* rnasesq - 1 - Replicate similarity heatmap.pdf
  + Summary: A heatmap of replicate similarity, including all RNA-seq datasets
  + Method
    - Replicate similarity statistic is a Spearman coefficient; values range from 0 to 1 (indicating most similar)
    - In RNA-seq experiments, you generally want a coefficient of > 0.9 for biological replicates
    - Clustering of samples was accomplished using agglomerative hierarchical clustering, whereby each sample begins in it's own cluster. Clusters are subsequently merged in a "step-up" fashion, where the most similar samples are merged at each step. Clustering is indicated by the ordering of the symmetric x- and y-axes and the dendrogram.
  + Conclusions
    - Samples cluster strongly by timepoint, and more weakly by condition.
    - The three overexpression lines cluster together by timepoint, though replicates do not cluster together. The most likely explanation is that that the technical and/or random variance between replicates is larger than biological variance between conditions.
    - I wouldn't take this as evidence that the three-overexpression conditions are identical. It's very possible that a a large number of small, random fluctuations in gene expression levels is disproportionately affecting the correlation statistic.
    - The Gro LoF samples strongly cluster independently from the overexpression and wild-type lines, though also primarily by timepoint.
      * The 1.5-4 hr Gro LoF lines still cluster with the other 1.5-4 hr samples, though they are more similar to the later timepoint Gro LoF datasets than to any of the 1.5-4 hr overexpression or wild-type data.
      * The later timepoint Gro Lof data sets cluster together, with an intermediate similarity to both the 4-6.5 hr and 6.5-9 hr data. They are both more similar to each other than to any other data sets.
        + This could indicate that large perturbations in gene expression of these embryos has started to stall their development (we don't see the large changes in transcription program we see when moving from the middle to late timepoint in the wild-type and overexpression lines)
* rnasesq - 2 - Replicate similarity PCA.pdf
  + Summary: Principal component analysis of RNA-seq data sets reveals similarity by sample and timepoint
  + Method
    - Principal component analysis was generated from the transcription profiles of each sample
    - The shortest distance between two points is a measure of their similarity, and should be roughly proportional to the difference in the Spearman coefficient of the two samples (the difference in intensity of each block in the heatmap)
  + Conclusions
    - PCA is potentially more informative than a heatmap (though it does not including any clustering analysis, that being difficult to represent with 2D data)
    - The axes have no pre-determined meaning. They capture the two axes of greatest variability between all the datasets.
    - In this case, the x-axis seems to capture much of the variability between timepoints
    - The y-axis captures variability based on condition
      * Gro LoF samples cluster near the bottom, wild-type samples in the middle, and the overexpression lines near the top
* rnasesq - 3A/B/C - Differential expression scatterplot - timepoint 1/2/3.png
  + Summary: A plot of the significant log2 fold-changes in gene expression between all samples at each timepoint
  + Method
    - For each gene that is significantly differentially expressed in both datasets being compared is plotted
      * This ignores the genes that are significant in one condition but not the other
      * Significance defined as having a log2 fold-change of greater 0.5 in magnitude and an adjusted p-value of < 0.01)
  + Conclusions
    - Significant changes in gene expression among the overexpression lines are largely similar (log2 fold-changes are linear)
      * Latter timepoint samples exhibit higher variance
    - Significant changes in gene expression between the Gro LoF samples compared with overexpression lines are more heterogeneous
      * Genes exhibiting the same magnitude of response are more dissimilar in magnitude (a fanning out of points in the upper-right and lower-left quadrants)
      * A large number of genes show a response opposite in magnitude between conditions
        + There is a potentially interesting shift in pattern following the progression of timepoints

At early timepoints, there are a greater number of genes repressed by Gro LoF and activated by Gro OE than the opposite. This trend is reversed by the last timepoint (more genes are both activated by Gro LoF and repressed by OE than vice-versa)

* rnasesq - 4 - Groucho expression timecourse.pdf
  + Summary: Groucho transcript levels change during each timepoint
  + Method
    - Normalized Groucho transcript levels are plotted versus timepoint
    - I need to regenerate this graph with error bars
  + Conclusions
    - Groucho levels match expectations
      * Overexpression lines show high levels initially, with excess transcript being largely cleared by the second and third timepoints.
      * Gro Lof are lower than wild-type through all timepoints. The transcript is potentially being eliminated by nonsense mediated decay(?).
* nascent - 1 - Replicate similarity heatmap with nascent.pdf
  + Summary: Same as the previous heatmap, but includes nascent-seq data
  + Method
  + Conclusions
    - The nascent-seq data clusters independently from all other data sets, indicating that all nascent datasets are more similar to each other than to any RNA-seq datasets
* nascent - 1B - Replicate similarity PCA.pdf
  + Summary: Principal component analysis of nascent-seq data with RNA-seq data
  + Method
  + Conclusions
    - Unlike the previous analysis, the x-axis here is capturing "nascent-ness", with the nascent -seq data sets far on the right, compressing all RNA-seq samples together.
    - THe y-axis is capturing differences based on timepoint. Both RNA-seq and nascent-seq data sets segregate appropriately based on timepoint.
    - No axis separates samples based on condition (wild-type/overexpression/loss-of-function)
* nascent - 2 - Top expression deviating genes heatmap.pdf
  + Summary: A heatmap of deviation from mean expression of the top 50 genes with the largest difference between wild-type and nascent expression profiles
  + Method
    - Average expression is calculated across all timepoints and samples
    - Yellow represents above average expression, blue represents below average
    - Significant difference is calculated with DESeq2. The statistical underpinnings are the same as a more straightforward differential expression test (i.e. overexpression vs wild-type), but the compared conditions take into account all three timepoints, so no meaningful fold-change is calculated
  + Conclusions
    - Multiple clusters of expression profiles
      * Top cluster are genes highly expressed at the first timepoint in both samples, which continue to be above-average in the nascent samples at the middle timepoint.
      * Bottom cluster are genes highly expressed at all timepoints, but which do not accumulate to significant levels. These could represent mRNAs with high turnover rates.
* nascent - 3A/B/C - pair\_plot\_1.5/4.0/6.5.pdf
  + Summary: A pair plot of normalized expression levels between genes in wild-type and nascent datasets at each timepoint
  + Method
    - Identical data is plotted in scatter plots in the lower-left and density graphs in the upper-right
  + Conclusions
    - Density graphs are useful here as the scatter plots have significant overplotting
    - Replicates show relatively good correspondance; nascent samples have higher variance; many of the outliers in the nascent samples are rRNA
    - Two clusters are apparent from the density graph. A region of genes with zero or very little expression, and a more highly-expressed region
    - Comparing total mRNA with nascent mRNA, many genes exhibit higher nascent counts than total (fanning of the plots upwards). These possibly represent genees that are actively expressed at each timepoint, but have not reached a steady-state
* multiple - 1A/B/C/D - \* vs Gro ChIP.pdf
  + Summary: Venn diagrams of overlap between Groucho-associated and differentially expressed genes
  + Method
    - The nearest genes to each Gro bind region were identified from the chip-seq data
    - Overlap between these gene sets were compared to differentially expressed genes from the RNA-seq data
    - The p-value represents the probability of having an identical or greater overlap if both sets of genes were chosen at random (Fisher's Exact Test)
  + Conclusions
    - Most overlaps are significant
    - Overexpression overlaps are highly significant at all timepoints
    - Gro LoF overlaps are less significant, most likely due to the fact that many more genes are differentially expressed in these datasets, so the probability of arriving at the same degree of overlap is overall higher
* multiple - 2A/B - RNAseq Down/Up Regulated vs ChIP.pdf
  + Summary: The same data represented as bar plots, but separated by significantly activated and repressed genes
  + Method
    - Percent overlap represents the percentage of Groucho-associated genes (from the chip-seq data) that are either up or down regulated at each timepoint
  + Conclusions
    - There's little difference between the overlap of up-regulated genes and Gro-associated genes in the different samples
    - Looking at down-regulated genes only, a greater percentage of such genes are associated with Gro binding in the loss-of-function data than the overexpression data. This is expected if these are in fact Groucho targets
* multiple - 3 - 1.5-4hr regulatory network.png
  + Summary: A network of Groucho direct and indirect targets generated from the chip-seq and rna-seq data
  + Method
    - Generated using a regulatory inference software ChIP-Array
    - Yellow circles are direct targets (appear as differentially expressed in rna-seq data and have nearby binding regions in the chip-seq data)
    - Pink circles are direct targets (as above) that are also annotated transcription factors
    - Grey circles are putative indirect targets; meaning they are differentially expressed, but not near Groucho binding regions. And are known targets of the other regulators shown (pink circles)
    - Jaspar was used for the database of transcription factors and their annotated targets
  + Conclusions
    - The software picked up a large number of genes that are Groucho regulated but not associated with other transcription factors. Presumably a limitation of the database.
    - It's interesting which secondary TFs it picks up, and which it doesn't. Dorsal is absent, but twist is represented.
    - I need to look into the TFs Groucho is regulating, as I'm not familiar with several. I should check if the transcription factors are activators or repressors, how they are differentially expressed, and see if that correlates with changes in secondary target expression.