Title
Sequencing mRNA from cryosliced Drosophila embryos to screen genome-wide patterning changes

Author
Combs, Peter Acuña

Publication Date
2015

Peer reviewed|Thesis/dissertation

Sequencing mRNA from cryosliced *Drosophila* embryos to screen genome-wide patterning changes

by

Peter Acuña Combs

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Biophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael B. Eisen, Chair
Professor Steven Brenner
Professor Han Lim
Professor Lior Pachter

Spring 2015

Sequencing mRNA from cryosliced *Drosophila* embryos to screen genome-wide patterning changes

Abstract

Sequencing mRNA from cryosliced *Drosophila* embryos to screen genome-wide patterning changes

by

Peter Acuña Combs

Doctor of Philosophy in Biophysics

University of California, Berkeley

Professor Michael B. Eisen, Chair

A complex, spatially and temporally dynamic network of gene expression underlies proper metazoan development, yet methods have not previously existed to assay this network in an efficient, systematic manner. Brute-force application of *in situ* imaging has been used to painstakingly assay the network, but is impractical to apply to assay the complete gene expression network in multiple mutant conditions. Sequencing, in contrast, is fast and genome-wide, but is generally applied to homogenized tissues, discarding any spatial information.

In the first phase of my research, I have developed methods for performing mRNAseq to globally profile gene expression in dissected subsets of single *Drosophila* embryos. The patterns I measure are consistent between replicates, and also faithfully reproduce patterns already known from *in situ* hybridization experiments. However, the global nature of this experiment also reveals spatial patterning in many genes that have not been successfully assayed, even in relatively large-scale projects. Furthermore, I demonstrate that this can be used across samples from different developmental time points to produce a genome-wide developmental atlas of zygotic gene activation.

Because the small sample size of slices of single embryos limited the ability to use standard cDNA library preparation protocols, I next assayed a number of previously published protocols that were specifically designed for minute samples. Previous literature had not addressed whether these protocols would be suitable for reconstructing spatial patterns, and I was particularly concerned that pre-amplification steps would introduce unacceptable non-linearities in the data. Upon analyzing the results, however, I determined that all the protocols I tried were acceptable, and all approximately equally good. I also investigated a few modifications to one of the protocols that reduces the cost of library preparation such that it is no longer the primary limiting factor in terms of number of samples to be sequenced.

Finally, with these methods well refined, I have sliced and sequenced embryos with severe genetic perturbations to maternally provided factors at the head of the patterning network. Comparing embryos deficient in the key pioneer factor *zelda*, I have revealed that this lethal mutant is nevertheless able to correctly produce the majority of gene expression patterns

in the wild-type embryo. Embryos lacking the maternally provided activator *bicoid* show a greater loss of proper patterning, including the surprising result of ubiquitous over-expression of a number of genes. Simple models of local action of *bicoid* cannot easily account for this change. Surveying *bicoid* over-expression mutants and *hunchback* knock-downs has also yielded a large number of unexpected patterning changes. All of these changes recapitulate previously measured patterning changes, but also highlight new avenues to investigate.

If it says TACACACATATCCTCGT
Then you'll probably wish that you didn't know
— *Jonathan Coulton, "That Spells DNA"*

To my family—both the family I was born with, and especially the family I've chosen—who have supported me through all of this.

Contents

List of Figures

List of Tables

Acknowledgments

It gives me great pleasure to particularly acknowledge the advice, insight, reassurances, and assistance of the following kind people:

Aisha Ellahi, Kate Chase, Matt Davis, Aaron Hardin, Steven Kuntz, Susan Lott, Elizabeth Roeske, Devin Scannell, Kelly Schiabor, Josh Shaevitz, Holli Weld, Jackie Villalta.

Chapter 1

Introduction

One of the central goals of modern genetics is, given the genome sequence combined with some suitable set of initial conditions, to be able to predict the pattern of expression of genes over space and time. Being able to effectively "decode" the genome would offer a wide range of possible applications, from improved diagnostics to more reliably designing complex biosynthetic pathways. Even absent the myriad practical applications, this understanding (even within a single organism) would be a massive step forward towards understanding the full beauty of life.

There are good reasons to hope that this goal, while certainly ambitious, is achievable. First and foremost is the fact that complex organisms are able to faithfully reproduce themselves from only a single fertilized egg. This reliability means that the information encoded in the genome must be able to be read consistently with only the material present in the fertilized egg. While this is merely an "existence proof" that some kind of system is capable of making sense of genomic information, the recent pace of scientific discovery and development of new techniques has allowed us to probe cells with ever greater speed, breadth, and resolution. With each technological achievement, we thus step ever closer to understanding the complete genomic program.

Because rapid and heritable phenotypic changes are not common between generations, we can surmise that the information that describes the morphology, physiology, and behavior of an organism is encoded in a relatively stable form. The genome is certainly one of the forms that this information is stored in, but I would argue that the information is also stored in the laws of physics themselves. Much as a computer's software does not fully describe its operation, but relies on assumptions about how the machine will respond to its instructions, so too is the program of the genome interpreted by a physical machine—the organism itself—which is embedded in the real world[Citation needed].

The analytical tools and frameworks from physics have long provided an attractive lens to bring to bear on questions in biology in general, and in molecular genetics in particular. Throughout the development of molecular biology over the last three quarters of a century, physicists have provided numerous key insights, and I have taken particular inspiration from several of these figures. Max Delbrück contributed valuable insights and rigorous

mathematical formalism to an experiment that provided strong experimental evidence for random—rather than directed—mutation (Luria & Delbrück 1943). Even understanding the double-helical structure of DNA by erstwhile physicist Francis Crick (with data by physicists Rosalind Franklin and Maurice Wilkins, and some insight from biologist James Watson) suggested the mode of information transfer across successive generations (Watson & Crick 1953). In my own education, the work of Bill Bialek, Michael Elowitz, Uri Alon, and Alexander van Oudenaarden among many others (in addition to those who came from non-physics backgrounds) have shaped my thinking in terms of the kinds of questions from biology that physics is well suited to answer (See Gregor, Tank, *et al.* 2007; Elowitz *et al.* 2002; Ozbudak *et al.* 2002; Mangan & Alon 2003, as a few examples).

Even in the early embryo, physics provides a number of both limitations and implicit instructions that inform the patterning of development. The terms "limitations" and "instructions" imply that there is some goal that the cells are either aided or hindered in their quest to achieve. This amount of anthropomorphization is convenient, but ultimately one must recognize that it is a metaphor to aid *us* in understanding. No such intentionality is genuinely present in the system[1]. I will briefly describe two particularly relevant examples that highlight the importance of physics in understanding biology.

The formation of the ventral furrow is the first step of gastrulation. During this process, the blastula-stage embryo (a nearly hollow ball of cells) undergoes a relatively large scale involution, with the single-layered ball producing two layers, the primordia of the mesoderm and ectoderm (Sweeton *et al.* 1991; Campos-Ortega & Hartenstein 1997). The primary method by which this is accomplished is mediated through subtle differences in tension on the apical surface of the cells of the blastoderm (Polyakov *et al.* 2014). The tension builds until reaching a critical point, when gastrulation begins and proceeds as a single, smooth motion. While each cell does not have a concrete program of where to go or how to get there, the boundary conditions of the embryo impose a path that is consistent across the conditions under which the embryo develops. Thus, the proteins around the rim of the basal surface of the cell have effectively offloaded some of the program of their destinations onto the physics of the embryo.

While in the previous example, the embryo exploits physics to make the process of its rearrangement more elegant, there are also circumstances where physics makes problems harder than one might naïvely assume. Proper gene expression relies upon nuclei "reading out" the concentration of various internal and external factors, then responding to them by producing more factors. While our intuition of measuring the concentration of a molecule may be informed by a career in the lab, where we can simply plug our molecule into some device, the machinery inside a cell is not so lucky. Given the concentration of certain transcription factors in the developing embryo, and the amount of time that the embryo evidently takes to respond to those factors, it becomes clear that the accuracy of the measurement of the concentration is primarily limited by the fact that these factors are actually diffusing

[1]The question of what kinds of systems are capable of such intentionality is perhaps better left to the Philosophy department...

(Gregor, Tank, *et al.* 2007). While the extent to which gene regulatory systems are normally at the edges of expression fidelity is debatable (Krotov *et al.* 2014), there seems to be a non-trivial set of error-correcting machinery that is able to tolerate some amount of noise in the input.

Even in the relatively deterministic situation of developmental transcription, it is important to be aware of these limitations. It has recently been demonstrated that redundant enhancers are prevalent throughout the fly genome (Perry, Boettiger & Levine 2011). While the precise purpose of these conserved enhancers are still somewhat mysterious, one hypothesized function is that they allow for precise and robust expression of their target genes (Lagha *et al.* 2012). Effectively, they may compensate for noise, allowing nuclei multiple chances to "get it right". Alternatively, the very similar (yet not precisely identical) patterns might cooperate to yield the proper pattern (Barolo 2012).

The first draft of "the" human genome[2] was completed in 2000, and the entire genome officially completed in 2003. This was accompanied by a number of predictions that with the genome, we would be able to understand, and ultimately cure, a large number of human diseases. Despite these optimistic predictions, the last dozen years have not seen a massive increase in the number of genetically informed treatments available, and those that we have seen (for instance, a number of cancer treatments like Gleevec and BRCA testing) have largely built on work that was begun even before the availability of the complete genome[3]. Indeed, much ink has been spilled in the medical genetics field searching for the source of "missing heritability", the idea that known genetic markers explain little of the heritable variation of traits (Reviewed in Eichler *et al.* 2010).

This project yielded a number of surprises. Building on a long history of humans believing that they are categorically different from the rest of the animal kingdom, the predicted number of coding genes based on the relative organismal complexity was much higher than the complement of 20-25,000 that were found to be actually present (International Human Genome Sequencing Consortium 2004). Thus, much of the complexity actually present in the system[4] seems not to be present in the coding genes themselves. In fact, we shouldn't even necessarily expect that to be the case. Borrowing the "DNA as blueprint and parts list" analogy, in constructing a more complicated building, for the most part the same nails, rivets, and beams are used in an 5-story office building as in a skyscraper. Instead of having

[2]Obviously, every individual has a distinct genome (Even identical twins should have dozens of somatic mutations that distinguish them (Bruder *et al.* 2008; Roach *et al.* 2010)). On the one hand, this inter-individual variation is a major factor in the inheritance of variation. However, from another perspective, each of these changes is not so extreme that those carrying them are not able to form a relatively complete human being. Even many genetically inherited birth defects affect a small number of pathways (important though they may be).

[3]Surely some of this is due to a conservative regulatory system, but I am unconvinced that knowing a genome sequence will be especially predictive for most "interesting" health traits. Also, where's my hoverboard?

[4]I think it fair to say that humans are categorically different from all other animals, and will continue to be seen so until one of those animals writes a doctoral dissertation. How much that is due to complexity is both debatable, and well beyond the scope of this dissertation.

fundamentally different parts, the majority of the complexity lies in the arrangement. In biological systems, genes are largely conserved, but there is an increased complexity of when and where those genes are expressed.

Genome in hand, the next major phase in human genetics began with the ENCODE (**EN**cyclopedia **O**f **D**NA **E**lements) project, a large, multi-center collaboration to identify the "function"[5] of every segment of DNA in the human genome. Across many tissues, on the order of 80% of the genome is subject to some sort of differential activity, although the meaning of this activity is unclear. Even with the bulk of the data generation of that project complete, a general predictive understanding of gene regulation has not emerged.

Experimental System

It is more clear than ever that a well-characterized model system is crucial for understanding the function of these DNA elements. The fruit fly in particular is an attractive choice for many reasons. Having been studied in the lab for more than 100 years, we have a thorough understanding of the developmental genetics in numerous sub-systems. As one of the first animals to have its genome sequenced, gene annotations are effectively complete[6](Adams *et al.* 2000). It is also inexpensive and straightforward to raise large cultures, and work on invertebrates is unencumbered by IRB oversight concerns[7]. Nevertheless, it has a complex life cycle, and many differentiation processes are homologous to those in vertebrates (See L. H. Wu & Lengyel 1998; Muhr *et al.* 2001; Dush & Martin 1992; Panganiban *et al.* 1997, for just a few examples).

The work presented here is performed entirely in the early *Drosophila* embryo, at the pre-blastoderm and blastoderm stages. The period studied is approximately 1 hour long, and takes place within 3 hours of the fertilized egg being laid. This system offers a number of advantages for the purpose of attempting to infer the correlation between sequence and expression.

First and foremost, the early *Drosophila* embryo is among the best studied systems in developmental genetics. Fruit flies have been used for more than 100 years as tractable model organisms, with a wealth of inherited morphological markers. More recently, Wieschaus, Nusslein-Völlhard, and Lewis (among many others) brought the development of embryonic phenotypes to the forefront, for which they were awarded the Nobel Prize in 1995.

The *bicoid* gradient, in particular, has been a fruitful subject in exploring developmental patterning. This maternally deposited activator is localized to the anterior of the embryo. This anterior localization is achieved through maternal deposition of the mRNA, which is

[5]I'm not going to touch this one, but I do recommend Graur *et al.* (2013).

[6]If anything, the FlyBase annotation has erred on the side of being too complete, including splice isoforms with relatively tenuous experimental support. However, these are labelled clearly, and can be removed if necessary.

[7]The crucial issue to me seems to be whether invertebrates are deserving of our ethical consideration. While I have spent comparatively little time truly contemplating the richness of a lab fly's life, it seems to me at least as comfortable as one in the wild

then translated into a diffusible protein that forms an approximately exponential gradient. While a number of studies have demonstrated that a simple Synthesis-Diffusion-Degradation model cannot completely account for the shape of the gradient, the approximation works surprisingly well (Grimm & E. Wieschaus 2010; Grimm, Coppey, *et al.* 2010; Drocco *et al.* 2012). Despite these higher order effects, the relatively simple model is useful for making mechanistic predictions of the response to *bicoid* (Gregor, Tank, *et al.* 2007; Gregor, E. F. Wieschaus, *et al.* 2007; Gregor, Bialek, *et al.* 2005).

Another attractive feature of the early *Drosophila* embryo is that the system is in a relatively naïve state. Recent experiments examining the deposition of chromatin marks at this stage have suggested that most histones are unmodified throughout the first 13 syncytial divisions, and only acquire activating or repressive marks during Stage 5, roughly concomitant with the first major wave of zygotic genome activation (X.-y. Li, Harrison, *et al.* 2014).

In the past three decades of work, the *Drosophila* research community has teased apart many of the key factors in the network that directs primary anterior-posterior patterning. Consistent with the design of circuits in other gene expression systems (Shen-Orr *et al.* 2002; Mangan & Alon 2003), the anterior-posterior patterning network is heavily enriched for feed-forward loops. In this architecture, genes at the first layer directly affect gene expression in (at least) two subsequent layers. This arrangement of interactions has a number of theoretical properties that allow for robust, tightly spatially positioned expression.

In general, most fine patterns are directed by a modular system of cis-regulatory modules (CRMs)—regions of sequence on the order of 1 kilobase that are accessible to transcription factors (TFs). These CRMs are often called enhancers, and the two terms are used somewhat interchangeably—though in general CRMs can also down-regulate the expression of a gene or set of genes via repressive means. In the relatively compact *Drosophila* genome, the majority of enhancers are within a few 10s of kbase of the genes they drive, though distances of nearly 1 megabase have been reported in other organisms (Jeong *et al.* 2006). Given this distance, it is widely assumed that enhancers loop to form physical contact with the genes they drive. Different subsets of the same pattern can be driven by different enhancers, and there are even reported cases of two or more enhancers that drive effectively identical patterns in the same gene (Perry, Boettiger & Levine 2011). Nevertheless, it is unclear how these CRMs make contact with the core promoters they drive.

The most extensively studied enhancer in *Drosophila*—if not all of animal development—is the one that drives the second stripe of the *even-skipped* gene, commonly called the "*eve* stripe 2 enhancer". A series of experiments in the late 1980s and early 1990s (Frasch & Levine 1987; Harding *et al.* 1989; Kraut & Levine 1991; Small, Kraut, *et al.* 1991; Small *et al.* 1992) worked out the molecular details of the system. In brief, broad activation domains driven by the TFs *bicoid* and *hunchback* pervade the anterior of the embryo, bind to the stripe 2 enhancer, and activate expression. Sharp borders are established through the repressive actions of *giant* in the anterior and *Kruppel* at the posterior. This motif of broad distribution of activators and sharp borders being established by repressors is also present in other enhancers found in the early fly embryo (Small *et al.* 1996).

In part driven by the abundance of data, early genes in this system have become a popular choice for mathematical modeling as well (Jaeger, Blagov, *et al.* 2004; Janssens *et al.* 2006). While they currently require an abundance of training data and do not yet take into account the full range of possible regulatory mechanisms, thermodynamic modeling has shown some promise in predicting the expression patterns (Samee & Sinha 2014). More recently, multi-factor logistic regression has been used to study the *eve* stripes in cellular-resolution detail, and to distinguish between alternative models of TF action (Ilsley *et al.* 2013; Staller, Vincent, *et al.* 2015). As a step towards understanding gene expression networks as abstract, dynamical systems, several groups have sought to understand key network features like criticality and dynamical attractors (Krotov *et al.* 2014; Jaeger & Monk 2014).

Despite the extensive work to understand this system, the majority of gene interactions are only understood at an abstract level—for instance, gene X acts on gene Y to form such-and-such a pattern. Only in a handful of cases do we truly understand the precise mechanisms by which these activation and repression steps are achieved in the context of TF binding to well-understood sites. Furthermore, while we are able to perform Chromatin Immuno-Precipitation (ChIP) experiments to directly assay the binding sites of a wide variety of transcription factors, it is not generally possible to identify from sequence alone whether any given potential binding site will be bound, to what level, and why (MacArthur *et al.* 2009; Kaplan *et al.* 2011; Cheng *et al.* 2013; Zabet & Adryan 2015). Furthermore, even knowing the locations and activities of CRMs is not sufficient to predict gene expression, since enhancers can loop across genes that separate a gene from its regulatory sequence (Wilczynski *et al.* 2012)

We also lack a distinct understanding of how the regions of DNA that encode these CRMs "know" to be functionally active. There exists a deep literature in a variety of systems of the particular kinds of histone marks and chromatin structure that distinguish active regulatory regions from parts of the genome that are not capable of driving patterns; typically this consists of histone H3K4 mono-methylation and un-compact DNA that is accessible to DNase and other DNA modifying and binding enzymes (Nègre, C. D. Brown, Shah, *et al.* 2010; Nègre, C. D. Brown, Ma, *et al.* 2011; Bogdanovic *et al.* 2012; Filion *et al.* 2010; Thomas *et al.* 2011; X.-y. Li, Thomas, *et al.* 2011; Giresi *et al.* 2007; Buenrostro *et al.* 2013).

While there are certainly new facets of enhancer-dependent regulation that have been illuminated only relatively recently (Shadow enhancers in Perry, Boettiger, Bothma, *et al.* 2010; and dynamic range tuning in H. G. Garcia *et al.* 2013), it seems likely that we have a handle on enough of the biology to be able to predict a large number of expression patterns. What is missing, however, is a sufficiently large set of training examples. Historically, we have painstakingly generated these examples by testing possible CRM constructs in a range of embryos mutant for putative TFs that drive that CRM. However, this approach is labor intensive, expensive, and relatively slow (taking several months between design of a construct and having the lines in hand to test them). Furthermore, the transgenic CRMs that have been tested thus far are not in the endogenous locus, introducing potential artifacts.

The goal of the project presented in this dissertation is to study the perturbations of

every gene in response to a broad range of conditions. Though clearly not the be-all-end-all solution to the general problem of predicting gene expression, I will demonstrate that such a dataset is useful in revealing new patterns, highlighting changes in patterned genes after perturbation, and making progress towards understanding the function of non-coding DNA.

Technical methods

The work I will present depends heavily on massively parallel short-read sequencing assays[8]. In particular, I use Illumina sequencing-by-synthesis to profile millions of distinct DNA molecules in parallel.

For profiling messenger RNA populations, there are several major steps. First, the mRNA is converted to DNA, which is more stable and has a well-developed collection of commercial available enzymes that perform useful modifications. We then selectively convert only the mRNA into cDNA through reverse-transcription, because the bulk of the RNA molecules in a cell are actually part of the ribosome or other cellular machinery, and uninteresting for the types of analyses we are interested in. Next, appropriate sequencing adapters are ligated onto the cDNA for use in the sequencers.

The library of adapter-ligated cDNA molecules is then delivered to the sequencing center. Here, the molecules are hybridized at low concentration to probes complementary for the adapter sequence on the surface of a glass slide; the low concentration ensures that most molecules will not be close to another, different molecule. The single molecules are then amplified in a process called bridge-PCR, whereby each single-molecule spot is copied to a strictly nearby probe. Thus, the single molecule becomes a cluster of effectively identical products. Finally, the sequence is actually generated: for each base, a fluorescent-labelled nucleotide with a terminator is incorporated on the complementary strand, imaged to call the base that was incorporated, then the terminator moiety is removed. Through this process, the sequence of hundreds of millions of spots can be determined over just a few days (Bentley *et al.* 2008).

Overview of the approach

The work described in this dissertation generated a genome-wide atlas of gene expression patterns using high-throughput sequencing of cDNAs. Sequencing is relatively fast—turnaround time from embryos to quantified global expression levels is as little as 3 weeks (though usually longer due to the desire for multiplexing and the variable length of the sequencing queue). Moreover, it is inexpensive compared to any other conceivable method for generating simi-

[8]The collection of approaches are sometimes called Next-Gen Sequencing (NGS), though there is actually a third generation of technologies that are starting to emerge, all of which use un-amplified single DNA molecules to generate their output. These single molecule sequencing approaches are still relatively new, and on a cost per base comparison are not yet competitive with NGS approaches, though they do have a number of attractive features which make them useful complements.

lar coverage of the genome[9]. Using careful sectioning of *Drosophila* embryos, RNA can be relatively easily recovered in sufficient quantities to perform mRNA-seq experiments.

In effect, any biological experiment is a slice across a multi-dimensional problem space. The nature of a slice is that it keeps information from one direction, while discarding information from others. An *in situ* hybridization micrograph, for instance, keeps information in (for instance) the X and Y directions[10], while discarding information along Z. However, in addition to the spatial dimensions represented, the image also implicitly keeps information only a single time point (the time dimension), and only from one gene out of all of the genes being expressed (the "genome dimension"). When we discard information from other dimensions, there is the option of either averaging across those dimensions, or in taking just a small subset. For instance, when looking at a 2-dimensional image of a 3-dimensional object, the Z direction is averaged over some more or less limited range (typically the depth of focus of the image, which can be much greater than the X-Y resolution of the image). Finally, different slices can be combined to gain a sense of the change across space, time, or genome. The key to designing informative experiments, then, is to choose which dimensions we are capable of looking at, and which ones we must average across.

In this work, I extend and improve upon genome-wide methods to provide spatially resolve slices. *In situ* hybridization experiments, for instance, keep X and Y, average over Z, and discard the large majority of the genes expressed in any given individual. By contrast, typical RNAseq studies (J. B. Brown *et al.* 2014; Tan *et al.* 2013; Graveley *et al.* 2011; Mortazavi *et al.* 2008, to name a very few) with many individuals will keep the full coverage of genes, while averaging over X, Y, Z, and time. The tradeoff that I set out to make is to keep information in the X, genome, and time directions, while only averaging over Y and Z.

Such an atlas of gene expression is not without certain limitations. First, the spatial resolution is a good deal lower than in imaging based studies. Second, because we don't fully understand the relationship between all genes and the enhancers that pattern them, our gene-level atlas is not capable of distinguishing the use of different enhancers that, in aggregate, produce relatively complex spatial patterns, even though each is restricted only to a subset of the embryo. Nevertheless, compared to examining *in situ* hybridization patterns of the thousands of genes bound by a given TF, this approach is fast and inexpensive.

[9]There is a common idea that sequencing is an expensive experiment overall. I do not disagree with this statement, but the costs are not generally where one would expect. While a lane of sequencing does cost on the order of $1000, plus sequencing reagents, advances in sequencing technology allow multiplexing of approximately 30-50 different samples to a suitable depth of coverage, and a skilled user can generate these libraries in a week. A similar qPCR experiment, at say $1/gene/samples (probably quite generous), would only be able to assay ~ 30 genes. Instead, I believe that the majority of the true cost of a sequencing experiment lies in the researcher's time taken to analyze the samples. Since both graduate students and postdocs cost a grant on the order of $60,000/year, one extra week of time spent analyzing the data wipes out any potential savings. A good RNAseq dataset should be able to test lots of hypotheses, so it will cost more to analyze it fully, but also generate much more science per dollar.

[10]Throughout this work, I will adopt the convention from decades of *Drosophila* embryology, where Anterior is to the left (X), dorsal is up (Y), and the left-right axis comes out of the page (Z).

Here I show progress in determining genome-wide spatial patterns in single *Drosophila* embryos by:

1. the development and demonstration of methods in wild-type *Drosophila melanogaster* (Chapter 2),
2. the improvement of those methods for generation of sequencing libraries from very small inputs at a reasonable cost (Chapter 3), and
3. the application of those methods towards a number of patterning mutants in *D. melanogaster* (Chapter 4).

Chapter 2

Sequencing mRNA from Cryo-Sliced *Drosophila* Embryos to Determine Genome-Wide Spatial Patterns of Gene Expression

Note that the bulk of this chapter was published as:

Combs, P. A. & Eisen, M. B. Sequencing mRNA from cryo-sliced Drosophila embryos to determine genome-wide spatial patterns of gene expression. *PLoS ONE* **8,** e71820 (2013).

2.1 Abstract

Complex spatial and temporal patterns of gene expression underlie embryo differentiation, yet methods do not yet exist for the efficient genome-wide determination of spatial expression patterns during development. *In situ* imaging of transcripts and proteins is the gold-standard, but it is difficult and time consuming to apply to an entire genome, even when highly automated. Sequencing, in contrast, is fast and genome-wide, but is generally applied to homogenized tissues, thereby discarding spatial information. To take advantage of the efficiency and comprehensiveness of sequencing while retaining spatial information, we cryosectioned individual blastoderm stage *Drosophila melanogaster* embryos along the anterior-posterior axis and developed methods to reliably sequence the mRNA isolated from each 25μm slice. The spatial patterns of gene expression we infer closely match patterns previously determined by *in situ* hybridization and microscopy. We applied this method to generate a genome-wide timecourse of spatial gene expression from shortly after fertilization through gastrulation. We identified numerous genes with spatial patterns that have not yet been described in the several ongoing systematic *in situ* based projects. This simple experiment demonstrates the potential for combining careful anatomical dissection with high-throughput sequencing to obtain spatially resolved gene expression on a genome-wide

scale.

2.2 Introduction

Analyzing gene expression in multicellular organisms involves a tradeoff between the spatial precision of imaging and the efficiency and comprehensiveness of genomic methods. RNA *in situ* hybridization (ISH) and antibody staining of fixed samples, or fluorescent imaging of live samples, provides high resolution spatial information for small numbers of genes (Fowlkes *et al.* 2008; Tomancak *et al.* 2007; Lécuyer *et al.* 2007). But even with automated sample preparation, imaging, and analysis, *in situ* based methods are difficult to apply to an entire genome's worth of transcripts or proteins. High throughput genomic methods, such as DNA microarray hybridization or RNA sequencing, are fast and relatively inexpensive, but, at least for the small species worked with in most labs, the amount of input material they require has generally limited their application to homogenized samples, often from multiple individuals. Methods involving the tagging, sorting, and analysis of RNA from cells in specific spatial domains have shown promise (Steiner *et al.* 2012), but remain non-trivial to apply systematically, especially across genotypes and species.

Recent advances in DNA sequencing suggest an alternative approach. With increasingly sensitive sequencers and improved protocols for sample preparation, it is now possible to analyze small samples without amplification. Several years ago we developed methods to analyze the RNA from individual *Drosophila* embryos (Lott *et al.* 2011). As we often recovered more RNA from each embryo than was required to obtain accurate measures of gene expression, we wondered whether we could obtain good data from pieces of individual embryos, and whether we could obtain reliable spatial expression information from such data. To test this possibility, we chose to focus on anterior-posterior (A-P) patterning in the early *D. melanogaster* embryo, as the system is extremely well-characterized and the geometry of the early embryo also lends itself to biologically meaningful physical dissection by simple sectioning along the elongated A-P axis.

2.3 Results

To test whether we could consistently recover and sequence RNA from sectioned *D. melanogaster* embryos, we collected embryos from our laboratory stock of the line CantonS (CaS), aged them for approximately 2.5 hours so that the bulk of the embryos were in the cellular blastoderm stage, and fixed them in methanol. We examined the embryos under a light microscope and selected single embryos that were roughly halfway through cellularization (mitotic cell cycle 14; developmental stage 5). We embedded each embryo in a cryoprotecting gel, flash-froze it in liquid nitrogren, and took transverse sections along the anterior-posterior axis. For this initial trial we used 60μm sections, meaning that we cut each approximately

350μm embryo into six pieces. We placed each piece into a separate tube, isolated RNA using Trizol, and prepared sequencing libraries using the Illumina Tru-Seq kit .

In early trials we had difficulty routinely obtaining good quality RNA-seq libraries from every section. We surmised that we were losing material from some slices during library preparation as a result of the small amount (approximately 15ng) of total RNA per slice. To overcome this limitation, after the initial RNA extraction we added RNA from a single embryo of a distantly related *Drosophila* species to each tube to serve as a carrier.

We used RNA as a carrier, instead of a standard carrier like salmon sperm DNA or linear acrylamide, so that the carrier was present throughout the experiment, and we used RNA from multiple *Drosophila* species in particular so that the sequence reads from the carrier RNA would not be wasted. In this first experiment the carrier RNA was part of an experiment examining gene expression in early embryos of other *Drosophila* species. We only used embryos from species that were fully sequenced and sufficiently diverged from *D. melanogaster* to allow us to readily separate reads derived from the *D. melanogaster* slice and the carrier species computationally after sequencing.

With the additional approximately 100ng of total RNA from the carrier in each sample, library preparation became far more robust. We independently sliced three CaS embryos, prepared libraries from the sliced RNA using the standard TruSeq RNA kit, and sequenced them using an Illumina HiSeq 2000, obtaining approximately 40 million 50 bp paired-end reads for each slice+carrier sample. We aligned these reads to the *D. melanogaster* and carrier genomes using TopHat(Langmead & Salzberg 2012; Kim *et al.* 2013), and identified between 1.7 and 31.4 percent of reads as having come unambiguously from *D. melanogaster* (see Table 2.1). We then used Cufflinks(Trapnell, Roberts, *et al.* 2012) to infer expression levels for all annotated mRNAs using the *D. melanogaster* reads alone.

The data for each slice within an embryo were generally highly correlated (Supplemental Figure 2.S1), reflecting the large number of highly expressed genes with spatially uniform expression patterns. The data for equivalent slices of embryos 2 and 3 were also highly correlated, while the slices for embryo 1 were systematically less well matched to their counterparts in embryos 2 and 3 (Supplemental Figure 2.S2), suggesting that it may have been sampled at a slightly different developmental stage.

To examine how well our data recapitulated known spatial profiles, we identified a panel of genes with known anterior-posterior patterns of gene expression and compared our data to their published expression patterns. Figure 2.1A shows RNA in-situ hybridization patterns from the Berkeley Drosophila Genome Project (BDGP) (Tomancak *et al.* 2007) alongside the expression data for that gene from our sliced embryos, demonstrating a close qualitative agreement between the visualized expression patterns and our sliced RNA-seq data.

In order to more quantitatively compare our data to existing patterns, we constructed a reference set of spatial expression patterns along the A-P axis using three-dimensional "virtual embryos" from the Berkeley Drosophila Transcription Network Project, which contain expression patterns for 95 genes at single-nucleus resolution (Fowlkes *et al.* 2008). We transformed the relative expression levels from these images into absolute values (FPKM) using genome-wide expression data from intact single embryos (Lott *et al.* 2011). We compared the

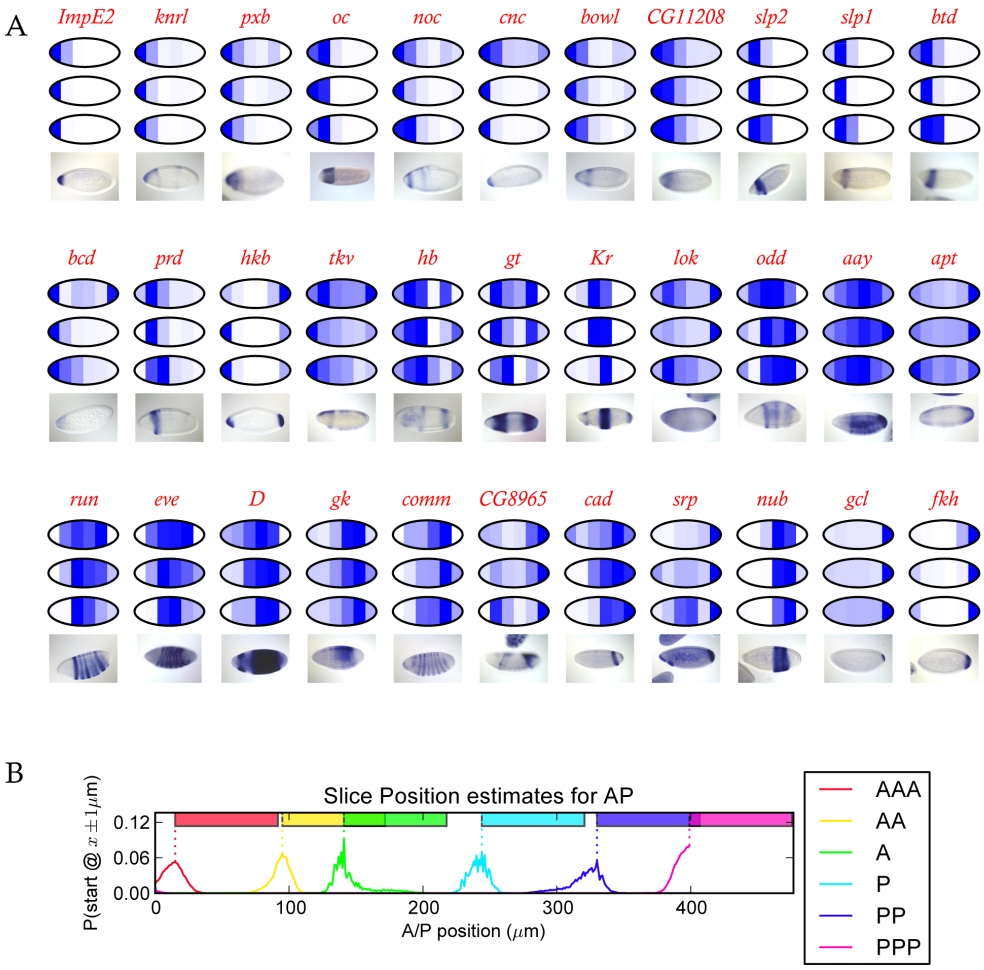


Figure 2.1. Expression in the slices closely matches published expression data. (A) Published *in situ* patterns for 33 genes are shown alongside reconstrucred spatial patterns for these genes from each of the three 60μm sliced CaS embryos. The reconstructed patterns were each scaled to the slice with the highest expression level for each embryo individually. (B) To evaluate the overall quality of our reconstructed spatial expression patterns, we compared expression levels of 98 genes from each slice in our 60 μm data (averaged across the three embryos) to all possible 60μm sections from a cellular resolution spatial atlas of gene expression from the Berkeley Drosophila Transcription Network Project (Fowlkes *et al.* 2008) with absolute expression levels computed using data from Lott *et al.* 2011. We computed the posterior probability that a slice from our data corresponded to a slice from the BDTNP atlas using a simple Bayesian procedure that compares the level of each gene in a slice to the level of that gene in sections of the atlas. The line graphs are the posterior probabilities that each slice started at a given position in the atlas. Each slice has a clear peak and the ordering of the peaks corresponds to the ordering of the slices, as expected. The colored bars show the portion of the embryo spanned by the slice assuming it begins at the peak in the posterior probability distribution.

Table 2.1. Sequencing statistics for sliced single-stage wild-type mRNA-Seq samples Counts are for read ends. Discordant read ends are always classed as ambiguous, but failure of one end to map does not disqualify the other.

Replicate	Slice	Carrier Species	Bar-code Index	Total Reads	Uniquely mapped *D. mel* reads (%)	Ambiguous Reads (%)
1	1	*D. per*	1	69,339,972	2,284,228 (3.2%)	1,634,055 (2.3%)
1	2	*D. per*	2	73,632,862	3,706,630 (5.0%)	1,603,444 (2.1%)
1	3	*D. per*	3	82,076,328	6,002,034 (7.3%)	1,774,485 (2.1%)
1	4	*D. per*	4	73,437,708	6,401,565 (8.7%)	1,592,665 (2.1%)
1	5	*D. per*	5	75,922,812	4,951,178 (6.5%)	1,559,097 (2.0%)
1	6	*D. per*	6	78,623,784	1,355,079 (1.7%)	1,574,067 (2.0%)
2	1	*D. wil*	7	59,813,036	4,066,295 (6.7%)	878,476 (1.4%)
2	2	*D. wil*	8	90,961,338	15,212,716 (16.7%)	1,301,095 (1.4%)
2	3	*D. wil*	9	73,201,902	14,855,374 (20.2%)	911,768 (1.2%)
2	4	*D. wil*	10	75,754,772	23,858,301 (31.4%)	1,136,031 (1.4%)
2	5	*D. wil*	11	84,497,566	10,026,713 (11.8%)	1,080,910 (1.2%)
2	6	*D. wil*	12	66,316,952	13,122,508 (19.7%)	898,776 (1.3%)
3	1	*D. moj*	13	75,847,986	12,496,248 (16.4%)	3,615,452 (4.7%)
3	2	*D. moj*	14	72,497,660	4,005,714 (5.5%)	803,381 (1.1%)
3	3	*D. moj*	15	77,532,368	11,138,154 (14.3%)	772,446 (0.9%)
3	4	*D. moj*	16	83,400,882	8,227,562 (9.8%)	861,839 (1.0%)
3	5	*D. moj*	18	83,608,454	2,630,069 (3.1%)	795,169 (0.9%)
3	6	*D. moj*	19	85,823,784	2,239,493 (2.6%)	829,382 (0.9%)

observed expression for these 95 genes from an average of each of our slices to all possible 60μm slices of these virtual embryos (Figure 2.1B). High scores for most slices fell into narrow windows, with the best matches for each slice falling sequentially along the embryo with a spacing of about 60μm , the same thickness as the slices.

We next used the program Cuffdiff (Trapnell, Hendrickson, *et al.* 2013) to identify 85 genes with statistically significant differences in expression between slices (Dataset S2; this is a very conservative estimate). We compared these genes to those examined by the BDGP, the most comprehensive annotation of spatial localization in *D. melanogaster* development that we are aware of (Tomancak *et al.* 2007). Of our differentially expressed genes, 21 had no imaging data available, and 33 were annotated as present in a subset of the embryo (the annotation term meant to capture patterned genes); the remaining 31 genes showed either clear patterns that were not annotated with the most general keyword, or no clear staining (Supplemental Figure 2.S3). There were 194 genes tagged by the BDGP as patterned that were not picked up as having statistically significant patterns in our data. However, most of these had primarily dorsal-ventral patterns, faint patterns, later staging in the images used for annotation, or had good qualitative agreement with our data but fell above the cutoff for statistical significance (Supplemental Figure 2.S4).

As a more sensitive approach to finding patterned genes, we applied k-means clustering

to our data. We first filtered on expression level (at least one slice in one embryo with FPKM > 10) and agreement between replicates (average Pearson correlation between embryos of > 0.5), then clustered based on normalized expression ($k = 20$, centroid linkage; 20 was chosen empirically as smaller k's merged genes with different patterns and larger k's provided no additional useful information) (de Hoon *et al.* 2004). We identified several broad classes of expression, including localization to each of the poles, and five different gap gene-like bands along the AP axis (Figure 2.2 and Supplemental Figure 2.S5). Of the 745 genes, only 349 had images in the BDGP set(Tomancak *et al.* 2007). Staining for these genes is sometimes undetectable and well-matched stages are often missing from the databases, but where comparisons were possible, the BDGP image data agrees with our RNA-seq patterns (Supplemental Figure 2.S6).

To extend our dataset, we collected individual embryos from seven different time points based on morphology—stage 2, stage 4, and 5 time points within stage 5—and sliced them into 25μm sections, yielding between 10 and 15 contiguous, usable slices per embryo. For these embryos we used total RNA from the yeasts *Saccharomyces cerevisiae* and *Torulaspora delbruckii* as carrier, which are so far diverged as to have fewer than 0.003% of reads ambiguously mapping.

These finer slices are better able to distinguish broad gap-gene domains, with several slices of relatively low expression between the multiple domains of *hb*, *kni*, and *gt*. Excitingly, we can also distinguish the repression between stripes of pair-rule genes like *eve* as well (Figure 2.3). Given the non-orthogonal orientation of the anterior-most and posterior-most *eve* stripes relative to the AP axis, we do not expect to see all 7 pair-rule stripes, but at least three can be unambiguously observed.

Putting the 60μm and 25μm slice datasets together, we find a large number of genes with reproducible patterns in the 60μm slices whose formation over time can be clearly seen in the timed 25μm slices, including many without previously described early patterns (Supplemental Figure 2.S7).

2.4 Discussion

The experiments reported here demonstrate that slicing and sequencing animal embryos is a practical and effective method to systematically characterize spatial patterns of expression. While we are by no means the first to dissect samples and characterize their RNAs—Ding & Lipshitz 1993 pioneered this kind of analysis twenty years ago—to our knowledge we are the first to successfully apply such a technique to report genome-wide spatial patterns in a single developing animal embryo.

Given the degree to which the *D. melanogaster* embryo has been studied, and the presence of at least two large *in situ* based studies whose goals were to systematically identify and characterize genes with patterned expression in the embryo, we were surprised by the large number of genes we find as clearly patterned that had not been previously described as such. We note in particular a large number of genes with expression restricted to the poles, most

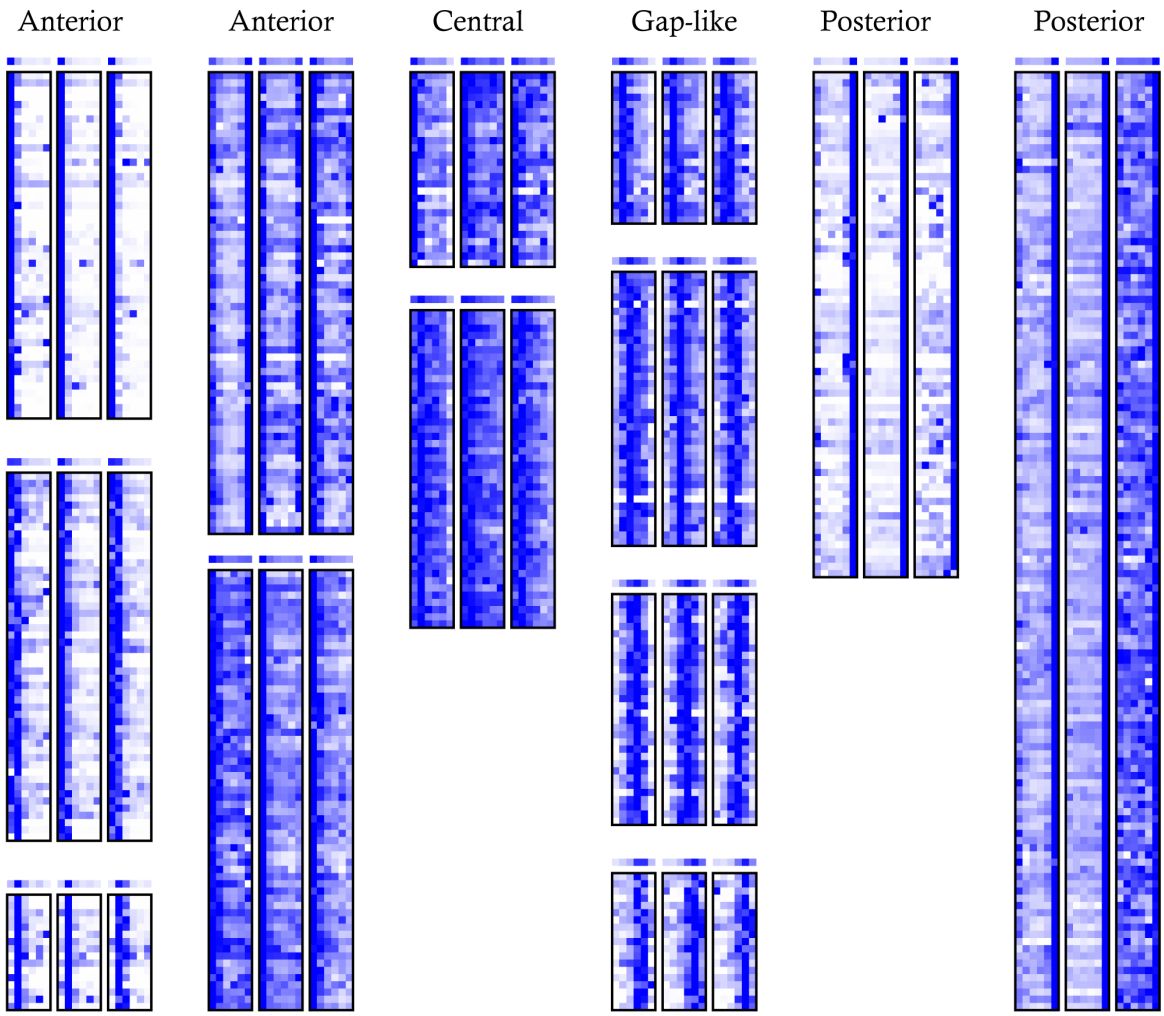


Figure 2.2. Heat maps of gene expression clusters. Of the $k = 20$ clusters, 13 with non-uniform patterns are shown. The expression levels for each gene was normalized for clustering and display so that the maximum expression of each gene in each embryo is dark blue. The plot above each cluster is the mean normalized expression level in that cluster. All clusters are listed in Dataset S3.

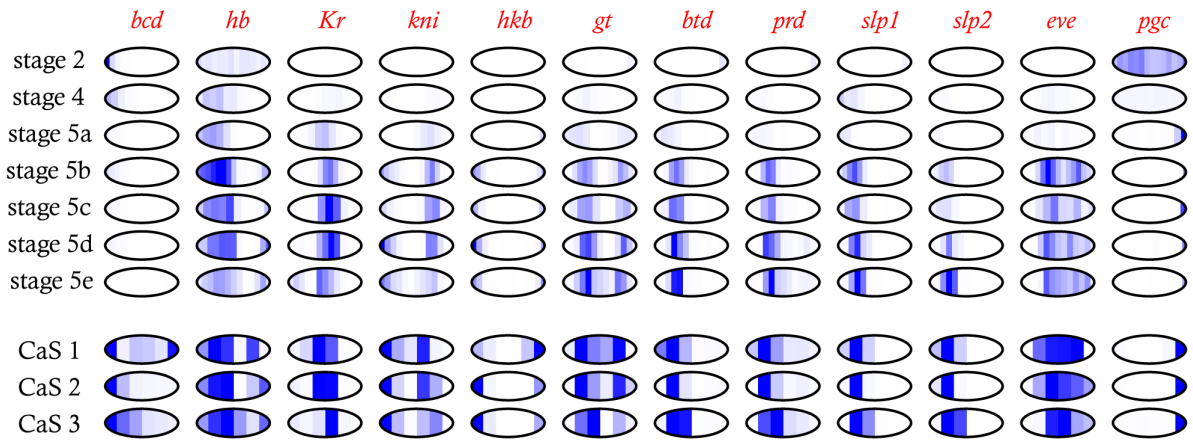


Figure 2.3. Expression of key patterning genes across early development. Expression levels in the 25μm timeseries are normalized to the highest expression level at any time pioint. For slices with poor quality data (timepoint 4, slice 10; timepoint 6, slice 6; timepoint 7, slice 7; and timepoint 7, slice 8) data imputed from neighboring slices is shown. Expression levels for the 60μm slice samples are normalized to the highest level in each embryo.

with no known role in either anterior patterning or pole cell formation or activity. This emphasizes the potential for sequencing-based methods to replace *in situ* based studies in the systematic analysis of patterned gene expression, as they are not only simpler, cheaper, and easier to apply to different species and genetic backgrounds, but appear to be more sensitive.

The data we present here are far from perfect - the relatively small number of reads per slice (due to the presence and sequencing of carrier RNA) means that the slice by slice data are somewhat noisy. However the consistency between replicates and the agreement between the 25μm and 60μm data demonstrate that the experiment clearly worked, and additional sequencing depth and better methods for working with small samples should greatly reduce the noise as we move forward.

Obviously, to truly replace *in situ* based methods, sequencing based methods will need to achieve greater resolution than presented here. One can envision several basic approaches to achieving the ultimate goal of determining the location of every RNA in a spatially complex tissue. Sequencing RNAs in place in intact tissues would obviously be the ideal method, and we are aware of several groups working towards this goal. In the interim, however, methods to isolate and characterize smaller and smaller subsets of cells are our only alternative. One possibility is to combine spatially restricted reporter gene expression and cell sorting to purify and characterize the RNA composition of differentiated tissue—c.f. Steiner *et al.* 2012. While elegant, this approach cannot be rapidly applied to different

genetic backgrounds, requires separate tags for every region/tissue to be analyzed, and will likely not work on single individuals.

Sectioning based methods offer several advantages, principally that they can be applied to almost any sample from any genetic background or species, and allow for the biological precision of investigating single individuals. The 60μm and 25μm slices we used here represent reasonable tradeoffs between sequencing depth and spatial resolution given the current limits of sample preparation and sequencing methods, but with methods having been described to sequence the RNAs from single cells, and with sequencing costs continuing to plummet, it should be possible to obtain far better resolution in the near future. A rough estimate suggests that a single embryo contains enough RNA to sequence over 700 samples to a depth of 20 million reads. Thus it is theoretically possible to dice an embryo into 20μm cubes and sequence each one to obtain genome-wide three-dimensional expression data, although this presents several difficult but likely solvable technical challenges, especially handling and tracking hundreds or thousands of tiny samples.

2.5 Materials and Methods

Fly Line, Imaging, and Slicing

We raised flies on standard media at 25° in uncrowded conditions, and collected eggs from many 3 to 10-day old females from our *Canton-S* lab stocks. We washed and dechorionated the embryos, then fixed them according to a standard methanol cracking protocol. Briefly, we placed embryos in 20ml glass vials containing 10ml of heptane and 10ml of PEM (100mM PIPES, 2mM EGTA, 1mM MgSO4) and mixed gently. We then removed the aqueous phase, added 10ml of methanol, shook vigorously for 15-30 seconds, and collected the devitellinized embryos, which we washed several times in methanol to remove residual heptane. We then placed the fixed embryos on a slide in halocarbon oil, and imaged on a Nikon 80i with DS-5M camera. After selecting embryos with the appropriate stage according to depth of membrane invagination and other morphological features, we washed embryos with methanol saturated with bromophenol blue dye (Fisher, Fair Lawn NJ), aligned them in standard cryotomy cups (Polysciences Inc, Warrington, PA), covered them with OCT tissue freezing medium (Triangle Biomedical, Durham, NC), and flash froze them in liquid nitrogen.

We sliced frozen embryos on a Microm HM 550 (Thermo Scientific, Kalamozoo, MI) at a thickness of 60μm or 25μm . We adjusted the horizontal position of the blade after every slice to eliminate the possibility of carry-over from previous slices, and used a new blade for every embryo. We placed each slice in an individual RNase-free, non-stick tube (Life Technologies, Grand Island, NY).

RNA Extraction, Library Preparation, and Sequencing

We performed RNA extraction in TRIzol (Life Technologies, Grand Island, NY) according to manufacturer instructions, except with a higher concentration of glycogen as carrier (20 ng) and a higher relative volume of TRIzol to the expected material (1mL, as in Lott *et al.* 2011). For the 60μm slices, we pooled total RNA from each slice with total RNA from single *D. persimilis*, *D. willistoni*, or *D. mojavensis* embryos, then made libraries according to a modified TruSeq mRNA protocol from Illumina. We prepared all reactions with half-volume sizes to increase relative sample concentration, and after AmpureXP cleanup steps, we took care to pipette off all of the resuspended sample, leaving less than 0.5 μL, rather than the 1-3 μL in the protocol. Furthermore, we only performed 13 cycles of PCR amplification rather than the 15 in the protocol, to minimize PCR duplication bias.

Libraries were quantified using the Kapa Library Quantification kit for the Illumina Genome Analyzer platform (Kapa Biosystems) on a Roche LC480 RT-PCR machine according to the manufacturer's instructions, then pooled to equalize index concentration. Pooled libraries were then submitted to the Vincent Coates Genome Sequencing Laboratory for 50bp paired-end sequencing according to standard protocols for the Illumina HiSeq 2000. Bases were called using HiSeq Control Software v1.8 and Real Time Analysis v2.8.

Mapping and Quantification

Reads were mapped using TopHat v2.0.6 to a combination of the FlyBase reference genomes (version FB2012_05) for *D. melanogaster* and the appropriate carrier species genomes with a maximum of 6 read mismatches (McQuilton *et al.* 2012; Trapnell, Pachter, *et al.* 2009). Reads were then assigned to either the *D. melanogaster* or carrier genomes if there were at least 4 positions per read to prefer one species over the other. We used only the reads that mapped to *D. melanogaster* to generate transcript abundances in Cufflinks.

Data and Software

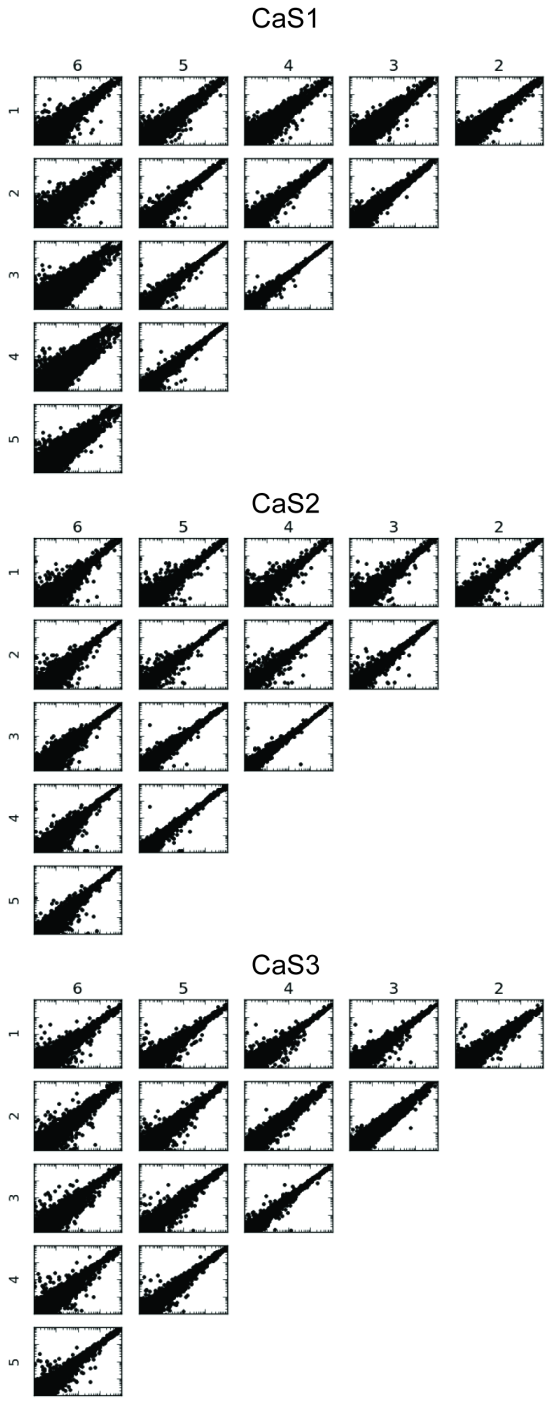
We have deposited all reads in the NCBI GEO under the accession number GSE43506. The processed data are available at the journal website (Dataset S1) and at `http://eisenlab.org/sliceseq` with a search feature for the 25μm dataset. All custom analysis software is available `https://github.com/eisenlab/SliceSeq`, and is primarily written in Python (Van Rossum & Drake 2003; Cock *et al.* 2009; Hunter 2007; Jones *et al.* 2001; Perez & Granger 2007). Commit b0b115a was used to perform all analyses in this paper.

Acknowledgments

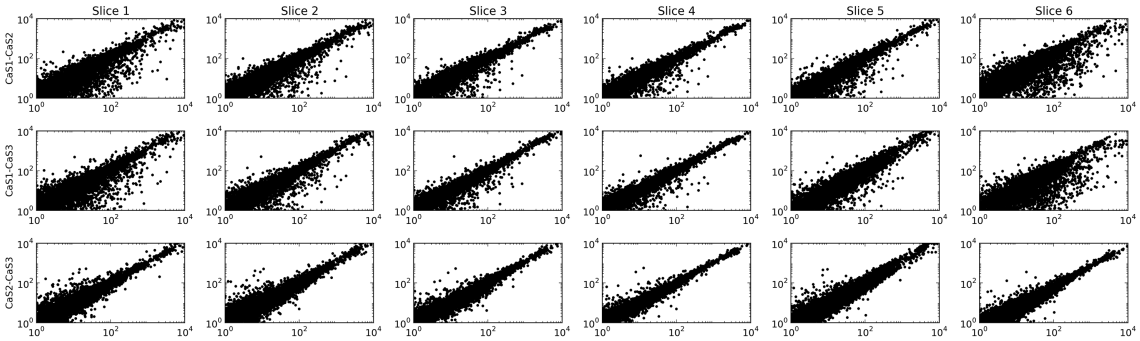
We thank peer reviewer Boris Adryan for helpful comments, and many readers who contributed feedback on a preprint of the manuscript posted on MBE's blog and the arXiv. We

also thank members of the Eisen lab for their assistance, especially Susan Lott and Jackie Villalta.

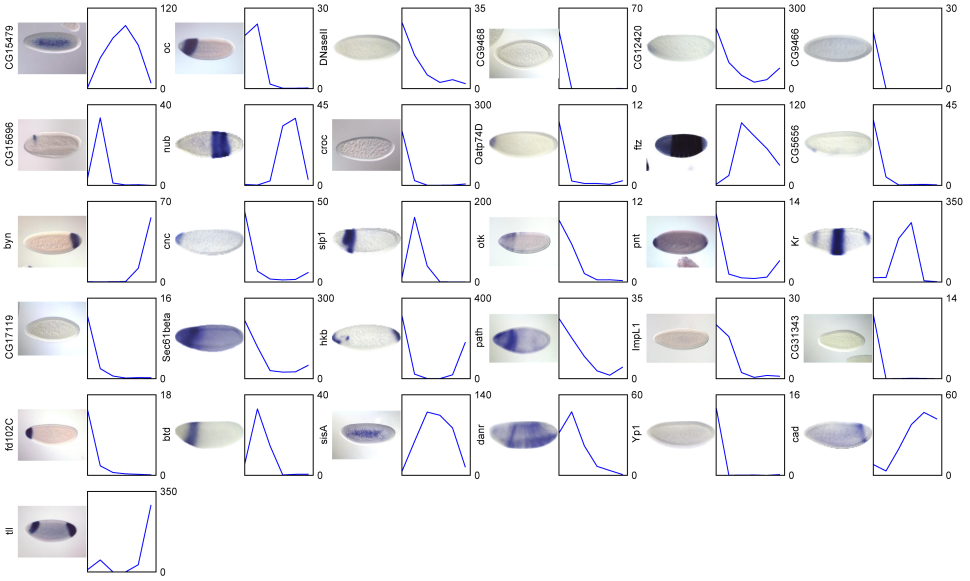
2.6 Supplemental Figures



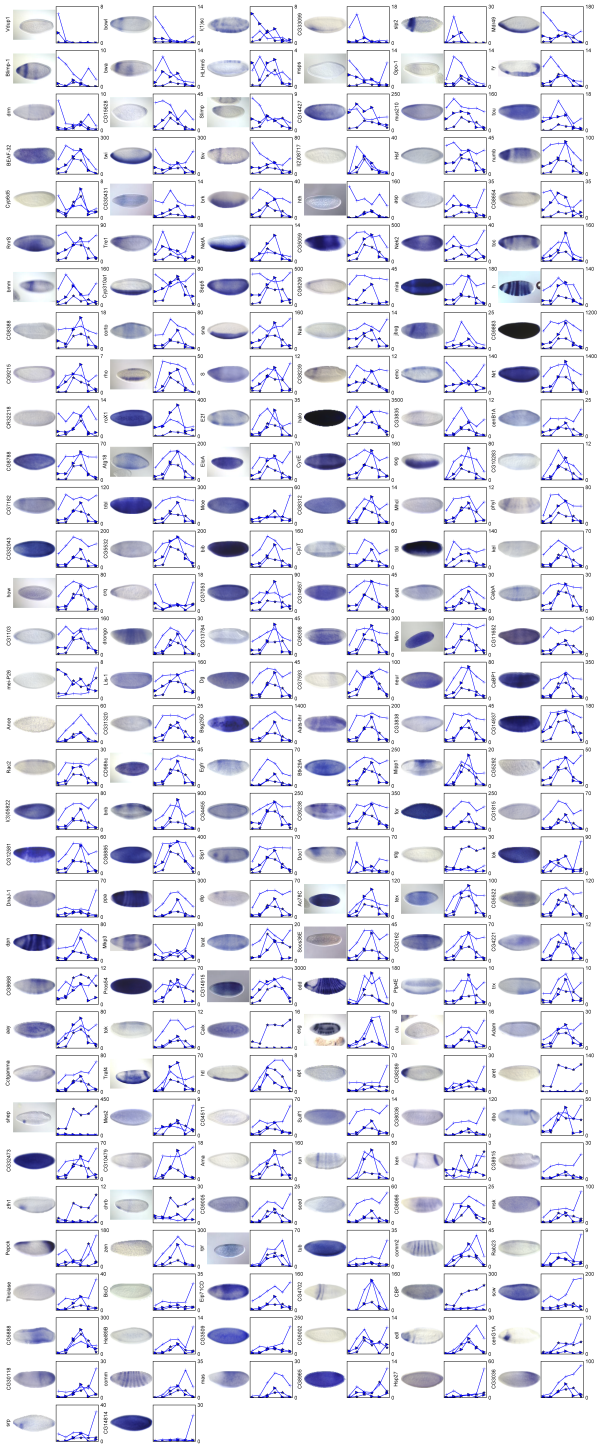
Supplemental Figure 2.S1. Correlation of slices within embryos. Log-log plots of FPKM values between slices within each of the three 60μm sliced embryos.



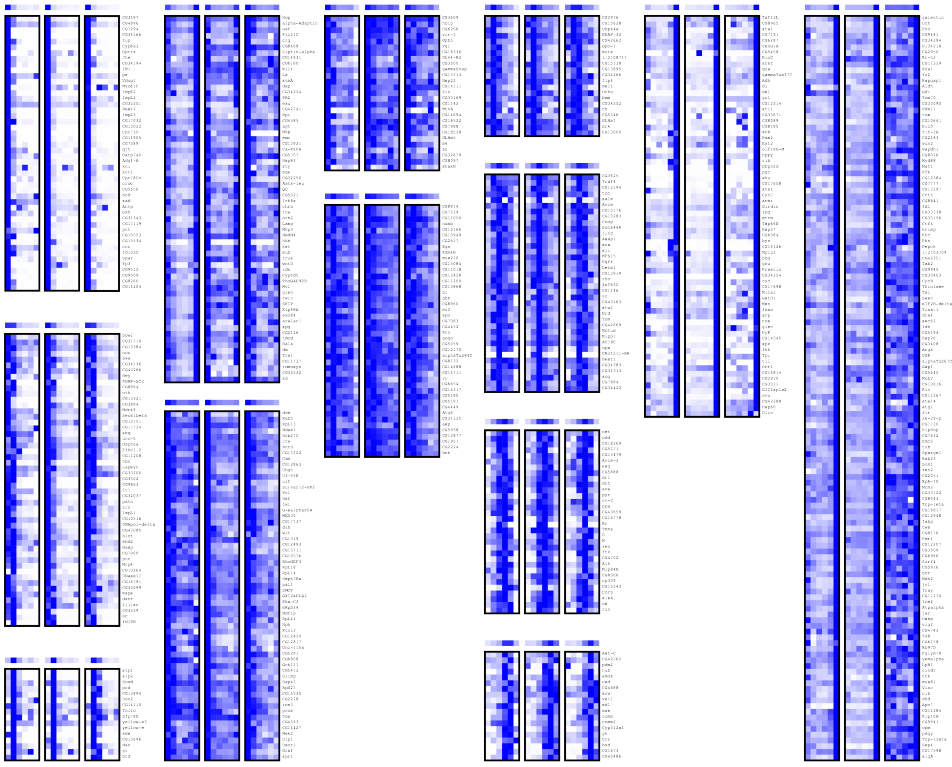
Supplemental Figure 2.S2. Correlation of slices between embryos. Log-log plots of FPKM values of corresponding slices between each of the three 60μm sliced embryos.



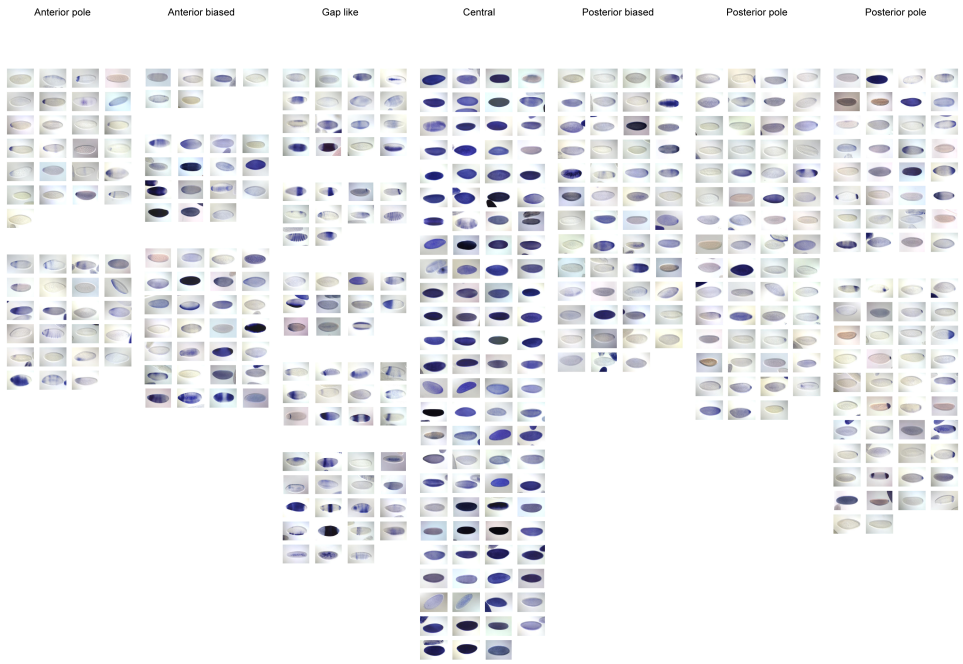
Supplemental Figure 2.S3. Genes called as patterned by Cuffdiff lacking subset tag in BDGP database. Images are from BDGP; graphs are average of three CaS embryos. Many of these are known patterned genes, highlighting the incompleteness of available annotations.



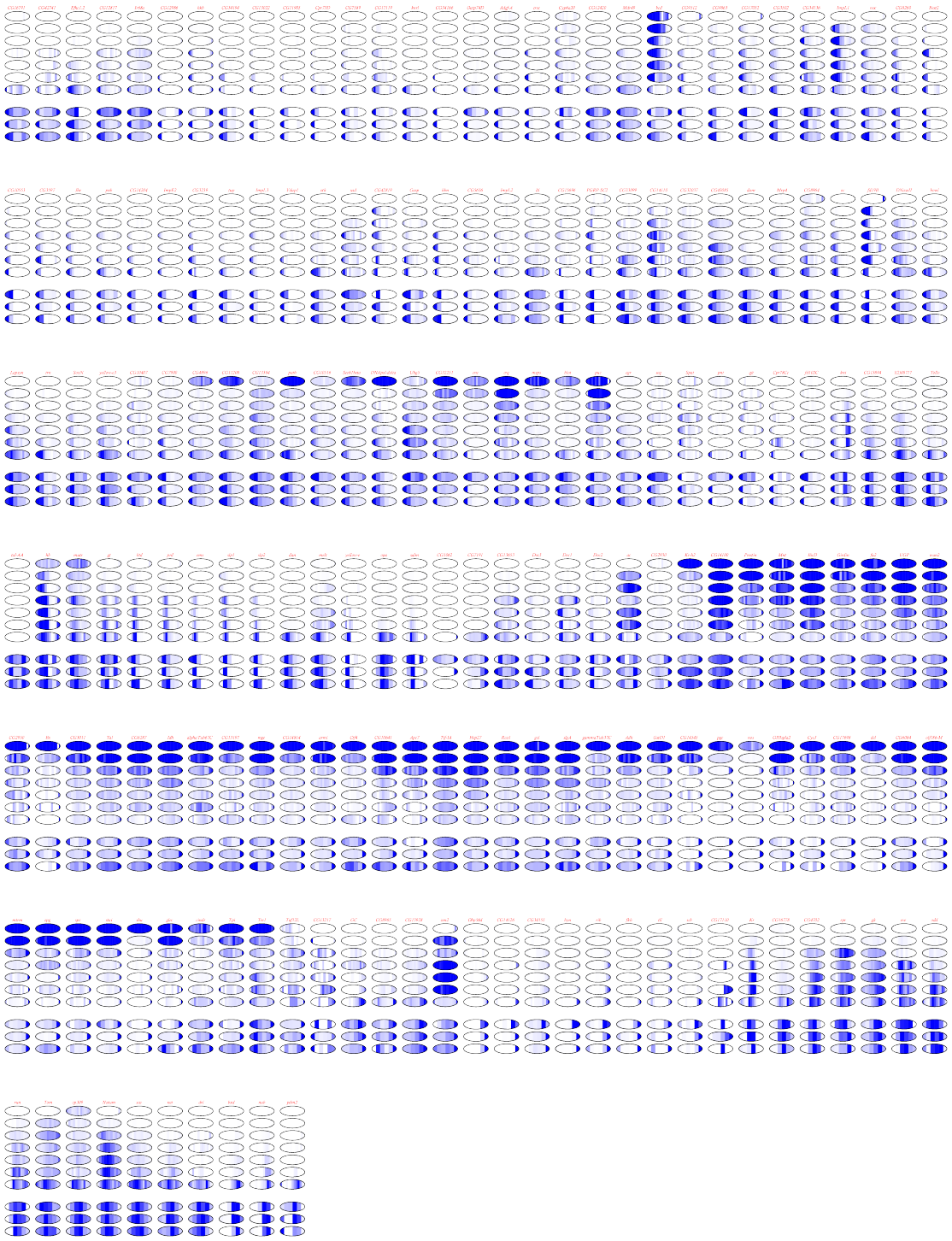
Supplemental Figure 2.S4. Genes with subset tag in BDGP not called as patterned by Cuffdiff.



Supplemental Figure 2.S5. Figure 2.2 with gene names.



Supplemental Figure 2.S6. Images from BDGP for genes in clusters shown in Figure 2.2.



Supplemental Figure 2.S7. Data from 25μm timecourse and 60μm embryos for a large number of genes with manually curated patterns.

Chapter 3

Low-cost, low-input RNA-seq protocols perform nearly as well as high-input protocols

Note that the bulk of this chapter was published as:

Combs, P. A. & Eisen, M. B. Low-cost, low-input RNA-seq protocols perform nearly as well as high-input protocols. *PeerJ* **3,** e869 (2015)

3.1 Abstract

Recently, a number of protocols extending RNA-sequencing to the single-cell regime have been published. However, we were concerned that the additional steps to deal with such minute quantities of input sample would introduce serious biases that would make analysis of the data using existing approaches invalid. In this study, we performed a critical evaluation of several of these low-volume RNA-seq protocols, and found that they performed slightly less well in per-gene linearity of response, but with at least two orders of magnitude less sample required. We also explored a simple modification to one of these protocols that, for many samples, reduced the cost of library preparation to approximately $20/sample.

3.2 Introduction

Second-generation sequencing of RNA (RNA-seq) has proven to be a sensitive and increasingly inexpensive approach for a number of different experiments, including annotating genes in genomes, quantifying gene expression levels in a broad range of sample types, and determining differential expression between samples. As technology improves, transcriptome profiling has been able to be applied to smaller and smaller samples, allowing for more powerful assays to determine transcriptional output. For instance, our lab has used RNA-seq on

single *Drosophila* embryos to measure zygotic gene activation (Lott *et al.* 2011) and medium-resolution spatial patterning (Combs & Eisen 2013). Further improvements will allow an even broader array of potential experiments on samples that were previously too small.

For instance, over the past few years, a number of groups have published descriptions of protocols to perform RNA-seq on single cells (typically mammalian cells) (Tang *et al.* 2009; Ramsköld *et al.* 2012; Sasagawa *et al.* 2013; Hashimshony *et al.* 2012; Islam *et al.* 2011). A number of studies, both from the original authors of the single-cell RNA-seq protocols and from others, have assessed various aspects of these protocols (such as the lower limit of detection, strand specificity, and uniformity of coverage), both individually and competitively (Levin *et al.* 2010; Bhargava *et al.* 2014; A. R. Wu *et al.* 2014; Marinov *et al.* 2013). One particularly powerful use of these approaches is to sequence individual cells in bulk tissues, revealing different states and cellular identities (Buganim *et al.* 2012; Treutlein *et al.* 2014).

However, we felt that published descriptions of single-cell and other low-volume protocols did not adequately address whether a change in concentration of a given transcript between two samples would result in a proportional change in the FPKM (or any other measure of transcriptional activity) between those samples. While there are biases inherent to any protocol, we were concerned that direct amplification of the mRNA would select for PCR compatible genes in difficult to predict, and potentially non-linear ways. For many of the published applications of single cell RNA-seq, this is not likely a critical flaw, since the clustering approaches used are moderately robust to quantitative changes. However, to measure spatial and temporal activation of genes across an embryo, it is important that the output is monotonic with respect to concentration, and ideally linear. A linear response allows for more easily interpretable experimental results, without necessarily relying on complicated transformations of the data.

While it is possible to estimate absolute numbers of cellular RNAs from an RNA-seq experiment, doing so requires spike-ins of known concentration and estimates of total cellular RNA content (Mortazavi *et al.* 2008; Lin *et al.* 2012). However, many RNA-seq experiments do not do these controls, nor are such controls strictly necessary under reasonable, though often untested, assumptions of approximately constant RNA content. While ultimately absolute concentrations will be necessary to fully predict properties such as noise tolerance of the regulatory circuits (Gregor, Tank, *et al.* 2007; Gregor, Bialek, *et al.* 2005), many current modeling efforts rely only on scaled concentration measurements, often derived from *in situ*-hybridization experiments (M. Garcia *et al.* 2013; Ilsley *et al.* 2013; He *et al.* 2010). Given that, we felt it was not important that different protocols should necessarily agree on any particular expression value for a given gene, nor are we fully convinced that absolute expression of any particular gene can truly reliably be predicted in a particular experiment.

In order to convince ourselves that data generated from limiting samples would be suitable for evaluating the spatial distribution of gene expression or other experiments where a linear response is necessary for proper interpretation of the data, we evaluated several protocols for performing RNA-seq on extremely small samples. We also investigated a simple modification to one of the protocols that reduced sample preparation cost per library by more than 2-fold. This study provides a single, consistent comparison of these diverse approaches, and

shows that in fact all data from the low-volume protocols we examined are usable in similar contexts to the earlier bulk approach.

3.3 Methods

RNA Extraction, Library Preparation, and Sequencing

We performed RNA extraction in TRIzol (Life Technologies, Grand Island, NY) according to manufacturer instructions, except with a higher concentration of glycogen as carrier (20 ng) and a higher relative volume of TRIzol to the expected material (1 mL, as in Lott *et al.* (2011) and Combs & Eisen (2013)). We quantified RNA concentrations using a fluorometric Qubit RNA HS assay (Life Technologies).

TruSeq libraries were prepared with the "TruSeq RNA Sample Preparation Kit v2" (Illumina Cat.#RS-122-2001) according to manufacturer instructions, except for the following modifications. All reactions were performed in half the volume of reagents. We find that this increases the effective concentration of RNA and cDNA. We performed all reactions and cleanups in 8-tube PCR strip tubes, which allowed us to reduce the volume of Resuspension Buffer to minimize volume left behind after each cleanup.

Clontech libraries were prepared with the "Low Input Library Prep Kit" (Clontech Cat.#634947). We generated cDNA by using TruSeq reagents until the cDNA synthesis step. Then, we used the Low Input Library Prep Kit to modify the cDNA into sequencing-competent libraries. We assume that a similar cDNA synthesis could be performed using oligo dT Dynabeads, RNA fragmentation reagents, and Superscript II (Life Technologies), for an approximate cost per sample of $15, but have not directly tested this with the Clontech reagents.

TotalScript libraries were prepared with the "TotalScript RNA-Seq Kit" and "TotalScript Index Kit" (Epicentre Cat.#TSRNA1296 and TSIDX12910). We followed the manufacturer's instructions, and used the oligo dT priming option. We performed the mixed priming option in parallel, which yielded approximately 4-fold more library, but did not sequence them due to concerns of ribosomal contamination.

SMARTseq2 libraries were prepared according to the protocol in Picelli, Faridani, *et al.* 2014. Because we had already extracted and mixed the RNA, we began at step 5 with 3.7 μL of dNTPs and 1 μL of 37 μM oligo dT primer, yielding the same concentration of primer and oligo as originally reported. We used 18 cycles for the preamplification PCR in step 14, added 1ng of cDNA to the Nextera XT reactions in step 28, and used 6 and 8 cycles for the final enrichment in step 33 (experiments 2 and 3, respectively).

Libraries were quantified using a combination of Qubit High Sensitivity DNA (Life Technologies) and Bioanalyzer (Agilent Technologies, Sunnyvale, CA) readings. Total yield Y in $femtomoles$ was estimated using Qubit concentration C measured in $(ng/\mu L)$, total volume V in μL, average size S in bp,

$$Y[\text{fmole}] = C[ng/\mu L] \cdot 10^{-9}[g/ng] \cdot V[\mu L] \cdot 10^{15}[\text{fmole/mole}] \div 608.9[g/\text{mole}] \div S[bp] \quad (3.1)$$

We then pooled libraries to equalize index concentration before sequencing.

Due to a pooling error in experiment 2 where non-concentration normalized tubes were mistakenly used instead of the normalized samples, the TruSeq libraries were included at much higher abundance. Pooled libraries were then submitted to the Vincent Coates Genome Sequencing Laboratory for 50bp single-end sequencing according to standard protocols for the Illumina HiSeq 2500. Bases were called using HiSeq Control Software v1.8 and Real Time Analysis v2.8.

Mapping and Quantification

Reads were mapped using STAR (Dobin *et al.* 2013) to a combination of the FlyBase reference genome version 5.54 for *D. melanogaster* and *D. virilis* (McQuilton *et al.* 2012). We randomly sampled the mapped reads to use an equal number in each sample compared. We used HTSeq (command line options `htseq-count --idattr='gene_name' --stranded=no --sorted=pos`) to count absolute read abundance per gene (Anders *et al.* 2014).

Simulation of Experiment 2

We wrote a Python script that simulated Experiment 2 assuming only uncorrelated counting noise in the number of reads per gene. The read counts from the sample with 20% *D. virilis* and the TruSeq protocol was used to generate the base probabilities. *D. virilis* gene probabilities were adjusted downwards, and the remaining probability was assigned evenly to the *D. melanogaster* genes. The `SciPy` function `stats.multinomial` was used to simulate read counts, assuming an equal number of reads as in the original experiment. Gene expression levels were normalized using equation 3.2, as in the actual experiment.

3.4 Results

Experiment 1: Evaluation of Illumina TruSeq

In our hands, the Illumina TruSeq protocol has performed extremely reliably with samples on the scale of 100ng of total RNA, the manufacturer recommended lower limit of the protocol. However, attempts to create libraries from much smaller samples yielded low complexity libraries, corresponding to as much as 30-fold PCR duplication of fragments (Appendix A). Anecdotally, less than 5% of libraries made with at least 90ng of total RNA yielded abnormally low concentrations, which we observed correlated with low complexity (Data not shown). To determine the lower limit of input needed to reliably produce libraries, we attempted to make libraries from 40, 50, 60, 70, and 80 ng of *Drosophila* total RNA, each in triplicate.

We considered the two libraries with lower than usual concentration to be failures. Although there is detectable material post-amplification, the size distributions as measured by

Table 3.1. Total TruSeq cDNA library yields made with a given amount of input total RNA. Yields measured by Nanodrop of cDNA libraries resuspended in $25\mu L$ of EB. The italicized samples were near the lower limit of detection, and when analyzed with a Bioanalyzer, showed abnormal size distribution of cDNA fragments.

Amount Input RNA	Replicate A	Replicate B	Replicate C
40 ng	*57 ng*	425 ng	672 ng
50 ng	435 ng	768 ng	755 ng
60 ng	*115 ng*	663 ng	668 ng
70 ng	300 ng	593 ng	653 ng
80 ng	468 ng	550 ng	840 ng

Bioanalyzer of these libraries is significantly different than known good libraries and manufacturer provided documentation (Supplemental Figure S1). In our experience, sequencing libraries with much lower than usual yield and abnormal size distributions has yielded libraries with low complexity and poor correlation to replicates.

While a failure rate of approximately 1 in 3 might be acceptable for some purposes, we ultimately wanted to perform RNA sequencing on precious samples, where a failure in any one of a dozen or more libraries would necessitate regenerating all of the libraries. Furthermore, due to the low sample volumes involved (less than approximately 500pg of poly-adenylated mRNA), common laboratory equipment is not able to determine the particular point in the protocol where the failures occurred.

It is clearly possible to use less than the manufacturer suggested amount of input RNA. Thus, we consider 70 ng of total RNA to be the conservative lower limit to the protocol. While this is about 30% smaller than the manufacturer suggests, it is still several orders of magnitude larger than we needed it to be. We therefore considered using other small-volume and "single-cell" RNA-seq kits, which often use a pre-amplification step that is known to influence estimation of absolute levels (Picelli, Björklund, *et al.* 2013).

Experiment 2: Competitive Comparison of Low-volume RNAseq protocols

We first sought to determine whether the low-volume RNAseq protocols available faithfully recapitulate linear changes in abundance of known inputs, even if absolute levels are not directly comparable to other protocols. We generated synthetic spike-ins by combining *D. melanogaster* and *D. virilis* total RNA in known, predefined proportions of 0, 5, 10, and 20% *D. virilis* RNA. For each of the low-volume protocols, we used 1ng of total RNA as input, whereas for the TruSeq protocol we used 100ng.

Although pre-defined mixes of spike-in controls have been developed and are commercially available (Jiang *et al.* 2011), we felt it was important to ensure that a given protocol would

function reproducibly with natural RNA, which almost certainly has a different distribution of 6-mers, which could conceivably affect random cDNA priming and other amplification effects. Furthermore, our spike-in sample more densely covers the approximately 10^5 fold coverage typical of RNA abundances. It should be noted, however, that our sample is not directly comparable to any other standards, nor is the material of known strandedness. We assumed that the majority of each sample is from the standard annotated transcripts, but did not verify this prior to library construction and sequencing.

We then estimated yield by measuring concentration in ng/μL with a Qubit High Sensitivity DNA assay and average fragment size with a Bioanalyzer High Sensitivity chip. The different protocols had a variation in yield of libraries from between 6 fmole (approximately 3.6 trillion molecules) and 2,400 femtomoles, with the TruSeq a clear outlier at the high end of the range, and the other protocols all below 200 fmole (Table 3). While the number of PCR cycles in the final enrichment steps can be adjusted, all of these quantities are sufficient to generate hundreds of millions of reads—far more than is typically required for an RNA-seq experiment. We pooled the samples, attempting equimolar fractions in the final pool; however, due to a pooling error, we generated significantly more reads than intended for the TruSeq protocol, and correspondingly fewer in the other protocols. Unless otherwise noted, we therefore sub-sampled the mapped reads to the lowest number of mapped reads in any library in order to provide a fair comparison between protocols.

We were interested in the fold-change of each *D. virilis* gene across the four libraries, rather than the absolute abundance of any particular gene. Therefore, after mapping and gene quantification, we normalized the abundance A_{ij} of every gene i across the $j = 4$ libraries by a weighted average of the quantity Q_j of *D. virilis* in library j, as show in equation 3.2. Thus, within a given gene, a linear fit of $\hat{A}_{ij}$ vs Q_j should have a slope of one and an intercept of zero. As expected, this normalized abundance increased with increasing *D. virilis* concentration (Supplemental Figure S2).

$$\hat{A}_{ij} = A_{ij} \div \frac{\sum_j Q_j A_{ij}}{\sum_j (Q_j)^2} \tag{3.2}$$

We then filtered the *D. virilis* genes for those with at least 20 mapped fragments in the library with 20% *D. virilis*, then calculated an independent linear regression and the Pearson correlation coefficient between the expected and measured concentration of *D. virilis* for each of those genes. As shown in Figure 1A and Supplemental Figure S3, this can be thought of as plotting the measured and known values, then fitting a line for each gene. As expected, for every protocol, the mean slope was 1 (t-test, $p < 5 \times 10^{-7}$ for all protocols). Similarly, the average intercepts for all protocols was 0 (t-test, $p < 5 \times 10^{-7}$ for all protocols). Also unsurprisingly, the TruSeq protocol had a noticeably higher mean correlation coefficient (0.98 ± 0.02) than any of the other protocols (0.95 ± 0.06, 0.92 ± 0.09, and 0.95 ± 0.06 for Clontech, TotalScript, and SMART-seq2, respectively). The mean correlation coefficient was statistically and practically indistinguishable between the Clontech libraries and the SMART-seq2 libraries (t-test $p = .11$, Figure 1). Taken together, all of these measures

indicate that the TruSeq protocol is better able to capture the linear trend in increasing transcript number.

While the TruSeq protocol clearly performed better than the low-volume kits (Figure 1), we wondered how well an ideal RNA-seq protocol could perform. We simulated an experiment with known levels of *D. virilis* spike in and assuming a multinomial distribution of read counts, and repeated the simulation 1,000 times to estimate the distribution of relevant quality metrics (Supplemental Figure S4). Surprisingly, the mean correlation coefficient for the TruSeq protocol was higher than the mean correlation coefficient of every repetition of the simulation, though indistinguishable for practical purposes (0.984 vs 0.982). The slopes were equally well clustered around 1, with an interquartile range of 0.0864 for the TruSeq protocol compared to 0.0843, the mean of all simulations; 13% of simulations had a higher IQR. We thus conclude that the major limiting factor for the TruSeq protocol to generate a linear response in the data is likely the sequencing depth, whereas the other protocols all contain additional biases.

Although there is some variation in the precise shape of the distributions of fit parameters, these were relatively small compared to the difference between any of these and the conventional TruSeq protocol. Indeed, the major differentiator we found among the low-volume protocols we compared was cost. For only a handful of libraries, the kit-based all inclusive model of the Clontech and TotalScript kits could be a significant benefit, allowing the purchase of only as much of the reagents as required. By contrast, the Smart-seq2 protocol requires the a la carte purchase of a number of reagents, some of which are not available or more expensive per unit for smaller quantities. Furthermore, there could potentially be a "hot dogs and buns" problem, where reagents are sold in non-integer multiples of each other, leading to leftovers. Many of these reagents are not single-purpose, however, so leftovers could in principle be repurposed in other experiments.

Table 3.2. Summary of protocols used in experiments 2 and 3. Cost is estimated per library assuming a enough libraries to consume all reagents at US catalog prices as of May 2014, and includes $2 for TRIzol RNA extraction, but not experimenter labor, sample QC, labware, or sequencing. Difference in prices in the Smart-seq2 protocols entirely due to scaling in cost of Nextera reagents.

Protocol	Shorthand	Cost/library
TruSeq	TruS	$45
Clontech	CT	$105
Epicentre TotalScript	TotS	$115
Smart-seq2, standard protocol	SS	$55
Smart-seq2, 2.5 fold dilution	SS—2.5×	$28
Smart-seq2, 5 fold dilution	SS—5×	$20

Table 3.3. Sequencing summary statistics for libraries. Protocols are the shorthands used in table 2. Reads indicates the total number of reads, and Mapped the total number of reads that mapped at least once to either genome. Experiments 2 and 3 were run in a single HiSeq lane each. Yield estimates were generated by adjusting Qubit High Sensitivity DNA readings by the average fragment size as measured by Bioanalyzer.

Expt	Protocol	% *D. virilis*	Yield	Total Reads	Mapped Reads
2	CT	0%	6.5 fmole	3,803,843	3,374,520 (89%)
2	”	5%	15.7 fmole	4,372,738	4,164,781 (95%)
2	”	10%	47.4 fmole	10,013,087	9,527,023 (95%)
2	”	20%	17.8 fmole	4,781,463	4,317,101 (90%)
2	TotS	0%	176.8 fmole	3,281,134	2,930,058 (89%)
2	”	5%	170.2 fmole	2,498,134	2,237,330 (90%)
2	”	10%	102.5 fmole	5,777,523	5,424,366 (94%)
2	”	20%	119.9 fmole	6,068,996	5,740,496 (95%)
2	TruS	0%	2,401.0 fmole	67,560,511	64,024,881 (95%)
2	”	5%	2,001.1 fmole	23,370,854	22,589,083 (97%)
2	”	10%	2,174.2 fmole	39,454,390	38,093,763 (97%)
2	”	20%	2,379.2 fmole	35,265,536	34,304,792 (97%)
2	SS	0%	34.3 fmole	2,439,518	2,297,087 (94%)
2	”	5%	59.6 fmole	2,550,023	2,419,889 (95%)
2	”	10%	67.9 fmole	2,534,628	2,444,568 (96%)
2	”	20%	39.8 fmole	2,504,340	2,389,850 (95%)
3	SS—2.5×	0%	104.4 fmole	15,769,915	14,393,959 (91%)
3	”	1%	124.7 fmole	21,349,748	20,084,131 (94%)
3	”	5%	113.0 fmole	17,047,120	16,329,641 (96%)
3	”	10%	103.5 fmole	23,762,232	22,372,562 (94%)
3	”	20%	123.8 fmole	20,809,781	20,041,548 (96%)
3	SS—5×	0%	59.4 fmole	19,214,155	17,324,598 (90%)
3	”	1%	58.6 fmole	23,832,274	22,364,220 (94%)
3	”	5%	65.4 fmole	18,149,452	17,157,450 (95%)
3	”	10%	28.8 fmole	15,821,419	14,869,864 (94%)
3	”	20%	57.2 fmole	22,466,345	21,620,603 (96%)

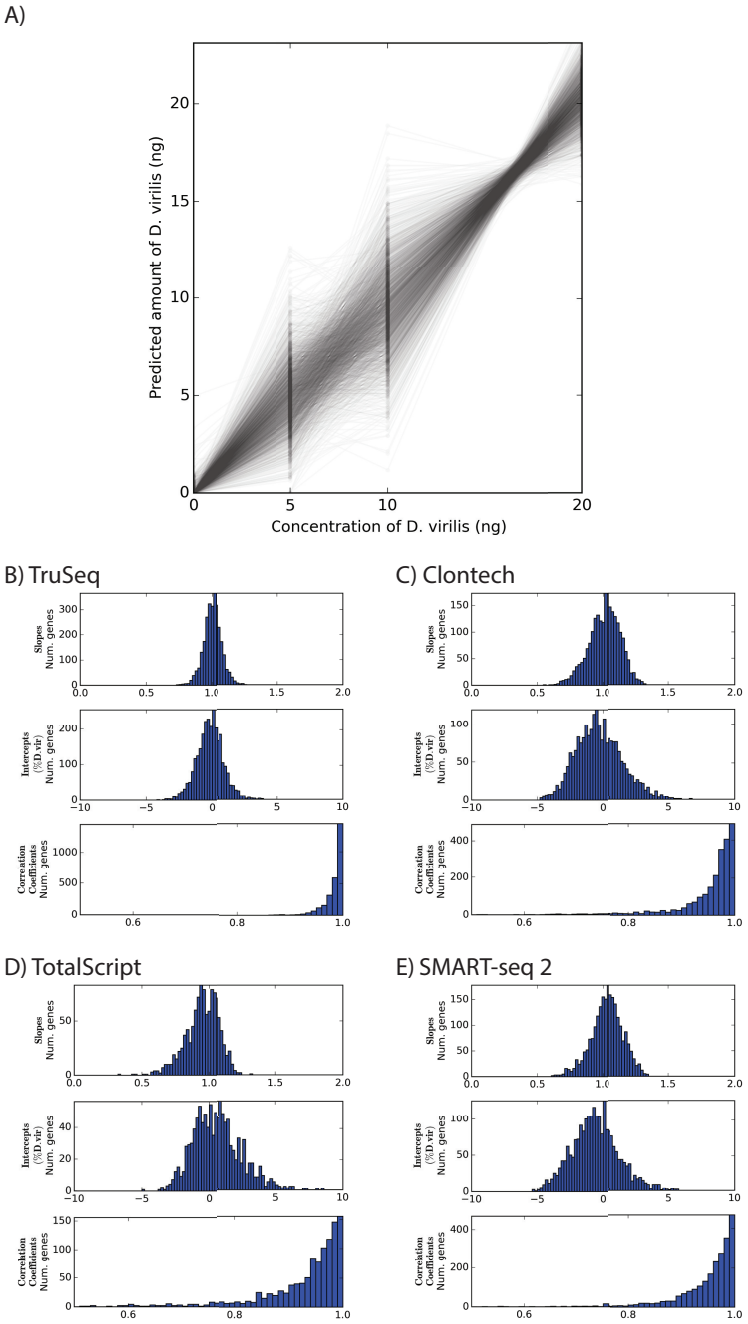


Figure 3.1. Comparison of linearity between different RNA-seq protocols. A) Normalized levels of gene expression $\hat{A}$ across libraries using the TruSeq protocol, where each line is for a different gene. B-E) Distributions of slopes, intercepts, and correlation coefficient for linear regressions of the abundance of each gene, as in panel A.

Table 3.4. Distribution of fit parameters. A simple linear fit, $\hat{A}_{ij} = m \cdot Q_j + b$ was computed for each gene i, and a correlation coefficent r calculated. For brevity, $\bar{x}$ is the mean of some variable x, and σ_x is its standard deviation.

Protocol	$\bar{m}\pm\sigma_m$	$\bar{b}\pm\sigma_b$	$\bar{r}\pm\sigma_r$
TruSeq	1.01±0.0698	-0.108±1.05	0.98±0.019
Clontech	1.01±0.12	-0.217±1.79	0.95±0.061
Epicentre TotalScript	0.952±0.129	0.715±1.93	0.93±0.094
Smart-seq2	1.03±0.121	-0.506±1.82	0.95±0.057
Smart-seq2, 2.5 fold dilution	0.996±0.111	0.0623±1.67	0.96±0.053
Smart-seq2, 5 fold dilution	1.01±0.111	-0.173±1.66	0.96±0.049

Experiment 3: Further modifications to the SMART-seq2 protocol

Although the SMART-seq2 was the cheapest of the protocols when amortized over a large number of libraries, we wondered whether it could be performed even more cheaply without compromising data quality. This would enable us to include more biological replicates in the future experiments for which we are evaluating these protocols. In the original protocol, we noticed that roughly 60% of the cost came from the Nextera XT reagents. Thus, reducing the cost of tagmentation was the obvious goal to target.

We made additional libraries, again starting with 1ng of total RNA. We amplified a single set of spike-in libraries with 0, 5, 10, and 20% *D. virilis* total RNA as in experiment 2, and made a single an additional sample with 1% *D. virilis* RNA. Starting at the point in the SMART-seq2 protocol where tagmentation was started, we performed reactions in volumes 2.5× and 5× smaller, using proportionally less cDNA as well. Due to the low total yield, we increased the number of enrichment cycles from 6 to 8 (see methods).

When normalized to the same number of reads as in experiment 2, the protocols with diluted Nextera reagents performed effectively identically: for instance, the mean correlation coefficients were in both cases 0.96 ± 0.05 (Fig. 2 and Table 4). This is despite the additional cycles of enrichment, which improved yield.

Because we used a common set of pre-amplified cDNA samples that was performed in a distinct pre-amplification from experiment 2, we can estimate the contribution of that pre-amplification to the overall variation. If, in fact, the pre-amplification is a major contributor to the variation, then we would expect to find that the correlation between, for instance, the slopes of two runs of the same experiment with different pre-amplifications would be significantly lower than the correlation between the slopes of two runs using the same pre-amplified cDNA pools.

Unsurprisingly, the sets of samples that used the same preamplification were more correlated with each other than with the set of samples that used a separate pre-amplification (Fig. 3). By analogy to dual-reporter expression studies such as Elowitz *et al.* (2002), we

term variation along the diagonal "extrinsic noise" ($\eta_{ext} = \text{std}(m_1 + m_2)$), and variation perpendicular to the diagonal "intrinsic noise" ($\eta_{int} = \text{std}(m_1 - m_2)$), being intrinsic to the pre-amplification step. Using that metric, the intrinsic noise is lower for the samples with the same pre-amplification ($\eta_{int} = 0.09$) than for the samples with different pre-amplifications ($\eta_{int} = 0.16$). Somewhat surprisingly, the extrinsic noise is higher for the samples with the same pre-amplification ($\eta_{ext} = 0.20$ vs $\eta_{ext} = 0.16$), perhaps due to the 2 additional cycles of PCR enrichment.

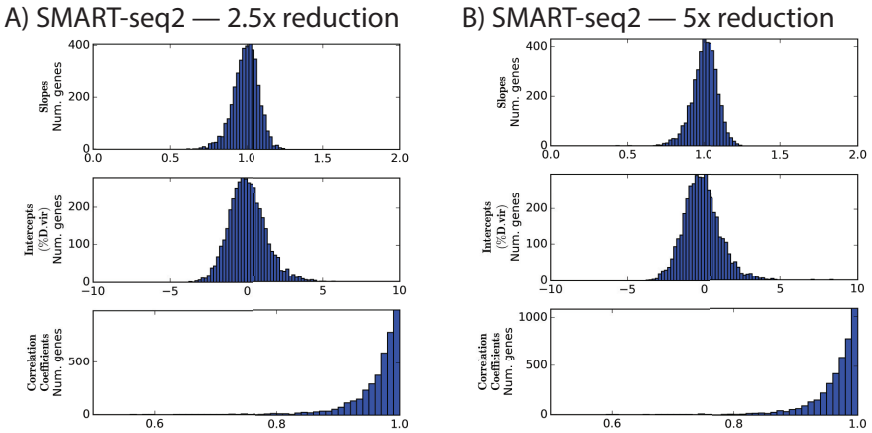


Figure 3.2. Distributions of slopes, intercepts, and correlation coefficients for experiment 3. Nextera XT reactions were reduced in volume by the indicated amount.

3.5 Discussion

When sample size is not the limiting factor, it is clear that using well-established protocols that involve minimal sequence-specific manipulation of the sample yields the best results, both in terms of reproducibility and linearity of response. However, if it is not practical to collect such relatively large samples, experiment 2 shows that any of the "single-cell" protocols we have tested should perform similarly to each other, and can be used as a drop-in replacement. While preamplification steps do introduce some detectable variance, it is not vastly detrimental to the data quality, and does not introduce obvious sequence-specific biases.

Such methods should be strongly preferred if it is feasible to collect a suitably homogenous sample. While bulk tissues may be a mixture of multiple distinct cell types, this may or may not affect the particular research question an RNAseq experiment is designed to answer. In our hands, the lower limit of reliable library construction using the Illumina TruSeq kit is approximately 70ng of total RNA and we have used this amount of RNA in as-yet unpublished experiments on dissected slices of embryos. With non precious samples, the

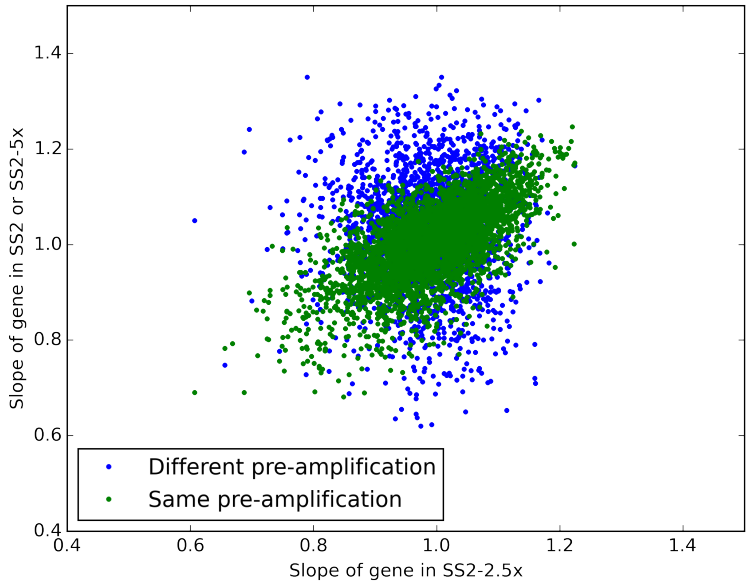


Figure 3.3. Estimating the source of preamplification noise. Plotted are the estimated slopes for each gene between experiments. The blue, "Different pre-amplification" compares the 2.5× diluted and full sized reactions, whereas the green "same pre-amplification" points compare the 2.5× and 5× dilution samples, which used the same preamplified cDNA but different tagmentation reactions.

practical limit is likely to be even lower. The manufacturer suggested 100ng is almost certainly safe, and we can think of relatively few experiments where it is not practical to collect more RNA than this. Although we have anecdotally observed significant user-to-user variation within our lab, it seems unreasonable to expect order-of-magnitude improvements are possible in techniques for precious samples. We suggest that this limit may be related to cDNA binding to tubes or purification beads, but since the quantities are lower than the detection threshold of many standard quality control approaches, we cannot directly verify this.

Compared to the regimes these protocols were designed for, we used a relatively large amount of input RNA—1 ng of total RNA—corresponding to approximately 50 nuclei of a mid-blastula transition *Drosophila* embryo. Previous studies have shown that this amount of RNA is well above the level where stochastic variation in the number of mRNAs per cell will strongly affect the measured expression of a vast majority of genes (Marinov *et al.* 2013). It is nevertheless a small enough quantity to be experimentally relevant. For instance, we have previously dissected single embryos into approximately 12 sections, yielding approximately 10ng per section(Combs & Eisen 2013), and one could conceivably perform similar experiments on imaginal discs or antennal structures, which contain a similar amount

of cells (Klebes *et al.* 2002; Hansson & Anton 2000).

One of the more striking results is that costs can be significantly reduced by simply performing smaller reactions, without noticeably degrading data quality. We do not suspect this will be true for arbitrarily small samples, such as from single cells. Instead, it is likely only true for samples near the high end of the effective range of the protocol. We have not explored where this result breaks down, and strongly caution others to verify this independently using small pilot experiments before scaling up.

3.6 Conclusions

The selection of protocols for performing RNAseq depends on the amount of material available to be profiled. We found that high quality libraries can be generated with slightly less than the manufacturer's recommended minimum using the standard Illumina TruSeq protocol. When sufficient material is available to use the TruSeq protocol, we find that this produces data with a better linear response to the increasing concentration of any given gene than a number of different "single cell" protocols, which have roughly comparable performance in this metric. Finally, we found that at least one of these protocols, SMART-seq2 could be easily modified to significantly reduce the cost of library preparation, without compromising data quality.

3.7 Acknowledgements

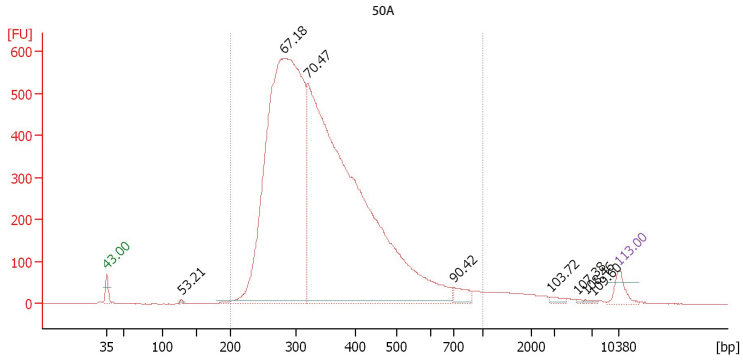
We would like to thank Lior Pachter for suggesting the simulation experiment, and reviewers Angela Wu and Scott Hunicke-Smith for their detailed comments and criticisms.

Data Deposition

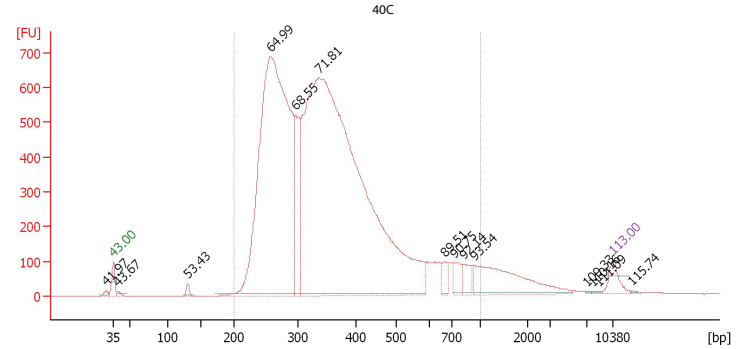
We have deposited all reads in the NCBI GEO under the accession number GSE64673. The processed data is available at http://eisenlab.org/lowvolume. All custom analysis software is available at https://github.com/eisenlab/SliceSeq, and is primarily written in Python (Van Rossum & Drake 2003; Cock *et al.* 2009; Hunter 2007; Jones *et al.* 2001; Perez & Granger 2007). Commit `9fc810e7` was used to perform all analyses in this paper.

3.8 Supplemental Material

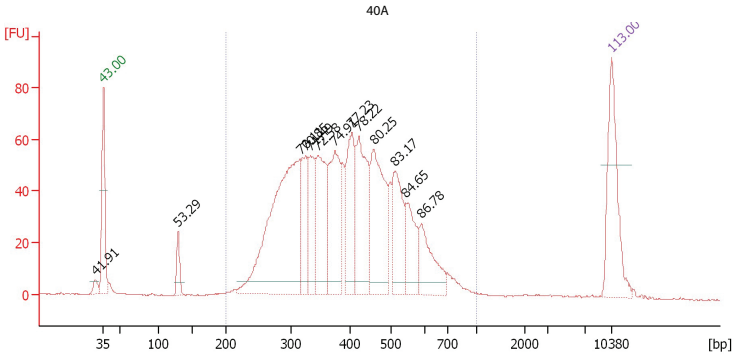
Appendix A Estimating complexity of RNAseq libraries.



Typical library distribution. The peak size can be adjusted somewhat with varying fragmentation time.

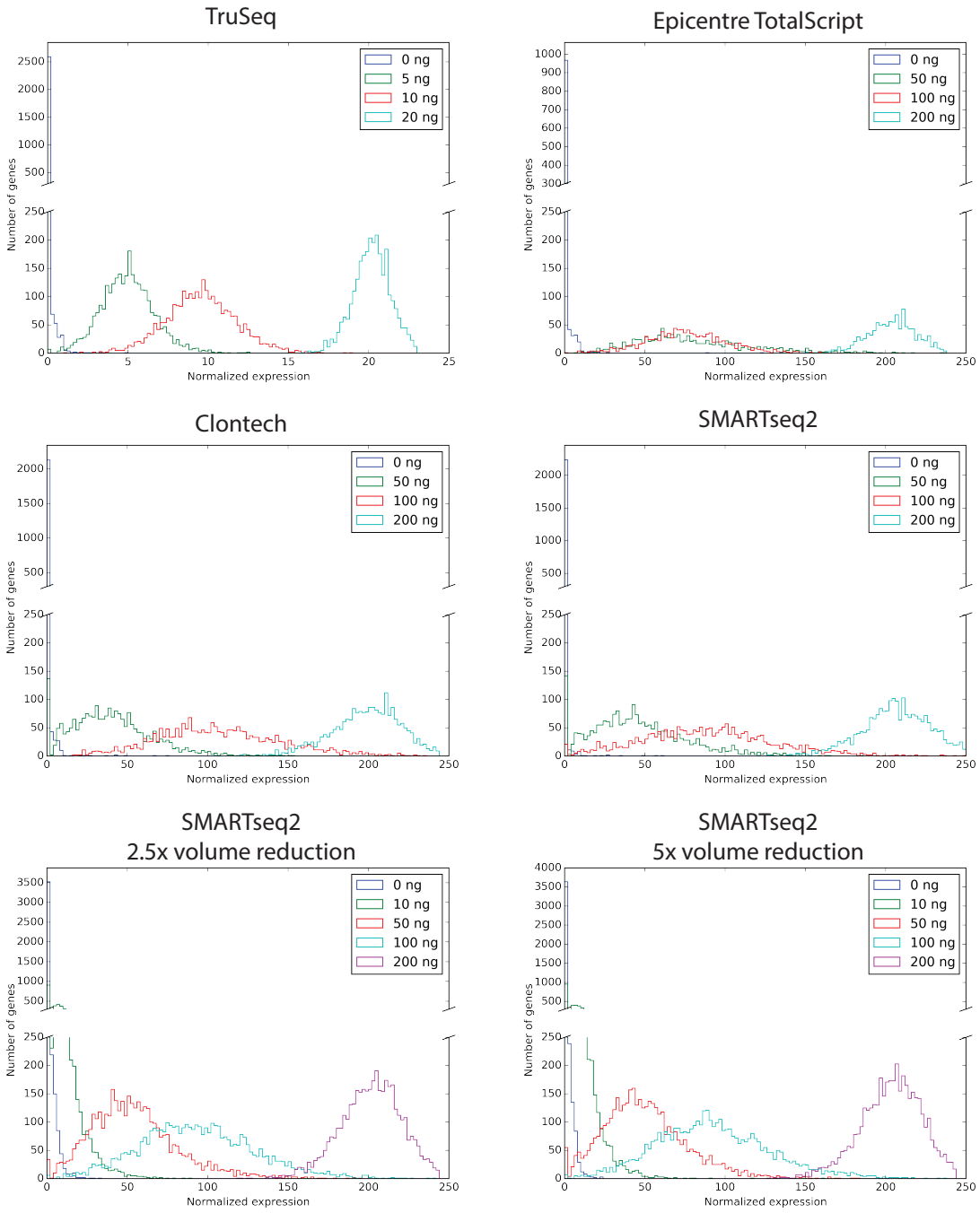


Also a reasonably good library. The secondary peak around 350bp seems to be an overloading artifact, and can be corrected by dilution prior to running the bioanalyzer.

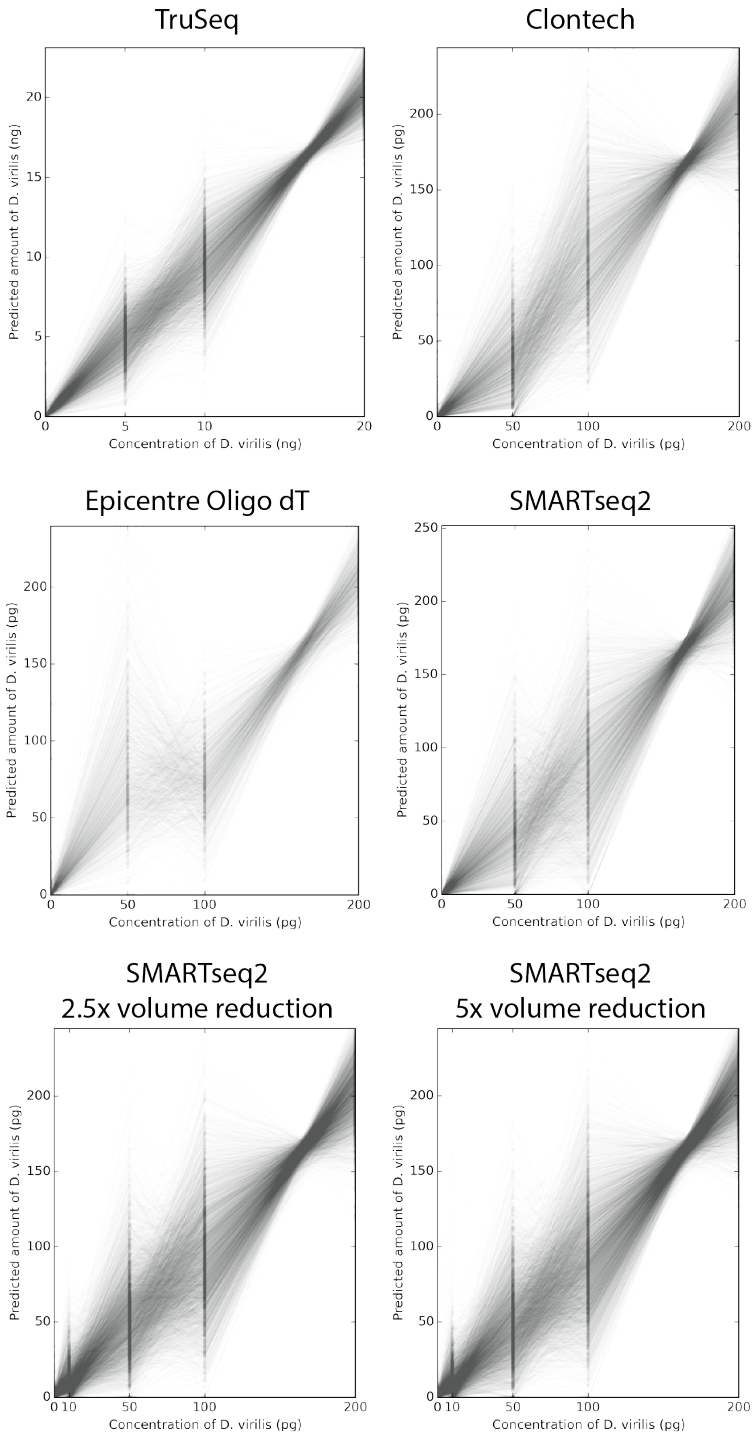


Poor library. Note the much lower concentration (compared to marker peaks), more rounded shape, and jagged peaks at the larger end of the distribution.

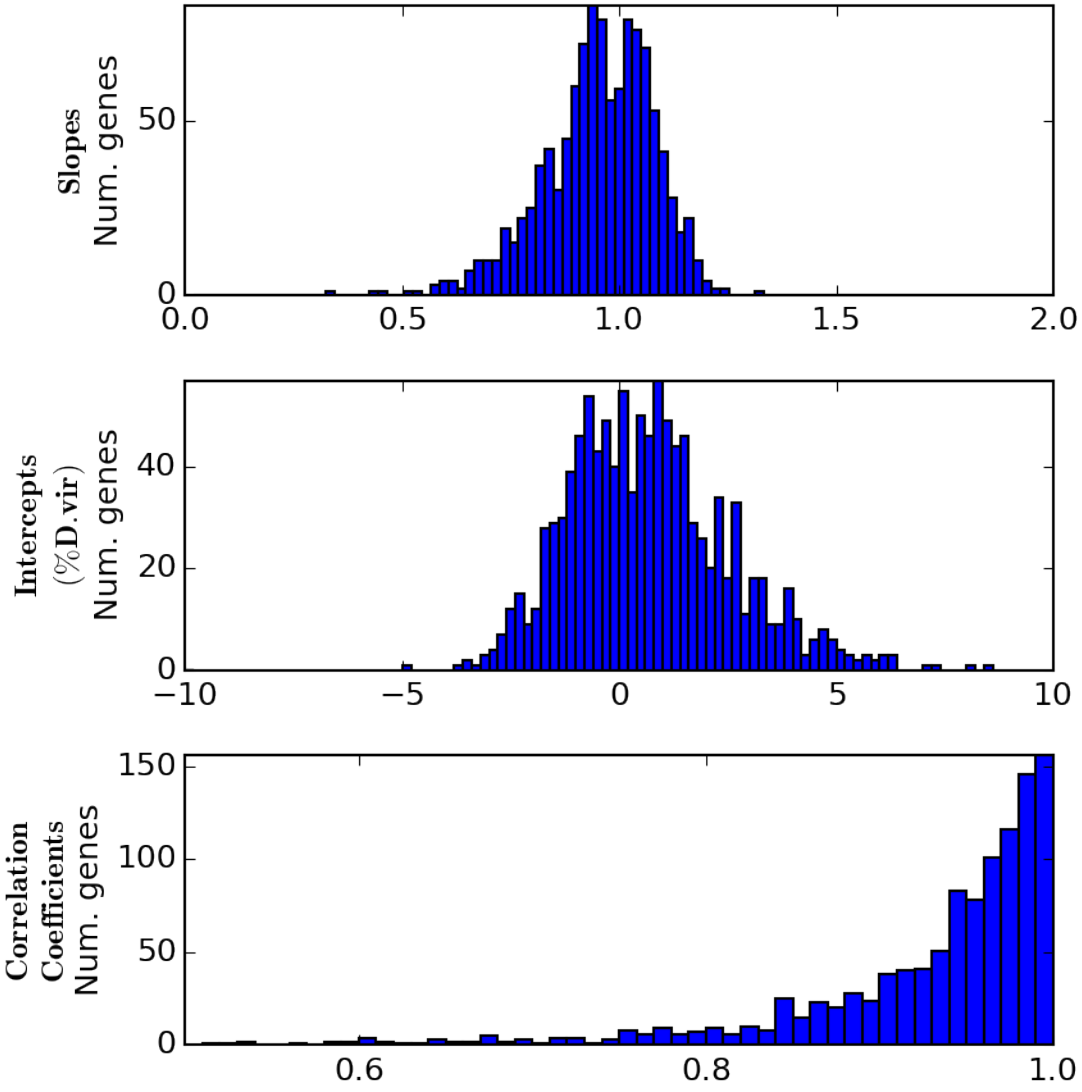
Supplemental Figure 3.S1. Representative Bioanalyzer traces of libraries with both good and abnormal size distributions.



Supplemental Figure 3.S2. Histograms of distributions of expression levels for all the protocols we used. Increasing the amount of *D. virilis* in each library increases the average abundance of *D. virilis* genes.



Supplemental Figure 3.S3. Similar to figure 1A, we have plotted the abundance of each gene individually for all of the protocols we used.



Supplemental Figure 3.S4. Fit parameters for one iteration of a multinomial simulation of Experiment 2. The distributions are practically indistinguishable from the corresponding distribution for the TruSeq data, in 1B.

Chapter 4

Genome-wide convergent phenotypes in diverse *Drosophila melanogaster* early patterning mutants

Abstract

Genome sequencing has become commonplace, but the understanding of how those genomes ultimately specify cell fate during development is still elusive. Extrapolating insights from deep investigation of a handful of developmentally important Drosophila genes to understanding the regulation of all genes is a major challenge. The developing embryo provides a unique opportunity to study the role of gene expression in pattern specification; the precise and consistent spatial positioning of key transcription factors essentially provides separate transcriptional-readout experiments at a critical point in development.

We cryosectioned and sequenced mRNA from single Drosophila melanogaster embryos at the blastoderm stage to screen for spatially-varying regulation of transcription. Expanding on our previous screening of wild type embryos, here we present data from dosage mutants for key maternally provided regulators, including depletion of zelda and hunchback and both over-expression and depletion of bicoid. These data recapitulate all of the expected patterning changes driven by these regulators; for instance, we show spatially-confined up-regulation of expression in the bicoid over-expression condition, and down-regulation of those genes in the bicoid knock-down case, consistent with bicoid's known function as an anterior-localized activator.

Our data highlight the role of combinatorial regulation of patterning gene expression. When comparing changes in multiple conditions, genes responsive to one mutation tend to respond to other mutations in a similar fashion. Furthermore, genes that respond differently to these mutations tend to have more complex patterns of TF binding.

Introduction

Proper animal development relies on complex, highly coordinated gene expression patterns in both space and time. In *Drosophila* (and many other long germ-band insects), this is achieved through nuclei nearly simultaneously reading out a number transcription factors (TFs) that have been maternally deposited in a spatially dependent manner, which establishes the dorsal-ventral and anterior-posterior axes. In many eukaryotes, this readout of regulatory signals is mediated through enhancers and other *cis*-regulatory elements that are located anywhere from several kilobases to megabases away from the promoters of the genes they regulate.

The classical reverse genetics approach involves mutating a gene, then investigating the phenotypic consequences of that mutation. This approach has been extremely successful in helping to elucidate the mechanisms of transcriptional regulation for a number of key loci. However, historically *in situ* hybridization has been used to assay these consequences. By necessity, previous studies have focused only on the patterning changes of a handful of genes, since assaying more genes is impractical. While these changes are often rationally selected—including known transcription factors that, in turn, drive other patterns—the scale of previous experiments means that they have only illustrated the common types of changes observed, rather than completely cataloging a given TF's effects.

For the TFs I have investigated in this study (*bicoid*, *hunchback*, and *zelda*), there are high quality chromatin immunoprecipitation (ChIP) experiments already in the literature for wild-type flies (MacArthur *et al.* 2009; Harrison *et al.* 2011). These experiments can suggest potential targets, but due to the large number of binding sites in the genome as well as the dense, interconnected regulatory network that can potentially buffer changes, simply performing *in situs* on every gene with a binding site is a prohibitively large experiment, and unlikely to perform radically better than chance. Strong binding is suggestive, but not conclusive, for functional effect. TFs bind to hundreds of sites throughout the genome, and can show strong ChIP signal at many hundreds more where no binding is found. The true gold-standard for assaying function of a particular binding site must therefore demonstrate some downstream effect that is different when that binding is removed.

In this study, I extend my previous method for sequencing mRNA from spatially restricted *Drosophila melanogaster* samples and apply it to a number lines that are mutant for TFs known to be important in establishing proper spatial patterns of expression. Our goal was to determine which patterned genes show distinct, TF dependent spatio-temporal regulation at the blastoderm stage. We identify a large number of genes with both expected and unexpected patterning changes, and through integrated analysis of previous ChIP data for a large number of anterior-posterior patterning factors, I highlight common themes in genes with complex dependence on these TFs.

Results

Genome-wide atlases of the blastoderm stage of multiple dosage mutants

I sliced embryos and sequenced the resulting mRNA from 4 mutant genotypes (fig. 4.1A): a *zld* germline clone, an RNAi knockdown for *bcd*, a knockdown for *hb*, and an overexpression line for *bcd* with approximately 2.4× wildtype expression. I chose two time points: cycle 13 (determined using nuclear density of either DAPI stained embryos or of the Histone-RFP present in the *zld* line) and mid-to-late cycle 14 (determined using 50-65% membrane invagination at stage 5) (fig. 4.1B). Genes expressed in cycle 13 are towards the end of the early round of genome activation and are enriched for *ZLD* binding (Tadros & Lipshitz 2009; Lott *et al.* 2011; Harrison *et al.* 2011), but are early enough that the majority of patterning disruptions are likely to be direct effects of the mutants. By contrast, I chose the stage 5 time point in order to highlight the full extent of the patterning changes across the network.

In order to show the range of patterning differences observed, I generated quilt plots of all the gene expression present in the dataset (Figure 4.2) (Wand *et al.* 2014). Of the 7104 genes with at least 15FPKM in at least one slice, approximately 3000 had uniform expression in all the wild-type embryos that was not greatly perturbed in any of the mutants. The total number of expressed genes is very consistent with previous estimates of the number of maternally deposited and zygotically transcribed genes (Lott *et al.* 2011).

The set of genes with anterior or posterior localization recapitulate the known literature (Staller, Fowlkes, *et al.* 2015) and general expectations in the *bcd-* case: those expressed in the anterior typically lose expression (Figure 4.3A), and those in the posterior also frequently gain an expression domain in the anterior (Figure 4.3B). Surprisingly, most of these patterns are qualitatively unaffected in the other mutants. In the absence of *zld*, most of these genes are able to retain the proper anterior patterning (although they may have differences in expression levels). Similarly, these genes seem not to be strongly dependent on maternal *hb* for patterning information, with most genes retaining a distinct anterior expression domain. As described in Liang *et al.* (2008), there are some genes that are normally ubiquitously expressed in the wild-type that become localized to the poles in the *zld-* embryo (Figure 4.3C).

Most spatial patterns are robust to mutation

I next compared expression patterns from each of the mutant lines at late cycle 14 to similar expression patterns in wild-type. Because there are a different number of slices both between the wild type and mutant flies, and between replicates of the mutant flies, I decided to use Earth Mover Distance (EMD) to compare patterns (Rubner *et al.* 1998). This metric captures intuitive notions about what kinds of patterns are dissimilar, yielding higher distances for dissimilar distributions of RNA, and zero for identical distributions. Patterns were normalized to have the same maximum expression, in order to highlight changes in positioning of patterns, rather than changes in absolute level. In contrast to traditional RNA-seq dif-

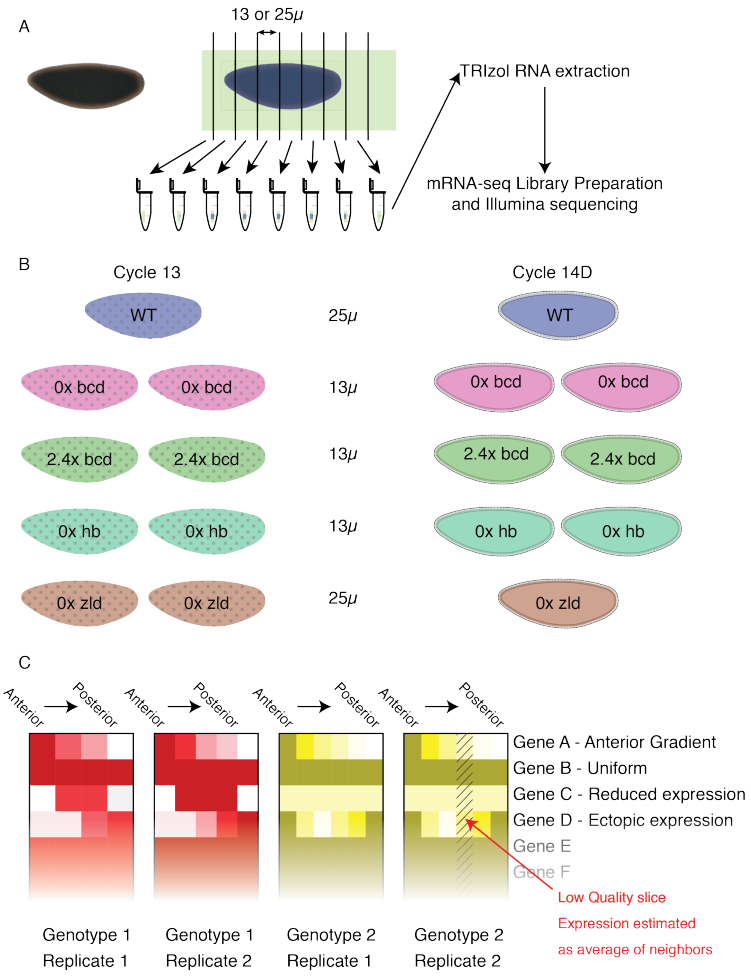
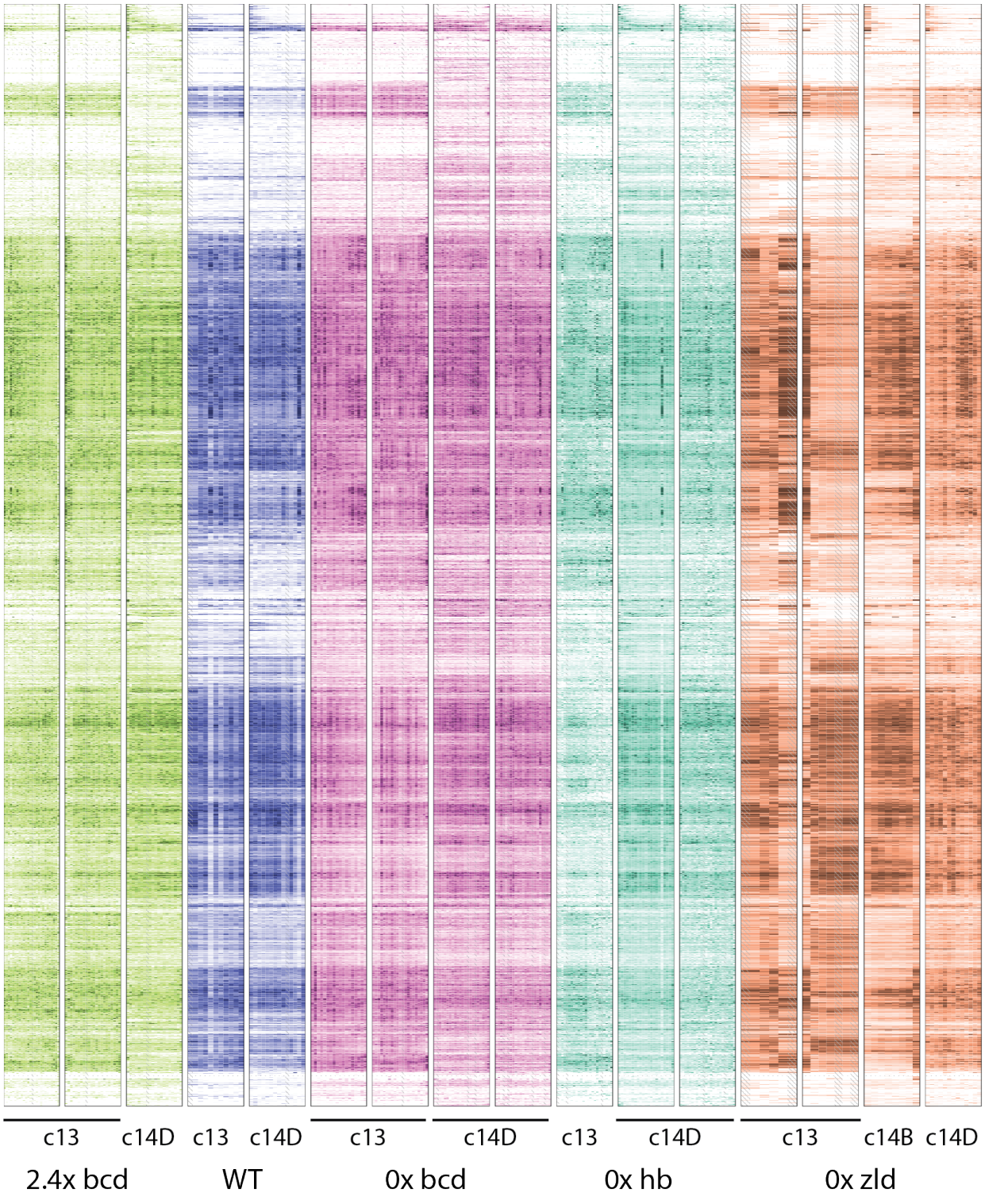


Figure 4.1. Schematic of experimental approach. A) We fixed embryos in methanol, then selected individual embryos at the correct stage, aligned them in sectioning cups, and sliced to the indicated thickness. We extracted RNA from individual slices, prepared barcoded libraries, then pooled them prior to sequencing. B) Overview of the mutant genotypes used. Two replicates per time point at two time points, based on nuclear density and morphology. C) Cartoon of heatmaps. Each genotype is assigned its own color (matching those in B), with darker colors representing higher expression and white representing no expression detected in that slice. Each boxed column represents a single individual, and within that column, slices are arranged posterior to the left and anterior to the right.


c13
c14D
c13
c14D
c13
c14D
c13
c14D
c13
c14B
c14D
2.4x bcd
WT
0x bcd
0x hb
0x zld

ferential expression metrics, this approach takes advantage of the spatial nature of the data, and with the fine slices, adjacent slices are able to function as "pseudo-replicates". Adjacent slices are, on average, much more similar than those from farther away in the same embryo (Figure 4.S1).

The overall level of divergence in pattern across all genes is, in most cases, slightly larger than when comparing nearby time-points in wild-type or replicates of the same genotype and time point (Figure 4.4). Notably, the *zld*- mutant is more similar to wild-type than the mutants of the other, spatially distributed transcription factors. This suggests that *zld* is a categorically different TF, consistent with its role as a pioneer factor rather than a direct activator. However, the low level of divergence is a reflection of the fact that the majority of genes are not dynamically expressed in either time or space.

In order to demonstrate that these mutants are more likely to affect already known *BCD* regulatory systems, I examined genes that were close to 64 *BCD* dependent enhancers previously identified (H. Chen *et al.* 2012; Ochoa-Espinosa *et al.* 2005; Schroeder *et al.* 2004; Biemar *et al.* 2005; Hartmann *et al.* 2001; Kantorovitz *et al.* 2009; Riddihough & Ish-Horowicz 1991). Although the bulk of these enhancers do not have validated associations with particular genes, I assumed that they would be relatively close to the genes that they drive. Of the 66 genes whose TSS's were the closest in either direction and within 10kb of the center of the tested CRM, only 32 were expressed at greater than 10FPKM in at least one slice of any of the wild-type embryos. Of these, only 10 had an obvious anterior localization bias (31%), with the majority of the rest being approximately uniformly expressed across the embryo (Figure 4.5). The majority of genes with ubiquitious or central localization did not radically change in either the bicoid overexpression or knockdown conditions. As expected, genes with anterior localization suffered a loss of patterning in the depletion mutant, and a posterior shift in the over-expression condition. I assume that genes that are not localized to the anterior are either driven by multiple enhancers, such that loss of expression from one does not severely affect the overall expression, or that they are merely close to the enhancer, but unrelated.

Figure 4.2 *(preceding page)*. Heatmaps of gene expression patterns for all expressed genes. Each individual embryo is represented by one boxed column in the heatmap. Within a column, slices are arranged anterior to the left, and posterior to the right. Each embryo is colored according to genotype, with green for the *bcd* over-expression, blue for wild-type, magenta for *bcd* knockdown, cyan for *hb* knockdown, and orange for the *zld* mutant. Within a genotype, darker colors correspond to higher expression and white to zero expression, on a linear scale normalized for each gene separately to the highest expression for that gene in the embryo or to 10FPKM, whichever is greater. Slices that did not match quality control standards are replaced by averaging the adjacent slices, and are marked with hash marks. Rows are arranged by using Earth Mover Distance to perform hierarchical clustering, so that genes with similar patterns across all of the embryos are usually close together.

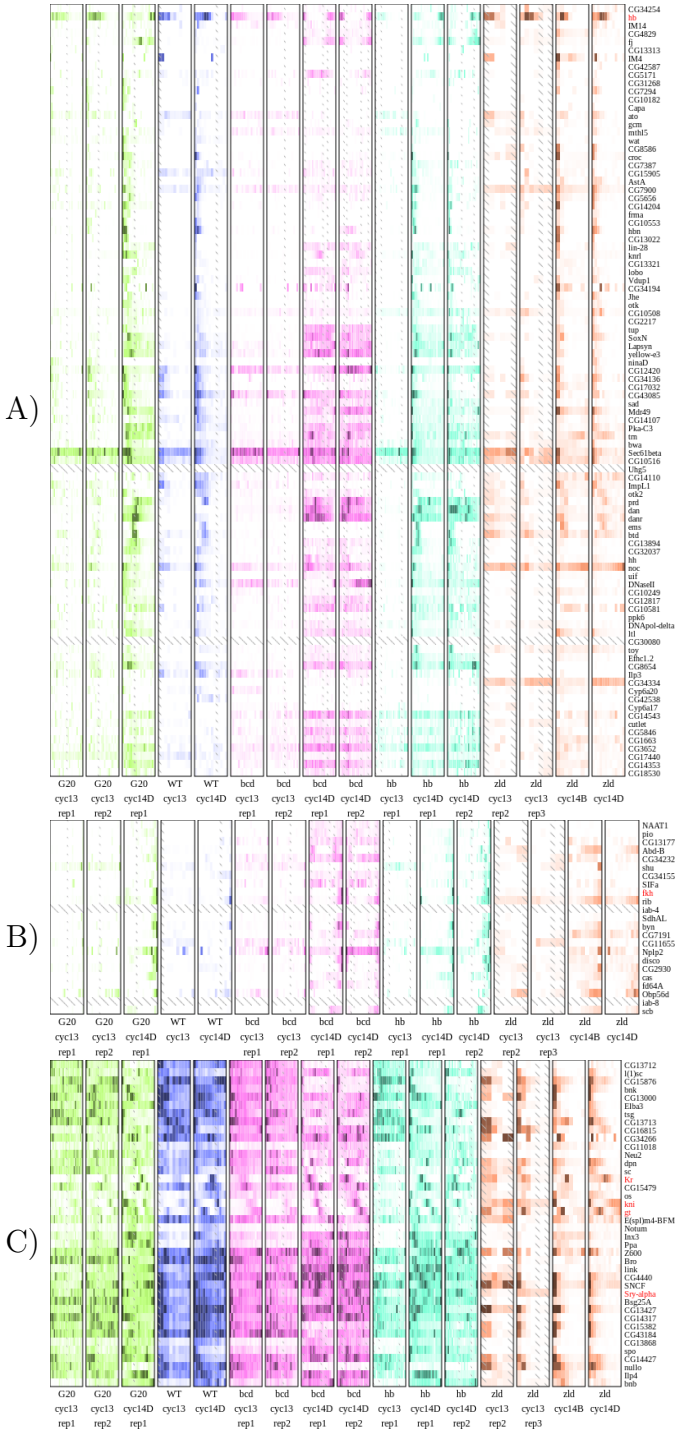

A)
B)
C)
G20 cyc13 rep1
G20 cyc13 rep2
G20 cyc14D rep1
WT cyc13
WT cyc14D
bcd cyc13 rep1
bcd cyc13 rep2
bcd cyc14D rep1
bcd cyc14D rep2
hb cyc13 rep1
hb cyc14D rep1
hb cyc14D rep2
zld cyc13 rep2
zld cyc13 rep3
zld cyc14B
zld cyc14D

Table 4.1. TF depletion is more likely to make a non-uniform pattern uniform than vice versa.

	Low Expr to Patterned	Patterned to Uniform	Patterned to Low Expr	Uniform to Patterned
g20	343	32	25	64
bcd	96	69	40	12
hb	348	36	20	43
zld	71	34	61	28

I measured EMD for each gene at cycle 14D in each genotype compared to a uniform distribution. I considered genes uniform if they had an EMD<0.04, and non-uniform if they have an EMD>0.08. I then considered genes with at least 15 FPKM in at least one slice in both wild-type and the mutant line.

Effects of TF depletion on patterned genes

I next sought to demonstrate that the technique of cryoslicing mutants is useful for identifying the effects of these early patterning genes. In comparison to fig. 4.3, where we looked for known patterning changes that we would expect from the literature, we also want to make sure that the largest and most common patterning changes that naturally arise from the data recapitulate the known literature. For each mutant genotype, we

Unsurprisingly, depletion of TFs known to be important for patterning are likely to make an otherwise non-uniform pattern more so (Table 4.1). Of the 465 genes that have clearly non-uniform patterning in the wild-type at cycle 14D, 12-20% are affected in each depletion mutant, either losing expression entirely or becoming uniform. The over-expression line is at the low end of this range, also at 12%.

However, this is not always simply abrogating expression—a large number of genes seem to have higher expression everywhere. In the case of *bcd* depletion, approximately a third of these cases are genes that are restricted to the anterior in wild-type that become approximately uniform throughout the embryo (Figure 4.7). While some of these are due to genes

Figure 4.3 *(preceding page)*. Heatmaps of gene expression patterns for anterior and posterior genes recapitulate expected patterning changes. Each individual embryo is represented by one boxed column in the heatmap. Within a column, slices are arranged anterior to the left, and posterior to the right. Each embryo is colored according to genotype, with green for the *bcd* over-expression, blue for wild-type, magenta for *bcd* knockdown, cyan for *hb* knockdown, and orange for the *zld* mutant. Within a genotype, darker colors correspond to higher expression and white to zero expression, on a linear scale normalized for each gene separately to the highest expression for that gene in the embryo or to 10FPKM, whichever is greater.

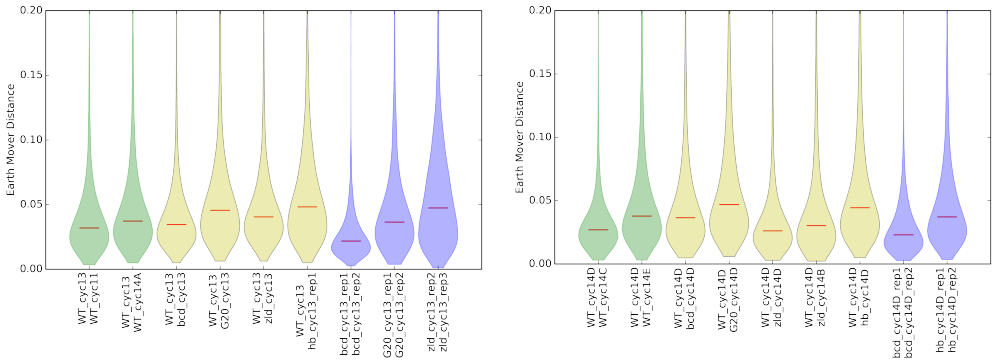


Figure 4.4. Distributions of patterning differences show that mutants have wide-spread subtle patterning effects and more genes with large patterning differences than replicates. Adjacent time points from the wild-type dataset in Combs & Eisen (2013) are colored green, and replicates of the same genotype and time point are colored blue. Median distances are marked in red.

Table 4.2. Patterning changes are strongly associated with increased TF binding.

	Low Expr to Patterned	Patterend to Uniform	Patterned to Low Expr	Uniform to Patterned
g20	bcd,gt,kni,hkb,tll	—	—	bcd
bcd	gt,kni,tll	bcd,cad,gt	—	bcd,kni
hb	bcd,gt,kni,tll	—	—	bcd
zld	—	—	bcd,cad,gt,kni,tll	bcd,cad,gt,kni,tll

Using the genes with identified patterning changes in Table 4.1, I performed a χ^2 test with a Bonferroni-corrected p-value of 0.05.

with an early uniform pattern that fails to properly resolve into spatially restricted domains, approximately half are true ectopic expression (Figure 4.S2).

As a first step to identifying likely regulatory motifs, I used binding data for 9 non-pair-rule AP TFs (MacArthur *et al.* 2009) and for *zld* (Harrison *et al.* 2011) to search for factors with differential rates of binding among the sets of genes with patterning changes (Table 4.2). This analysis highlights that *zelda* operates in a qualitatively different manner from the other transcription factors—in it's absence other TFs are likely to continue expression, though in abnormal patterns. Additionally, *zld* is crucial for maintaining patterned expression, as the most common change is from patterned genes integrating one or more AP factors to minimal overall expression.

Furthermore, *bicoid* stands out as a major factor involved in AP patterning. In all of the

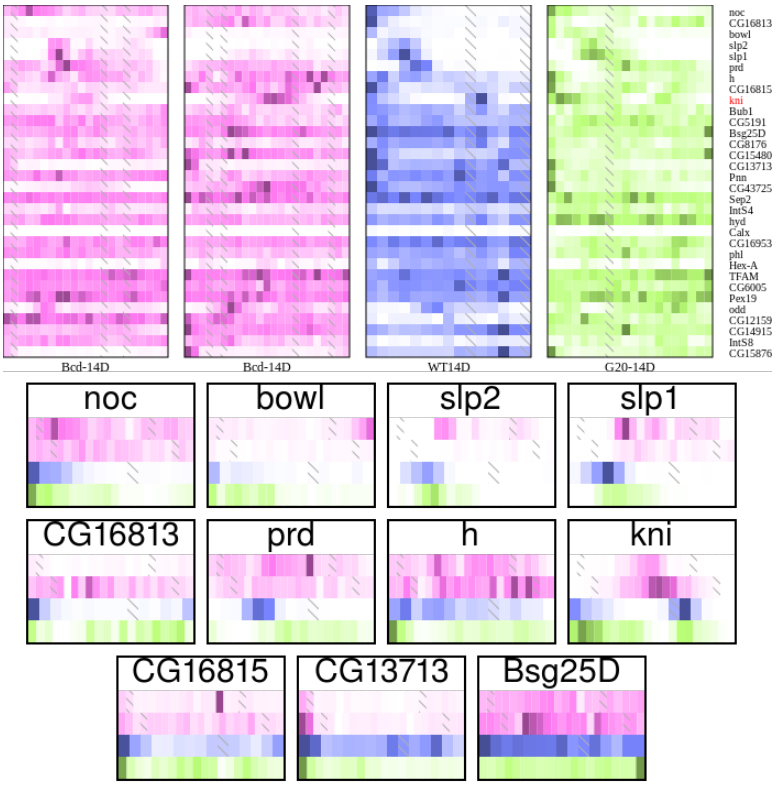


Figure 4.5. Patterning changes of genes near *BCD*-dependent enhancers in *bcd* knockdown and overexpression are clearly visible in anterior-localized genes. A) Each individual is represented in its own heatmap. The magenta heatmaps are from the *bcd*- embryos, blue from wild-type, and green from 2.4× *bicoid*. B) Each gene with anterior localized expression in WT, with data from each individual as its own row, to highlight position changes across the mutant genotypes.

mutant conditions except *bcd* and *zld* depletion, having a *BCD* binding site is associated with an increase in patterned expression. In all of the conditions, a *BCD* binding site is associated with a ubiquitous becoming patterned, and this pattern is often anterior expression.

In addition to patterning changes, some genes with ubiquitous localization actually showed the same response in absolute level as a result of both *bcd* depletion and overexpression. Of these genes, 1002 showed at least 1.5 fold higher expression on both conditions, and 414 showed a 1.5 fold decrease in expression. Such a scenario suggests that these genes are, at wild-type levels, tuned to a particular level of bicoid expression.

It is difficult to reconcile increases of expression in the posterior with any local model of transcription factor action. *BCD* protein is only present at approximately 5nM at 50% embryo length, and negligible levels more posterior (Gregor, Tank, *et al.* 2007). It is conceivable that *bcd* activates a repressor gene somewhere in the anterior, which then diffuses more rapidly than *BCD* to cover at least some of the posterior of the embryo. Nevertheless, there

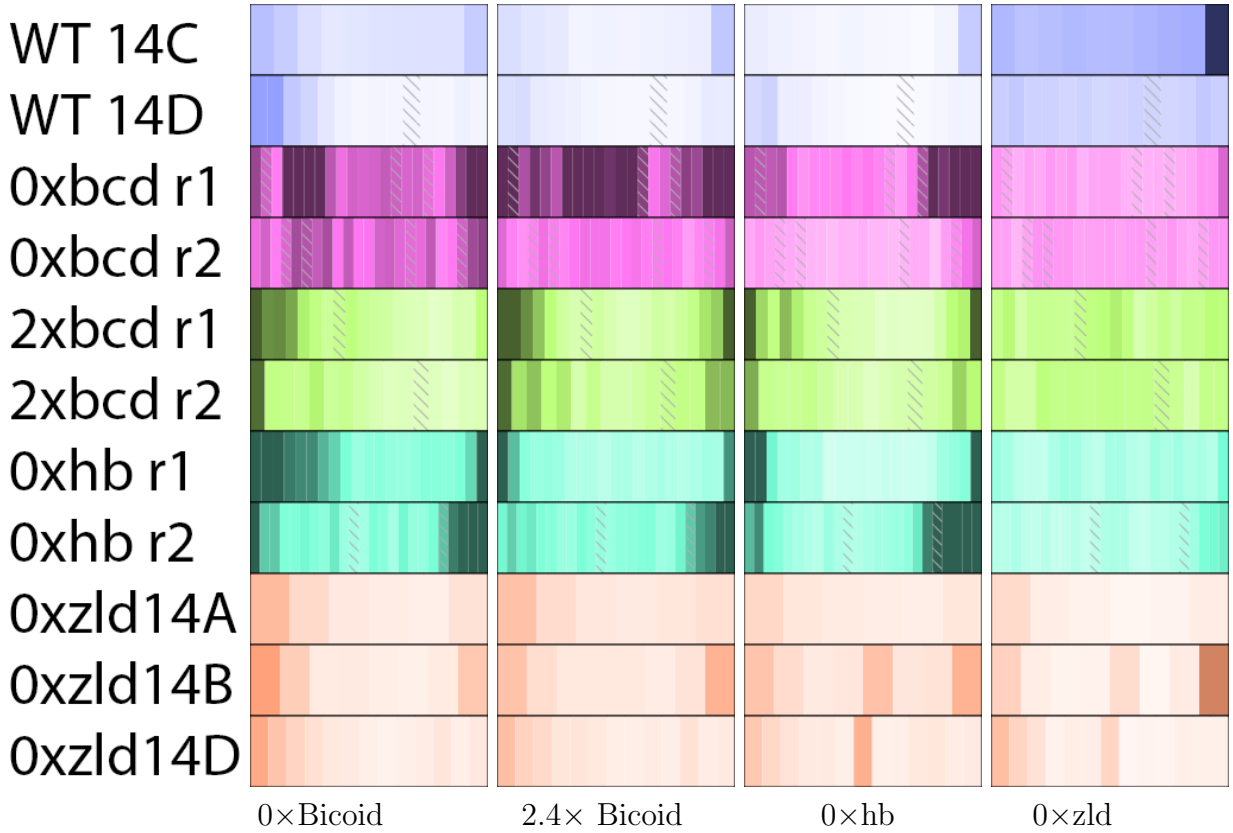


Figure 4.6. Averaging patterning changes in each genotype recapitulates known gene localization and function. Each individual is represented in its own heatmap. The magenta heatmaps are from the *bcd-* embryos, blue from wild-type, and green from 2.4× *bicoid*.

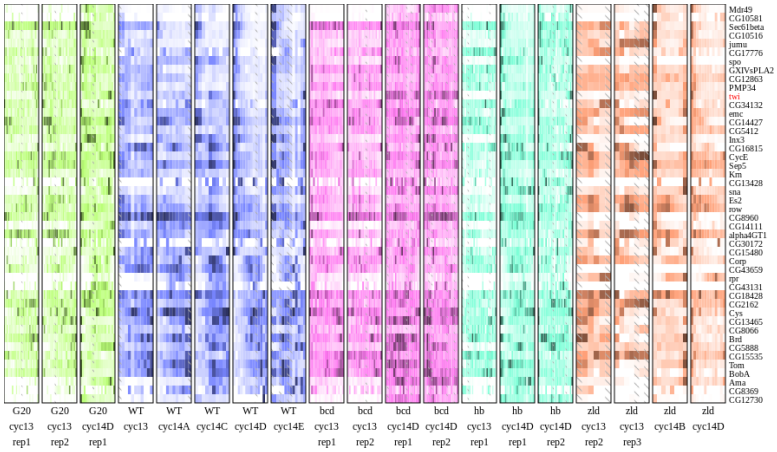


Figure 4.7. Patterned genes in wild-type that become uniformly expressed are widespread in *bcd-*. Each embryo is normalized independently.

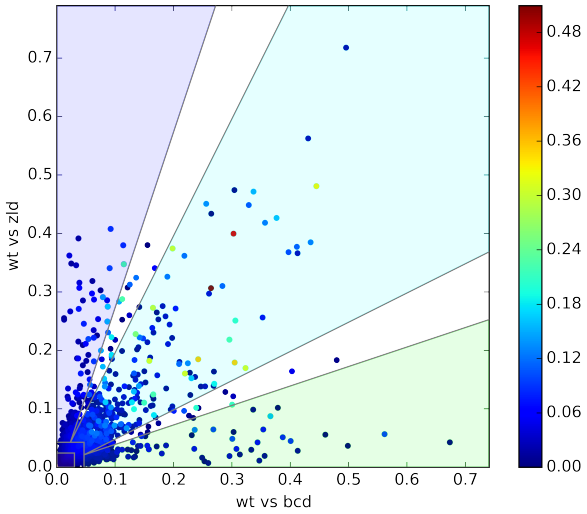


Figure 4.8. Genes that change in *bcd-* are likely to change in the same way in *zld-*, and *vice-versa*. Change versus the wild-type is plotted on the x and y axes. Each point is colored according to its ΔD score, calculated in eq. (4.1), in order to highlight genes that change differently between the two conditions.

have previously been hints that *bicoid* can function as far to the posterior as *hairy* stripe 7 (La Rosée *et al.* 1997).

Genes are likely to change in similar ways in different mutant conditions

I next asked whether patterning changes in one genotype could be used to predict whether the pattern changes in another. Therefore, I plotted the EMD between wild-type and the bicoid RNAi line on the X axis, and wild-type to the zelda GLC on the Y axis (Figure 4.8). Unsurprisingly, the majority of genes did not change, but of those that did, only a small fraction of them changed in one condition but not the other (the blue and green regions near the axes). I grouped genes according to whether they were in the top 20% of the EMD distribution for each genotype independently, then performed a Pearson's χ^2 test of independence of change in *bcd-* versus *zld-*. The result was highly significant ($p<1\times10^{-100}$), with the largest overrepresentation coming from the case where both changed. Repeating this across all combinations of wild-type and two other mutant genotypes yielded the same results: in every case, there were between 2.2 to 2.7 times as many genes that changed in both categories as would be expected (Figure 4.S3).

Of these genes that do change in both conditions, the majority changed in effectively identical ways. I computed a modified EMD that down-weights genes that are very similar

Table 4.3. TF binding is enriched near differentially changing genes between WT, *bcd-*, and *zld-*.

	odds ratio	base freq	p-value
bcd	4.69	17.40%	1.2e-09
kni	4.19	3.01%	0.000176
gt	3.79	19.00%	2.81e-08
cad	3.17	31.10%	6.73e-08
kr	3.1	56.58%	1.6e-07
tll	2.97	7.64%	0.000597
hb	2.21	39.53%	0.000143
hkb	2.1	23.86%	0.00115

χ^2 test results for TF binding within 10kb of the TSS for the wild-type/*bcd-*/*zld-* three-way comparison. I examined the top 50 genes by ΔD, compared to the 200 genes closest to the median ΔD of genes that change in response to both mutations. Base frequency indicates the fraction of genes expressed at this time point with at least one ChIP peak for that TF. In this comparison, only Dichaete and zelda binding were not significant at a Bonferroni-corrected p-value of 0.05.

to wild-type in at least one mutant genotype:

$$\Delta D = \text{EMD}(M1, M2) - |\text{EMD}(M1, WT) - \text{EMD}(M2, WT)| \tag{4.1}$$

where EMD(x, y) is the Earth Mover Distance between identically staged embryos of genotype x and genotype y. Even among only the set of genes that change in both conditions, ΔD is small (mean of 3.5%, 95th percentile of 11.9%)—equivalent to a shift of the entire pattern by about 1 or 2 slices in either direction. However, there are 13 genes that change differently between wild-type, *bcd-*, and *zld-* (ΔD>20%). These genes have noticeably different patterns in all three genotypes (Figure 4.S4).

Differential response to mutation is strongly associated with transcription factor binding

I sought to understand what is different about genes with a high ΔD, compared to those that change in response to wild-type, but have a low ΔD (that is, those that change in the same way in response to distinct mutant conditions). I found that genes with a high ΔD score were strongly enriched for a number of TF binding sites (Tables 4.S1 and 4.3).

Next, I binned genes by ΔD score, then examined trends in combinatorial transcription factor binding. As ΔD score increases, genes are more likely to be bound by multiple TFs (Figure 4.9). Due to the high background rate of binding, assaying the presence of at least

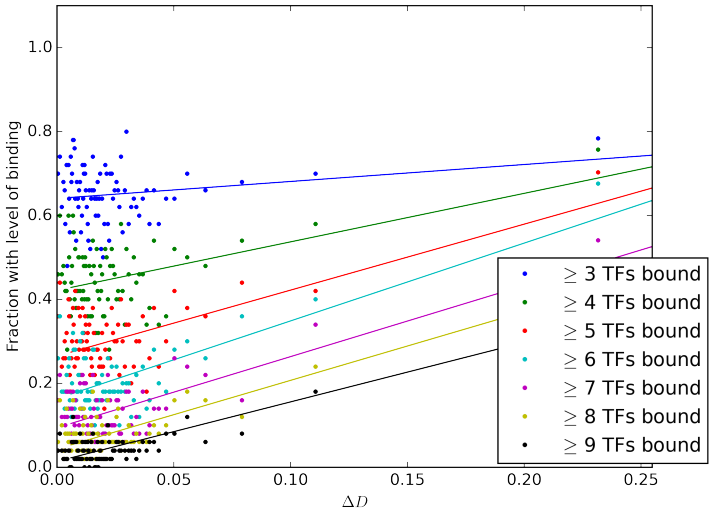


Figure 4.9. Higher ΔD scores are correlated with increased combinatorial binding. I grouped genes into non-overlapping windows of 50 genes by ΔD score, and calculated the fraction of those genes with at least 3,4, ... *etc.* of the 10 early AP TFs bound (including *zld*). I also plotted a simple linear regression on the binned points.

3 factors is not readily able to distinguish between genes with high and low ΔD's, as nearly 70% of all genes expressed have at least 3 TFs bound. Assaying for the presence of more factors is better able to identify which genes are likely to change, and the top 50 genes all have at least 8 factors bound.

I sought to understand the extent to which genes with the same pattern of upstream regulators had the same responses to perturbation. I grouped genes according to the complement of ChIP-validated TF binding sites near that gene, then examined the patterning changes. Although with 10 different TFs there are potentially over one thousand distinct combinations of binding patterns, in practice the dense, combinatorial patterns found around patterning enhancers reduces this set to a much more manageable 157 different combinations, of which only 52 had at least 30 genes.

Within these sets of genes with similar TF binding profiles, I then asked whether the distribution of patterning changes was, overall, any different from the distribution of patterning changes overall. I performed a KS-test between the distribution of summed EMD scores for the 2.4×*bcd*, *bcd*-, and *zld*- between the set of genes with a given binding pattern, and for all genes overall. I found only 2 binding patterns with a Bonferroni-corrected p-value less than .05. Both of these sets were highly bound, and they were also very similar to each other in their binding, differing only in the presence of a *KNI* site (Figure 4.10).

Despite the similar binding patterns near these genes, there is a wide range of responses. The wild-type expression patterns run nearly the complete gamut, including uniform ex-

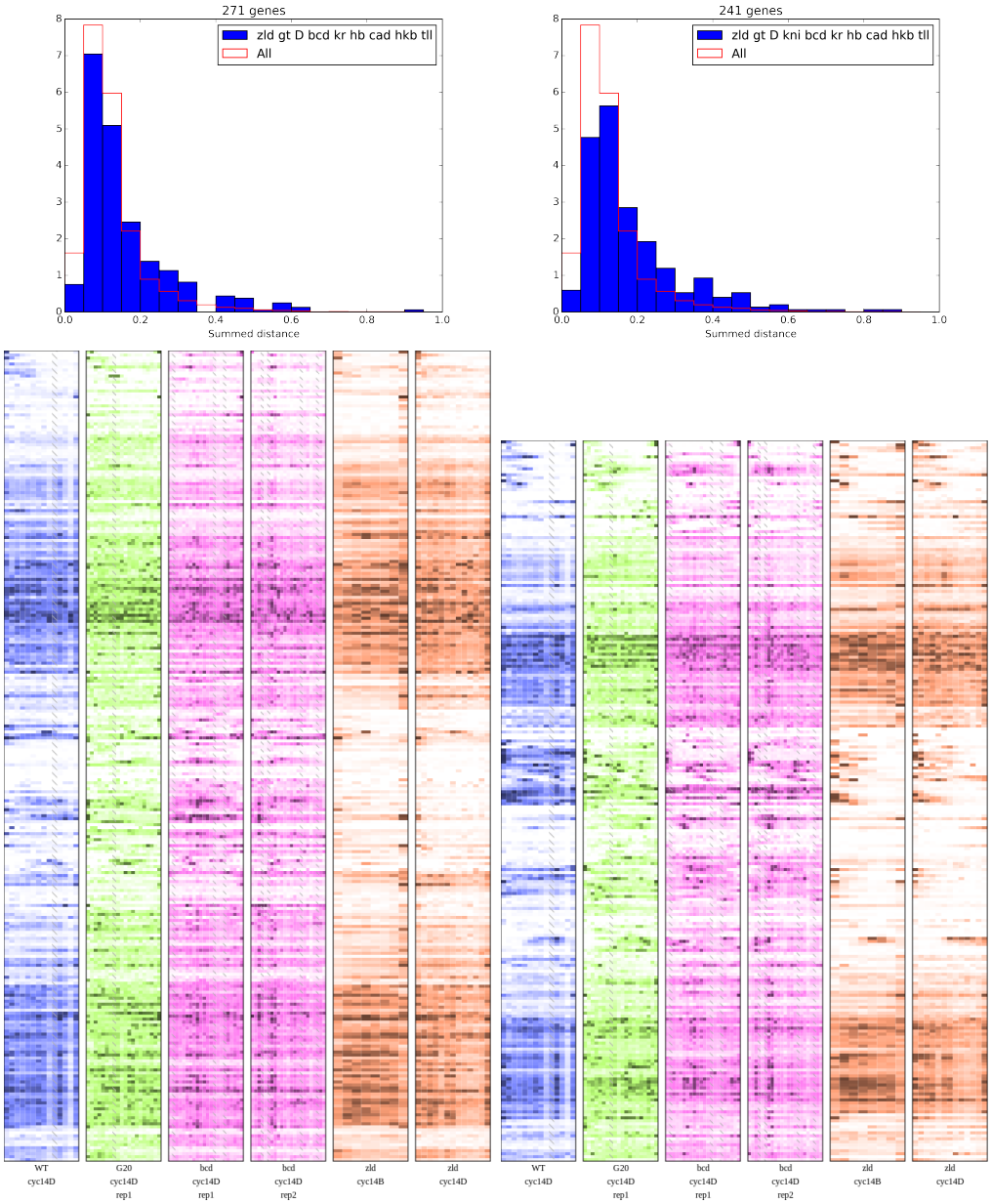


Figure 4.10. Identical binding patterns have a wide range of patterned responses.

pression, anterior stripes, posterior stripes, and central expression domains. Additionally, the presence of a *KNI* site seems to yield an increased number of genes with an anterior expression domain.

Discussion

I have generated a dataset that is unparalleled in its coverage assaying patterning changes in mutant conditions. When these patterning mutants have been described previously, either major morphological readouts like cuticle staining or *in situ* hybridization has been used to illustrate the effects on downstream target genes (Driever & Nüsslein-Volhard 1988; Liang *et al.* 2008; Staller, Fowlkes, *et al.* 2015). However, *in situ* hybridization suffers from a strong selection bias in the genes that are chosen. By assaying spatial differences in the patterning of every gene in the genome, I demonstrate the full effect that these TFs have on developmental gene expression networks.

Despite the importance of the factors I chose for establishing spatially and temporally correct patterning, only a relatively small number of genes have significant expression pattern changes. Many of the targets that do show clearly abnormal expression patterns are, themselves, key transcription factors. This suggests that, even though key, maternally provided patterning factors bind to thousands of places throughout the genome (MacArthur *et al.* 2009), many of those binding sites are not functional in any meaningful sense. Certainly some of this binding is due to artifacts in the ChIP data, and even reproducible, non-artifactual binding should not be confused with function(Graur *et al.* 2013; Teytelman *et al.* 2013). However, the fact that genes near binding sites for multiple factors tend to have more complicated responses to mutation suggests that there is some truth to the idea that gene regulation in complex animals tends to be combinatorial, even if the ChIP data are imperfect.

I was surprised how much proper *bicoid* expression seems to be required for proper patterning at all points along the embryo, not just in the anterior. The fact that *bcd* is normally understood to be an activator, while the plurality of genes with higher, ubiquitous expression in the mutant are normally localized to the anterior in wild-type suggests that this is normally mediated through one or more repressors that depend on *bcd*. As one of three TFs overrepresented at genes with this phenotype, *gt* is likely to be involved in this global derepression, but since it is itself neither ubiquitous throughout the embryo nor universally bound at the genes that change, it is likely not the only player.

The mutants I examined seemed to produce very similar changes in their downstream targets, despite the wild-type TFs having widely varying spatial distributions. Our initial expectation was that there would be many more ways to fail to properly pattern expression, and that different mutations would have different average effects from each other. Indeed, relying on different mutations having different responses has been the key to genetic analysis of fine scale patterns such as the *eve* stripes (Frasch & Levine 1987; Frasch, Warrior, *et al.* 1988; Small *et al.* 1996; Andrioli *et al.* 2002). Although averaging across the most different

genes in a mutant genotype does yield different patterns (Figure 4.6), for any given gene excursions from the correct spatial expression pattern seems to be largely canalized (Figure 4.8). This seemingly-canalized expression change may be a consequence of the types of genes we can easily measure patterning changes among—we cannot resolve individual pair-rule stripes, for instance—so genes with coarser patterns may be more likely to have a single "failure" phenotype, as compared to those with finer patterns, which have more layers of regulation to perturb.

We do recognize a number of distinct limitations of this data set towards predicting gene expression change as a function of mutation. The spatial resolution is still much coarser than *in situ* hybridization based experiments. This is especially concerning near regions where there are fine stripes of expression, which cannot be resolved between adjacent slices, or at regions where there is a transition between expression domains, where it is possible that the slicing axis is not perfectly aligned with the domain border. Finally, it is worth remembering that especially in the later stages examined, the gap gene positions will also be perturbed, so any observed changes in pattern positioning is likely to be a combination of direct effects and downstream effects of the original mutation.

A number of recent studies have used various technical or experimental techniques to improve the resolution of RNA-seq maps of gene expression in developing embryos. Iterated sectioning of different embryos in all three dimensions can be deconvolved to yield estimates of the original pattern (Junker *et al.* 2014). Similarly, sequencing mRNA from dissociated nuclei allows for the maximum possible spatial resolution, assuming the original location of those nuclei can be estimated (Satija *et al.* 2015; Achim *et al.* 2015). While these approaches are worthwhile for establishing a baseline map of expression patterns in wild-type embryos, the expense of sequencing still makes single-dimensional studies worthwhile. Furthermore, the single-cell approaches in Satija *et al.* (2015); Achim *et al.* (2015) require some prior knowledge of spatial gene expression, which may be significantly perturbed in patterning mutants. Other approaches for multiplexed *in situ* profiling of mRNA abundance have been described, but are not yet cheap or reliable enough to be readily useful for screening mutants (Lee *et al.* 2014; K. H. Chen *et al.* 2015).

Additionally, the time and expense required for a single individual necessarily means that I have profiled only a small number of individuals. I was therefore careful to choose only highly penetrant mutations for analysis, and to choose individuals at as similar staging as I could. However, even for genes with a consistent, precise time-dependent response between individuals, the differences in staging are likely to be a significant contributor to variation. Furthermore, I only examined two relatively distant time points in this study (approximately 45 minutes apart), making comparisons across time fraught at best.

Nevertheless, this experiment suggests a number of genes for more detailed follow up studies. As our predictive power for relatively well-studied model systems, such as the *eve* stripes improves, it will be especially important to take these insights to other expression patterns in the embryo. The risk of over-fitting increases with the depth of study of any particular model system, even if any given study is relatively well controlled. Therefore, by demonstrating that particular insights hard-won in these model systems are broadly

applicable, we can gain some confidence in the results, and we approach having a rigorous, broadly applicable predictive model of gene regulation.

Ultimately, I believe more datasets addressing chromatin state in response to different conditions will be necessary for accurate prediction of spatial responses to mutation. In a ChIP-seq dataset on embryos with different, uniform levels of *bicoid* expression, hundreds of peaks seem to vary with differing affinities to *BCD* protein (Colleen Hannon and Eric Wieschaus, personal communication, March 2015). The zygotically expressed genes near these differential peaks also have different spatial localization in wild-type, and different average responses to the mutants presented here. In addition to spatially resolved expression measurements, spatially resolved binding and chromatin accessibility data will likely be necessary. While ChIP-seq experiments currently require several orders of magnitude more input material than can be reasonably collected from spatially resolved samples, recent methods developments in measuring chromatin accessibility have shown that it is possible to collect data from as few as 500 mammalian nuclei (Buenrostro *et al.* 2013). A similar amount of DNA is present in a single *Drosophila* embryo, which suggests that spatially resolved chromatin accessibility data may be achievable.

Materials and Methods

Fly lines, imaging, and slicing

Zelda germline clone flies (w zld- FRT/FM7a; His2Av RFP) were a gift of Melissa Harrison, and were mated and raised as described previously. Embryos were collected from mothers 3-10 days old.

The construction of the *bcd* and *hb* RNAi flies has been described previously (Staller, Yan, *et al.* 2013) and were obtained from the DePace Lab at Harvard Medical School. Briefly, we generated F1s from the cross of maternal tubulin Gal4 mothers (line 2318) with UAS-shRNA-bcd or UAS-shRNA-hb fathers (lines GL00407 and GL01321 respectively), then collected embryos from the sibling-mated F1s. In order to take advantage of the slowed oogenesis and resulting greater RNAi efficiency, we aged the F1 mothers for approximately 30 days at 25°C.

The *bcd* overexpression lines were a generous gift of Thomas Gregor at Princeton University. We used line 20, which has 2.4× wild-type levels of eGFP-bcd fusion. Flies were kept in uncrowded conditions, and embryos were collected at 25°C from 3-7 day old mothers.

We washed, dechorionated, and fixed the embryos according to our standard protocol (see (Combs & Eisen 2013)), incubated in 3 µM DAPI for 5 minutes, washed twice with PBS, and then imaged on a Nikon 80i microscope with a Hamamatsu ORCA-Flash4.0 CCD camera. We did not DAPI stain the *zld-* embryos because they had a histone RFP marker. After selecting embryos with the appropriate stage according to density of nuclei in histone-RFP or DAPI staining and membrane invagination for the cycle 14 embryos, we washed embryos with methanol saturated with bromophenol blue (Fisher), aligned them in standard cryotome

cups (Polysciences Inc), covered them with VWR Clear Frozen Section Compound (VWR, West Chester, PA), and froze them at -80C.

We sliced the embryos as in Combs & Eisen (2013). Single slices were placed directly in non-stick RNase-free tubes (Life Technologies), and kept on dry ice until storage at -80C.

RNA Extraction, Library Preparation, and Sequencing

We performed RNA extraction in TRIzol as previously (Combs & Eisen 2013). All RNA quality was confirmed using a BioAnalyzer 2100 RNA Pico chip (Agilent).

We generated libraries of the *zld-* embryos using the TruSeq mRNA unstranded kit (Illumina). As described previously, we added in 70ng of yeast total RNA as a carrier and performed reactions in half-sized volumes to improve concentration (Combs & Eisen 2013).

We generated libraries from the RNAi and overexpression embryos using the SMARTseq2 protocol; we skipped the cell lysis steps because RNA had already been extracted (Picelli, Faridani, *et al.* 2014; Combs & Eisen 2015). As described previously, tagmentation steps were performed at 1/5th volume to reduce costs (Combs & Eisen 2015).

Data analysis and deposition

All data was compared to FlyBase genome version r6.03 (`2014_6`). Mapping was performed using RNA-STAR v2.3.0.1 (Dobin *et al.* 2013), and expression estimates were generated using Cufflinks v2.2.1 on only the *D. melanogaster* reads(Trapnell, Roberts, *et al.* 2012). Reads from Combs & Eisen (2013) were re-mapped to the new genome version. When carrier RNA was used (data from Combs & Eisen (2013) and the *zld-* embryos), we discarded as ambiguous reads with 3 or fewer mismatches to prefer one species or the other. The vast majority of mapped reads (>99.99%) of were unambiguous as to the species of origin. After mapping, we removed samples that had fewer than 500,000 *D. melanogaster* reads and samples with less than a 70% mapping rate when no carrier RNA was used; no other filtering or corrections were performed.

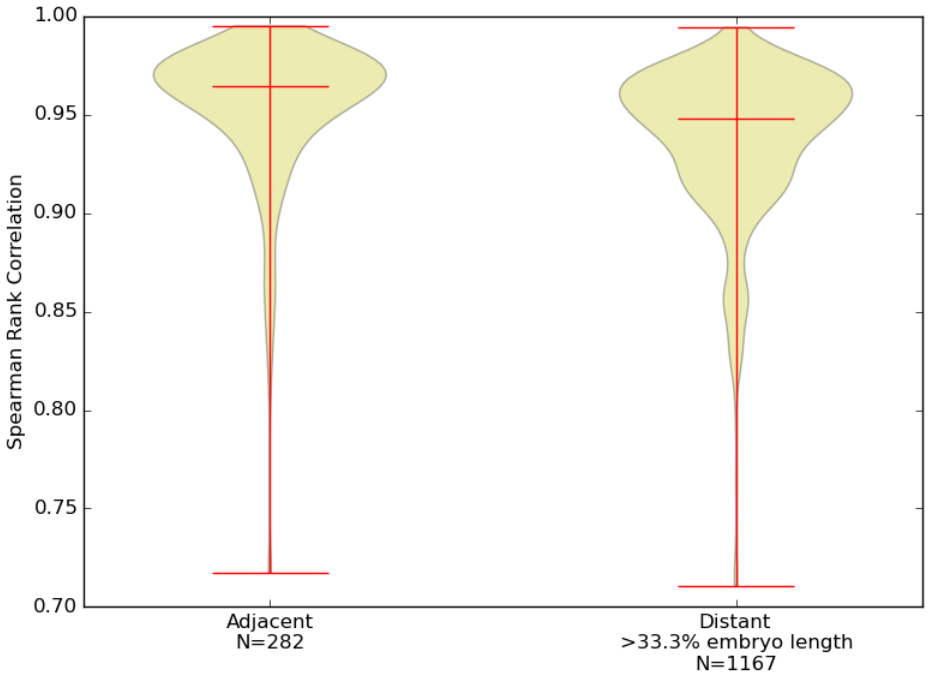
Specific analysis code was custom-written in Python. Custom analysis code is available from `https://github.com/petercombs/EisenLab-Code`. All analyses presented here and all data figures were made using commit `2c144be`.

Newly generated sequencing reads will be deposited at the Gene Expression Omnibus. Mapped reads, additional files, and a searchable database will be available at `http://eisenlab.org/mutantslices`.

In situ hybridization

Probe templates were generated by PCR amplifying cDNA with gene specific T7 promoter fusion primers to cover at least 500 bp of the transcripts. We then made RNA probes using a DIG-labelling kit (Roche), and resuspended in formamide.

Hybridization was performed as described previously (Kosman *et al.* 2004). All genotypes for each gene were processed in parallel, with hybridization and development times varying by less than 3 minutes.

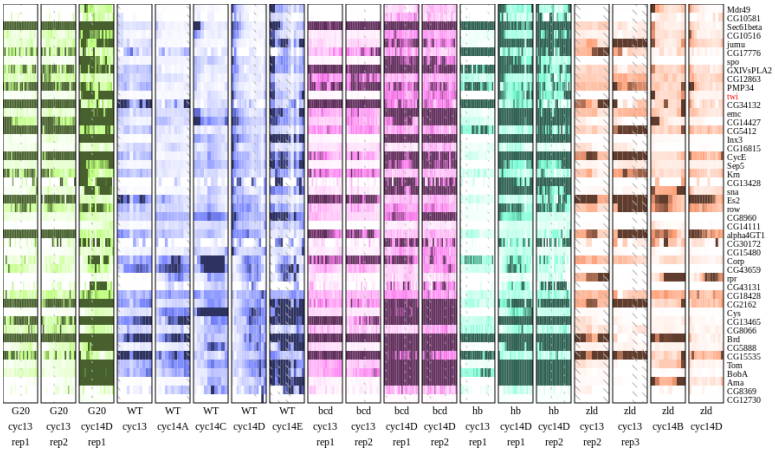


Supplemental Figure 4.S1. Adjacent slices are more similar than distant ones. Violin plots of the Spearman Rank correlations between adjacent slices and pairs of slices separated by more than one third of the embryo length.

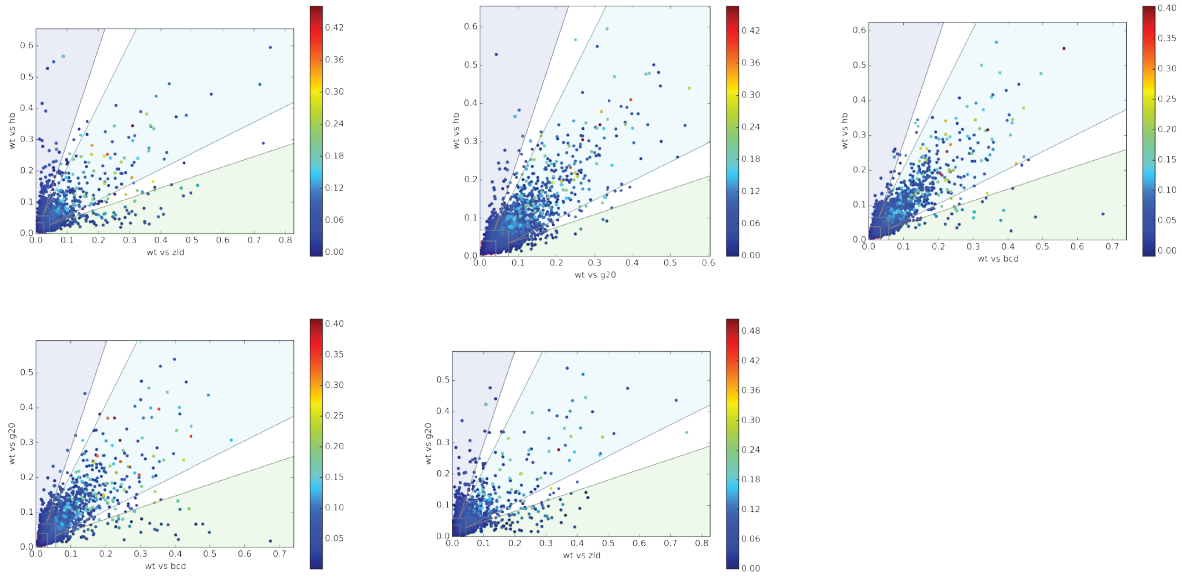
Supplemental Table 4.S1. TF binding is enriched near differentially changing genes across all three-way comparisons.

	odds ratio	base freq	p-value
WT vs *bcd-* vs 2.4×*bcd*			
kni	16.4	3.28%	8.93e-07
tll	5.5	8.08%	7.72e-05
bcd	4.1	17.58%	9.24e-05
gt	3.47	19.55%	0.00053
hkb	3.1	24.15%	0.0016
cad	3.09	31.98%	0.00067
WT vs *bcd-* vs *hb-*			
kni	11.0	3.34%	0.00297
tll	5.31	8.15%	0.000671
bcd	3.29	17.67%	0.00183
D	0.281	90.70%	0.00251
WT vs *zld-* vs 2.4×*bcd*			
kni	12.3	3.05%	3.63e-07
tll	11.3	7.67%	3.91e-11
cad	7.05	31.20%	1.62e-08
bcd	6.74	17.55%	7.02e-09
gt	4.56	18.99%	5.95e-06
kr	4.56	56.67%	8.79e-05
hb	3.95	39.74%	4.62e-05
WT vs *zld-* vs *hb-*			
kni	9.33	3.04%	0.000162
bcd	7.82	17.51%	6.34e-10
gt	6.14	19.01%	7.57e-08
tll	5.05	7.67%	0.000156
cad	4.77	31.01%	2.05e-06
kr	3.41	56.61%	0.00102
hb	2.84	39.68%	0.00186
WT vs 2.4×*bcd* vs *hb-*			
tll	18.5	8.16%	1.19e-07
kni	8.95	3.32%	0.00169
gt	4.27	19.57%	3.32e-05
bcd	3.78	17.76%	0.000173
hb	2.9	40.61%	0.00147
cad	2.87	31.72%	0.00152

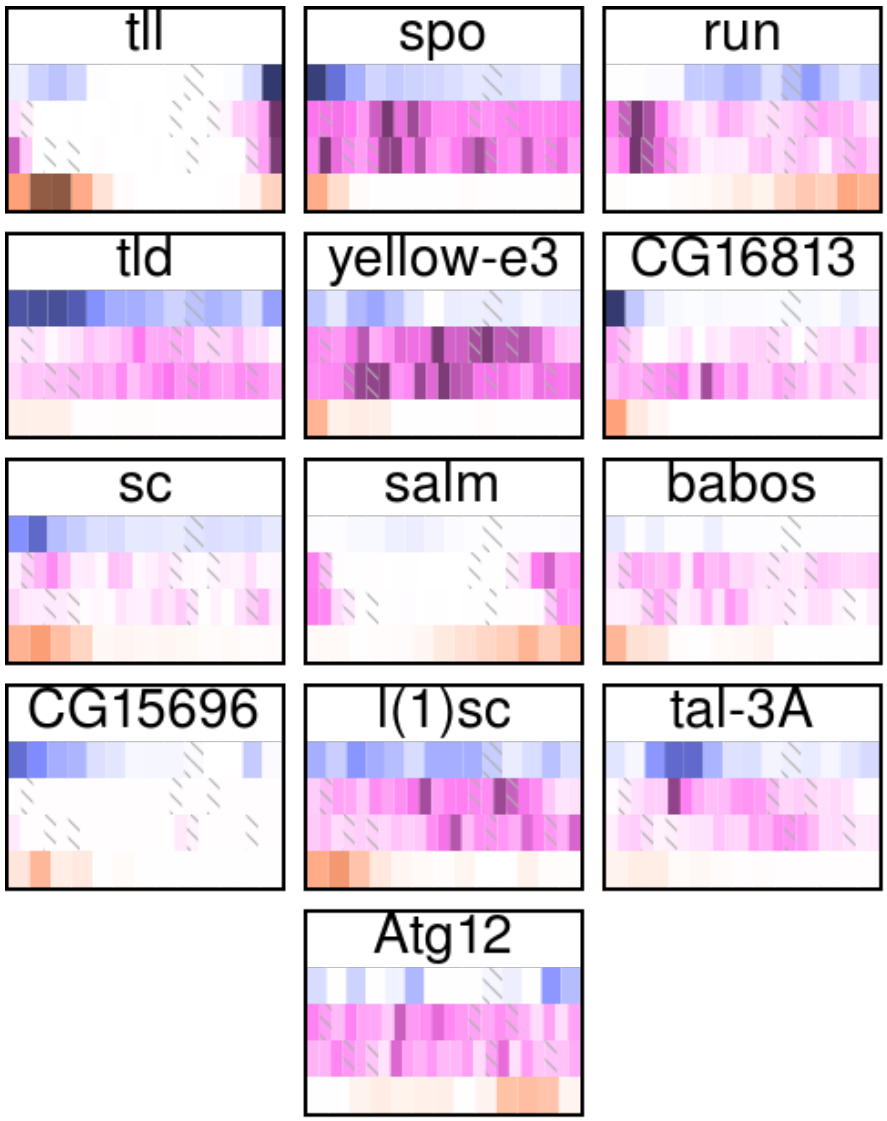
χ^2 test results for TF binding within 10kb of the TSS for the indicated three-way comparison. I examined the top 50 genes by ΔD, compared to the 200 genes closest to the median ΔD of genes that change in response to both mutations. Base frequency indicates the fraction of genes with at least one ChIP peak for that TF and that are expressed at this time point in all three conditions.



Supplemental Figure 4.S2. Figure 4.7 normalized to expression in wild-type cycle 14D highlight absolute expression level changes. Slices with higher expression are clipped to the maximum expression in wild-type.



Supplemental Figure 4.S3. Genes that change tend not to change in only one condition. Three-way comparisons, as in Figure 4.8, between wildtype and the remaining combinations of *bcd* depletion, *bcd* overexpression, *hb* depletion, and *zld* depletion.



Supplemental Figure 4.S4. Only a handful of genes change differently between the different conditions. In the WT vs *bcd*- vs *zld*- three-way comparison, only 13 genes had a ΔD score above 20%. Thumbnails indicate wild-type pattern in blue, *bcd*- pattern in both replicates in pink, and *zld*- pattern in orange. All expression is scaled to the highest in each individual.

Chapter 5

Future Directions and Outlook

Perhaps the most remarkable feature of the experiments I have presented here is the lack of a number of features of the system that I would have predicted. I searched quite extensively for genes that seemed to show differential splicing across space, but could not find any convincing evidence of this phenomenon. While the relatively shallow sequencing made this somewhat more complicated, there is clear evidence of changes in alternative promoter use across time. The gene *hunchback* has two promoters, one which is primarily used for maternal transcripts and one for zygotic transcripts (Margolis *et al.* 1995). Comparing *hb* expression in the early and late embryos from the wild-type time course, there is a marked shift in the usage of these two promoters (Figure 5.1). Thus, if there were a very strong signal of spatially varying incorporation of exons, it ought to be detectable.

The next feature that I could not find any clear evidence for was the differential production of eRNAs across space. A number of studies over the last half-decade have detected bi-directional, poly-adenylated transcripts from enhancer regions (van Bakel *et al.* 2010; Wang *et al.* 2011; Djebali *et al.* 2012; Mousavi *et al.* 2013). While the purpose, if any, of these eRNAs is hotly debated, the null hypothesis has to be that they are merely artifacts of the transcriptional machinery. Functional or not, they would nevertheless be exciting to detect, as such transcription could indicate which CRM was responsible for patterning, and for patterning changes in mutants. However, I could never detect any convincing evidence of differential use of different, previously annotated and validated CRMs (Gallo *et al.* 2011). This is not to say that there was no transcription—there are a handful of enhancers both distal to the gene and in introns that seem to have some differential usage across space. However, in those few cases there was not another enhancer for the same gene that had clear expression in a different region. There were some marginal cases (see, for instance, figure 5.2), but even the clearest of them is not convincing.

This is in sharp contrast to how enhancers are understood to work to generate complex patterns. Perhaps the best studied example, *eve* is driven by at least 4 different CRMs, each of which drives one or two of the seven total stripes. The failure to observe this effect in the eRNAs in no way challenges this idea, though. While there are a number of protocols to actively seek out nascent RNAs (Core *et al.* 2008), in the protocols used here eRNAs

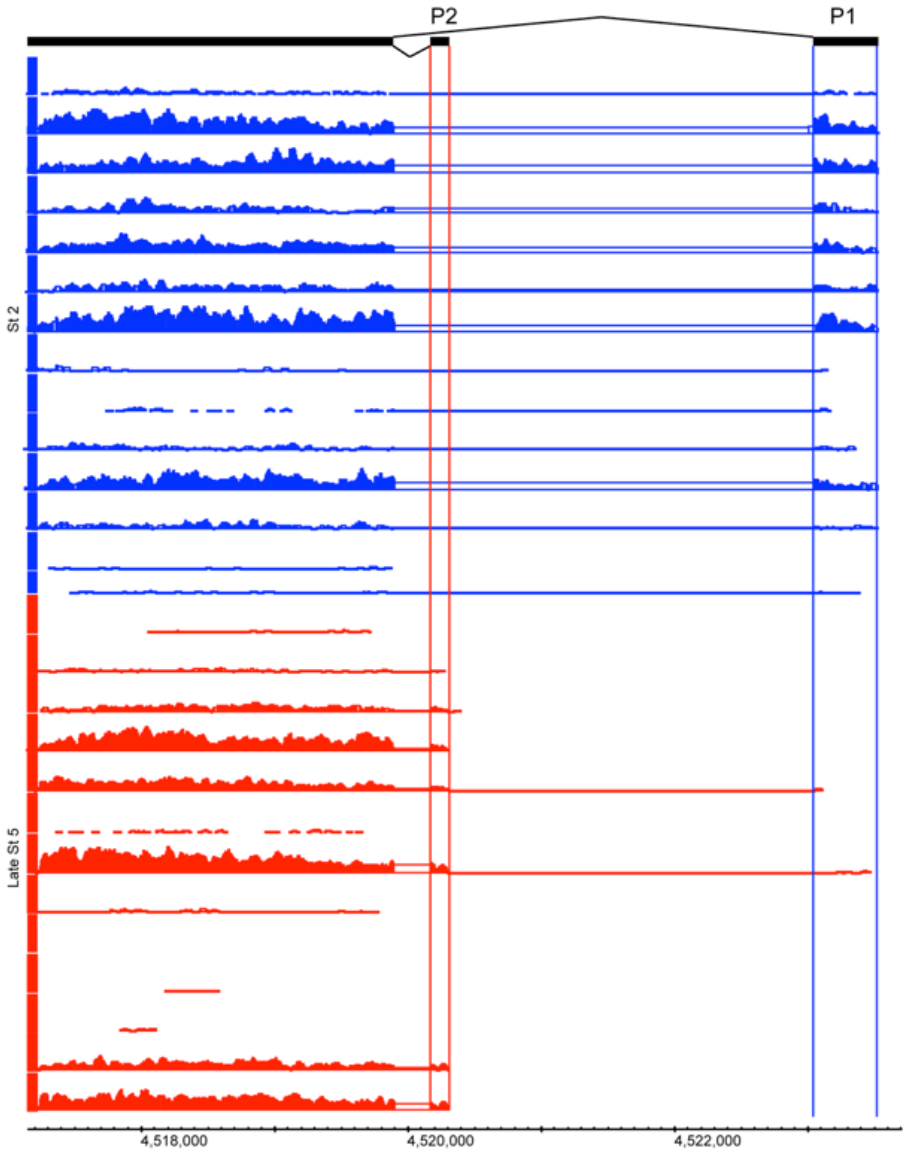


Figure 5.1. Clear differences in alternate promoter usage across time, but not space. The GBrowse track is centered to display the two alternate promoters of *hunchback* (only the 5' most end of the second exon, which is included in both *hb* isoforms, is shown). There is clearly enough read depth to distinguish alternate promoter usage between the early (blue) and late (red) stages, but no convincing evidence of alternative promoter usage is evident across different slices of the same time point.

ought to be considered off-target effects, even if they are actually poly-adenylated. Thus, a lack of a spatially varying signal could be due to any number of technical artifacts, and not to a bonafide lack of differential transcription[1]. It is also not clear that the protocols to more specifically search for eRNA signal are easily applicable to the early embryo—the closed nature of the system makes it difficult to use clever molecular biology techniques such as labelled nucleotides.

Figure 5.2. eRNA FPKM of six *Ubx* enhancers. Each of the six rows represents coverage over a disjoint enhancer region known to drive *ubx*, and the columns are arranged by time and AP order. Although there are slight differences in the rows, they are not convincingly different from noise overlaid on the expression of the gene itself.

Finally, I was greatly surprised by how limited the patterning effects were of altering levels of key patterning transcription factors. While I should perhaps have predicted that there would be systems that attempt to mitigate deleterious effects, I was nevertheless surprised by how few genes really were affected. This could be a great boon for future experiments, however. If one could predict which genes are likely to have their patterning altered, then generating a complete inventory by *in situ* hybridization should be relatively quick. While performing thousands of stains is impractical, assaying a few hundred patterns is ambitious, but entirely doable.

These gene expression data are useful in targeting the search for perturbed patterns, but are not in and of themselves capable of fully unravelling the response. As shown in the experiments in chapter 4, it is important to use the RNAseq data to target further, higher resolution studies on the same mutants. A single in situ hybridization experiment can quickly and cheaply show the full shape of the gene expression pattern and the change in the pattern over time.

5.1 Additional Experiments

I would, with the gentle reader's forbearance, also like to briefly describe two additional experiments in the vein of the already presented research that I believe would be illuminating with regards to issues of the regulation of gene expression. These are feasible with relatively little additional technology development required, and of a small enough scale that the sequencing experiments for each could be completed in less than six months[2].

[1]It is also possible that at this stage, there truly are no spatially patterned eRNAs, despite patterning of the associated genes. If that were truly the case, it would be a remarkable finding. However, absence of evidence is not evidence of absence, and my data does not significantly alter the posterior probability.

[2]Data analysis is another question entirely, and such far-reaching datasets are likely to be fruitful for years of analysis.

Evolution: a grand mutation experiment

In the work presented in Chapter 4, I examined in detail the effect of perturbing the transcriptional network on patterning, either by increasing or decreasing the concentration of key transcription factors. None of the mutations were guaranteed to inflict tolerable effects to the network, and indeed three of the four genotypes examined were lethal, some as early as cellularization, like the *zld* mutant. The responses of the network to these mutations are unlikely to be representative of a "typical" perturbation, such as a different growth temperature (Indeed Kuntz & Eisen 2014, have shown that a broad range of temperatures are tolerable to most flies, with few detectable changes aside from developmental rate).

The *Drosophila* phylogeny, however, is more than 40 million years old. Between distant branches of the tree, there is a total of 4.1 substitutions per neutral site, approximately similar to the evolutionary distance between humans and platypodes (Drosophila 12 Genomes Consortium *et al.* 2007; Stark *et al.* 2007). During that time, mutation will have allowed the network to explore at least the local neighborhood in parameter space. The constraints of natural selection have meant that as this exploration has proceeded, only those parameters that are capable of producing a valid fruit fly are tolerated. While a few detailed theoretical treatments have looked into the critical range where these parameters can exist, certain factors can tolerate an order of magnitude variation in expression level.

Here I propose examining the patterning networks of a range of *Drosophila* species using cryosectioned embryos. This study should certainly include *D. yakuba, D. pseudoobscura*, and *D. virilis*, for which ChIP experiments have already been performed (Paris *et al.* 2013), but ideally would cover the tree more broadly. The most obvious variations in adult phenotype among these species has to do with wing size and shape, body size, and pigmentation. There are not major differences in segmentation or the body plan—the flies all look like flies—which are the features that many of the patterning genes expressed before gastrulation (such as the pair-rule, segment polarity, and hox genes) critically control. Additionally, examining hybrid embryos of *D. melanogaster* with it sister species *D. simulans* is likely to provide insights into transcriptional regulation. If there are any genes with effectively identical patterns in both *D. melanogaster* and *D. simulans*, but different patterns in the hybrid, that would indicate parallel evolution in the cis-regulatory modules that control those genes' expression.

The goal of such a study would be two-fold: first, it would help to place empirical limits on the amount of expression variation that can be tolerated. It is unlikely that these systems are sufficiently well studied that these limits could be deduced from first principles and examining only the *D. melanogaster* network alone. The second goal of the study would be to better understand how cis-regulatory grammars operate. It has been demonstrated that enhancers with wildly varying binding site organization can drive nearly identical expression patterns (Hare *et al.* 2008), but such examples are few and far between. I assert that if we were presented by an oracle with every possible enhancer sequence that could drive a given expression pattern, it would be possible to deduce from that set the rules that govern the system. In this way, one can imagine the system as analogous to the games "Mastermind" or "Black Box", where one player (the "solver") is able to probe a system with defined inputs,

and must use the outputs to deduce the hidden, internal state (Strom & Barolo 2011). These sorts of games usually impose a limit on the number of probes the solver is allowed to make, which is suggestive of the fact that probing all possible states makes the game too easy. As a caveat, in Black Box it is possible to construct multiple solutions consistent with a given input-output mapping; by analogy, even with the complete set of enhancers it might not be possible to completely rule out certain models. Even this worst-case scenario, however, would be a great leap over the current state of the art.

The spatial organization of chromatin accessibility

It is well known that transcription factors both depend on proper chromatin state for successful binding and also modify chromatin state through the displacement of nucleosomes after they have bound. However, due to the relatively large amount of sample required for most chromatin experiments, very little is known about whether TFs have different affinity in different parts of the embryo nor whether chromatin state is spatially as well as temporally dynamic. It seems very likely to me that the answer to both of these questions is yes, but the precise identities of the binding sites and nucleosomes that are most dynamic will have a huge impact on gene expression prediction.

As yet unpublished experiments from the Eisen lab, inspired in part by the feasibility of performing RNA-seq on subsets of a single embryo, have examined whether transcription factor *binding* varies in a spatially dependent manner. The TF *giant*, for instance, has two expression domains, one in the anterior of the embryo and one in the posterior. In this conceptually simple (yet technically demanding) experiment, Colin Brown looked at ChIP-seq data in these two separate domains. While a majority of binding peaks are present in both the anterior and posterior, a handful of were present in only one half. These data jibe conceptually with other unpublished data from an orthogonal experiment conducted by Colleen Hannon in Eric Wieschaus's lab, wherein *bicoid* ChIP-seq peaks are assayed in embryos with a varying, uniform concentration of *bicoid*. It seems likely that peaks with responses to different *bicoid* concentrations are also highlighting the presence of other TFs in the embryo. While it is perhaps not surprising that binding of one factor should be different among different milieux of various other transcription factors, such a result has not been demonstrated elsewhere in similarly close tissues[3]. These experiments have provided, at the very least, suggestive evidence of widespread spatial patterns not only of gene expression, but also in the organization of the transcriptional apparatus.

In somewhat surprising contrast with techniques measuring RNA levels from ever smaller samples, techniques for reducing the quantities required for profiling chromatin have been slower in being developed. The ATACseq protocol, recently published by the Greenleaf lab, is to my knowledge the only approach for measuring chromatin accessibility that is applicable

[3]In the trivial case, one can look at the binding of a factor across different tissues—and the ENCODE project has done this in exhausting (if not yet exhaustive) detail—but these are in the majority of cases terminally differentiated tissues, and perhaps not informative to the process of differentiation that is actively playing out in the early embryo.

to samples smaller than $\sim 10^5$ nuclei. This relatively simple approach, using tagmentation enzymes to directly ligate Illumina sequencing adapters to native chromatin, is quick and relatively easy to perform. More excitingly, however, the approach is claimed to work for samples with as few as 500 nuclei; if extensible to *Drosophila* embryos, one could conceivably assay chromatin accessibility at 10 different points along any given axis of a stage 5 embryo.

The point of such an experiment would be to limit the space of possibilities for how the nucleus "knows" that a particular region of DNA is an enhancer, while another is not. If some amount of sequence-specific transcription factor binding is able to displace nucleosomes, then more permanent chromatin marks are added nearby, we would expect that in parts of the embryo very distant from the enhancer's region of activity, where none of the sequence-specific factors that actually drive that enhancer are expressed, DNA would primarily be closed chromatin, rather than the open chromatin characteristic of enhancers.

If not, then that would require a ubiquitous enhancer-marking machinery. One particular hope was that *zelda* would be just such a global enhancer marker, though transgene experiments in *zelda* depleted embryos have so far been unable to confirm that hypothesis. However, the gold-standard experiment of adding or removing *zelda* binding sites in endogenous enhancers has not been performed. Fortunately, the advent of genome editing techniques makes such an experiment feasible. Presumably CRISPR-Cas9 would be the preferred method for its low cost and easy retargetability.

I have performed preliminary experiments using ATACseq on single *Drosophila* embryos, and have generated sequencing libraries whose fragment size distribution is consistent with good libraries. While I have not yet actually sequenced these libraries, it seems reasonable to believe they might recapitulate chromatin accessibility from larger experiments. If this is the case, then a push could be made towards developing dissection techniques that preserve the chromatin structure suitably well for spatial patterns of differential chromatin accessibility to be assayed.

Unfortunately, there are a number of limitations to directly applying the techniques developed for RNA towards profiling chromatin. Both freezing and fixing samples tend to create noisier data (B. Wu and J. Buenrostro, personal communication). Fixing seems to be the more detrimental process, but this is not strictly required for the slicing. Fixing the embryos serves a few purposes in my experiments with RNA: first and foremost, it simplifies the process of staging embryos, since the embryos are no longer developing the time pressure for staging accurately is alleviated; second, by permeabilizing the embryo, dyes can be used that allow only the slices from the frozen block of cryosectioning media that contain a piece of the embryo to be retained; finally, it is believed to provide some more structural integrity to the manipulation required to align the sample in the block, though this has not been verified or assayed.

Stopping the development for staging is dispensable. In practice, while it does take a few minutes to properly align the embryo in the cup, working efficiently should minimize the staging error. There are also alternative ways to dye the embryo. While a major function of the egg shell is to keep out foreign molecules, it is possible wash the embryos with a gentle detergent, then apply a non-toxic stain (Rand *et al.* 2010). This could also allow

the introduction of a non-toxic nuclear marker, such as Hoechst 33342. The most likely response, if any, is an increase in stress-response pathways. This could be assayed through bulk embryo RNAseq, and at any rate would be unlikely to also affect the ATACseq data around patterning enhancers.

Bibliography

1. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology* (Apr. 2015).
2. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)* **287,** 2185–2195 (Mar. 2000).
3. Anders, S., Pyl, P. T. & Huber, W. HTSeq A Python framework to work with high-throughput sequencing data. doi:10.1101/002824. ¡http://dx.doi.org/10.1101/002824¿ (Feb. 2014).
4. Andrioli, L. P. M., Vasisht, V., Theodosopoulou, E., Oberstein, A. & Small, S. Anterior repression of a Drosophila stripe enhancer requires three position-specific mechanisms. *Development* **129,** 4931–4940 (Nov. 2002).
5. Barolo, S. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *BioEssays* **34,** 135–141 (Feb. 2012).
6. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40,** e72 (May 2012).
7. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (Nov. 2008).
8. Bhargava, V., Head, S. R., Ordoukhanian, P., Mercola, M. & Subramaniam, S. Technical Variations in Low-Input RNA-seq Methodologies. *Scientific reports* **4,** 3678 (2014).
9. Biemar, F. *et al.* Spatial regulation of microRNA gene expression in the Drosophila embryo. *Proceedings of the National Academy of Sciences* **102,** 15907–15911 (Nov. 2005).
10. Bogdanovic, O. *et al.* Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Research* **22,** 2043–2053 (Oct. 2012).
11. Brown, J. B. *et al.* Diversity and dynamics of the Drosophila transcriptome. *Nature* **512,** 393–399 (Aug. 2014).

12. Bruder, C. E. G. *et al.* Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *American journal of human genetics* **82,** 763–771 (Mar. 2008).

13. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10,** 1213–1218 (Dec. 2013).

14. Buganim, Y. *et al.* Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* **150,** 1209–1222 (Sept. 2012).

15. Campos-Ortega, J. A. & Hartenstein, V. The Embryonic Development of Drosophila Melanogaster, 1–415 (Oct. 1997).

16. Cheng, Q. *et al.* Computational Identification of Diverse Mechanisms Underlying Transcription Factor-DNA Occupancy. *PLoS Genetics* **9,** e1003571 (Aug. 2013).

17. Chen, H., Xu, Z., Mei, C., Yu, D. & Small, S. A system of repressor gradients spatially organizes the boundaries of Bicoid-dependent target genes. *Cell* **149,** 618–629 (Apr. 2012).

18. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (New York, N.Y.)* (Apr. 2015).

19. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)* **25,** 1422–1423 (June 2009).

20. Combs, P. A. & Eisen, M. B. Low-cost, low-input RNA-seq protocols perform nearly as well as high-input protocols. *PeerJ* **3,** e869 (2015).

21. Combs, P. A. & Eisen, M. B. Sequencing mRNA from cryo-sliced Drosophila embryos to determine genome-wide spatial patterns of gene expression. *PLoS ONE* **8,** e71820 (2013).

22. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.)* **322,** 1845–1848 (Dec. 2008).

23. Daines, B. *et al.* The Drosophila melanogaster transcriptome by paired-end RNA sequencing. *Genome Research* **21,** 315–324 (Feb. 2011).

24. De Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics (Oxford, England)* **20,** 1453–1454 (June 2004).

25. Ding, D. & Lipshitz, H. D. A molecular screen for polar-localised maternal RNAs in the early embryo of Drosophila. *Zygote (Cambridge, England)* **1,** 257–271 (Aug. 1993).

26. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (Sept. 2012).

27. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29,** 15–21 (Jan. 2013).

28. Driever, W. & Nüsslein-Volhard, C. The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell* **54,** 95–104 (July 1988).

29. Drocco, J. A., Wieschaus, E. F. & Tank, D. W. The synthesis-diffusion-degradation model explains Bicoid gradient formation in unfertilized eggs. *Physical biology* **9,** 055004 (Oct. 2012).

30. Drosophila 12 Genomes Consortium *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450,** 203–218 (Nov. 2007).

31. Dush, M. K. & Martin, G. R. Analysis of mouse Evx genes: Evx-1 displays graded expression in the primitive streak. *Developmental Biology* **151,** 273–287 (May 1992).

32. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics* **11,** 446–450 (June 2010).

33. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science (New York, N.Y.)* **297,** 1183–1186 (Aug. 2002).

34. Filion, G. J. *et al.* Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143,** 212–224 (Oct. 2010).

35. Fowlkes, C. C. *et al.* A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell* **133,** 364–374 (Apr. 2008).

36. Frasch, M. & Levine, M. Complementary patterns of even-skipped and fushi tarazu expression involve their differential regulation by a common set of segmentation genes in Drosophila. *Genes & Development* **1,** 981–995 (Nov. 1987).

37. Frasch, M., Warrior, R., Tugwood, J. & Levine, M. Molecular analysis of even-skipped mutants in Drosophila development. *Genes & Development* **2,** 1824–1838 (Dec. 1988).

38. Gallo, S. M. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Research* **39,** D118–23 (Jan. 2011).

39. Garcia, H. G., Tikhonov, M., Lin, A. & Gregor, T. Quantitative Imaging of Transcription in Living Drosophila Embryos Links Polymerase Activity to Patterning. *Current biology : CB* (Oct. 2013).

40. Garcia, M., Nahmad, M., Reeves, G. T. & Stathopoulos, A. Size-dependent regulation of dorsal-ventral patterning in the early Drosophila embryo. *Developmental Biology* **381,** 286–299 (Sept. 2013).

41. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research* **17,** 877–885 (June 2007).

42. Graur, D. *et al.* On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution* **5,** 578–590 (2013).

43. Graveley, B. R. *et al.* The developmental transcriptome of Drosophila melanogaster. *Nature* **471,** 473–479 (Mar. 2011).

44. Gregor, T., Bialek, W., de Ruyter van Steveninck, R. R., Tank, D. W. & Wieschaus, E. F. Diffusion and scaling during early embryonic pattern formation. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 18403–18407 (Dec. 2005).

45. Gregor, T., Tank, D. W., Wieschaus, E. F. & Bialek, W. Probing the limits to positional information. *Cell* **130,** 153–164 (July 2007).

46. Gregor, T., Wieschaus, E. F., McGregor, A. P., Bialek, W. & Tank, D. W. Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell* **130,** 141–152 (July 2007).

47. Grimm, O., Coppey, M. & Wieschaus, E. Modelling the Bicoid gradient. *Development* **137,** 2253–2264 (July 2010).

48. Grimm, O. & Wieschaus, E. The Bicoid gradient is shaped independently of nuclei. *Development* **137,** 2857–2862 (Sept. 2010).

49. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38,** e131 (July 2010).

50. Hansson, B. S. & Anton, S. Function and morphology of the antennal lobe: new developments. *Annual review of entomology* **45,** 203–231 (2000).

51. Harding, K., Hoey, T., Warrior, R. & Levine, M. Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. *The EMBO journal* **8,** 1205–1212 (Apr. 1989).

52. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genetics* **4,** e1000106 (June 2008).

53. Harrison, M. M., Li, X.-y., Kaplan, T., Botchan, M. R. & Eisen, M. B. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genetics* **7,** e1002266 (Oct. 2011).

54. Hartmann, B., Reichert, H. & Walldorf, U. Interaction of gap genes in the Drosophila head: tailless regulates expression of empty spiracles in early embryonic patterning and brain development. *Mechanisms of Development* **109,** 161–172 (Dec. 2001).

55. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell reports* (Aug. 2012).

56. He, X., Samee, M. A. H., Blatti, C. & Sinha, S. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Computational Biology* **6** (2010).

57. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* **9,** 90–95 (July 2007).

58. Ilsley, G. R., Fisher, J., Apweiler, R., DePace, A. H. & Luscombe, N. M. Cellular resolution models for even skipped regulation in the entire Drosophila embryo. *eLife* **2,** e00522 (2013).

59. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (Oct. 2004).

60. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research* **21,** 1160–1167 (July 2011).

61. Jaeger, J., Blagov, M., *et al.* Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. *Genetics* **167,** 1721–1737 (Aug. 2004).

62. Jaeger, J. & Monk, N. Bioattractors: dynamical systems theory and the evolution of regulatory processes. *The Journal of physiology* **592,** 2267–2281 (June 2014).

63. Janssens, H. *et al.* Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. *Nature Genetics* **38,** 1159–1165 (Oct. 2006).

64. Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M. & Epstein, D. J. A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* **133,** 761–772 (Feb. 2006).

65. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Research* **21,** 1543–1551. ISSN: 1088-9051 (Sept. 2011).

66. Jones, E., Oliphant, T., Peterson, P., *et al. SciPy: Open source scientific tools for Python* (2001).

67. Junker, J. P. *et al.* Genome-wide RNA Tomography in the zebrafish embryo. *Cell* **159,** 662–675 (Oct. 2014).

68. Kantorovitz, M. R. *et al.* Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. *Developmental cell* **17,** 568–579 (Oct. 2009).

69. Kaplan, T. *et al.* Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genetics* **7,** e1001290 (2011).

70. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14,** R36 (Apr. 2013).

71. Klebes, A., Biehs, B., Cifuentes, F. & Kornberg, T. B. Expression profiling of Drosophila imaginal discs. *Genome Biology* **3,** RESEARCH0038 (July 2002).

72. Kosman, D. *et al.* Multiplex detection of RNA expression in Drosophila embryos. *Science (New York, N.Y.)* **305,** 846 (Aug. 2004).

73. Kraut, R. & Levine, M. Mutually repressive interactions between the gap genes giant and Krüppel define middle body regions of the Drosophila embryo. *Development* **111,** 611–621 (Feb. 1991).

74. Krotov, D., Dubuis, J. O., Gregor, T. & Bialek, W. Morphogenesis at criticality. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 3683–3688 (Mar. 2014).

75. Kuntz, S. G. & Eisen, M. B. Drosophila embryogenesis scales uniformly across temperature in developmentally diverse species. *PLoS Genetics* **10,** e1004293 (Apr. 2014).

76. Lagha, M., Bothma, J. P. & Levine, M. Mechanisms of transcriptional precision in animal development. *Trends in genetics : TIG* **28,** 409–416 (Aug. 2012).

77. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (Apr. 2012).

78. La Rosée, A., Häder, T., Taubert, H., Rivera-Pomar, R. & Jäckle, H. Mechanism and Bicoid-dependent control of hairy stripe 7 expression in the posterior region of the Drosophila embryo. *The EMBO journal* **16,** 4403–4411 (July 1997).

79. Lécuyer, E. *et al.* Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **131,** 174–187 (Oct. 2007).

80. Lee, J. H. *et al.* Highly Multiplexed Subcellular RNA Sequencing in Situ. *Science (New York, N.Y.)* **343,** 1360–1363 (Feb. 2014).

81. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* **7,** 709–715 (Aug. 2010).

82. Liang, H.-L. *et al.* The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature* **456,** 400–403 (Nov. 2008).

83. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079 (Aug. 2009).

84. Lin, C. *et al.* Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell* **151,** 56–67. ISSN: 0092-8674 (Sept. 2012).

85. Li, X.-y., Harrison, M. M., Villalta, J. E., Kaplan, T. & Eisen, M. B. Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition. *eLife* **3** (2014).

86. Li, X.-y., Thomas, S., *et al.* The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biology* **12,** R34 (2011).

87. Lott, S. E. *et al.* Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-seq. *PLoS Biology* **9,** e1000590 (2011).

88. Luria, S. E. & Delbrück, M. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28,** 491–511 (Nov. 1943).

89. MacArthur, S. *et al.* Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biology* **10,** R80 (2009).

90. Mamanova, L. *et al.* FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nature Methods* **7,** 130–132 (Jan. 2010).

91. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences* **100,** 11980–11985 (Oct. 2003).

92. Margolis, J. S. *et al.* Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element. *Development* **121,** 3067–3077 (July 1995).

93. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Research* **24,** 496–510 (Dec. 2013).

94. McQuilton, P., St Pierre, S. E., Thurmond, J. & FlyBase Consortium. FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Research* **40,** D706–14 (Jan. 2012).

95. modENCODE Consortium *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science (New York, N.Y.)* **330,** 1787–1797 (Dec. 2010).

96. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (July 2008).

97. Mousavi, K. *et al.* eRNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci. *Molecular Cell* **51,** 606–617 (Sept. 2013).

98. Muhr, J., Andersson, E., Persson, M., Jessell, T. M. & Ericson, J. Groucho-mediated transcriptional repression establishes progenitor cell pattern and neuronal fate in the ventral neural tube. *Cell* **104,** 861–873 (Mar. 2001).

99. Nègre, N., Brown, C. D., Ma, L., *et al.* A cis-regulatory map of the Drosophila genome. *Nature* **471,** 527–531 (Mar. 2011).

100. Nègre, N., Brown, C. D., Shah, P. K., *et al.* A comprehensive map of insulator elements for the Drosophila genome. *PLoS Genetics* **6,** e1000814 (Jan. 2010).

101. Ochoa-Espinosa, A. *et al.* The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 4960–4965 (Apr. 2005).

102. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genetics* **31,** 69–73 (May 2002).

103. Panganiban, G. *et al.* The origin and evolution of animal appendages. *Proceedings of the National Academy of Sciences* **94,** 5162–5166 (May 1997).

104. Paris, M. *et al.* Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. *PLoS Genetics* **9,** e1003748 (2013).

105. Perez, F. & Granger, B. E. IPython: a system for interactive scientific computing. *Computing In Science & Engineering* (2007).

106. Perry, M. W., Boettiger, A. N., Bothma, J. P. & Levine, M. Shadow enhancers foster robustness of Drosophila gastrulation. *Current biology : CB* **20,** 1562–1567 (Sept. 2010).

107. Perry, M. W., Boettiger, A. N. & Levine, M. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 13570–13575 (Aug. 2011).

108. Picelli, S., Björklund, A. K., *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10,** 1096–1098 (Nov. 2013).

109. Picelli, S., Faridani, O. R., *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9,** 171–181 (Jan. 2014).

110. Polyakov, O. *et al.* Passive mechanical forces control cell-shape change during Drosophila ventral furrow formation. *Biophysical journal* **107,** 998–1010 (Aug. 2014).

111. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* **30,** 777–782 (Aug. 2012).

112. Rand, M. D., Kearney, A. L., Dao, J. & Clason, T. Permeabilization of Drosophila embryos for introduction of small molecules. *Insect biochemistry and molecular biology* **40,** 792–804 (Nov. 2010).

113. Riddihough, G. & Ish-Horowicz, D. Individual stripe regulatory elements in the Drosophila hairy promoter respond to maternal, gap, and pair-rule genes. *Genes & Development* **5,** 840–854 (May 1991).

114. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science (New York, N.Y.)* **328,** 636–639 (Apr. 2010).

115. Rubner, Y., Tomasi, C. & Guibas, L. J. A metric for distributions with applications to image databases, 59–66 (1998).

116. Samee, M. A. H. & Sinha, S. Quantitative modeling of a gene's expression from its intergenic sequence. *PLoS Computational Biology* **10,** e1003467 (Mar. 2014).

117. Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biology* **14,** R31 (Apr. 2013).

118. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* (Apr. 2015).

119. Schroeder, M. D. *et al.* Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biology* **2,** E271 (Sept. 2004).

120. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* **31,** 64–68 (May 2002).

121. Small, S., Blair, A. & Levine, M. Regulation of even-skipped stripe 2 in the Drosophila embryo. *The EMBO journal* **11,** 4047–4057 (Nov. 1992).

122. Small, S., Blair, A. & Levine, M. Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. *Developmental Biology* **175,** 314–324 (May 1996).

123. Small, S., Kraut, R., Hoey, T., Warrior, R. & Levine, M. Transcriptional regulation of a pair-rule stripe in Drosophila. *Genes & Development* **5,** 827–839 (May 1991).

124. Staller, M. V., Fowlkes, C. C., *et al.* A gene expression atlas of a bicoid-depleted Drosophila embryo reveals early canalization of cell fate. *Development* **142,** 587–596 (Jan. 2015).

125. Staller, M. V., Vincent, B. J., *et al.* Shadow enhancers enable Hunchback bifunctionality in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 785–790 (Jan. 2015).

126. Staller, M. V., Yan, D., *et al.* Depleting gene activities in early Drosophila embryos with the "maternal-Gal4-shRNA" system. *Genetics* **193,** 51–61 (Jan. 2013).

127. Stark, A. *et al.* Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450,** 219–232 (Nov. 2007).

128. Steiner, F. A., Talbert, P. B., Kasinathan, S., Deal, R. B. & Henikoff, S. Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome Research* **22,** 766–777 (Apr. 2012).

129. Strom, A. R. & Barolo, S. Using the game of Mastermind to teach, practice, and discuss scientific reasoning skills. *PLoS Biology* **9,** e1000578 (2011).

130. Sweeton, D., Parks, S., Costa, M. & Wieschaus, E. Gastrulation in Drosophila: the formation of the ventral furrow and posterior midgut invaginations. *Development* **112,** 775–789 (July 1991).

131. Tadros, W. & Lipshitz, H. D. The maternal-to-zygotic transition: a play in two acts. *Development* **136,** 3033–3042 (Sept. 2009).

132. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6,** 377–382 (May 2009).

133. Tan, M. H. *et al.* RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. *Genome Research* **23,** 201–216 (Jan. 2013).

134. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 18602–18607 (Nov. 2013).

135. Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biology* **12,** R43 (2011).

136. Tomancak, P. *et al.* Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biology* **8,** R145 (2007).

137. Trapnell, C., Hendrickson, D. G., *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31,** 46–53 (Jan. 2013).

138. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* **25,** 1105–1111 (May 2009).

139. Trapnell, C., Roberts, A., *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (Mar. 2012).

140. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (May 2014).

141. Van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most "dark matter" transcripts are associated with known genes. *PLoS Biology* **8,** e1000371 (May 2010).

142. Van Rossum, G. & Drake, F. L. *Python language reference manual* (2003).

143. Wand, H., Iversen, J., Law, M. & Maher, L. Quilt plots: a simple tool for the visualisation of large epidemiological data. *PLoS ONE* **9,** e85047 (2014).

144. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474,** 390–394 (June 2011).

145. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (Apr. 1953).

146. Wilczynski, B., Liu, H., Yeo, Z. X. & Furlong, E. E. M. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Computational Biology* **8,** e1002798 (2012).

147. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods* **11,** 41–46 (Jan. 2014).

148. Wu, L. H. & Lengyel, J. A. Role of caudal in hindgut specification and gastrulation suggests homology between Drosophila amnioproctodeal invagination and vertebrate blastopore. *Development* **125,** 2433–2442 (July 1998).

149. Xu, H. *et al.* FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS ONE* **7,** e52249 (2012).

150. Zabet, N. R. & Adryan, B. Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Research* **43,** 84–94 (Jan. 2015).

Appendix A

Estimating Complexity of RNAseq Libraries

The presence of a PCR amplification step in most RNA-seq protocols has the potential to introduce a significant number of duplicated reads that arise from the same cDNA fragment (Benjamini & Speed 2012; Xu *et al.* 2012). Although there are experimental approaches to mitigate this effect (Mamanova *et al.* 2010), another computational option is to simply remove reads that map to identical locations(H. Li *et al.* 2009; Xu *et al.* 2012). However, this has the potential to remove bonafide duplicates—those that came from different cDNA fragments, but due perhaps to the high expression level of a gene or deep sequencing for a library, came from the same location along a transcript. Moreover, we have found that libraries with high levels of PCR duplication tend to be less reliable when compared to replicates, and so prefer to simply discard those libraries, regenerating them from the initial RNA when possible.

However, there is not to our knowledge an accepted method to determine whether a sample has a high level of duplication. While examining GBrowse tracks for "blockiness" can qualitatively identify problems (as in Figure A1), it is both time consuming and not particularly rigorous. Noting the fraction of reads that map to unique locations is imperfect, as it will be sensitive to the depth of sequencing, and therefore potentially difficult to compare across libraries.

Comparing the number of unique read start sites to the total number of bases in a gene offers one potential way to quantify the amount of PCR duplication actually present in the library. We simulated drawing unique, independent positions from a 1.5kb transcript (the average *Drosophila* transcript size, according to Daines *et al.* (2011))., and noting the fraction of unique start sites. We recognize that this simulation is highly idealized: there is a position-dependent bias, most often favoring the 5′ end of the read (Mortazavi *et al.* 2008; Hansen *et al.* 2010; Picelli, Faridani, *et al.* 2014); not all read start sites are equally likely, as fragmentation followed size selection can lose the reads closest to either end; and in a stranded protocol (which we have not tested here) the "forward" read must, necessarily, come before the "reverse" read. Nevertheless, this captures the essence of the problem, and

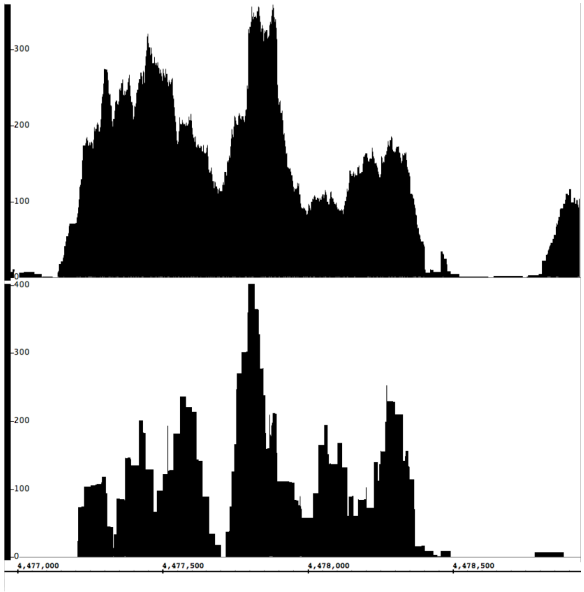


Figure A1. GBrowse views of the same 2kb region of Chromosome 3R in two libraries with different levels of duplication. Despite the similar number of reads in each sample, the lower library has much less information, due to a relatively high level of PCR duplication.

we will show that it closely matches real data.

For each gene, we calculated f, the fraction of bases in that gene that had a read start at that position. We plotted this quantity in figure A2 against the number of reads mapping to that gene divided by the gene length (that is, the total coverage, c). As expected, as the average coverage for each gene increased, the fraction of unique start sites increased as well, until the coverage approached 1x, at which point the available start sites became saturated.

Crucially, when we plotted these on a log-log plot, the region below about 10% of the occupied start sites was approximately linear for at least 3 logs below that. The fit equation was:

$$\log_{10} f = m_{sim} \cdot \log_{10} c + b_{sim} \approx .986 \log_{10} c - 0.031 \tag{A.1}$$

or

$$f = 10^{b_{sim}} x^m \tag{A.2}$$

We expect the slope m to be slightly less than 1, to indicate that increasing the coverage should increase the fraction of start sites occupied, but with some chance of multiple, independent reads coming from the same location, even in the absence of duplication.

This leads to an easy interpretation of the score $B = 10^{b_{sim}-b}$ as the average level of PCR duplication. A lower intercept corresponds to a shift to the right on the plot, which in turn means that the coverage to yield a given number of unique sites must be higher.

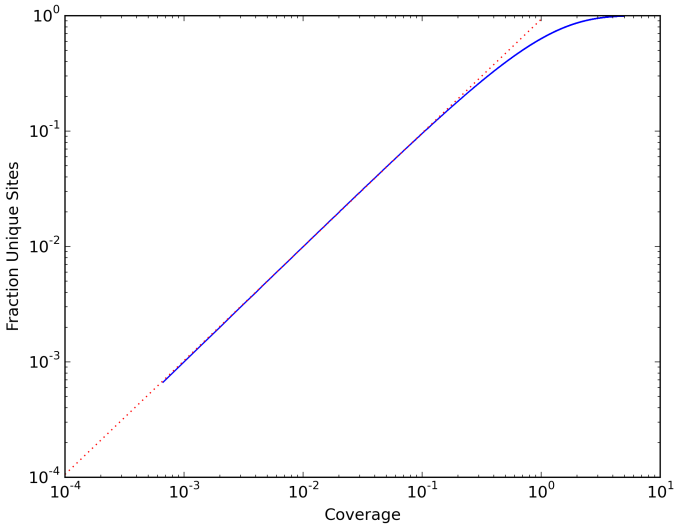


Figure A2. Log-log plot of simulated unique sites as a function of coverage. Simulated values are in blue, and the fit to the portion of the graph below 10% of the unique start sites is the dotted red line.

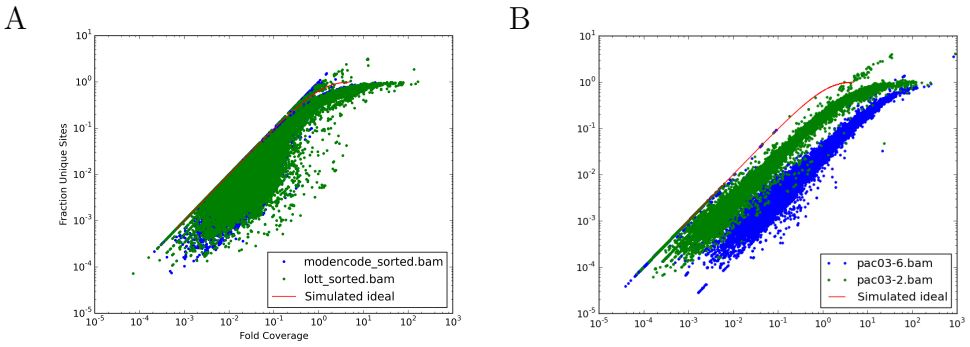


Figure A3. Log-log plot of actual of fractions of unique sites vs coverage. A) Previously published RNAseq data. Green points from modENCODE Consortium *et al.* 2010, with a B-score of 1.3. Blue points from Lott *et al.* 2011, with a B-score of 2.6. B) Unpublished, low-quality RNAseq data. Green points have a B-score of 5.7, blue points have a B-score of 36.5

Protocol	% Virilis	b	m	B score
TruS	0	-0.6546	0.9195	4.2030
TruS	5	-0.4405	1.0584	2.5676
TruS	10	-0.5567	0.9998	3.3550
TruS	20	-0.5245	1.0224	3.1149
CT	0	-0.9669	0.8645	8.6285
CT	5	-0.5320	1.0386	3.1701
CT	10	-0.3784	1.1027	2.2254
CT	20	-0.5421	1.0335	3.2442
TotS	0	-0.2840	1.0871	1.7907
TotS	5	-0.2583	1.0769	1.6877
TotS	10	-0.4882	1.0199	2.8659
TotS	20	-0.4940	1.0202	2.9046
S2	0	-0.2139	1.1206	1.5238
S2	5	-0.1748	1.1303	1.3927
S2	10	-0.1586	1.1324	1.3416
S2	20	-0.1607	1.1302	1.3483
S2—2.5×	0	-0.5398	1.0036	3.2275
S2—2.5×	1	-0.6005	0.9771	3.7118
S2—2.5×	5	-0.5448	1.0038	3.2645
S2—2.5×	10	-0.6117	0.9713	3.8084
S2—2.5×	20	-0.5532	0.9987	3.3284
S2—5×	0	-0.5944	0.9815	3.6599
S2—5×	1	-0.6373	0.9620	4.0397
S2—5×	5	-0.5707	0.9842	3.4650
S2—5×	10	-0.5695	0.9951	3.4557
S2—5×	20	-0.5831	0.9930	3.5658

Table A1. Fit parameters and estimate duplication rate (B-score) of libraries used in Chapter 3

When we calculated the B-score for samples from previously published datasets, both from our lab (Lott *et al.* 2011) and the modENCODE consortium (modENCODE Consortium *et al.* 2010), we found that the scores were all less than 3 (Figure A3A). By contrast, previous libraries from our lab that have been unpublished due to our lack of confidence that the data was not highly enriched for duplicates (including the lower panel of Figure A1) had B-scores in excess of 5 and up to 30 (Figure A3B). RNAseq data used in the main text of this study had B-scores as shown in table 1 in this document. Although in many cases they are higher than ideal, they are comparable to previously published data, and only one is greater than 4.3.

We also simulated the fits for various sizes between 100bp and 10kb. Although the

fit parameters did have a clear, increasing trend in response to increasing the simulated transcript size (Figure A4), the increase was small compared to the actual values. A variation of 0.005 in the intercept, which is used to calculate the B-score, corresponds to an actual difference of about 1%. We are thus not concerned about the choice of 1.5kb to simulate the ideal scenario.

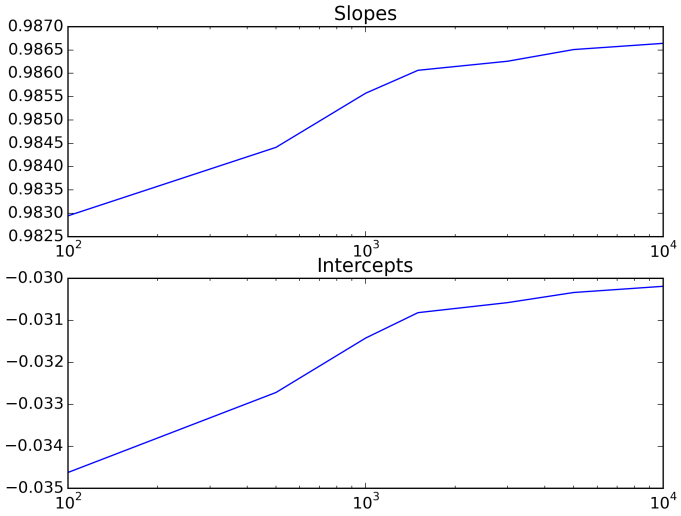


Figure A4. Regression coefficients vs size of transcript. Simulated at 100, 500, 1000, 1500, 3000, 5000, and 10,000 bp transcript sizes.

A.1 Simulation Code

```
from __future__ import division
from scipy import random
from numpy import (zeros, zeros_like, arange,
                   unique, mean, log10)
from scipy.stats import linregress
from progressbar import ProgressBar

avg_size = 1500.
n_reps = 100

regs = []
all_sizes = [100., 500., 1000., 1500., 3000., 5000., 10000.]
```

```
for avg_size in all_sizes:
    cs = arange(1, 5*avg_size)
    fs = zeros_like(cs)

    pb = ProgressBar()

    all_fracs = []
    for c in pb(cs):
        fracs = zeros(n_reps)
        for i in range(n_reps):
            fracs[i] = len(unique(
                random.randint(0, avg_size, c))
                          )/avg_size
            # Use randint to generate random read positions,
            # then count the number of unique start sites
            # and normalize by size of the transcript
            all_fracs.append((c/avg_size, fracs[i]))
        fs[c-1] = mean(fracs)

    print('-'*30)
    print(avg_size)
    print('-'*30)
    regs.append(linregress(log10(cs[fs<.1]/avg_size),
                           log10(fs[fs<.1])))
    print(regs[-1])
```