**Figure 2-a. 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, 2012 #308}. **(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

//PB

**Figure 2-b. Peaks present in both replicates were preserved for further analysis** 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 in its matchign replicate. Samples exhibited a replicate overlap rate of approximately 35% of all peaks, for the first and third timepoint, and 61% for the middle timepoint.

//PB

**Figure 2-c. A large number of Groucho binding regions are preserved throughout all stages analyzed.** **(A)** While a significant fraction of Groucho binding sites are unique to each of the three timepoints analyzed, several 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. **(B)** Approximately 1500 peaks were called in both anti-Gro and input libraries. These are most likely the result of DNA isolation, library preparation, or read mapping effects and were removed from further analyses.

TODO: Remake (A) venn, removing Input overlapping peaks

//PB

**Figure 2-d. A larger number of genes are associated with a smaller grouping of genes at each timepoint.** The most probably Groucho regulated gene associated with each binding region was identified by identifying the closest TSS site to that gene. This method has been shown to be instructive in predicting TF/gene interactions, though due to the complexity of the Drosophila genome and both the numerousness and distant location of enhancer regions, this technique is prone to high type I and II error.

//PB

**Figure 2-f. Groucho is preferentially recruited to gene bodies at all timepoints.** 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.

TODO: Cut 2-f or remove 2-e//B

//PB

**Figure 2-g. 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.

//PB

**Figure 2-h. Average Groucho peak widths suggest spreading is an 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.

//PB

**Figure 2-i. Groucho-regulated genes are significantly enriched for Groucho binding within introns.** Genes that become differentially expressed in Groucho LoF or overexpressing embryos are specifically enriched for Groucho binding within introns. This effect is most strongly seen in the 2x and 4x overexpression lines, in which genes that become less transcribed are enriched 4- to 5-fold for Groucho binding compared to genes exhibiting no response.

//PB

**Figure 2-j. The three Groucho overexpression lines show similar patterns of altered gene expression, though significant differences in the magnitude of gene expression changes are evident.**

TODO: This should go in an other section.

//PB //PB

**Figure 2-k. 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.

//PB

**Figure 2-l. Clustering of embryonic transcriptomes across Gro levels and timepoints and between replicates.** Pair-wise Spearman correlation coeffients 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 as well as potentially stalling of any further developmental change. Abbreviations: OE1 - 2x Gro overexpression, OE2 - 4x Gro overexpression, dGro - Gro maternal loss-of-function, dSP - Gro

//PB

**Figure 2-m. 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, 2011 #3044}, 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.

//PB

**Figure 2-n. 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, where the expression level of each gene constitutes a dimension. It can be conceptualized as a simplifying rotation and scaling of each data-set (represented as a point in ~15,000 dimensional space) such that all points fall onto a two-dimensional plane. 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 timepoint, while the y-axis captures to 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.

//PB

**Figure 2-o. 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.

//PB

**Figure 2-p. A small subset of early-embryo Groucho-bound sites are utilized later in development.** 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, 2001 #3033}.

//PB

**Figure 2-q. Groucho binds to the two 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, 1998 #3036}. 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 similarily to *zen* is represssed ventrally by Groucho and Dorsal activity {Dubnicoff, 1997 #2366} 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.

//PB

**Figure 2-r. 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, 1997 #2366}. **(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.