UTRs is enriched. This pattern of occupancy is at odds with the traditional view that Groucho is primarily a long-range corepressor {Dubnicoff, 1997 #2366} {Barolo, 1997 #2365}, capable of repressing genes several kilobases away from its recruitment site, though it has also been shown to be capable of short-range repression through recruitment by additional transcription factors, such as knirps {Payankaulam, 2009 #2955}. Our data indicates that it preferentially associates near TSSs and within genes. Groucho sites exhibit a strong preference for binding within genes, with approximately 50% of peaks occurring within gene bodies across all timepoints (Fig. 2-e//B). A small fraction of binding overlaps start sites (3 – 10%), and intergenic binding is nearly evenly split between binding upstream and downstream of the nearest feature. Due to the complexity of enhancer-gene interactions, it is difficult to assign these intergenic binding sites accurately to a regulatory target, if such a target even exists for each peak.

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-g//A). 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 {Bradnam, 2008 #3034}.

However, it is not known whether Groucho peaks arising inside genes are the result of direct recruitment of Groucho to those sites by interaction with sequence-specific transcription factors, or if Groucho is recruited to intergenic regulatory regions which are brought into contact with these introns through looping or a similar rearrangement of the local topology, resulting in immunoprecipation of these regions during ChIP-seq. However, it does appear that the interaction of Groucho with these introns is specific due to the restricted size of the intronic binding sites. The median width of intronic sites is identical to intergenic sites (403 vs 402 bp, respectively). Therefore, some property of the intron is directing Groucho to associate with specific points within. Whether that property is a protein interaction with another intron-associated factor, or a topological property of the surround chromatin, is not known.

Motif analysis of the intronic Groucho recruitment sites identifies a small number of transcription factor biding motifs enriched at each timepoint, including some know Groucho-associated factors, including twist (twi), tinman (tin), and pannier (pnr), and previously unassociated factors, including vielfältig (vlf)

*Identification of Groucho Targets by Developmental Stage*

In order 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). Additionally, we analyzed the transcriptome of embryos lacking maternally-contributed functional Groucho. These embryos are derived from female germline clones of a Gro allele that give rise to a severely truncated and therefore non-functional Gro gene product. 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-m//A). Groucho loss-of-function embryos failed to accumulate Gro transcripts to any significant degree across all timepoints. Wild-type embryos exhibit the normal pattern of initially high levels of maternally-deposited transcript, which are graduadly reduced as development proceeds (Fig. 2-m//B).

Clustering of RNA-seq samples reveals the transcriptomes cluster by timepoint, then by Groucho dosage (Fig 2-l). 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 together, independently from all other samples at those two timepoints, indicating that accumulation of differences in gene regulation has put these embryos on a unique and non-viable developmental trajectory (Fig. 2-l, 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-n). 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 cluster independently at all timepoints, exhibiting as 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-k//A). The Groucho loss-of-function phenotype was more severe than that obtained from overexpression, with over 10% of genes exhibiting changes in expression level at each timepoint, with the greatest effect seen in the second, 4 to 6.5 hour stage (Fig. 2-k//B). Overexpression samples exhibit a smaller yet still significant proportion of differentially expressed genes, 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 across samples, with this effect holding across all timepoints (Fig. 2-j). 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 constrict the list of potential Groucho targets in two ways. First, we focus on genes that exhibit a response of an opposite magnitude in the loss-of-function and one and more 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-s). Secondly, we narrowed this list to only those genes associated with 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 2x and 4x Gro overexpression lines (Fig 2-t & Supplemental Table 1).

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}.

The significantly enriched gene ontology groups for predicted Groucho-repressed genes (n = 162) contain several groups indicative of transcription regulation (GO:0006355, n = 37) and developmental processes (GO:0032502, n = 81). (*Fig: GO groups)* Of the 146 predicted Groucho-activated genes, no gene ontologies were significantly enriched, leading us to hypothesize that these genes are potentially the result of random noise in the gene expression data and do not represent direct Groucho targets.

*Groucho-regulated genes are enriched for stalled RNA polymerase*

Promotor-proximal pausing of RNA Polymerase II has been identified as a crucial step in gene regulation. Pausing has been primarily characterized in *Drosophila* at multiple heat-shock genes, presumably to facilitate rapid induction of gene expression upon receipt of an appropriate regulatory signal{Lis:1993uk}. Since this discovery, polymerase stalling has been found to be a ubiquitous regulatory mechanism{Conaway:2000un}, with strong peaks of PolII present in the promoter regions of a diverse array of genes throughout the *Droosphila* genome.

To explore whether Groucho regulation potentially promotes the stalling of polymerase, we undertook to compare Groucho-regulated genes with publically available genome-wide PolII localization data. 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. Comparing genes exhibiting change in expression levels under Groucho loss-of-function conditions, we see a strong correlation between genes repressed by Groucho and PolII pausing ( 179 genes, *p* < 10-20), and limited correlation between genes activated by Groucho exhibiting pausing (68 genes, *p* > 0.05). Conversely, genes activated by Groucho are enriched for active PolII (315 genes, *p <* 10-20), while Gro repressed genes are not (174 genes, *p* > 0.01). Together, this provides strong evidence that, at least at early timepoints, a significant fraction of Groucho-associated genes exhibit characteristics of PolII pausing. The retention or prevention of PolII from transitioning to an active complex is a potential mechanism of Groucho-dependent repression.

## Discussion