Title

The conserved regulatory basis of mRNA contributions to the early Drosophila embryo differs between the maternal and zygotic genomes

Journal

PLOS Genetics, 16(3)

ISSN

1553-7390

Authors

Omura, Charles S
Lott, Susan E

Publication Date

2020

DOI

10.1371/journal.pgen.1008645

Peer reviewed

PLOS GENETICS

RESEARCH ARTICLE

The conserved regulatory basis of mRNA contributions to the early *Drosophila* embryo differs between the maternal and zygotic genomes

Charles S. Omura*, Susan E. Lott*

Department of Evolution and Ecology, University of California, Davis, Davis, California, United States of America

* csomura@ucdavis.edu (CSO); selott@ucdavis.edu (SEL)

OPEN ACCESS

Citation: Omura CS, Lott SE (2020) The conserved regulatory basis of mRNA contributions to the early *Drosophila* embryo differs between the maternal and zygotic genomes. PLoS Genet 16(3): e1008645. https://doi.org/10.1371/journal.pgen.1008645

Editor: Melissa M. Harrison, University of Wisconsin Madison School of Medicine and Public Health, UNITED STATES

Received: September 13, 2019

Accepted: February 3, 2020

Published: March 30, 2020

Data Availability Statement: All raw and processed data are available at NCBI/GEO under accession number GSE112858.

Funding: This work was supported by the National Institutes of Health, National Institute of General Medical Sciences, grant number R01GM111362 to SEL and by start-up funds from the University of California, Davis to SEL. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The gene products that drive early development are critical for setting up developmental trajectories in all animals. The earliest stages of development are fueled by maternally provided mRNAs until the zygote can take over transcription of its own genome. In early development, both maternally deposited and zygotically transcribed gene products have been well characterized in model systems. Previously, we demonstrated that across the genus *Drosophila*, maternal and zygotic mRNAs are largely conserved but also showed a surprising amount of change across species, with more differences evolving at the zygotic stage than the maternal stage. In this study, we use comparative methods to elucidate the regulatory mechanisms underlying maternal deposition and zygotic transcription across species. Through motif analysis, we discovered considerable conservation of regulatory mechanisms associated with maternal transcription, as compared to zygotic transcription. We also found that the regulatory mechanisms active in the maternal and zygotic genomes are quite different. For maternally deposited genes, we uncovered many signals that are consistent with transcriptional regulation at the level of chromatin state through factors enriched in the ovary, rather than precisely controlled gene-specific factors. For genes expressed only by the zygotic genome, we found evidence for previously identified regulators such as Zelda and GAGA-factor, with multiple analyses pointing toward gene-specific regulation. The observed mechanisms of regulation are consistent with what is known about regulation in these two genomes: during oogenesis, the maternal genome is optimized to quickly produce a large volume of transcripts to provide to the oocyte; after zygotic genome activation, mechanisms are employed to activate transcription of specific genes in a spatio-temporally precise manner. Thus the genetic architecture of the maternal and zygotic genomes, and the specific requirements for the transcripts present at each stage of embryogenesis, determine the regulatory mechanisms responsible for transcripts present at these stages.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Early development in animals is a unique period of time, as it is controlled by gene products from two different genomes: that of the mother and that of the zygote. The earliest stages of development are directed by maternal mRNAs and proteins that are deposited into the egg, and only later does the zygote take over the transcription of its own genome. In this paper, we use data from 11 fruit fly species characterizing all the genes transcribed by the mother and by the zygote, to investigate how transcription is regulated in the maternal and zygotic genomes. While we find some conserved regulatory elements at both stages, regulation of maternal transcription is much more highly conserved across species. We present evidence that maternal transcription is controlled in large co-regulated chromatin domains, while zygotic transcription is much more gene-specific. These results make sense in the context of where these genes are being transcribed, as maternal transcripts are generated in support cells which churn out a large amount of mRNA during oogenesis, while zygotic genes are often transcribed in a particular time and place in the embryo.

Introduction

Development is a sequential process, where each step builds on the one before it. The earliest stages of embryonic development are therefore critical, as processes such as cleavage cycles and the beginnings of axial patterning become the basis for all subsequent developmental processes. Regulation of these important tasks is controlled by mRNAs and proteins, and perhaps unsurprisingly then, mRNA levels in *Drosophila* are found to be precisely controlled during early embryogenesis [1,2]. This precise control of transcript levels is especially remarkable, however, given that the transcripts at early stages of development come from two different genomes, that of the mother and that of the zygote [3–5]. The regulatory mechanisms responsible for this precise control of transcript levels across both genomes are not yet fully understood.

During oogenesis, the oocyte itself is mostly transcriptionally silent [6]. Instead, support cells called nurse cells synthesize RNA, proteins, and organelles which are transported into the oocyte [7]. These maternally produced mRNAs are responsible for many of the critical events of early embryogenesis, such as the rapid cleavage cycles, the establishment of body axis, and the coordination of the handoff of control to the zygotic genome. This handoff of developmental control from mother to zygote, known as the maternal to zygotic transition (MZT), is complex from a regulatory standpoint. Critical housekeeping genes retain a steady transcript level, despite changing the genome of origin. New transcripts must be synthesized from the newly activated zygotic genome, and maternal transcripts must be degraded, in a highly regulated and time-specific manner [8]. This transition is well studied in model systems such as *Drosophila melanogaster*, where maternal mRNA degradation regulators such as *smaug* (*smg*) [8] and regulators critical to the activation of the zygotic genome such as *zelda* (*zld*) [9,10] have been identified. When the transition of developmental control between the two genomes is complete, the zygotic genome must be poised to carry out the rest of development in a precise manner. One well-studied process that exemplifies the precision required at the handoff to the zygotic genome is segmentation along the anterior-posterior axis in *Drosophila*. This process begins with broad maternal gradients which control transcription of early zygotic gap genes, and later pair-rule genes, at precise locations within the embryo at specific developmental times [11,12].

Regulation of transcription and transcript levels in development has been the subject of considerable study in *D. melanogaster*. Much of this previous work has been focused around the process of the MZT or other important events in early development. For example, a number of regulators of maternal transcript degradation at or prior to the MZT have been identified [13–16]. Zygotic transcription activation has also been the subject of considerable study, and has implicated critical transcription factors such as *zelda* and *grainy head* [4]. How transcripts are transported into eggs has been the subject of some study [7,17,18], as has how those maternal transcripts are regulated post-transcriptionally [3,19–23]. Post-transcriptional mRNA regulation is especially crucial at the maternal stage as new transcripts cannot be produced after the completion of oogenesis. However, how transcript production is regulated in the nurse cells is largely unknown. As transcript pools at both the maternal and zygotic stages are highly conserved over evolutionary time [24], we employed a comparative approach to investigate gene regulation at these stages.

In this study, we uncover regulatory elements that are associated with transcription in the early *Drosophila* embryo, from both maternally deposited and zygotically transcribed genes. We use motif analysis to compare regulation of maternal versus zygotic transcription, and also investigate how regulation at these two stages is different across *Drosophila* species. To this end, we used a previously generated RNA-seq dataset from Atallah and Lott, 2018, which sampled embryos from a developmental stage where all transcripts are maternal (stage 2 [25,26]) and a stage after zygotic genome activation (end of stage 5, or end of blastoderm stage), across 14 species, representing ~50 million years of divergence time. Here, we used the transcript abundance data from 11 of these species (due to limitations in genome annotation quality, see Methods), representing the same span in divergence time, to examine putative regulatory regions of maternally deposited or zygotically transcribed genes. Through comparisons of these sequences and associated gene transcription levels, we identified a number of sequence motifs as being enriched in either maternally deposited or early zygotically expressed genes. We found a high similarity between motifs across all species, suggesting a high level of conservation for regulation of transcription within each genome (maternal and zygotic). At the stage controlled by maternal transcripts, we found a number of motifs that bind to proteins annotated with insulator function or that have previously been associated with boundaries between topologically associating domains (TADs). Our findings suggest that maternal transcription is largely controlled through regulation of chromatin state, and not through gene-specific mechanisms. Many transcription factors predicted to bind the identified motifs were found to be enriched in ovaries [27]. After zygotic genome activation (end of stage 5), we find many of the motifs known to be associated with early zygotic transcription, such as the binding site for the pioneer transcription factor, Zelda, reinforcing many previously identified aspects of transcriptional regulation at this stage. We also find a larger number of motifs with less significant enrichment at this stage, with evidence that points to these motifs regulating a smaller subset of genes. This study provides evidence for global control of maternal transcription at the level of chromatin, while zygotic transcription is regulated in a more gene-specific manner. This is especially striking considering that the maternal transcript pool is more highly conserved than that of the early zygote [24].

Results

Discovered maternal-associated motifs bind architectural proteins; discovered zygotic-associated motifs bind to known zygotic regulators

To examine the regulatory basis of maternal and zygotic transcription, we surveyed the genomes of 11 *Drosophila* species for regulatory elements. These species represent the

evolutionary divergence of the *Drosophila* genus, encompassing divergence times from 250,000 to 50 million years [28]. The RNA-seq datasets produced from Atallah and Lott (2018) [24] were used. These data sampled two developmental stages, one where all transcripts present are maternally derived (stage 2, Bownes' stages [25,26]) and the other after zygotic genome activation (the end of stage 5, or the end of blastoderm stage). The transcript abundance data was used to classify each gene as being on or off at both stage 2 and stage 5 for each species (see Methods). For each gene, we extracted sequences at likely locations for proximal regulatory elements (see Methods). To accommodate the varying annotation quality of the various species, this search encompassed introns, exons, and a 2kb region upstream of the gene.

To identify motifs associated with maternally deposited genes, we employed HOMER [29]. For most species, a characteristic pattern emerged where the most enriched motifs were present in the upstream region of the maternally deposited genes, with less enriched motifs appearing in exons (S1 Fig). Some motifs, possibly representing repressor binding sites, were enriched in the upstream and intron region of genes that were not maternally deposited as compared to the genes that were maternally deposited (S1 Fig).

Analyzing regulatory elements at the post-zygotic genome activation stage (end of stage 5) presents a challenge, as it is difficult to distinguish newly transcribed zygotic mRNAs from residual maternally deposited mRNAs. At this stage, roughly half of the transcripts present are maternal transcripts that have not yet been degraded [8,30–32]. Therefore, to interrogate regulatory elements associated with zygotic transcription, we restricted our search to genes that do not have transcripts present at stage 2 but do have transcripts present by stage 5 (see Methods). We refer to these genes as zygotic-only. Because of these stricter requirements for zygotically transcribed genes, there were far fewer genes in the dataset (66,206 genes in the stage 2 dataset combined from all species, compared to 10,215 total genes in the stage 5 dataset for all species), resulting in a reduction in statistical power. However, without these assumptions, we fail to identify signals associated specifically with zygotic transcription amongst the signal of maternal transcription.

To determine which proteins are likely to bind to maternal or zygotic motifs, we used Tomtom [33] to evaluate the similarity of the discovered motifs to several motif databases for *D. melanogaster* (Fig 1). The motifs found in maternally deposited transcripts are similar to those discovered previously in two different contexts: those associated with topologically associated domains (TADs) [34], and those associated with housekeeping promoters [35,36]. This is consistent with existing data showing that functions of maternally deposited genes are enriched for genes with housekeeping activities [36,37]. In order to determine whether the motifs associated with maternal transcripts in our data were simply due to the inclusion of promoter elements from housekeeping genes, we measured the enrichment of these motifs in maternally deposited genes that are not housekeeping genes (see Methods). We found that our motifs are strongly enriched ($p < 1 \times 10^{-34}$) in maternally deposited genes even when excluding housekeeping genes (S2 Fig). This indicates that these motifs are having a strong effect outside that of those contained in housekeeping genes during this stage. Thus, we hypothesize that the regulatory mechanisms responsible for generating TADs [34] are also responsible for maternal transcripts, and that maternal transcription may be regulated by the establishment of TADs. TADs are genomic regions where the chromatin on one side of the boundary interacts substantially less than expected with the chromatin on the other side, and interactions of DNA elements within the domains can be promoted. While TADs are generally thought to be associated with transcription [34], there is some controversy as to the nature and magnitude of the effect of TADs on gene expression [38], as disruption of TADs has not been found to be sufficient to alter transcription in some cases.

The motifs associated with maternally deposited genes are predicted to bind several different insulators or architectural proteins. An insulator is a regulatory element that suppresses

	Logo	-log p-value	% Target	% Background	Best Match	Description
stage 2		2091	35.49%	12.32%	DREF	• "Master key-like factor for cell proliferation" (Akio Matsukage et al. 2008) • Shares binding site with BEAF-32
		2091	35.49%	12.32%	BEAF-32	• Insulator (Yang, Ramos, and Corces 2012; Nègre et al. 2010) • Shares binding site with DREF
		1591	27.85%	9.42%	M1BP	• Causes PolII to pause on the gene (Li and Gilmour 2013)
		934	27.35%	12.73%	ZIPIC	• Recruits insulator CP190 (Maksimenko et al. 2015).
		843	40.25%	23.98%	Ohler-6	• Commonly found between TAD boundaries (Ramirez et al 2018)
		692	20.13%	9.05%	E-box	• Regulates gene expression
stage 5		163	18.45%	8.77%	Zld	• "Master regulator of genome activation"
		84	35.74%	26.05%	Trl	• Required for embryogenesis • Known to regulate developmental genes
		56	48.41%	39.9%	Trl	• Required for embryogenesis • Known to regulate developmental genes

Fig 1. A summary of the top ranked motifs. HOMER was used to find motifs enriched in the 2kb windows upstream of maternally deposited genes (stage 2) and zygotically transcribed genes (stage 5). Sequence logo shows the consensus motif where the probability of each base is proportional to its representative character. P-value is given by HOMER. %target represents the percent of either maternally deposited or zygotically expressed genes that contain at least one instance of the motif. %background indicates the percent of all genes that contain this motif. Best match indicates protein with a previously identified binding site that mostly closely matches the discovered motif (see Methods).

https://doi.org/10.1371/journal.pgen.1008645.g001

the interactions of other regulatory elements with genes, or prevents the spread of chromatin state. An architectural protein is a protein that organizes and regulates chromatin structure. The most prominent motif by q-value binds to DNA replication-related element factor (DREF), a known architectural protein and the "master key-like factor for cell proliferation"

[39]. It is required for normal progression through the cell cycle. It is known to occur in the promoters of many cell proliferation genes and to interact with chromatin remodeling proteins. Interestingly, the DREF binding site overlaps with the binding site for BEAF-32, another well-studied protein that acts as an insulator [40,41] and often appears between head-to-head genes (genes with adjacent promoters that are transcribed in opposite directions). Another identified motif is predicted to bind ZIPIC, which is known to bind and recruit CP190, an insulator. A previous study provides evidence for the co-localization of ZIPC and BEAF-32 [42], which likely work together with CP190 to perform insulator functions. Thus of the most enriched motifs in maternal genes (DREF, BEAF-32, ZIPC), many have previously identified roles as insulators or in other ways regulating chromatin state.

Another maternal motif identified is predicted to bind M1BP (motif-1 binding protein), which causes RNA polymerase II (Pol II) to pause on the gene [43]. Pol II pausing is critical to early zygotic expression [36,44] but its function in producing the maternal transcriptome is unknown. Several functions have been suggested for this Pol II pausing behavior, including maximizing transcription speed once certain conditions are met, synchronizing with RNA processing machinery, reacting to other developmental or environmental signals, keeping chromatin accessible, and acting as an insulator [36,43,44]. Given that M1BP is both maternally deposited at high levels and has increased expression in the early embryo, it is possible that M1BP has multiple functions at different time points. During oogenesis, pausing to wait for external signals or RNA processing machinery seems counterproductive to maximizing transcription in the ovary, but the other function of maintaining a state of open chromatin and solidifying TAD boundaries may be very important. In contrast, at stage 5 it may be much more important to maximize expression in response to certain signals.

In searching for motifs associated with zygotic-only expression, we recovered motifs for well-known regulators of the zygotic genome (Fig 1). We only identified a small number of highly enriched motifs at this stage, and thus were able to predict a much smaller number of predicted factors binding to these motifs, including Trl (or GAGA factor) and Zelda. Trl is a known early zygotic activator and chromatin remodeler [45–47] and Zelda is known as a "master key regulator" to early developmental genes [9,48] and appears to be a pioneer transcription factor that establishes the initial chromatin landscape of the zygotic genome [49, 50]. In addition to these high-quality motifs, we found a large number of motifs with lower quality scores (S1 Table) at the zygotic stage. These motifs may regulate spatio-temporal specific genes that we observe in the early embryo, and thus have a lower enrichment score due to our whole-embryo approach being ill-equipped to finding such specific patterns.

Similar motifs appear in different species

To quantify the conservation of the discovered motifs across the 11 species in our study, we used Tomtom[33] to measure the similarity between the sets of motifs discovered in different species. For a motif to be considered conserved between two species, we required that it be discovered by HOMER in both species and for Tomtom to report a statistically significant alignment score (see Methods). At the maternal stage, we found that high quality (q-value $< 1\times10^{-100}$ by HOMER, see Methods) motifs tended to be well-conserved (Fig 2A) with a large percentage of the total discovered motif content shared across species. We observed that sister species *D. pseudoobscura* and *D. persimilis* are unique in that they have the highest number of motifs that are either species-specific or are only shared with each other, and have the fewest number of motifs shared with the rest of the species. This is especially noteworthy considering that this lineage is roughly in the middle of the distribution of divergence times from most of the other species, and thus many more distantly related species comparisons have a higher

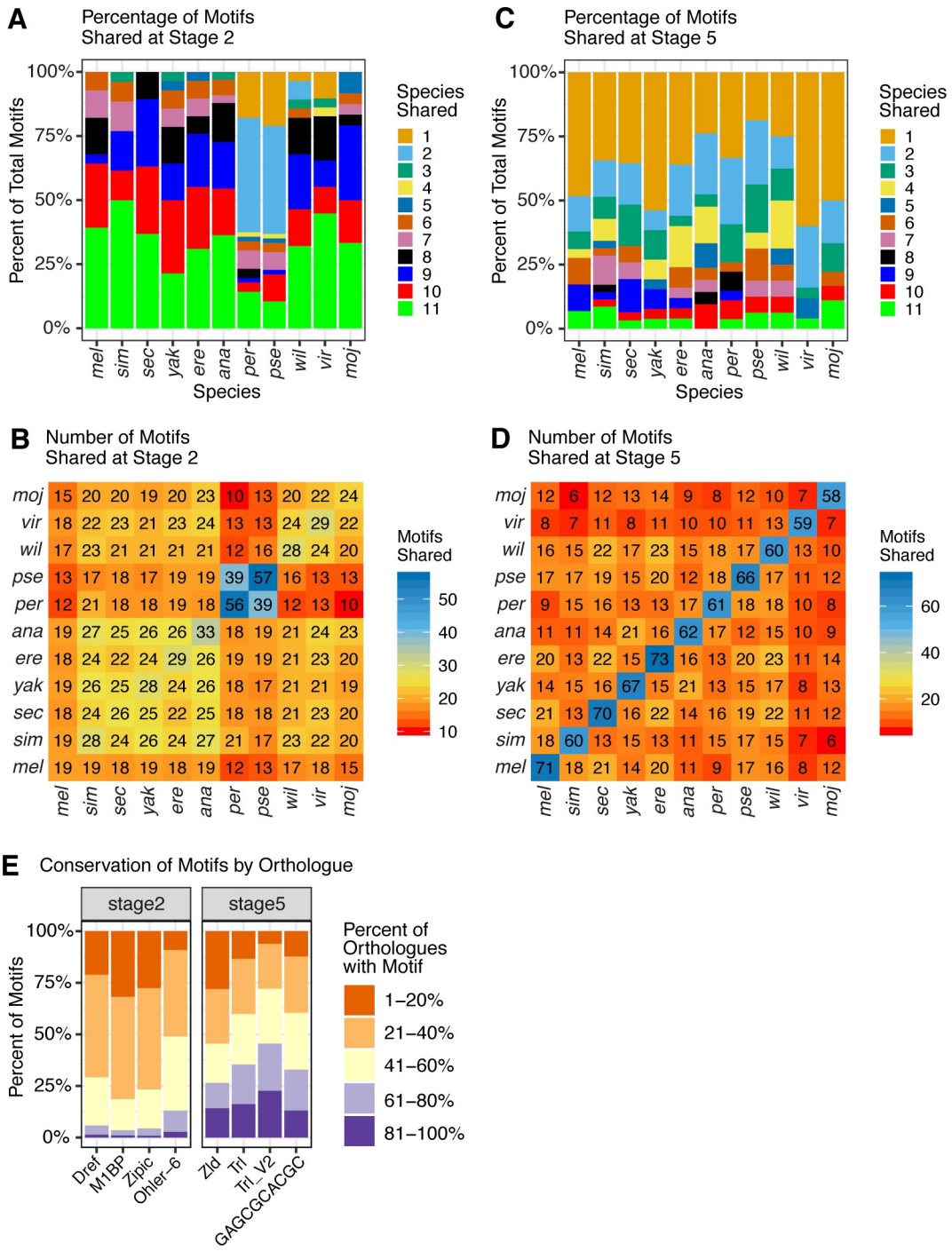

Fig 2. Motifs associated with maternal deposition are largely shared across species, zygotic motifs are likely to be species-specific. For each analysis represented in A-D, motif enrichment was determined for each group of genes at each stage (all maternally deposited genes at stage 2; or zygotic-only genes at stage 5) separately in each species, then lists of enriched motifs at each stage were compared across species. For stage 2 motifs, we required motifs to have a -log qvalue > 100, while for stage 5 motifs we required motifs to have a -log qvalue > 10 (see Methods). (A, C) Percent of motif content in the upstream region that is found to be shared between species at stage 2 and stage 5, respectively. The number of species that share each motif is indicated by the color of the bar. Note that in stage 2, a large majority of motifs are shared in all (11 species) or almost all (9 or 10 species), with the exception of *D. pseudoobscura* and *D. persimilis*, sister species that share common motifs between themselves but are different from the rest of the species. Zygotic motifs identified at stage 5 are much more likely to be species specific or shared by only a couple of species. (B, D) Number of motifs shared between each pair of species at stage 2 and stage 5, respectively. Comparisons of a species to itself indicate the total number of motifs that fit quality criteria discovered in that

species. Comparing the number of shared motifs between pairs of species, there is some signal of the phylogeny in stage 2 (B), with *D. melanogaster* subgroup species sharing more motifs in common with one another than they do with the more distantly related species, and *D. pseudoobscura* and *D. persimilis* with the highest number of motifs in common but the most differences from the remaining species. For stage 5 (D), apparent patterns include both the number of species-specific motifs (diagonal) and less apparent phylogenetic structure. (E) Conservation of top motifs in orthologous genes across species. Y-axis indicates all of the instances of the motif of interest within the upstream region. Coloration represents how many species' orthologs also contain that motif. In general, top motifs at the zygotic stage (stage 5) are more likely to be conserved in orthologous genes at this stage. This sets up a contrast with parts A-D, where maternal deposition is broadly associated with a shared set of motifs across species, but part E shows that orthologous maternal genes are less likely to share a specific motif.

https://doi.org/10.1371/journal.pgen.1008645.g002

degree of motif conservation than do any comparisons with these two species. This is consistent with previous results [24] that this lineage has a disproportionately high number of changes in transcript abundance for its phylogenetic position, and suggests that these large number of changes in transcript abundance may be due to the large scale changes in regulation in these species as observed here. When comparing the rest of the species, we found a relatively higher number of conserved motifs shared between pairs of species within the *Drosophila melanogaster* species group (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*), and a slightly reduced number of conserved motifs between the *D. melanogaster* group species and the more distantly related species (*D. willistoni*, *D. mojavensis*, *D. virilis*) (Fig 2B). At stage 5, we do not observe a high percentage of conserved motifs between species, rather we observe many motifs that are significantly enriched in just one or two species. We also observe little phylogenetic signal in the data, with the only detectable pattern being that the species with the longest divergence time from the rest of the species, *D. virilis* and *D.mojavensis*, have slightly fewer motifs shared with other species (Fig 2C and 2D). If the unique motifs at either stage indeed represent newly evolved regulatory mechanisms, we expect that these motifs to be rare or to have a smaller frequency difference between transcribed and non-transcribed genes. Either of these effects would raise the false discovery rate as reported by HOMER, which makes the number of species-specific zygotic motifs identified all the more remarkable. Additionally, more highly conserved motifs should require less power to be discovered as they are by definition present across more species, and thus we should have more power to identify them than less-conserved motifs. It is still possible that there are more conserved motifs at the zygotic stage that we do not observe due to the lower number of genes used at this stage. Despite this, however, the dominant signal we find from the motifs we have power to detect is non-conserved. This is underscored by the observation that when we reduce our quality threshold for motifs at stage 5, we still observe motifs to be generally non-conserved across species (S3 Fig).

Motif conservation by gene

While these results show that some motifs are important to regulation in the genomes of multiple species, they do not speak to whether orthologous genes in different species tend to contain similar motifs. To investigate whether regulation was conserved at the level of individual genes, we compared the motif content of each *D. melanogaster* gene (see Methods) to the motif content of each of its orthologs from other species. We counted motifs as conserved between two species if the motif appeared in both orthologs. For both stage 2 and stage 5, we categorized motifs based on the percent of orthologs for which the motif was conserved (Fig 2E). Motifs have different levels of gene-specific conservation between stages, with maternal stage motifs appearing to have lower conservation across orthologs than zygotic stage motifs, where a larger proportion of orthologs possess the same motif. This is striking, as this seems to imply that while transcript levels and regulation are both highly conserved for maternal genes,

which genes are regulated by a particular regulator is not. It is possible that the genes that are missing motifs compared to their orthologs are regulated by different motifs, or that the same motifs that are in radically different positions in different species. As many different maternal motifs appear to be regulating transcription at the level of chromatin state, these motifs may be able to function interchangeably. Thus this environment may be more conducive to more motif turnover at this stage but with higher conservation of transcription overall [24], as compared to the zygotic stage.

Motif position

While similar binding motifs identified in multiple species implies that regulatory proteins with similar binding domains are acting in these species, we can also verify the similarity in the regulatory machinery by the relative positions of the binding sites relative to the genes they are regulating. To investigate whether the discovered motifs had the same positional relationship with the transcription start site (TSS) across all species, we generated position frequency data for each motif. For each gene, we examined each position starting from 2kb upstream of the TSS to the 3' end of the gene body, and whether there was a motif at that position. Many of the most prominent motifs shared a similar distribution pattern, characterized by a strong peak at -100bp, and sometimes a secondary peak at -340bp (S4 Fig). To quantify this similarity, we performed an Anderson-Darling test on each motif for each pair of species, which indicated that 65% (stage 2) and 91% (stage 5) of motif distributions are identical between species (percent of motifs for which $p < .05$). This suggests conservation of the relationship between binding to these motifs and initiation of transcription. The higher conservation of motif position in stage 5, which has fewer conserved motifs between species than stage 2, may be consistent with this stage having more gene-specific regulation, as discussed further below.

Motif strandedness

While some studies focus on finding motifs with a particular orientation relative to their proximal genes [51], there is some evidence that motifs do not behave in a strand-specific manner [52]. To evaluate the importance of the strandedness of the discovered motifs, we generated a regression to predict expression level that differentiated between forward and reverse versions of each motif (see Methods). This regression indicated a significant difference between the forward and reverse versions of many motifs. For example, we found the E-box motif affects the log-odds of maternal deposition by .192 in the forward orientation but only .115 in the reverse orientation (t-test, $p < .001$). For almost all motifs, different strands had the same qualitative effect on expression, but with different magnitudes, indicating that while motifs had the same effect regardless of orientation, their efficiency could be increased if the orientation was optimal.

While the strandedness of motifs may play a small role in their overall effect, we want to know if strandedness makes a qualitative difference to our motifs effects on transcript level, and if we can use motif strand to improve our model. To determine this, we ran HOMER exclusively on the same strand that the gene appeared on, rather than the default mode of scanning both strands. This resulted in the same set of motifs being discovered. This is consistent with the regression results that show that each motif, whether located on the positive strand or the negative strand relative to the transcription start site, has the same qualitative effect on gene expression, indicating that the direction of each motif had minimal effect on expression. To evaluate whether the strand the motif was located on relative to the gene was predictive in whether a gene was transcribed at a particular stage, we constructed another regression using only the data from the same-strand motifs. This regression performed less well than the

regression using motifs from both strands (AIC = 7915.8 for the unstranded regression, AIC = 8612.5 for the stranded regression for a representative species *D. ananassae*, see S1 Text and S2 Text). Overall, this suggests that motif binding elements need not bind in a strand specific manner to induce their effects, though the optimal orientation provides measurable increase in their effect on transcription. This result is the same at both stage 2 and stage 5.

GO analysis

After having identified a set of motifs that together seem to be responsible for early embryonic RNA content, we next asked if these motifs are likely to be regulating genes with specific types of functions. To this end, we performed gene ontology (GO) analysis on groups of genes, based on their motif content. To simplify this analysis, we chose to focus on the top 8 motifs as reported by HOMER, and for each of the 8 motifs, we performed GO analysis on the transcript pools at each stage as well as on each motif individually [53,54]. We initially performed a GO analysis on both the maternally deposited and zygotically transcribed transcript pools, disregarding motif content. When comparing stages, we observe no overlap between GO terms (Fig 3A), which is consistent with our expectations that the genes that are activated in the zygote have different functionality to those transcripts that are maternally deposited, especially as our definition of zygotically transcribed genes excludes genes present in stage 2. When examining genes containing specific motifs within each stage, we observe that many of the stage 2 motifs show a similar pattern of associated GO categories, with the strongest associations belonging to the DREF motif, which is also associated with most identified categories (Fig 3B). This could be an indication that there is a high degree of homogeneity in terms of the types of genes these motifs may regulate. In contrast, the stage 5 motifs present in zygotic-only genes show more variety in the GO terms of genes they are associated with (Fig 3C), which could be indicative of more specific regulation for these genes at this stage.

While the previous GO analysis indicated that the top motifs at stage 2 display significant overlap in associated GO categories, this does not exclude the possibility that specific GO categories are regulated by specific motifs. To search for more specific motifs, we performed motif analysis using HOMER to find overrepresented sequences in the top GO terms within maternally deposited genes, resulting in several motifs which are enriched in specific GO terms (S5 Fig), though very few of them are significantly enriched after multiple test correction. These motifs do not appear in other analyses, and do not have strong matches to proteins expressed in the ovary found in the literature. Because these motifs are associated with a small subset of genes, we hypothesized that these motifs confer specificity to transcription of specific genes with accessible chromatin. To determine whether these motifs are associated with increased expression at stage 2, we used linear models to measure the effect of the presence of these motifs, specifically in genes that already contain motifs that bind to architectural proteins, or whose adjacent genes are highly expressed. We did not find that the presence of these GO term-specific motifs increased the odds of maternal deposition (S5 Fig). It is possible that this result is due to the lack of statistical power surrounding these motifs, as these motifs are somewhat rare. This result could also reflect the underlying biology, where perhaps these motifs are functional at developmental stages other than stage 2.

Predicted maternal motif binding proteins are enriched in the ovary

Next, we investigated whether the potential motif binding proteins we identified were plausible regulators of maternal deposition. It is unclear whether the motifs we identified as enriched in maternally deposited genes are associated specifically with maternal deposition, given that chromatin regulators are important at all stages in all tissues. To investigate, we used

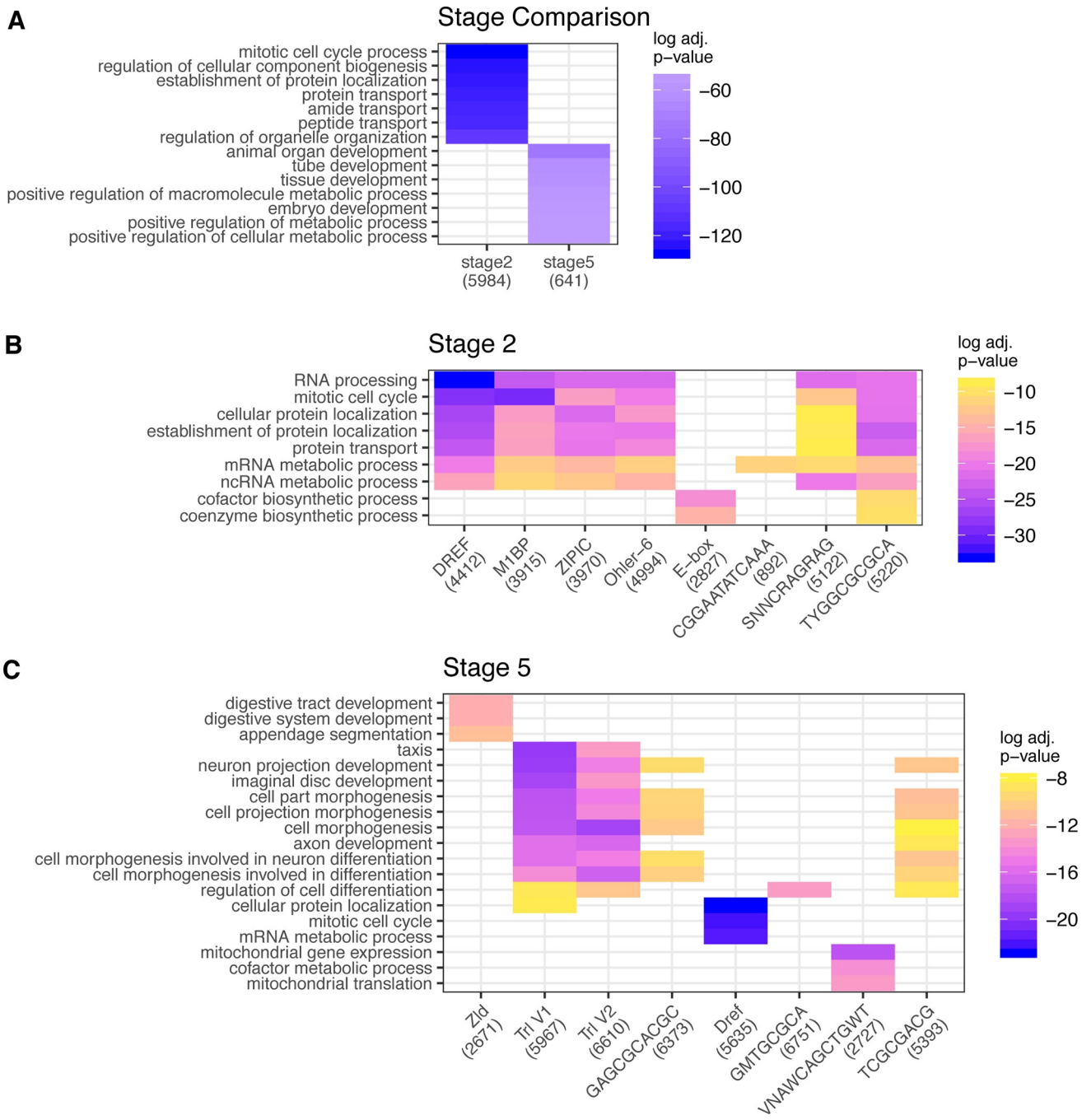

Fig 3. Top GO terms show that motifs regulate broader set of genes at the maternal stage, and a more specific set of developmentally associated genes at the zygotic stage. (A) GO terms associated with each stage. Note that the set of identified GO categories does not overlap between stages. (B) GO terms associated with top motifs in stage 2, where a majority of motifs are associated with similar broad GO categories (C) GO terms associated with top motifs in stage 5, some motifs are associated with the same categories, some appear to be more specialized, with identified categories showing more specificity than categories associated with stage 2.

https://doi.org/10.1371/journal.pgen.1008645.g003

modENCODE transcript abundance data [55] to compare the mRNA transcript levels for proteins predicted to bind our discovered motifs, and found increased expression in ovaries (Fig 4A) as compared to other tissues sampled. This pattern exists, though to a lesser extent, in the FlyAtlas 2 dataset [56], which is a tissue-specific database of transcript levels that utilizes RNA-

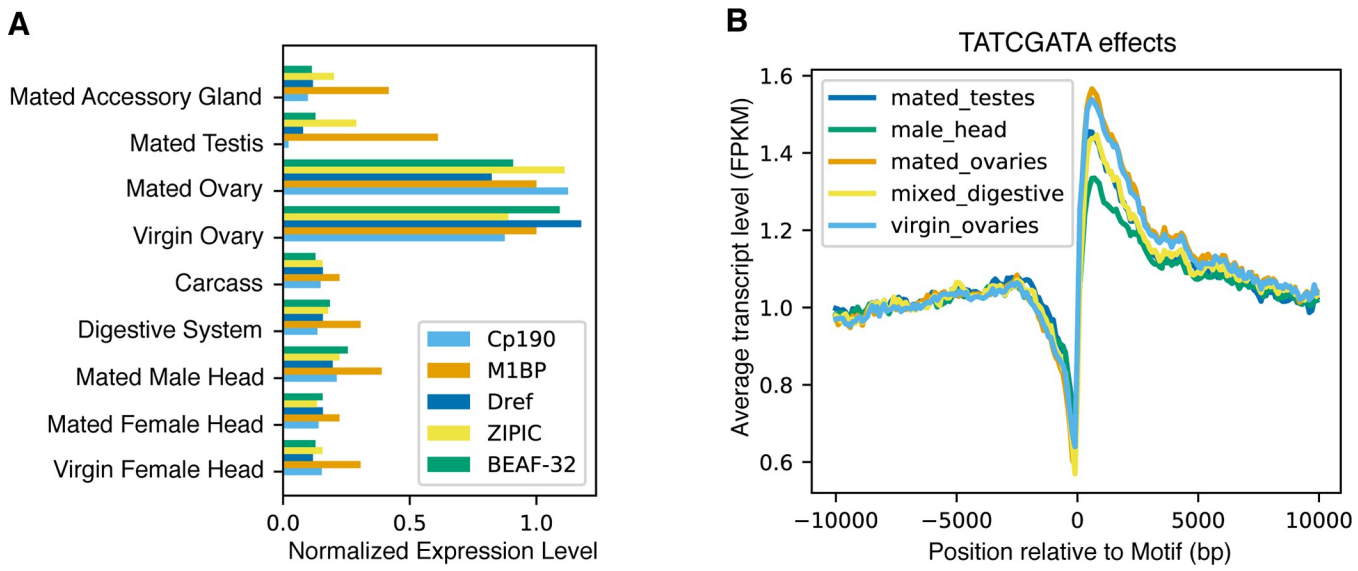

Fig 4. Identified maternal regulators are ovary-enriched, as is their effect on transcription. (A) RNA levels of putative binding proteins by tissue type. Transcript abundances within each gene have been normalized such that the average abundance in ovaries is equal to 1. While identified maternal regulators have regulatory functions in multiple tissue types, they are highly enriched in ovaries compared to other tissues. (B) Average normalized expression levels versus proximity to motif by tissue type. Normalization was performed by dividing each expression value by the average expression from 9.9-10kb away. While binding sites for identified maternal regulators are present in multiple tissues, the effect on gene expression is stronger in ovaries compared to other tissues. All data used were from tissues sampled from four-day-old adults.

https://doi.org/10.1371/journal.pgen.1008645.g004

seq data rather than microarray analysis. The discrepancy between the two datasets could be due to the differences in gene expression measurement method or in experimental methods. The transcripts for these proteins also show moderately high abundance in our own dataset (S3 Text). While it has been demonstrated that mRNA levels do not necessarily mirror protein levels [57], the enrichment of mRNA in ovaries compared to other tissues is evidence that these proteins are important in ovaries.

To investigate whether these proteins are acting to affect transcription in the ovaries specifically, we examined the expression profiles of RNA in various tissue types (referenced in Fig 4B) from existing RNA quantification datasets [55,58]. For each instance of a motif of interest, we extracted the transcript level from within a 20kb window surrounding the motif and measured the normalized relative transcript level for each position (an example of this is shown in Fig 4B). While the relative normalized transcript level changes in each of the measured tissues, the effect is strongest in ovaries, indicating that the presence of one of these binding sites is associated with a higher increase in transcript levels in the ovary compared to other tissues.

As the motifs associated with maternal transcription also act to some degree in other tissues, we next wanted to ask whether the motifs were more enriched in maternally deposited genes than in genes expressed in other tissues. To determine whether regulation in different tissue types were associated with different motifs, we ran HOMER in the same manner as with the maternal stage data to discover enriched motifs (see Methods) in transcripts present in other tissues, as identified from ModENCODE data [59]. We found that most other tissue types were also enriched in the same motifs discovered in transcripts present in stage 2 embryos. However, examining the frequency of motifs in specific genes revealed that the majority of those motifs were from genes that were shared between those tissue types and stage 2 embryos. When we exclude genes that are expressed in stage 2 embryos, HOMER fails to identify the original set of motifs as enriched in male larval gonads, male reproductive tract,

adult heads and adult midgut. Furthermore, HOMER detects the motifs at a lesser rate in larval ovaries, larval CNS, and intestinal tract. Despite being identified in fewer tissue types and at a lesser rate in other tissue types as compared to the stage 2 expression levels, the observation that these motifs may also have important functions in other tissue types is consistent with the literature. For example, DREF is known to be important for cell proliferation and chromatin regulation, and is active in many other tissues [60,61]. These motifs are likely associated with many housekeeping genes that are vital to a variety of tissue types.

Maternally deposited genes are physically clustered on the genome

Given that many of our discovered motifs bind architectural proteins, we hypothesize many effects may be linked to the physical location of genes on the chromosome. We examined the positional distribution of transcribed genes along chromosomes in various tissue types (Fig 5A). As previous papers utilizing the Hi-C method have shown correlation with active topologically associated domains (TADs) and gene expression [62,63], we predicted that any tissue type where regulation is dominated by architectural proteins to transcribe a set of genes physically clustered on the chromosome. To compare the physical gene clustering of transcription at the maternal stage with that of other tissue types, we acquired several RNA-seq datasets from NCBI/GEO [64] and performed a Wald–Wolfowitz runs test [65] on each tissue of the previously described tissue types. While all tissues examined showed a strong preference for physical groupings of transcribed genes on chromosomes, embryonic stage 2 samples were the most highly grouped (Fig 5B). This result was robust to changes in the threshold of what is considered to be expressed (see Methods). This pattern of physical co-expressed gene clustering on the chromosome is consistent with our model of regulation via architectural proteins.

While these results speak to the pattern of clustering of expression for maternal genes in terms of adjacent genes being on or off, they do not account for the distance between genes. To answer the question of whether this clustering phenomenon is dependent on distance, we examined the distance to adjacent genes. We observed a trend whereby proximity to an active promoter increases the odds of maternal deposition (Fig 5C). This effect was slightly affected by the strandedness of the two genes whereby genes that have an opposite orientation are more likely to have different expression. This is consistent with observations from previous studies [34] that consecutive genes on the same strand were more likely to show co-expression, while consecutive genes on opposite strands were more likely to differ in expression.

Many previous studies have observed that zygotic genes tend to be short in length [24,30,66,67]. In addition to affecting transcription speed, shorter gene lengths result in a smaller distance between transcriptional units along the chromosome, especially when considering which strand the gene is on. To explore gene length in maternal genes and the relationship between gene length and the position on the chromosome, we measured the maternal deposition rates with respect to gene length. We observed a trend that in most species, shorter genes are less likely to be maternally deposited. There are differences in the length of maternal genes across species, and this trend could be partly due to the bias for more highly annotated genomes to be enriched in shorter genes (Fig 5D). Additionally, chromatin context seems to heavily influence this effect: when the adjacent genes are off, gene length is much more important (Fig 5C) and very short genes are very likely to be off. This could be because shorter genes are more likely to be influenced by the regulatory machinery of a nearby gene. Alternatively, longer genes might be long enough to physically isolate themselves more effectively and establish their own unique regulatory environment.

Given that a number of motifs found in this study are bound by proteins annotated as insulators, and the motifs are similar to those that are associated with TADs, we asked where the

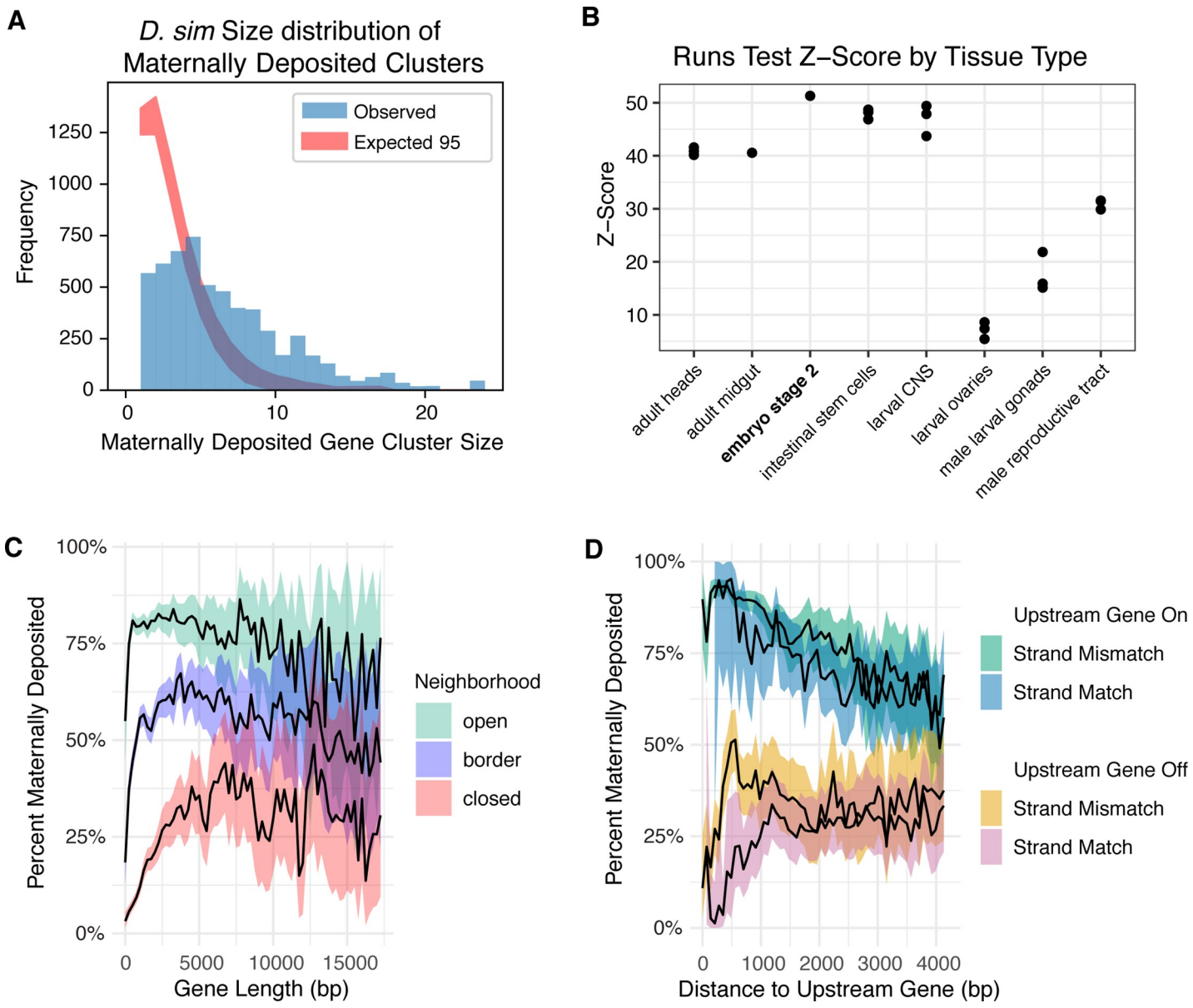

Fig 5. Maternal genes are found in co-expressed clusters. For the analyses in A and B, genes were categorized as either expressed or not expressed (see Methods) and adjacent expressed genes were considered to be clustered, with a cluster size equal to the number of constituent genes. (A) Physical clustering of maternally deposited genes along the chromosome, in a representative species (*D. simulans*). The shaded blue region represents the observed frequency of co-expressed maternal gene clusters of various sizes. The red region represents the 95% CI constructed with 10,000 bootstrap iterations. Maternal genes are co-expressed in clusters along the chromosome more often than expected, given the percent of the genome that is transcribed at this stage. (B) Physical clustering of co-expressed genes on chromosomes in various tissue types. In order to compensate for differing proportions of the genome that are expressed in each tissue type, physical clustering was measured by performing a Wald-Wolfowitz runs test and taking the z-score (see Methods). Maternally expressed genes, represented by stage 2 embryos, show the highest proportion of physical clustering of co-expressed genes, though other tissues such as intestinal stem cells and larval CNS also have highly physically clustered co-expressed genes. (C) Gene length by number of adjacent maternally expressed genes, "open" indicating both adjacent genes are expressed, "border" indicating that one is expressed, and "closed" indicating that neither are expressed. Genes that with more expressed neighbors are more likely to be maternally deposited, regardless of length. Genes without expressed neighbors are less likely to be maternally deposited, with the odds increasing as length increases. (D) Odds of maternal deposition versus distance to the nearest upstream gene by upstream expression and strand. Distance is measured by from transcription start site (TSS) to TSS. When the upstream gene is maternally deposited, odds of maternal deposition are high, but decrease with distance regardless of strand. When the upstream gene is not maternally deposited, odds of maternal deposition are low and have a strand-dependent relationship with distance.

https://doi.org/10.1371/journal.pgen.1008645.g005

motifs found in our dataset can be found relative to TAD boundaries. Previous results suggest that architectural proteins are prevalent in the centers of TADs as well as the boundaries [34], and may be involved in mediating interactions of the DNA within a TAD [38].To determine the location of motifs in the context of TADs, we assessed the transcription of nearby genes relative to the transcription of a gene with these identified motifs. For each regulatory region, the gene nearest to that regulatory region was examined, as well as two genes downstream and two upstream. The frequency of motifs was measured based on the transcript abundance pattern of these five genes. Many of the top motifs including Dref, M1BP, Zipic, and E-box, occur more frequently in the center of maternally deposited gene clusters, rather than on the edge of clusters. (t-test p-values 7×10^{-3}, 2×10^{-6}, 3×10^{-10}, and 1×10^{-4} respectively). This is consistent with previous results [34], and may suggest an important role for architectural proteins in promoting interactions within a TAD as well as potentially in establishing TAD boundaries in oogenesis.

Stage-specific genes are isolated on the genome

Given that maternally deposited genes are physically clustered together in the genome, we wanted to examine if this pattern held with the set of genes that were stage-specific. To determine if consecutively expressed cluster size is related to stage-specificity of transcript representation, we examined maternal-only (transcripts present at stage 2 and entirely degraded by stage 5) and zygotic-only genes (transcripts present at stage 5, not present at stage 2; for both stage-specific categories, see Methods for further definitions) and their frequencies in clusters of different sizes. We determined that for most species, in contrast to all maternally deposited genes, both maternal-only and zygotic-only genes are more likely to be in smaller (1–3 consecutive active genes) groups than in larger groups (more than 3 consecutive active genes) (Fig 6A and 6B). For these stage-specific genes, this could be an indication that control of stage-restricted genes is more specific, affecting single genes rather than larger clusters. Results for most other analyses of maternal-only genes were unable to be obtained due to the very low number of genes in this category (see Methods).

GC-content of upstream regions is predictive of maternal deposition

In *Drosophila*, transcription start sites are frequently associated with a spike in GC content. These spikes in GC content have been suggested to act as "genomic punctuation marks" to delineate functional regions, though their mechanisms of action are not clear [68]. To explore this phenomenon with respect to the two developmental stages we examined, we evaluated the average GC content of upstream regions for genes in stage 2 and stage 5. When comparing the GC-content of putative cis-regulatory sequences in maternally versus non-maternally deposited genes, we observed an increase in GC-content upstream of the TSS (S6 Fig), as well as a dip in GC content ~200bp upstream of these genes. In contrast, this modulation does not occur in genes that are off at both stage 2 and stage 5, nor in genes that are off at stage 2 but activated at stage 5. To determine whether this modulation of GC-content was predictive of maternal deposition, we constructed four generalized linear models using the GC-content, the motif data, and both the motif data and GC-content as data sources (see Methods, S4 Text). Adding the GC-content to the model that already included motif data improved the model (AIC: 185589 without GC content AIC:183079 with GC content), hence increased GC content upstream of TSS is somewhat predictive of maternal deposition, even when accounting for motif presence in this region.

The biological significance of this spike in GC content is unclear. Fluctuations in GC content have been observed in *Drosophila* previously [68], and there is evidence in humans that

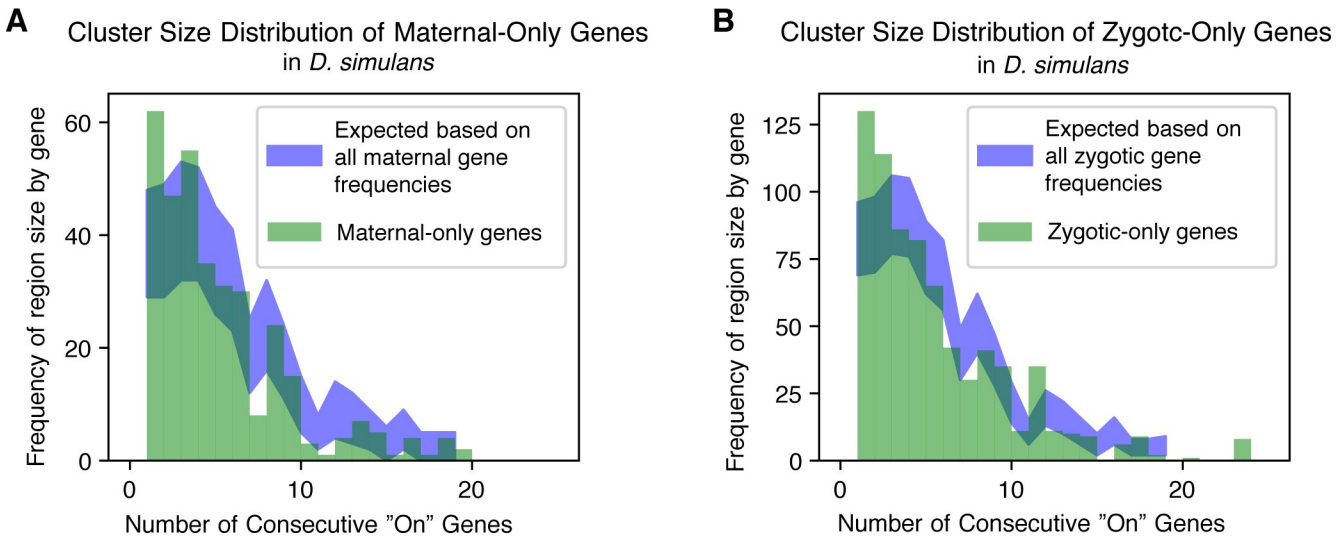

Fig 6. Stage-specific genes are more likely to be different from their chromatin neighborhood. *D. simulans* was chosen as a representative species. (A) Co-expressed cluster size distribution of maternal-only genes (green bars) compared with the expected frequencies based on the overall cluster size frequencies observed at stage two (blue region). The expected frequencies are based on the distribution in Fig 5A multiplied by a scale factor equal to the proportion of maternally deposited genes that are maternal-only, with the shaded region representing a 95% confidence interval. (B) Co-expressed cluster size distribution for zygotic-only genes (green bars) compared with the expected frequencies based on the overall cluster size frequencies observed at stage 5 (blue region) in a manner similar to Fig 6A. the shaded region represents a 95% confidence interval. For both stages, stage-specific genes are more likely to be the single gene (or one of a small number of genes) that are expressed where their neighboring genes are not, representing small numbers of "on" genes in an "off" chromatin environment.

https://doi.org/10.1371/journal.pgen.1008645.g006

spikes in GC content are associated with supercoiling [69]. DNA supercoils are generated in via transcription, and positive supercoils are observed to inhibit transcription [70]. In Drosophila negative supercoils have been associated with high transcriptional activity in polytene salivary gland cells [71], and GC content directly impacts the biochemistry of DNA with respect to torsional stress [72]. As the nurse cells where maternal transcripts are produced are polyploid with a high transcription rate, nurse cell chromosomes may be under similar torsional stress. This may explain why maternally deposited genes in particular are associated with this spike in GC content.

Discussion

The RNAs present in early embryogenesis are critical for surviving this developmental stage. Early embryos undergo many important developmental processes, such as and axial patterning [12], at the same time as the complement of transcripts in the egg are undergoing a precise and highly regulated turnover between those provided by the mother to those transcribed by the zygote [3,73]. Due to the critical nature of these gene products and the developmental processes they direct, the mechanisms behind regulation of early zygotic transcription have been under intense investigation for some time [8]. However, regulation of transcription of maternally deposited genes in the maternal genome has received surprisingly little attention. In this study, we leveraged a large comparative dataset to investigate a number of aspects of regulation of both the zygotically transcribed and maternally deposited transcriptomes.

Here, we identified a number of conserved transcription factor binding motifs associated with transcript abundance for both maternal and zygotic-only transcripts. At the maternal stage, there were a larger number of more highly conserved motifs than were found for the zygotic-only genes. This is consistent with a previous study that found that maternal transcripts themselves were more highly conserved than transcripts at the zygotic stage [24]. Given

this, surprisingly we also found less conservation of particular motifs at conserved genes transcribed at the maternal stage. As we identified a number of motifs involved in regulation at the level of chromatin at the maternal stage, perhaps different combinations of chromatin-regulating motifs can be utilized interchangeably without altering expression. This could provide robustness, permitting evolutionary changes in sequence without affecting gene expression of maternal genes. In contrast, while we find that the zygotic-only transcripts are associated with fewer conserved motifs overall, and more divergent lineage- and species-specific motifs, individual conserved genes are more likely to be regulated with the same motifs. This provides conservation of gene expression by a different mechanism for the zygotic-only genes that are functionally required across *Drosophila*. Why the two stages and genomes would have such different ways of activating conserved genes across the genus is likely due to the underlying biology of regulation at the two stages, as discussed in detail below.

Maternal regulation

We found that motifs associated with putative *cis*-regulatory regions of maternally deposited genes are predominantly annotated as insulator binding sites. An insulator is a type of regulatory element that can block the interactions of *cis*-regulatory elements with promoters or prevent the spread of chromatin state. Insulators are known to be important in creating and maintaining the gene expression patterns, ubiquitous in *Drosophila*, and are potentially a key factor for *Drosophila* to maintain such a high gene density [42]. Because the roles and mechanisms of factors annotated as insulators are not well understood, using the term "Architectural Protein" instead of insulator binding protein may be more appropriate [74]. Recently, these proteins have been studied using genome-wide chromatin organization methods, such as Hi-C, which detects regions of interacting chromatin known as Topologically Associated Domains (TADs) and identifies boundaries between them. Histone marks appear to be enriched in certain TADs but stop abruptly at TAD boundaries, supporting the idea that certain TADs are entirely transcriptionally silenced while others are expressed [34]. Furthermore, ChIP-seq has demonstrated that TAD boundaries in other tissues are enriched in architectural protein binding sites [34] , including several those that we identified in this study.

There is some disagreement on the effect that TADs have on gene expression, however. Ghavi-Helm et al [75] demonstrate that the disruption of TADs does not necessarily disrupt the expression of constituent genes. Instead, they suggest TAD boundaries acting to prevent interactions between TADs is rare or tissue specific. Others propose that TADs are acting to increase robustness of other regulatory mechanisms [76]. Because TAD-associated elements are associated with maternal deposition in our dataset, we hypothesize that these elements may be regulating maternal deposition via chromatin-level control. It is possible that there are other additional mechanisms that we do not detect.

To understand the connection between architectural proteins and maternal deposition, we need to examine where these transcripts are produced to understand the cellular context. In the ovary, nurse cells are responsible for the transcription of maternally deposited genes, and there is a considerable body of literature devoted to nurse cell biology. Much study has been directed towards elucidating how nurse cells transport their products into the oocyte and how post translational control mechanisms fine-tune protein levels from maternal transcripts [3,7,18–23,77]. However, despite this wealth of knowledge, the regulatory mechanisms by which the nurse cells specify which genes to transcribe are largely unknown. One unusual feature of nurse cells is that they are highly polyploid [78,79]. One of the major benefits of this could be an across-the-board increase in transcription rates necessary to provision the embryo with all necessary transcripts. These transcripts represent a large proportion of the genome,

with estimates ranging from 50–75%, depending on experimental conditions [3], and necessitate large amount of transcription overall in a short period of time. We extract >100ng total RNA from an embryo; this is an astonishingly large amount of RNA to be present in what is essentially at the time of fertilization a single, albeit a highly specialized, cell. One point of comparison can be found in Abruzzi et al. 2015 [80], who extracted 2-5pg RNA per *Drosophila* neuron. A transcriptional environment that is optimized to quickly transcribe huge numbers of genes might be more amenable to control via chromatin state.

Given the amount of overlap between the motifs enriched in the *cis*-regulatory regions of maternally deposited genes and the motifs associated with TAD boundaries, it is possible that these same architectural proteins are functioning to define which genes are maternally transcribed and then deposited into the embryo. We found that the maternally deposited genes are both highly clustered on the genome, and that the expression status of nearby genes is predictive of expression levels. We also identified a pattern whereby the relative strandedness of adjacent genes is indicative of whether they will be maternally deposited, which is a pattern that has been previously observed with insulators [34]. Each of these results is consistent with known behavior of architectural proteins, suggesting that expression at stage 2 is controlled locally on the chromosome by activating TADs rather than specific genes.

As architectural proteins are important in determining genome organization and regulating transcription to some degree in all tissues and stages, we investigated whether the regulatory patterns we observed for maternal genes were ovary-specific or shared across all stages and tissues. Many of the motif binding elements discovered in this analysis appear to be enriched in ovaries, although these proteins have important functions in other tissues as well. Some of the proteins predicted to bind our motifs have been noted for enrichment in the regulatory DNA of housekeeping genes [35], and as maternally deposited genes themselves are enriched in housekeeping genes, this result is perhaps unsurprising. A number of studies have suggested that in addition to the common architectural proteins shared across conditions and developmental stages, there may exist tissue-specific architectural proteins that integrate into the canonical protein complex to produce tissue-specific TAD patterns [81–83]. Perhaps this is the case with the ovary, and further study will reveal whether there are ovary-specific factors that may interact with the common architectural proteins whose binding sites we find enriched here. For example, the authors of Mataz et al. 2012 [84] suggest that Shep may be a tissue-specific factor interacting with architectural proteins in the central nervous system. Shep is less enriched in the central nervous system than CP190 (a known interaction partner of ZIPC, one of our maternal expression associated motifs) is in ovaries, suggesting that CP190 could also qualify as tissue-specific. Alternatively, the polyploid nature of nurse cells and the extensive and rapid transcription that occurs in these cells may instead provide a high level of enrichment for the common architectural proteins, without the need for stage or tissue specific architectural proteins. This hypothesis is supported by our finding that the architectural protein binding sites we identify are enriched in maternally deposited genes, whether or not they are also housekeeping genes (S2 Fig).

Given that chromatin level regulation would appear less precise than specific regulation, we are left with the question of how the stage 2 mRNA content is so highly conserved across species [24]. Perhaps regulatory control primarily at the level of chromatin provides redundancy to maintain transcription despite the gain or loss of individual binding sites. Alternatively, there could be other levels of regulatory control that we are unable to detect, with the signal from chromatin-level control being so strong during this time. The high level of conservation of maternal transcripts is also remarkable given the importance of post-transcriptional regulators at this stage [3,19,23,85], as it is not clear if conservation at the transcript level is necessary for conservation at the protein level.

Zygotic regulation

Our examination of motifs that are associated with zygotic mRNA expression revealed several previously discovered motifs, including those that bind Zelda and GAGA factor (Trl). Additionally, several motifs are likely binding sites for other well-characterized developmental proteins (S1 Table) which are sometimes highly localized in the embryo. If transcripts are produced in a spatially localized manner, they are necessarily not expressed in the entire embryo, and thus their signal may be more difficult to detect in our data from whole embryos. Overall, we observe few motifs at stage 5 that are conserved across species, in comparison to motifs for maternally deposited genes. However, the motifs that we do find at stage 5 tend to have higher conservation within specific genes than the motifs we discover at stage 2. This highlights that it may be more important for specific genes to have precise signals after zygotic genome activations.

Additionally, in our zygotic analysis, we focused only on transcripts that are present at stage 5 and do not have a maternal component, as many maternally deposited transcripts are still present at stage 5 (roughly half of maternal transcripts are still present at this stage [8,30–32]). Because many maternal transcripts are still present, analysis of the total stage 5 transcriptome would largely recapitulate the stage 2 results, especially as stage 5 transcripts are much more likely to be expressed in specific spatio-temporal patterns, which to our whole-embryo analysis would appear as low or noisy signal. Our decision to remove transcripts with maternal deposition highlights the signals that are unique to stage 5, but comes at the cost of an overall reduction in the number of genes available for analysis, resulting in higher false discovery rates for all motifs. However, the remaining dataset is sufficient to recover well-known regulators at this stage, as described above.

Conclusions

In this study, we examined regulatory elements associated with maternal transcripts present at stage 2 of embryogenesis and zygotic transcripts present at the end of stage 5 across species of *Drosophila*. At both stages, we found regulatory motifs that are conserved throughout the ~50 million years of divergence represented by these species. This provides evidence for a high level of conservation of regulatory mechanisms across the genus at each stage, and speaks to the critical nature of the complement of transcripts present to direct early embryogenesis. The differing patterns observed in the *obscura* group species (*D. pseudoobscura* and *D. persimilis*), and the regulatory basis of changes in transcript representation between species is the subject of ongoing study. At the maternal stage, we found many regulators that appear to be defining general regions of the genome to be transcribed via chromatin regulation through architectural proteins and likely at the level of TADs. Given the exceptionally high level of conservation of maternal transcript deposition, the relatively non-specific mechanism of maternal gene regulation and the lack of conservation of binding sites at orthologous genes appears contradictory. In contrast, we found zygotic regulatory elements to be considerably more highly conserved at the orthologous gene level. We expect these zygotic regulatory elements to affect smaller groups of genes, or singular genes. The different patterns of regulation for transcripts present at these two stages of embryogenesis is consistent with the specific transcriptional contexts of these two genomes, with the non-specific mechanism active in highly transcriptionally active polyploid nurse cells in oogenesis in the mother, and the gene-specific mechanism acting in the zygote where transcription is often localized in time and space.

Methods

Data acquisition

RNA-seq data utilized for this study was generated previously [24], and is available at NCBI/GEO at accession number GSE112858. This dataset contains RNA-Seq data from single

embryos. Embryos were collected either at stage 2, representing a time point before zygotic genome activation, and at the end of stage 5, representing a time point after widespread zygotic genome activation. Embryos were collected from 14 species, however we only used the data from 11 (*D. simulans*, *D. sechellia*, *D. melanogaster*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*) due to annotation deficiencies in the remaining 3. GTF files and references genomes from previously sequenced species [28] were downloaded from Flybase [86].

To determine whether a gene would be labeled as 'off' or as 'on', the overall distribution of FPKMs was analyzed. For all species, for both stage 2 and stage 5, a bimodal distribution appeared, with one peak at 0 and another at approximately $e^{3.5}$. The commonly used cutoff of FPKM = 1 [87,88] was chosen as it falls between these two distributions.

To determine which genes were orthologs, we used the FlyBase orthology table "gene_orthologs_fb_2014_06_fixed.tsv".

Sequence selection

Preliminary tests were performed to determine which regions were most likely to have regulatory elements. For each gene, several regions were extracted: 10kb upstream, 5kb upstream, 2kb upstream, 1kb upstream, 500bp upstream, 5' UTR, total introns, total exons, and 3' UTR. For each region, boundaries were obtained from the appropriate GTF and sequences were extracted using BioPython (Version 1.73, [89]). The 2kb upstream region showed the highest quality motifs (S1 Fig), and thus were used for matching motifs in external databases, measuring motif overlap between species, analyzing motif position distributions, and GO analysis. For these analyses, featured in Figs 1–4, UTRs were ignored as not every species had annotated UTRs.

Motif discovery

We used HOMER [29] to discover motifs in test sets using the background sets as control FASTA files, test and background sets are defined below. Deviations from the default settings include the use of the -fasta flag to specify a custom background file. For stage 2 queries, the test FASTA files included genes that had a FPKM $>=1$ at stage 2 while the control FASTA files included genes that had an FPKM <1. For the stage 5 queries, the test FASTA files contained genes where the stage 5 FPKM $>=1$ and the stage 2 FPKM <1, while the control FASTA files included genes whose stage 5 FPKM <1 and stage 2 FPKM <1. Additionally, we used the -p flag to utilize our computational resources more efficiently. We used -norevopp flag in the case of strand-specific searches. Motif quality was evaluated based on the HOMER-outputted q-values.

To validate the HOMER output files we used MEME [33] v4.12.0 and RSAT [90]. MEME was run using-mod zoops -nmotifs 2 -minw 8 -maxw 12 -revcomp. The RSAT analysis uses the purge-sequences tool, followed by oligo-analysis using the following parameters: -lth occ_sig 0 -uth rank 5000 -return occ,proba,rank -2str -noov -quick_if_possible -seqtype dna -l 8, followed by pattern-assembly using the following parameters: -v 1 -subst 1 -toppat 5000 -2str, followed by matrix-from-patterns using the following parameters: -v 1 -logo -min_weight 5 -flanks 2 -max_asmb_nb 10 -uth Pval 0.00025 -bginput -markov 0 -o purged_result.

Stage-specific gene analysis

For analyses of zygotic transcripts, such as the motif analysis, we defined genes as being zygotic-only if they were off at stage 2 (FPKM <1) and on at stage 5 (FPKM >1), for N = 10,215 genes across all species. It is necessary to impose such a restriction, as a large percentage (approximately 85%) of genes that are zygotically expressed were also maternally

deposited, and analysis of stage 5 regulatory mechanisms would be confounded the signal of stage 2 genes. For analyses of maternal-only transcripts, we define maternal only if they are on at stage 2 (FPKM >1) and off at stage 5 (FPKM <1). As the class of maternal-only genes is very small (N = 3194 across all species), we were unable to obtain results for some analyses such as the motif content detection and GO analyses for this group of genes.

Motif sharing

To determine weather motifs were shared between species, the HOMER-formatted motifs were converted to meme-formatted motifs using chem2meme from the MEME Suit [33]. Tomtom, also from the MEME Suit, was then used to find matching motifs, using default parameters. For a motif to be considered shared with another species, the Tomtom output threshold of $\alpha = .05$ was used. This technique was used to calculate the similarity of motifs found in different species, as well as to evaluate the similarity of different motif discovery strategies using MEME, RSAT, or HOMER with alternative parameters.

To refine the results of shared motifs, we applied an additional quality cutoff. For stage 2, motifs were first filtered for a q-value of less than 1×10^{-100}, and for stage 5, motifs were first filtered for a q-value of 1×10^{-10}. The difference in the cutoffs used at the two different stages was due to the differences in the overall distribution of q-values for these stages due to a reduced number of zygotic-only genes (see zygotic-only motifs above).

Because sharing was calculated on a by-species basis, it is possible that one species has a motif that meets the criteria for being shared among all other species while other species' version of that same motif failing to meet the criteria. This can occur, for example, when a motif is an intermediary version of two motifs that fall just outside the cutoff.

To find proteins that bind to the discovered motifs, we used Tomtom to query JASPAR and Combined Drosophila Databases using the default parameters [91].

Motif position and count

Motif position was determined by using the scanMotifGenomeWide tool to in the HOMER package. Queries were performed by scanning the discovered motifs against the FASTA files for each gene. The 5' boundary of the motif was used as the motif position. For the motif counts per gene used in many downstream analyses analysing motif position distributions, GO analysis, GC content analysis, and motif strand analysis. We used this output and counted the occurrence of a given motif in the target region. To quantify positional distribution similarity, we used the stats.anderson_ksamp function from the scipy library V1.2.1 [92]. Distributions were considered to be different at $\alpha = .05$ after Bonferroni correction.

Transcript enrichment by tissue

Expression data for various adult tissues was downloaded from modENCODE [59]. To compare enrichment for transcripts with different magnitudes of abundance, we applied an additional normalization. For each transcript, transcript levels in FPKMs were divided by a scaling factor equal to the average of the expression levels in ovaries. This normalization preserves the relative abundances within each transcript, but allows for visualization of transcript levels with dramatically different overall expression levels.

Housekeeping gene identification

To compare the enrichment of the discovered motifs in maternally deposited genes versus housekeeping genes, we identified housekeeping genes using modENCODE data [59].

Housekeeping genes were defined as having expression in each of the following tissue types: larval CNS, larval ovaries, male larval gonads, male reproductive tracts, adult midguts, adult heads. In addition, putative housekeeping genes needed expression levels of greater than 1 FPKM in our stage 2 and stage 5 dataset in *Drosophila melanogaster*.

Expression by position

D. melanogaster expression data by position was downloaded from modENCODE [59] for several tissue types. Positions for each motif was determined as previously described in the Motif Position and Count section above. For each instance of the motif of interest, we determined expression values in area from -10kb to +10kb. Transcript abundance in FPKMs were then normalized by the average FPKM reported on the track.

GO analysis

We used the R package clusterProfiler 3.10.1 [53] and the org.Dm.eg.db 3.7.0 [93] dictionary to perform gene ontology (GO) analysis. For the stage 2 comparison, we generated a test set of the *D. melanogaster* gene names for every gene in our dataset that was maternally deposited in at least any 7 of our species, and performed an enrichment analysis using enrichGO's default parameters using a background set of all *D. melanogaster* genes. For the stage 5 comparison, we generated a test set of the *D. melanogaster* gene names for which at least two orthologs in our dataset showed zygotic-only expression (see Stage-specific gene analysis section above for definition). This threshold approximates the percent of the genome that we observed to be zygotic-only. We then performed an enrichment analysis using enrichGO's default parameters using a background set of *D. melanogaster* genes that are not maternally deposited in at least two species. This analysis therefore specifically examines the zygotically activated genes in the context of genes that are "off" at stage 2 (FPKM$<$1 at this stage). For our analysis of stage 2 motifs, we generated a test set for each motif consisting of genes that contained that motif in at least two species and were maternally deposited (FPKM $>$ 1) in at least two species. We then performed an enrichment analysis using enrichGO's default parameters using a background set of all *D. melanogaster* genes. For our analysis of stage 5 motifs, we generated a test set for each motif using genes that were represented by transcripts $>$1 FPKM at stage 5 in at least two species and had the motif of interest in at least two species. We then performed an enrichment analysis using enrichGO's default parameters using a background set of *D. melanogaster* genes that were represented by transcripts $>$1 FPKM at stage 5. To visualize our results, we employed the dotplot method for enrichGO objects, also from the clusterProfiler package. For each motif, the top 3 GO terms were identified and added to the y-axis labels. Whenever any GO category from another motif was identified as statistically significant (α = .05), that GO category was shaded appropriately.

To discover motifs associated with particular GO categories, we generated a list of genes that were both maternally deposited and associated with each GO term of interest, as well as a list of genes that were maternally deposited but not associated with the GO term of interest. For each GO term, we ran HOMER using the same parameters as the initial motif discovery, using the genes associated with the GO term as the test list and the genes not associated with the GO term as the background. We restricted this analysis to the upstream regions of *Drosophila melanogaster* genes.

Model fitting

Logistic regression was performed using the "glm" function in R, using the logit link function. As inputs, we used the list of motifs generated from HOMER and their counts as described in

the "Motif Position and Count" section above. To avoid redundant motifs in our model, only motifs of size 10 were considered. To evaluate the strand-specificity of motifs, we compared two generalized linear models using the formulas indicated in S5 Text. To identify the most important motifs, the R function stepAIC from the MASS library 7.3–51.4 [94] was used to find generate an ordered list of motifs. The base model used contained no additional features (chromatin state, etc). StepAIC was run 8 steps to generate a short list of motifs for evaluation.

Analysis of physical clustering of co-expressed genes

To evaluate the effect of gene cluster size on expression, we iterated through each species for both stage 2 and stage 5 and assigned sizes of co-expressed gene clusters on the chromosome, based on how many adjacent genes were co-expressed, resulting in cluster size frequencies for each genome. Errors were calculated using 95% confidence interval for a two-tailed binomial distribution.

To compare the clustering of different datasets with varying percents of "on" genes, we employed the Wald–Wolfowitz runs test.

Tissue-specific RNA Levels

modENCODE tissue profiles [55] were downloaded from flybase.org. Flyatlas2 tissue profiles were downloaded from http://flyatlas.gla.ac.uk/FlyAtlas2/ [56].

Gene length and distance between genes

To determine gene length, we examined the relevant line of the appropriate .GFF file and took the difference between the end and the start positions. To determine the distance between genes, we look at the appropriate .GFF file and took the difference of positions between adjacent genes from transcription start site (TSS) to TSS.

Maternal deposition rates as compared to gene length, distance, and orientation

Genes were binned by category and by either distance or length. For Fig 5C, 250 bins of 70bp width were used. For Fig 5D, 60 bins of 70bp width were used and bins with fewer than 6 genes were disregarded. Confidence intervals were calculated using the binomial distribution with $\alpha = .05$ after Bonferroni.

GC content

GC content levels associated with each gene were evaluated by calculating the number of GC nucleotides within a sliding window of size 50bp for each of 1950 window positions to cover the upstream 2kb of each gene. To evaluate the first bin of each gene, the region from -1bp to -50bp was extracted, and the number of G and C nucleotides was counted. The result was divided by 50 to get the %GC for this window. To calculate the GC content for the next bin, this process was repeated on the region from -2bp to -51bp. Each bin had its GC content evaluated this way until the final bin of -451bp to -500bp. To evaluate how closely a particular upstream region resembled a maternally deposited-like distribution or a non maternally deposited-like distribution for the purposes of modeling, we calculated the average GC content for each position of maternally deposited, and not maternally deposited genes. Then for each gene, we measured the correlation between the GC content and that of both category averages. We used the difference in these correlations as a metric to evaluate similarity in GC content for each gene.

Supporting information

S1 Fig. Distribution of motif qualities by location in a representative species in each stage. *D. ananassae* was selected as a representative species. Motif qualities are given by the negative natural logarithm of the q-value outputted by HOMER. High quality motifs enriched for stage 2 (A) are most likely to be found in the 2kb upstream of a gene. Motifs for stage 5 (B) are generally less high quality by this metric, and while the highest quality tend to also be enriched 2kb upstream, some are enriched in 2kb upstream regions of non-expressed genes or enriched in exons.
(PDF)

S2 Fig. Identified motifs are more highly enriched within maternal genes than housekeeping genes. (A) Within non-housekeeping genes, the discovered motifs are much more common within maternally deposited genes. Error bars represent 95% confidence intervals by the binomial distribution. P-values are generated by the prop.test function in R. This shows that maternal genes that are not housekeeping genes are highly enriched for the identified motifs, thus the motifs are not solely being identified due to high proportions of housekeeping genes among maternally deposited genes. (B) Genes labeled as maternally deposited are more likely to contain the identified motifs than genes labeled as housekeeping. Effects were calculated by generating a generalized linear model in the form [presence of motif within genes] ~ [housekeeping or not] + [maternally deposited or not]. Error bars represent standard error. This provides additional evidence that the motifs are not being identified only due to their role in regulating housekeeping genes, but rather that they are more highly associated with maternally deposited genes than housekeeping genes.
(PDF)

S3 Fig. Low quality motifs are less likely to be shared across species. In a manner similar to Fig 1A and 1B, we discovered motifs for each species at both stage 2 and stage 5 and evaluated what percent of motifs were shared among species. Unlike the analysis described in Fig 1A and 1B, we did not apply a quality filter.
(PDF)

S4 Fig. Representative positional distributions of motifs. Distributions for both maternally deposited genes ("on") and non- maternally deposited genes ("off") are shown. (A) The positional distribution of the DREF motif, which follows the same pattern as M1BP, Zipic, Ohler-6, and E-box, and many motifs without identified factors that bind them. These motifs are found upstream of maternally deposited genes (red), with a higher frequency closer to the transcription start site. They are not found with any frequency in non-maternally deposited genes (blue). (B,C) Positional distribution patterns of some rare, undocumented motifs. In both, we see that the motif is more enriched in maternally deposited genes than in non-maternally deposited genes, but that the enrichment difference is less than those motifs represented by (A) above. In (B), this motif is most highly enriched upstream, less enriched around the transcription start site (TSS), and more highly enriched again downstream of the TSS (though less so than upstream). In (C), we see the highest enrichment downstream of the TSS, with a dip in enrichment around the TSS, and less enrichment upstream of the TSS than downstream.
(PDF)

S5 Fig. GO-term specific motifs exist, but are not predictive of maternal deposition. The effect and p-value column data are generated from a generalized linear models of the form [maternal deposition] ~ [motif presence], given a number of genes whose adjacent genes are

expressed. Although the effect is always positive, indicating a slight increase in maternal deposition rates for genes with this motif, the high p-values indicate that these results are not statistically significant.
(PDF)

S6 Fig. GC content of the region upstream of the TSS. GC content for each gene in a sliding window with 50bp width is summed for each gene in the category. (A) Maternally deposited genes. (B) Non-maternally deposited genes. (C) Zygotic-only genes. Note the high number of genes with higher GC content immediately upstream of maternally deposited genes, and the lower GC content upstream of this GC-enriched region.
(PDF)

S1 Table. A summary of the top ranked zygotic motifs. Motifs were selected if they were enriched in the combined upstream regions of all species with a q-value $< 1\times10^{-50}$ and a Tomtom match to any motif in an existing database with $q < 0.1$. If there were more than one, the best two matches to motifs in existing databases were reported in the Best Match column. Some motifs are plausible binding sites for known embryonic regulators.
(PDF)

S1 Text. A list of the motifs used in the model construction. This list of motifs is was found in representative species *D.ananassae* by searching for overrepresented motifs in maternally deposited genes, regardless of motif strand.
(TXT)

S2 Text. A list of the motifs used in the model construction. This list of motifs is was found in representative species *D.ananassae* by searching for overrepresented motifs in maternally deposited genes on only one strand using the -norevopp parameter in HOMER.
(TXT)

S3 Text. Transcript levels (in FPKM) for proteins of interest in both stages, as determined by Atallah and Lott, 2018.
(TXT)

S4 Text. A list of the motifs used in the model construction. This list of motifs was generated by searching for overrepresented motifs in maternally deposited genes in all species.
(TXT)

S5 Text. A description of the model generation process used in several sections.
(PDF)

Acknowledgments

We would like to thank Joel Atallah for his work on the original dataset, Gizem Kalay, Anna Feitzinger, and Emily Cartwright for comments on the manuscript, and all members of the Lott Lab and the UC Davis Superfly group for feedback.

Author Contributions

Conceptualization: Charles S. Omura, Susan E. Lott.

Formal analysis: Charles S. Omura.

Funding acquisition: Susan E. Lott.

Investigation: Charles S. Omura, Susan E. Lott.

Supervision: Susan E. Lott.

Visualization: Charles S. Omura.

Writing – original draft: Charles S. Omura, Susan E. Lott.

Writing – review & editing: Charles S. Omura, Susan E. Lott.

References

1. Driever W, Nüsslein-Volhard C. The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. Cell. 1988; 54: 95–104. https://doi.org/10.1016/0092-8674(88)90183-3 PMID: 3383245
2. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, et al. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell. 2007; 131: 174–187. https://doi.org/10.1016/j.cell.2007.08.003 PMID: 17923096
3. Vastenhouw NL, Cao WX, Lipshitz HD. The maternal-to-zygotic transition revisited. Development. 2019; 146. https://doi.org/10.1242/dev.161471 PMID: 31189646
4. Schulz KN, Harrison MM. Mechanisms regulating zygotic genome activation. Nat Rev Genet. 2019; 20: 221–234. https://doi.org/10.1038/s41576-018-0087-x PMID: 30573849
5. Ventos-Alfonso A, Ylla G, Belles X. Zelda and the maternal-to-zygotic transition in cockroaches. FEBS J. 2019. https://doi.org/10.1111/febs.14856 PMID: 30993896
6. Navarro-Costa P, McCarthy A, Prudêncio P, Greer C, Guilgur LG, Becker JD, et al. Early programming of the oocyte epigenome temporally controls late prophase I transcription and chromatin remodelling. Nat Commun. 2016; 7: 12331. https://doi.org/10.1038/ncomms12331 PMID: 27507044
7. Mische S, Li M, Serr M, Hays TS. Direct observation of regulated ribonucleoprotein transport across the nurse cell/oocyte boundary. Mol Biol Cell. 2007; 18: 2254–2263. https://doi.org/10.1091/mbc.E06-10-0959 PMID: 17429069
8. Tadros W, Westwood JT, Lipshitz HD. The mother-to-child transition. Dev Cell. 2007; 12: 847–849. https://doi.org/10.1016/j.devcel.2007.05.009 PMID: 17543857
9. Liang H-L, Nien C-Y, Liu H-Y, Metzstein MM, Kirov N, Rushlow C. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature. 2008; 456: 400–403. https://doi.org/10.1038/nature07388 PMID: 18931655
10. Harrison MM, Li X-Y, Kaplan T, Botchan MR, Eisen MB. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet. 2011; 7: e1002266. https://doi.org/10.1371/journal.pgen.1002266 PMID: 22028662
11. Akam M. The molecular basis for metameric pattern in the Drosophila embryo. Development. 1987; 101: 1–22.
12. Ingham PW. The molecular genetics of embryonic pattern formation in Drosophila. Nature. 1988; 335: 25–34. https://doi.org/10.1038/335025a0 PMID: 2901040
13. Tadros W, Goldman AL, Babak T, Menzies F, Vardy L, Orr-Weaver T, et al. SMAUG is a major regulator of maternal mRNA destabilization in Drosophila and its translation is activated by the PAN GU kinase. Dev Cell. 2007; 12: 143–155. https://doi.org/10.1016/j.devcel.2006.10.005 PMID: 17199047
14. Benoit B, He CH, Zhang F, Votruba SM, Tadros W, Westwood JT, et al. An essential role for the RNA-binding protein Smaug during the Drosophila maternal-to-zygotic transition. Development. 2009; 136: 923–932. https://doi.org/10.1242/dev.031815 PMID: 19234062
15. Laver JD, Li X, Ray D, Cook KB, Hahn NA, Nabeel-Shah S, et al. Brain tumor is a sequence-specific RNA-binding protein that directs maternal mRNA clearance during the Drosophila maternal-to-zygotic transition. Genome Biol. 2015; 16: 94. https://doi.org/10.1186/s13059-015-0659-4 PMID: 25962635
16. Bushati N, Stark A, Brennecke J, Cohen SM. Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in Drosophila. Curr Biol. 2008; 18: 501–506. https://doi.org/10.1016/j.cub.2008.02.081 PMID: 18394895
17. Becalska AN, Gavis ER. Lighting up mRNA localization in Drosophila oogenesis. Development. 2009; 136: 2493–2503. https://doi.org/10.1242/dev.032391 PMID: 19592573
18. Clark A, Meignin C, Davis I. A Dynein-dependent shortcut rapidly delivers axis determination transcripts into the Drosophila oocyte. Development. 2007; 134: 1955–1965. https://doi.org/10.1242/dev.02832 PMID: 17442699

19. Barckmann B, Simonelig M. Control of maternal mRNA stability in germ cells and early embryos. Biochim Biophys Acta. 2013; 1829: 714–724. https://doi.org/10.1016/j.bbagrm.2012.12.011 PMID: 23298642

20. Cui J, Sackton KL, Horner VL, Kumar KE, Wolfner MF. Wispy, the Drosophila homolog of GLD-2, is required during oogenesis and egg activation. Genetics. 2008; 178: 2017–2029. https://doi.org/10.1534/genetics.107.084558 PMID: 18430932

21. Benoit P, Papin C, Kwak JE, Wickens M, Simonelig M. PAP- and GLD-2-type poly(A) polymerases are required sequentially in cytoplasmic polyadenylation and oogenesis in Drosophila. Development. 2008; 135: 1969–1979. https://doi.org/10.1242/dev.021444 PMID: 18434412

22. Sallés FJ, Lieberfarb ME, Wreden C, Gergen JP, Strickland S. Coordinate initiation of Drosophila development by regulated polyadenylation of maternal messenger RNAs. Science. 1994; 266: 1996–1999. https://doi.org/10.1126/science.7801127 PMID: 7801127

23. Temme C, Simonelig M, Wahle E. Deadenylation of mRNA by the CCR4-NOT complex in Drosophila: molecular and developmental aspects. Front Genet. 2014; 5: 143. https://doi.org/10.3389/fgene.2014.00143 PMID: 24904643

24. Atallah J, Lott SE. Evolution of maternal and zygotic mRNA complements in the early Drosophila embryo. PLoS Genet. 2018; 14: e1007838. https://doi.org/10.1371/journal.pgen.1007838 PMID: 30557299

25. Bownes M. A photographic study of development in the living embryo of Drosophila melanogaster. J Embryol Exp Morphol. 1975; 33: 789–801. PMID: 809527

26. Campos-Ortega JA, Hartenstein V. The Embryonic Development of Drosophila melanogaster. Springer, Berlin, Heidelberg; 1985.

27. Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, et al. A cis-regulatory map of the Drosophila genome. Nature. 2011; 471: 527–531. https://doi.org/10.1038/nature09990 PMID: 21430782

28. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450: 203–218. https://doi.org/10.1038/nature06341 PMID: 17994087

29. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38: 576–589. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432

30. De Renzis S, Elemento O, Tavazoie S, Wieschaus EF. Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. PLoS Biol. 2007; 5: e117. https://doi.org/10.1371/journal.pbio.0050117 PMID: 17456005

31. Thomsen S, Anders S, Janga SC, Huber W, Alonso CR. Genome-wide analysis of mRNA decay patterns during early Drosophila development. Genome Biol. 2010; 11: R93. https://doi.org/10.1186/gb-2010-11-9-r93 PMID: 20858238

32. Lott SE, Villalta JE, Zhou Q, Bachtrog D, Eisen MB. Sex-specific embryonic gene expression in species with newly evolved sex chromosomes. PLoS Genet. 2014; 10: e1004159. https://doi.org/10.1371/journal.pgen.1004159 PMID: 24550743

33. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37: W202–8. https://doi.org/10.1093/nar/gkp335 PMID: 19458158

34. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018; 9: 189. https://doi.org/10.1038/s41467-017-02525-w PMID: 29335486

35. Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature. 2015; 518: 556–559. https://doi.org/10.1038/nature13994 PMID: 25517091

36. Chen K, Johnston J, Shao W, Meier S, Staber C, Zeitlinger J. A global change in RNA polymerase II pausing during the Drosophila midblastula transition. Elife. 2013; 2: e00861. https://doi.org/10.7554/eLife.00861 PMID: 23951546

37. Liu MM, Davey JW, Jackson DJ, Blaxter ML, Davison A. A conserved set of maternal genes? Insights from a molluscan transcriptome. Int J Dev Biol. 2014; 58: 501–511. https://doi.org/10.1387/ijdb.140121ad PMID: 25690965

38. Ghavi-Helm Y. Functional consequences of chromosomal rearrangements on gene expression: not so deleterious after all? J Mol Biol. 2019. https://doi.org/10.1016/j.jmb.2019.09.010 PMID: 31626801

39. Matsukage A, Hirose F, Yoo M-A, Yamaguchi M. The DRE/DREF transcriptional regulatory system: a master key for cell proliferation. Biochim Biophys Acta. 2008; 1779: 81–89. https://doi.org/10.1016/j.bbagrm.2007.11.011 PMID: 18155677

40. Yang J, Ramos E, Corces VG. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. Genome Res. 2012; 22: 2199–2207. https://doi.org/10.1101/gr.142125.112 PMID: 22895281

41. Nègre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, et al. A comprehensive map of insulator elements for the Drosophila genome. PLoS Genet. 2010; 6: e1000814. https://doi.org/10.1371/journal.pgen.1000814 PMID: 20084099

42. Maksimenko O, Bartkuhn M, Stakhov V, Herold M, Zolotarev N, Jox T, et al. Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin. Genome Res. 2015; 25: 89–99. https://doi.org/10.1101/gr.174169.114 PMID: 25342723

43. Li J, Gilmour DS. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. EMBO J. 2013; 32: 1829–1841. https://doi.org/10.1038/emboj.2013.111 PMID: 23708796

44. Levine M. Paused RNA polymerase II as a developmental checkpoint. Cell. 2011; 145: 502–511. https://doi.org/10.1016/j.cell.2011.04.021 PMID: 21565610

45. Benyajati C, Mueller L, Xu N, Pappano M, Gao J, Mosammaparast M, et al. Multiple isoforms of GAGA factor, a critical component of chromatin structure. Nucleic Acids Res. 1997; 25: 3345–3353. https://doi.org/10.1093/nar/25.16.3345 PMID: 9241251

46. Tsai S-Y, Chang Y-L, Swamy KBS, Chiang R-L, Huang D-H. GAGA factor, a positive regulator of global gene expression, modulates transcriptional pausing and organization of upstream nucleosomes. Epigenetics Chromatin. 2016; 9: 32. https://doi.org/10.1186/s13072-016-0082-4 PMID: 27468311

47. Granok H, Leibovitch BA, Shaffer CD, Elgin SC. Chromatin. Ga-ga over GAGA factor. Curr Biol. 1995; 5: 238–241. https://doi.org/10.1016/s0960-9822(95)00048-0 PMID: 7780729

48. Harrison MM, Botchan MR, Cline TW. Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes. Dev Biol. 2010; 345: 248–255. https://doi.org/10.1016/j.ydbio.2010.06.026 PMID: 20599892

49. Schulz KN, Bondra ER, Moshe A, Villalta JE, Lieb JD, Kaplan T, et al. Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early Drosophila embryo. Genome Res. 2015; 25: 1715–1726. https://doi.org/10.1101/gr.192682.115 PMID: 26335634

50. Sun Y, Nien C-Y, Chen K, Liu H-Y, Johnston J, Zeitlinger J, et al. Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophila zygotic genome activation. Genome Res. 2015; 25: 1703–1714. https://doi.org/10.1101/gr.192542.115 PMID: 26335633

51. Ohler U, Liao G-C, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. Genome Biol. 2002; 3: RESEARCH0087.

52. Lis M, Walther D. The orientation of transcription factor binding site motifs in gene promoter regions: does it matter? BMC Genomics. 2016; 17: 185. https://doi.org/10.1186/s12864-016-2549-x PMID: 26939991

53. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16: 284–287. https://doi.org/10.1089/omi.2011.0118 PMID: 22455463

54. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available: http://www.R-project.org/

55. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, et al. Diversity and dynamics of the Drosophila transcriptome. Nature. 2014; 512: 393–399. https://doi.org/10.1038/nature12962 PMID: 24670639

56. Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. Nucleic Acids Res. 2018; 46: D809–D815. https://doi.org/10.1093/nar/gkx976 PMID: 29069479

57. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012; 13: 227–232. https://doi.org/10.1038/nrg3185 PMID: 22411467

58. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, et al. The developmental transcriptome of Drosophila melanogaster. Nature. 2011; 471: 473–479. https://doi.org/10.1038/nature09715 PMID: 21179090

59. modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330: 1787–1797. https://doi.org/10.1126/science.1198374 PMID: 21177974

60. Bauke A-C, Sasse S, Matzat T, Klämbt C. A transcriptional network controlling glial development in the Drosophila visual system. Development. 2015; 142: 2184–2193. https://doi.org/10.1242/dev.119750 PMID: 26015542

61. Gurudatta BV, Yang J, Van Bortle K, Donlin-Asp PG, Corces VG. Dynamic changes in the genomic localization of DNA replication-related element binding factor during the cell cycle. Cell Cycle. 2013; 12: 1605–1615. https://doi.org/10.4161/cc.24742 PMID: 23624840

62. Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Res. 2016; 26: 70–84. https://doi.org/10.1101/gr.196006.115 PMID: 26518482

63. Hou C, Li L, Qin ZS, Corces VG. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol Cell. 2012; 48: 471–484. https://doi.org/10.1016/j.molcel.2012.08.031 PMID: 23041285

64. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013; 41: D991–5. https://doi.org/10.1093/nar/gks1193 PMID: 23193258

65. Bradley JV. Distribution-free statistical tests. Prentice-Hall; 1968.

66. Artieri CG, Fraser HB. Transcript length mediates developmental timing of gene expression across Drosophila. Mol Biol Evol. 2014; 31: 2879–2889. https://doi.org/10.1093/molbev/msu226 PMID: 25069653

67. Heyn P, Kircher M, Dahl A, Kelso J, Tomancak P, Kalinka AT, et al. The earliest transcribed zygotic genes are short, newly evolved, and different across species. Cell Rep. 2014; 6: 285–292. https://doi.org/10.1016/j.celrep.2013.12.030 PMID: 24440719

68. Zhang L, Kasif S, Cantor CR, Broude NE. GC/AT-content spikes as genomic punctuation marks. Proc Natl Acad Sci U S A. 2004; 101: 16855–16860. https://doi.org/10.1073/pnas.0407821101 PMID: 15548610

69. Naughton C, Avlonitis N, Corless S, Prendergast JG, Mati IK, Eijk PP, et al. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat Struct Mol Biol. 2013; 20: 387–395. https://doi.org/10.1038/nsmb.2509 PMID: 23416946

70. Pedone F, Filetici P, Ballario P. Yeast RNA polymerase II transcription of circular DNA at different degrees of supercoiling. Nucleic Acids Res. 1982; 10: 5197–5208. https://doi.org/10.1093/nar/10.17.5197 PMID: 6292834

71. Matsumoto K, Hirose S. Visualization of unconstrained negative supercoils of DNA on polytene chromosomes of Drosophila. J Cell Sci. 2004; 117: 3797–3805. https://doi.org/10.1242/jcs.01225 PMID: 15252118

72. Vlijm R, V D Torre J, Dekker C. Counterintuitive DNA Sequence Dependence in Supercoiling-Induced DNA Melting. PLoS One. 2015; 10: e0141576. https://doi.org/10.1371/journal.pone.0141576 PMID: 26513573

73. Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. Development. 2009; 136: 3033–3042. https://doi.org/10.1242/dev.033183 PMID: 19700615

74. Van Bortle K, Nichols MH, Li L, Ong C-T, Takenaka N, Qin ZS, et al. Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol. 2014; 15: R82. https://doi.org/10.1186/gb-2014-15-5-r82 PMID: 24981874

75. Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, Furlong EEM. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nat Genet. 2019; 51: 1272–1282. https://doi.org/10.1038/s41588-019-0462-3 PMID: 31308546

76. Despang A, Schöpflin R, Franke M, Ali S, Jerković I, Paliou C, et al. Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. Nat Genet. 2019; 51: 1263–1271. https://doi.org/10.1038/s41588-019-0466-z PMID: 31358994

77. Jambor H, Surendranath V, Kalinka AT, Mejstrik P, Saalfeld S, Tomancak P. Systematic imaging reveals features and changing localization of mRNAs in Drosophila development. Elife. 2015; 4. https://doi.org/10.7554/eLife.05003 PMID: 25838129

78. Dej KJ, Spradling AC. The endocycle controls nurse cell polytene chromosome structure during Drosophila oogenesis. Development. 1999; 126: 293–303. PMID: 9847243

79. Zhimulev IF, Belyaeva ES, Semeshin VF, Koryakov DE, Demakov SA, Demakova OV, et al. Polytene Chromosomes: 70 Years of Genetic Research. International Review of Cytology. Academic Press; 2004. pp. 203–275. https://doi.org/10.1016/S0074-7696(04)41004-3 PMID: 15548421

80. Abruzzi K, Chen X, Nagoshi E, Zadina A, Rosbash M. Chapter Seventeen—RNA-seq Profiling of Small Numbers of Drosophila Neurons. In: Sehgal A, editor. Methods in Enzymology. Academic Press; 2015. pp. 369–386. https://doi.org/10.1016/bs.mie.2014.10.025 PMID: 25662465

81. Liang J, Lacroix L, Gamot A, Cuddapah S, Queille S, Lhoumaud P, et al. Chromatin immunoprecipitation indirect peaks highlight long-range interactions of insulator proteins and Pol II pausing. Mol Cell. 2014; 53: 672–681. https://doi.org/10.1016/j.molcel.2013.12.029 PMID: 24486021

82. Phillips-Cremins JE, Corces VG. Chromatin insulators: linking genome organization to cellular function. Mol Cell. 2013; 50: 461–474. https://doi.org/10.1016/j.molcel.2013.04.018 PMID: 23706817

83. Matzat LH, Lei EP. Surviving an identity crisis: a revised view of chromatin insulators in the genomics era. Biochim Biophys Acta. 2014; 1839: 203–214. https://doi.org/10.1016/j.bbagrm.2013.10.007 PMID: 24189492

84. Matzat LH, Dale RK, Moshkovich N, Lei EP. Tissue-specific regulation of chromatin insulator function. PLoS Genet. 2012; 8: e1003069. https://doi.org/10.1371/journal.pgen.1003069 PMID: 23209434

85. Vardy L, Orr-Weaver TL. Regulating translation of maternal messages: multiple repression mechanisms. Trends Cell Biol. 2007; 17: 547–554. https://doi.org/10.1016/j.tcb.2007.09.002 PMID: 18029182

86. Gramates LS, Marygold SJ, Santos GD, Urbano J-M, Antonazzo G, Matthews BB, et al. FlyBase at 25: looking to the future. Nucleic Acids Res. 2017; 45: D663–D671. https://doi.org/10.1093/nar/gkw1016 PMID: 27799470

87. Brooks MJ, Rajasimha HK, Roger JE, Swaroop A. Next-generation sequencing facilitates quantitative analysis of wild-type and Nrl(-/-) retinal transcriptomes. Mol Vis. 2011; 17: 3034–3054. PMID: 22162623

88. Tao T, Zhao L, Lv Y, Chen J, Hu Y, Zhang T, et al. Transcriptome sequencing and differential gene expression analysis of delayed gland morphogenesis in Gossypium australe during seed germination. PLoS One. 2013; 8: e75323. https://doi.org/10.1371/journal.pone.0075323 PMID: 24073262

89. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25: 1422–1423. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

90. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, et al. RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Res. 2015; 43: W50–6. https://doi.org/10.1093/nar/gkv362 PMID: 25904632

91. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018; 46: D260–D266. https://doi.org/10.1093/nar/gkx1126 PMID: 29140473

92. Jones E, Oliphant T, Peterson P, Others. SciPy: Open source scientific tools for Python. Available: http://www.scipy.org/

93. Carlson M. org.Dm.eg.db: Genome wide annotation for Fly. 2018.

94. Venables WN, Ripley BD. Modern Applied Statistics with S. New York: Springer; 2002. Available: http://www.stats.ox.ac.uk/pub/MASS4