### Chapter 3: Investigating the dynamics of the embryonic transcriptome

#### Abstract

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

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

#### Introduction

*Drosophila* development involves the coordinated expression of a vast number of genes under strict temporal and spatial control {Brown, 2014 #3144}. Transcript levels in the embryo are therefore highly dynamic, undergoing rapid shifts in level dependent on altered rates of accumulation, processing, and degradation. In excess of ~7,000 unique transcripts, arising from at least 3,000 distinct genes are expressed from the earliest onset of zygotic transcription {Graveley, 2011 #3044} {De Renzis, 2007 #3149}. Of these, at least 1,000 are expressed in a spatially restricted manner {Tomancak, 2002 #1150}. Regulatory systems go to great lengths to minimize even relatively slight stochastic changes in expression, which can nevertheless have a negative effect on viability {Perry, 2010 #385}. Additional processes influencing mRNA abundance are also tightly controlled, with regulated degradation pathways resulting in significant and transcript-specific differences in the kinetics of mRNA decay in the embryo {Thomsen, 2010 #1350}. Ideally, accurate and quantitative measurements of protein abundance could be utilized to investigate the mechanics of development, but lacking such a technique, mRNA abundance is often substituted as a measurement providing a strong correlation with protein abundance {Fu, 2009 #3150}.

A complicating factor in the interpretation of transcriptome data from the early *Drosophila* embryo arises from the significant effects of maternally-contributed mRNAs and the timeline of activation of the zygotic genome. This latter process, the maternal-to-zygotic transition (MZT) is a common feature of animal development and encompasses a cascade of processes whereby widespread alterations to the chromatin landscape are engendered by the activity of pioneering transcription factors {Tadros, 2009 #3146} {Li, 2014 #3151}. In *Drosophila,* the zinc-finger transcription factor Zelda is a well-studied example of such a pioneering factor {Liang, 2008 #3153} {Harrison, 2011 #3152} {Xu, 2014 #1928}. In most animals, the MZT process consists of two distinct stages. An initial “minor wave” of activation becomes significant at ~1.5 hours post-fertilization in *Drosophila*, followed by a more rapid and synchronous “major wave” at ~2.5 hours {Pritchard, 1996 #3147}. During this time the majority of the maternally-contributed transcriptome is destabilized and undergoes coordinated degradation {Tadros, 2007 #3155;Benoit, 2009 #3154}.

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

Nascent-seq has been successfully applied to track the transcriptional changes in a number of biological contexts, including macrophages (Bhatt et al., 2012), where it was utilized to obtain a timeline of transcriptional changes following induction of an immune response, adult *D. melanogaster* tissues to analyze the prevalence of cotranscriptional splicing {Khodor, 2011 #2081}, and circadian transcript cycling {Rodriguez, 2013 #1782}, in which the authors saw significant differences in total mRNA and nascent mRNA levels between non-overlapping ninety minute embryo collections. We have adopted the method to developing embryos, using an established protocol for embryo nuclei isolation {Nechaev, 2010 #3006} followed by isolation of a chromatin-associated fraction from these nuclei. Purification of RNA from the chromatin fraction yields a RNA pool significantly enriched for nascent RNA.

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

#### Materials & Methods

*Chromatin-associated RNA isolation in embryos*

Wild-type fly embryos were collected in three 2.5 hour cohorts beginning 1.5 hours post-deposition. Between 3 to 5 grams of embryos were utilized for each fractionation. The chromatin-associated RNA isolation protocol was adapted from Nechaev et al. (2010) and Khodor et al. (2011). Embryos were dechorionated in 50% bleach for 90 sec and transferred to a chilled Dounce homogenizer. Embryos were then rinsed three times with 25 ml of homogenization buffer (15 mM HEPES-KOH pH 7.6; 10 mM KCl; 3 mM CaCl2; 2 mM MgCl2; 0.1% Triton X-100; 1 mM DTT; 0.1 mM PMSF; 0.1x RNAase inhibitor). Embryos were then suspended in homogenization buffer containing 0.3 M (15 ml) sucrose and dounced five times each with loose and tight pestles. Embryo lysate was filtered through 50-micron nylon cell strainer. Clarified lysate was layered over a sucrose cushion consisting of a layer of 1.7 M sucrose (15 ml) underneath a layer of 3 M sucrose (15 ml) in homogenization buffer.. The samples were centrifuged at 15,000 RCF for 10 min at 4˚C. Pelleted nuclei were resuspended in 250 µl of nuclear lysis buffer (10 mM HEPES-KOH pH 7.6; 100 mM KCl; 0.1 mM EDTA; 10% glycerol; 0.15 mM spermine; 0.5 mM spermidine; 0.1 mM NaF; 0.1 mM Na3VO4; 0.1 mM ZnCl2; 1 mM DTT; 0.1 mM PMSF; 1x RNAase inhibitor). While gently vortexing, an equal volume of NUN buffer (25 mM HEPES-KOH pH 7.6; 300 mM NaCl; 1M urea; 1% NP-40; 1 mM DTT; 0.1 mM PMSF) was added drop-by-drop over a period 5 minutes. Condensed chromatin became visible as a fluffy white precipitate. The solution was then incubated for 20 min on ice and centrifuged at 14,000 rpm for 30 min at 4˚C. The supernatant (primarily nucleoplasm) was discarded and the pellet was resuspended in Trizol reagent (Qiagen). RNA was then purified following the manufacturer’s protocol.

*rRNA removal*

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

*RNA-seq library construction and sequencing*

rRNA-depleted RNA was concentrated via ethanol precipitation. Size distribution of samples was determined via Agilent 2100 Bioanalyzer (Agilent Technologies). Indexed RNA-seq libraries were generated with the ScriptSeq v2 RNA-seq Library Preparation Kit (Epicentre). Sequencing was performed on Illumia HiSeq 2000 sequencing platform (High Throughput Sequencing Core Facility, Broad Stem Cell Research Center, UCLA). Reads were demultiplex via custom scripts and mapped to the BDGP5/dm3 *D. melanogaster* genome with Tophat2 (v2.1.0) {Kim, 2013 #1817} using the following parameters: -g 1 –solexa1.3-quals. A gene model annotation (iGenomes UCSC dm3) was provided as a mapping guide. Assignment of mapped reads to transcripts was performed with HTSeq {IAnders, 2015 #3027}.

*Data analysis*

Mean normalized transcript expression levels (FPKM) were generated with DESeq2 (v1.10.0) {Love, 2014 #3031}. Significant changes in transcript abundance were quantified with the same software by comparison with poly(A)+ RNA-seq from wild-type embryo data described in Chapter 2 of this thesis. RNA-seq read mapping density analysis was performed using PicardTools (http://broadinstitute.github.io/picard/). Additional metagene analysis was performed using the ‘metagene’ package of R/Bioconductor {Beauparlant, 2014 #3167}.

#### 

#### Results

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

Total RNA was extracted from chromatin isolated from *D. melanogaster* embryos collected over three time spans in early development and subjected to high-throughput sequencing. This chromatin-associated RNA is expected to be enriched for nascent transcripts, as well as additional RNA species which associate with chromatin in structural, catalytic, or regulatory capacities {Cernilogar, 2011 #3156}. Isolated RNA was affinity-depleted for polyadenylated RNA in order to further minimize the contribution of mature mRNA from analysis. The level of enrichment for nascent transcript was validated and quantified through various measures. The efficiency of chromatin isolation was confirmed through analysis of protein compartmental markers (Fig. 3-1A/B), confirming that the sequenced RNA was derived from a embryonic fraction enriched for histones and deficient for a cytoplasmic marker. Sequencing reads obtained from mature transcripts ideally map uniformly across genes, though this is heavily dependent on the quality of the RNA utilized for library generation. Non-uniformity generally manifests as a 3’ bias in mappable reads, as partially fragmented mRNA is purified by affinity selection to polyadenylation sites {Roberts, 2011 #3157}. Chromatin-associated RNA exhibits a significant 5’ bias at each developmental stage, and is partially depleted at the 3’ end (Fig. 3-2). The large enrichment of reads arising from the initial 15% of gene bodies may be indicative of promoter-proximal paused polymerase. The sharp decrease in read occupancy near the 3’ terminus may result from frequent polymerase pausing in terminal exons {Carrillo Oesterreich, 2010 #3159}. Pausing in terminal exons is thought to promote recognition of polyadenylation sites and transcriptional termination {Gromak, 2006 #3166}.

Chromatin-associated RNA is enriched for intronic sequence when compared to poly(A)+ libraries prepared from the same developmental stages (Fig. 3-3). Our data indicates that on average, 13% of poly(A)+ RNA-seq reads map to constitutive introns compared to 35% of chromatin-associated RNA reads. While 60-70% of gene sequence in *D. melanogaster* is annotated as intronic, the large majority of introns are believed to be cotranscriptionally spliced, with only 16% of introns exhibiting little or no splicing {Khodor, 2011 #2081} {Wuarin, 1994 #1471}. Therefore, an intronic content of between 13 and 60% should be expected for a library enriched for pre-mRNA.

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

Analysis of nascent pre-mRNA levels in multiple contexts has shown that the rate of accumulation of a particular transcript can be strongly uncoupled from the rate of transcript synthesis, owing to differential rates of accumulation, processing, and degradation {Khodor, 2011 #2081} {Rodriguez, 2013 #1782}. In developmental contexts, a significant proportion of the transcriptome is far from steady-state. Comparison of chromatin-associated RNA transcript profiles with mRNA profiles obtained from the same timepoints by principal components analysis indicates significant differences, with the majority of expressed genes exhibiting some deviance in expression rate and accumulation level (Fig. 3-4). Samples continue to cluster by developmental time stage, but segregate first by degree of “nascentness.” Comparison of the normalized expression levels of each gene indicates that many genes exhibit comparable levels of expression in poly(A)+ and nascent samples (Fig 3-5). A small number of genes are significantly enriched in the nascent population, however, while showing very little accumulation in the poly(A)+ RNA-seq data. A large fraction of these correspond to non-polyadenylated RNAs, including histones, snRNAs, and snoRNAs. These non-polyadenylated RNA species were removed from further analysis.

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

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

Promoter-proximal pausing of RNA Polymerase II has been identified as a crucial step in gene regulation. Pausing was originally characterized in *Drosophila* at multiple heat-shock genes, presumably to facilitate rapid induction of gene expression upon receipt of an appropriate regulatory signal {Lis, 1993 #2380}. Since this discovery, polymerase stalling has been found to be a ubiquitous regulatory mechanism in higher eukaryotes {Conaway, 2000 #2381}, with strong peaks of PolII present in the promoter regions of a diverse array of genes throughout the *Drosophila* genome. Expression of the majority of protein-coding genes in humans is regulated to some degree after the initiation of transcription {Guenther, 2007 #3118}, as is a large fraction of the *Drosophila* developmental genome {Zeitlinger, 2007 #3010}.

To explore whether Groucho regulation potentially promotes the stalling of polymerase *in vivo*, we undertook to compare Groucho-regulated genes with publically available genome-wide PolII localization data {Zeitlinger, 2007 #3010}. In this data set, the authors classified each gene into one of several states including the lack of detected PolII, active (elongation phase) PolII, or stalled PolII. Genes found to bind Gro internally or in adjacent intergenic regions were found to be significantly enriched for stalled PolII at each timepoint (Fig. 3-7A). We observe a significant correlation between PolII pausing within genes and those genes becoming up-regulated in Gro loss-of-function embryos (*p* < 10-10), while no significant correlation is observed between PolII pausing and down-regulation in Gro loss-of-function embryos (*p* > 0.05) (Fig. 2-7B). Conversely, down-regulated genes are enriched for active PolII (*p <* 10-10), while Gro repressed genes are not. Together, this provides evidence that, at least at early developmental timepoints, a significant fraction of Groucho-associated genes exhibit characteristics of PolII pausing, suggesting that retention or prevention of PolII from transitioning to an active elongation complex is a potential mechanism of Groucho-dependent repression.

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

Evidence presented in the previous section hypothesized that many Groucho-repressed genes possess significant levels of promoter-proximal stalled polymerase. Much of this was established using previously-published data from the 2 -4 hour *Toll10B* mutant embryos {Zeitlinger, 2007 #3010}. These embryos generate a more homogenous population of cells, as all portions of the embryo adopt a cell type representative of the presumptive mesoderm {Schneider, 1991 #3161}, and so simplified the embryo-wide classification of PolII pausing state. Our data allows us to quantify the accumulation of promoter-proximal nascent transcript at later stages of development, albeit in a more heterogeneous population of cell types. This heterogeneity limits the interpretation of Groucho’s involvement with promoter-proximal stalled polymerase, as we can determine whether a gene is regulated by Groucho and possesses stalled PolII at each developmental time span, but we cannot make definite conclusions as to whether those events are occurring in identical populations of cells. A correlation is still informative, as association of the two states potentially represents a program of regulation whereby Groucho either promotes stalling itself, or is recruited to repress genes that undergo stalling at the same developmental stage but in different tissues.

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

#### Discussion

Quantification of chromatin-associated pre-mRNA is a useful metric for the exploration of dynamic transcriptional systems such as the *Drosophila* embryo. The relatively high stability of the RNA Polymerase II ternary elongation complex facilitates the purification of nascent transcripts in a highly specific manner, thereby enabling us to more thoroughly characterize the dynamics of this transcriptional system and relate aspects of gene expression to the activity of Groucho. We observed that chromatin-associated pre-mRNA exhibits a modest 5’ bias throughout the gene body, with a density spike at the 5’ transcription start site likely corresponding to nascent transcript locked in stalled ternary complexes. Investigations of stalled PolII in the embryo have previously shown that in 2-4 hour embryos, 12% of all protein-coding genes have stalled promoter-proximal PolII {Zeitlinger, 2007 #3010}. Additionally, purification of chromatin-associated short RNA from *Drosophila* S2 cells predicted that 30% of protein-coding genes experienced some degree of PolII pausing {Nechaev, 2010 #3006}.

The manner in which PolII pausing is utilized to regulate transcription remains poorly understood, though multiple non-exclusive mechanisms have been proposed, (reviewed {Adelman, 2012 #3165}). One of these mechanisms posits that sustained or transient pausing facilitates the participation of additional regulatory elements in the determination of transcriptional activity {Nechaev, 2008 #3162}. This allows the expression level of a gene to be regulated through multiple, independent pathways, potentially at the behest of independent signaling pathways {Blau, 1996 #3163}. Combinatorial control of gene expression is a common regulatory motif in eukaryotes, so it is feasible that the capability to exert influence over expression both before the assembly of the PolII complex as well as after transcription has began would be useful in such scenarios. Members of the Rel family of transcription factors, of which the Groucho-interactor Dorsal is a member, have been found to promote both PolII pausing and release in mammals {Barboric, 2001 #3164}.

#### References