Sequencing mRNA from cryo-sliced *Drosophila* embryos to determine genome-wide spatial patterns of gene expression

Peter A. Combs[1,*], Michael B. Eisen[2,3]

1 Graduate Program in Biophysics, University of California, Berkeley, California, United States of America

2 Department of Molecular and Cell Biology, University of California, Berkeley, California, United States of America

3 Howard Hughes Medical Institute, University of California, Berkeley, California, United States of America

∗ **E-mail: peter.combs@berkeley.edu**

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 patterns of gene expression. *In situ* imaging of transcripts and proteins is the gold-standard, but is difficult and time consuming to apply to an entire genome, even when highly automated. Sequencing, in contrast, is fast and genome-wide, but generally applied to homogenized tissues, thereby discarding spatial information. At some point, these methods will converge, and we will be able to sequence RNAs *in situ*, simultaneously determining their identity and location. As a step along this path, we developed methods to cryosection individual blastoderm stage *Drosophila melanogaster* embryos along the anterior-posterior axis and sequence the mRNA isolated from each 60μm slice. The spatial patterns of gene expression we infer closely match patterns determined by *in situ* hybridization and microscopy, where such data exist, and thus we conclude that we have generated the first genome-wide map of spatial patterns in the *Drosophila* embryo. We identify numerous genes with spatial patterns that have not yet been screened in the several ongoing systematic in situ based projects, the majority of which are localized to the posterior end of the embryo, likely in the pole cells. 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.

Introduction

Analyzing gene expression in multicellular organisms has long involved a tradeoff between the spatial precision of imaging and the efficiency and comprehensiveness of genomic methods. RNA *in situ* hybridization and antibody staining of fixed samples, or fluorescent imaging of live samples, provides high resolution spatial information for small numbers of genes [1–3]. But even with automated sample preparation, imaging, and analysis, *in situ* based methods are difficult to apply to an entire genomes worth of transcripts or proteins. High throughput genomic methods, such as DNA microarray hybridization or RNA sequencing, are fast and relatively inexpensive, but 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 [4], 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 [6]. 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 patterning in the early embryo. The

system is extremely well characterized, and there are a large number of genes with well-characterized A-P patterns against which to compare our results. The geometry of the early embryo also lends itself to biologically meaningful physical dissection by simple sectioning along the elongated A-P axis.

Results

We collected *D. melanogaster* embryos, 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. We embedded each embryo in a cryoprotecting gel, flash-froze it in liquid nitrogren, and took transverse sections along the anterior-posterior axis. We used 60μm sections, meaning that we cut each approximately 350μm embryo into 6 pieces. We placed each piece into a separate tube, isolated RNA, and prepared sequencing libraries.

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 of RNA present—approximately 15ng of total RNA per slice. To overcome this limitation, after RNA extraction we added RNA from a single embryo of a divergent *Drosophila* species to each tube to serve as a carrier. As we only used distantly related and fully sequenced species as carriers, we could readily separate reads derived from the *D. melanogaster* slice and the carrier species computationally after sequencing. With the additional approximately 100ng of total RNA in each sample, library preparation became far more robust.

We sliced and sequenced three wild-type *D. melanogaster* embryos, with summary statistics for the mRNA sequencing results shown in Table 1. To ensure that our libraries faithfully recapitulated known spatial profiles, we manually examined a panel of genes with known spatial distributions, as detected in prior ISH studies [2, 5]. In these cases, there was close qualitative agreement between the visualized expression patterns and our sliced RNA-seq data (Figure 1A).

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 [1], transforming the relative expression levels from these images into absolute values (FPKM) using genome-wide expression data from single embryos [6]. We first compared the sum of FPKM values across all slices to the values for whole single embryos, which converged on the correct stage after only about 40 genes, and remained constant thereafter (Supplemental Figure 1). We then compared expression data from each slice to all possible virtual slices of 60μm , and identified likely positions of the slice along the A-P axis (Figure 1B). The positions estimated 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 examined all genes for differential expression between slices using Cufflinks and Cuffdiff [7] and identified 85 genes differentially expressed between slices (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 [2]. Of our differentially expressed genes, 21 had no imaging data available, and 33 were annotated as present in a subset of the embryo; the remaining 31 genes showed either clear patterns that were not annotated, or no staining (Supplemental Figure 2). There were an additional 194 expressed genes tagged as present in a subset of the embryo, but most of these had primarily dorsal-ventral patterns, faint patterns, later staging in the images used for annotation, or simply fell above the threshold of significance in our data, with generally good qualitative agreement (Supplemental Figure 3).

We next turned our attention to maternally deposited genes, of which we expected only a small subset to have patterned expression. We used previous data from our lab on embryos with a genetic background capable of distinguishing maternal patterns of expression [6]. When we sorted genes by the

center of expression mass, we found that many fewer of these maternal genes have spatially localized expression (Figure 2). More zygotic genes had the bulk of their expression in the anterior versus the posterior (330 vs 132, respectively, had the center of expression mass outside the central 60μm), whereas the maternal genes had essentially the opposite pattern (138 vs 370). The posterior-most slice shows a noticeably different pattern of expression among the maternal genes, which we believe is attributable to the population of pole cells in the posterior.

Discussion

We view the experiments reported here as a proof of principle for sectioning based methods to systematically characterize spatial patterns of expression. While we are by no means the first to dissect samples and characterize their RNAs—Ding and Lipshitz pioneered this kind of analysis twenty years ago [8]—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.

One can envision three 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. [4]. 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.

We believe 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 slices we used here represented the limit of sample preparation and sequencing methods available at the time, but with methods having been described to sequence the RNAs from "single" cells, it should be possible to obtain far better spatial resolution in the near future.

In principle, nothing limits our approach to *D. melanogaster*. While we used other *Drosophila* species for the carrier RNA, generating similar position-specific data for those related species could easily piggyback off of concurrent *D. melanogaster* work. Previous cross-species work on patterning has principally focused on only a handful of genes at once [9, 10], but similar data can help illuminate the sequence-expression connection by examining hundreds of genes simultaneously.

Finally, as sequencing costs continue to plummet, it should be possible to sequence greater numbers of increasingly small samples. According to our estimates, a single embryo contains enough RNA to sequence over 700 samples to a depth of 20 million reads. While this number of samples would necessitate more advanced sectioning techniques, the ultimate goal of knowing the localization of every single transcript is rapidly becoming feasible.

Materials and Methods

Fly Line, Imaging, and Slicing

We raised flies on standard media at 25° in uncrowded conditions, and collected eggs from many 3–10-day old females. Wild-type flies were *Canton-S* lab stocks. We washed and dechorionated the embryos, then fixed them according to a standard methanol cracking protocol. We 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, 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 . 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 [6]). We pooled Total RNA 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 [11, 12]. 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 which is available immediately. The processed data are available at the journal website and at eisenlab.org/sliceseq. All custom analysis software is available github.com/petercombs/Eisenlab-Code, and is primarily written in Python [13–17]. Commit b0b115a was used to perform all analysis in this paper.

Acknowledgments

References

1. Fowlkes CC, Hendriks CLL, Keränen SVE, Weber GH, Rübel O, et al. (2008) A Quantitative Spatiotemporal Atlas of Gene Expression in the Drosophila Blastoderm. Cell .

2. Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, et al. (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biology .

3. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, et al. (2007) Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. Cell .

4. Steiner FA, Talbert PB, Kasinathan S, Deal RB, Henikoff S (2012) Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. Genome Research .

5. Kumar S, Konikoff C, Van Emden B, Busick C, Davis KT, et al. (2011) FlyExpress: visual mining of spatiotemporal patterns for genes and publications in Drosophila embryogenesis. Bioinformatics (Oxford, England) .

6. Lott SE, Villalta JE, Schroth GP, Luo S, Tonkin LA, et al. (2011) Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-seq. PLoS Biology .

7. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, et al. (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology .

8. Ding D, Lipshitz HD (1993) A molecular screen for polar-localised maternal RNAs in the early embryo of Drosophila. Zygote (Cambridge, England) .

9. Lott SE, Kreitman M, Palsson A, Alekseeva E, Ludwig MZ (2007) Canalization of segmentation and its evolution in Drosophila. Proceedings of the National Academy of Sciences of the United States of America .

10. Gregor T, Bialek W, de Ruyter van Steveninck RR, Tank DW, Wieschaus EF (2005) Diffusion and scaling during early embryonic pattern formation. Proceedings of the National Academy of Sciences of the United States of America .

11. McQuilton P, St Pierre SE, Thurmond J, the FlyBase Consortium (2011) FlyBase 101 - the basics of navigating FlyBase. Nucleic Acids Research .

12. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England) .

13. Python language reference manual.

14. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England) .

15. Hunter JD (2007) Matplotlib: A 2D graphics environment. Computing In Science & Engineering .

16. SciPy: Open source scientific tools for Python.

17. Perez F, Granger BE (2007) IPython: a system for interactive scientific computing. Computing In Science & Engineering .

Figure Legends

Tables

Figure 1. Expression in the slices closely matches previous expression data. (A) Expression data broadly parallels ISH for individual genes. The genes shown are those with previously known A-P patterns. Y-axes are scaled to maximum FPKM values as reported by Cufflinks, and are different for each panel. (B) Expression data closely matches with previous quantitative data at the same stages. We used a Bayesian procedure to estimate the location of each 60 micron slice with reference to an ISH-based atlas, with absolute expression levels set using whole embryo RNA-seq data. The line graphs represent the distribution of position estimates for each slice, and the colored bars are one sixth the embryo width and placed at the position of greatest probability.

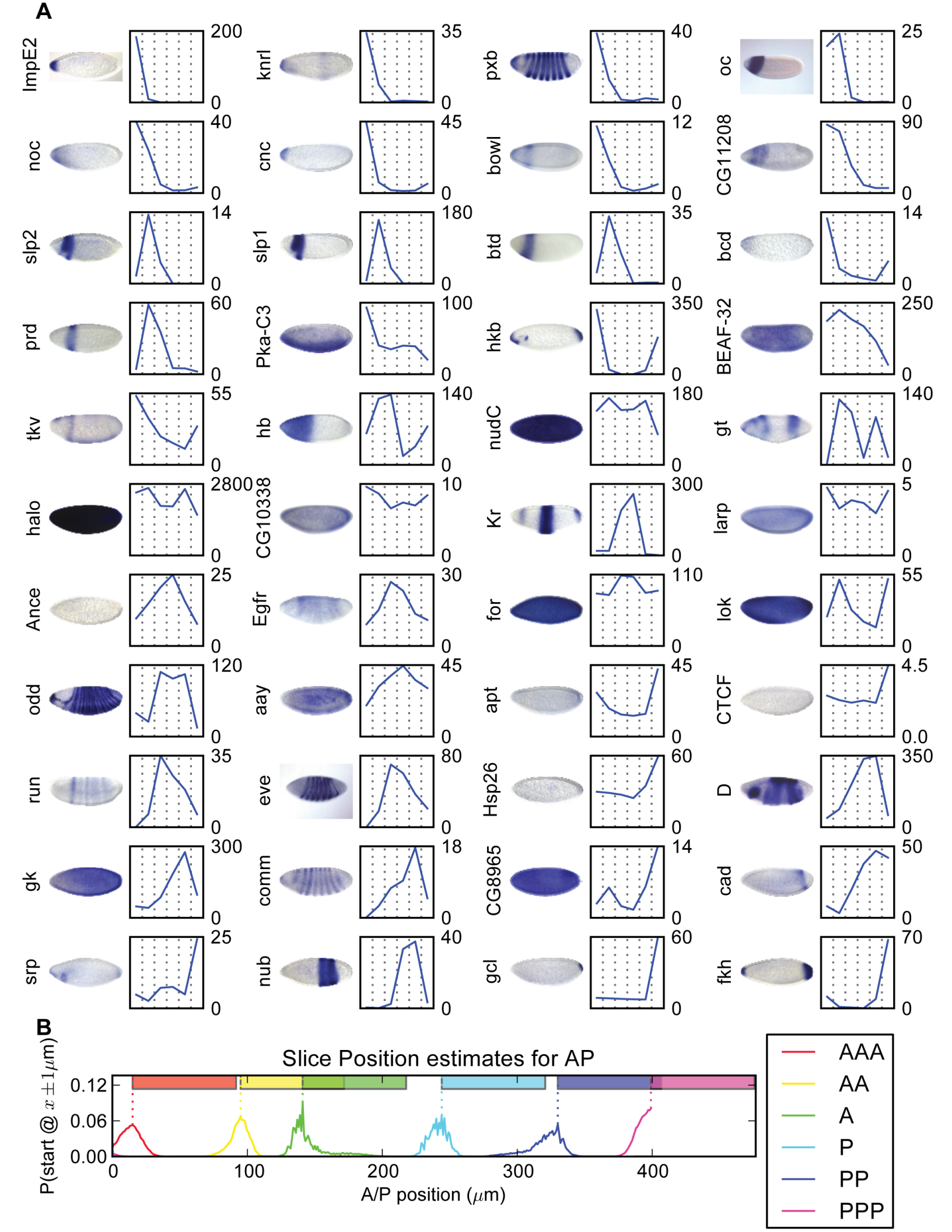
Figure 2. Heat map of wild-type gene expression sorted by center-of-mass of expression. Expression is normalized by gene, and maternal and zygotic genes are plotted separately.

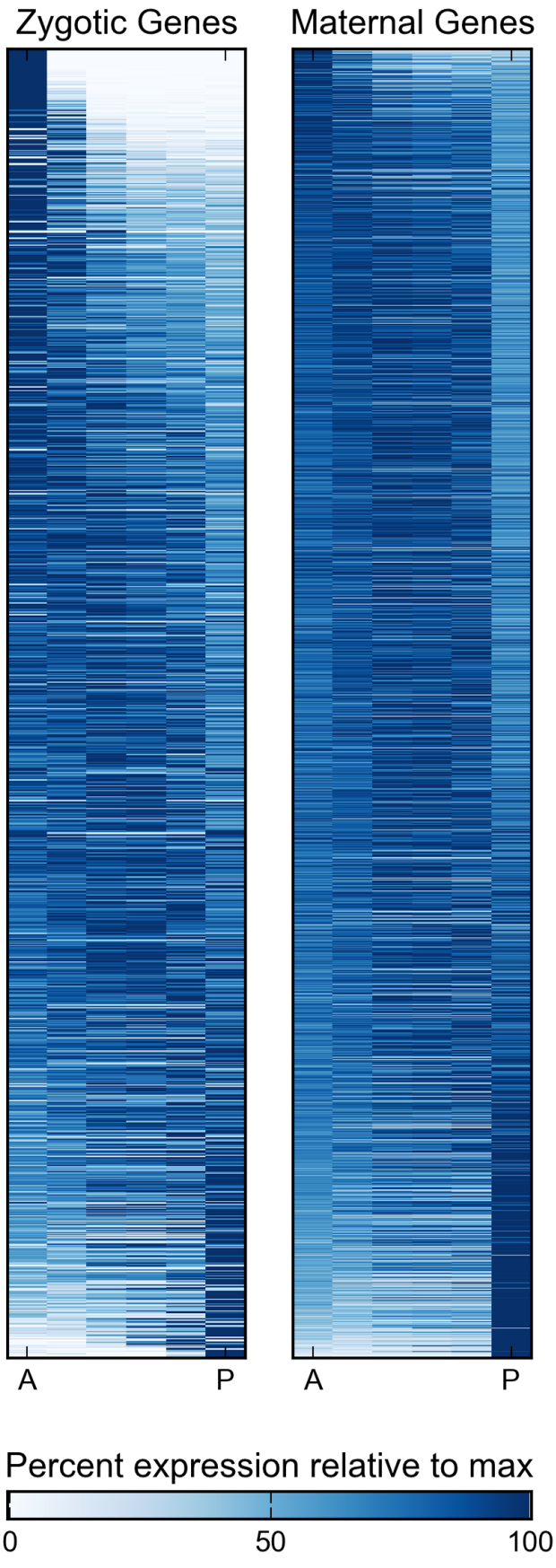
Table 1. Sequencing statistics for sliced wild-type mRNA-Seq samples

Replicate	Slice	Carrier Species	Barcode 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	1,249,6248 (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%)

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.

Figures


A
ImpE2
200
0
knrl
35
0
pxb
40
0
oc
25
0
noc
40
0
cnc
45
0
bowl
12
0
CG11208
90
0
slp2
14
0
slp1
180
0
btd
35
0
bcd
14
0
prd
60
0
Pka-C3
100
0
hkb
350
0
BEAF-32
250
0
tkv
55
0
hb
140
0
nudC
180
0
gt
140
0
halo
2800
0
CG10338
10
0
Kr
300
0
larp
5
0
Ance
25
0
Egfr
30
0
for
110
0
lok
55
0
odd
120
0
aay
45
0
apt
45
0
CTCF
4.5
0.0
run
35
0
eve
80
0
Hsp26
60
0
D
350
0
gk
300
0
comm
18
0
CG8965
14
0
cad
50
0
srp
25
0
nub
40
0
gcl
60
0
fkh
70
0
B
Slice Position estimates for AP
P(start @ $x \pm 1\mu$m)
0.12
0.06
0.00
0
100
200
300
400
A/P position (μm)
AAA
AA
A
P
PP
PPP


Zygotic Genes
Maternal Genes
A
P
A
P
Percent expression relative to max
0
50
100

Supplemental Figures

Available on request from peter.combs@berkeley.edu