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

Charles S. Omura* and Susan E. Lott*

Department of Evolution and Ecology, University of California, Davis CA 95616

*Corresponding Authors, e-mail: csomura@ucdavis.edu (CSO), selott@ucdavis.edu (SEL)

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 two genomes, maternal versus zygotic, are quite different. For maternally deposited genes, we uncovered many signals that are consistent with transcriptional regulation through control at the level of chromatin 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 spatiotemporally 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.

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, given that the transcripts at early stages of development come from two different genomes. The first set of transcripts are those deposited into the egg by the mother, while the second set are transcribed from the zygotic genome [3–5]. However, the regulatory mechanisms responsible for this precise control 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 process that

47 exemplifies the precision required at the handoff to the zygotic genome is segmentation in *Drosophila*.
48 This process begins with broad maternal gradients which control transcription of early zygotic gap
49 genes, and later pair-rule genes, at precise locations within the embryo at specific developmental times
50 [11,12].

51 Regulation of transcripts in development has been the subject of considerable study in *D.melanogaster*.
52 Much of this study has been focused around the process of the MZT or other important events in early
53 development, such as patterning along the anterior-posterior or dorsal-ventral axes. For example, a
54 number of regulators of maternal transcript degradation at or prior to the MZT have been identified
55 [13–16]. Zygotic transcription activation has also been the subject of considerable study, and has
56 implicated critical transcription factors such as *zelda* and *grainy head* [4]. How transcripts are
57 transported into eggs has been the subject of some study [7,17,18], as has how those maternal
58 transcripts are regulated post-transcriptionally [3,19–23]. Post-transcriptional mRNA regulation is
59 especially crucial at the maternal stage as new transcripts cannot be produced after the completion of
60 oogenesis. However, how transcript production is regulated in the nurse cells is largely unknown. As
61 transcript pools at both the maternal and zygotic stages are highly conserved over evolutionary time
62 [24], we employed a comparative approach to investigate gene regulation at these stages.

63 In this study, we uncover regulatory elements that are associated with transcription in the early
64 *Drosophila* embryo, from both maternally deposited and zygotically transcribed genes. We use motif
65 analysis to compare regulation of maternal versus zygotic transcription, and also investigate how
66 regulation at these two stages is different across *Drosophila* species. To this end, we used a previously
67 generated RNAseq dataset from Atallah and Lott, 2018, which sampled embryos from a developmental
68 stage where all transcripts are maternal (stage 2[25,26]) and a stage after zygotic genome activation
69 (end of stage 5 [23,24), across 14 species, representing ~50 million years of divergence time. Here, we
70 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 high 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 (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 are 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

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

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

108 Analyzing regulatory elements at the post-zygotic genome activation stage (stage 5) presents a
109 challenge, as it is difficult to distinguish newly transcribed zygotic mRNAs from residual maternally
110 deposited mRNAs. At this stage, roughly half of the transcripts present are maternal transcripts that
111 have not yet been degraded [8,30–32]. Therefore, to interrogate regulatory elements associated with
112 zygotic transcription, we restricted our search to genes that do not have transcripts present at stage 2
113 but do have transcripts present by stage 5. Because of these stricter requirements for zygotically
114 transcribed genes, there were far fewer genes in the dataset, resulting in less statistical power.
115 However, without these assumptions, we risk failing to identify signals associated specifically with
116 zygotic transcription amongst the signal of maternal transcription.

117 To determine which proteins are likely to bind to maternal or zygotic motifs, we used Tomtom [33] to
118 evaluate the similarity of the discovered motifs to several motif databases (Table 1) for *D. melanogaster*.
119 For motifs associated with maternally deposited transcripts, we found a striking similarity between the
120 motifs generated by Ramírez et al. in 2018[34], who discovered a number of motifs associated with
121 topologically associated domain (TAD) boundaries in *Drosophila* S2 cells. TAD boundaries are genomic
122 regions where the chromatin on one side of the boundary interacts substantially less than expected with
123 the chromatin on the other side. We hypothesize that the regulatory mechanisms responsible for
124 generating TAD boundaries in the Ramírez et al. (2018) paper are also responsible for maternal
125 deposition, and that maternal deposition is regulated by the establishment of TAD boundaries. This set
126 of motifs that are found to be associated with TAD boundaries in our study are also similar to the set
127 that were found to be associated with housekeeping core promoters by Zabidi et al. in 2015[35]. This is
128 consistent with existing data showing that functions of maternally deposited genes are enriched for
129 genes with housekeeping activities [36].

130 The motifs associated with maternally deposited genes are predicted to bind several insulators or
131 architectural proteins. An insulator is a regulatory element that suppresses the interactions of other
132 regulatory elements with genes, or prevents the spread of chromatin state. An architectural protein is a
133 protein that organizes and regulates chromatin structure. The most prominent motif by q-value binds to
134 DNA replication-related element factor (DREF), a known architectural protein and the "master key-like
135 factor for cell proliferation" [37]. It is required for normal progression through the cell cycle. It is known
136 to occur in the promoters of many cell proliferation genes and to interact with chromatin remodeling
137 proteins. Interestingly, DREF binding site overlaps with the binding site for BEAF-32, another well-
138 researched protein that acts as an insulator [38,39] that often appears between head-to-head genes
139 (genes with adjacent promoters that get transcribed in opposite directions). Another identified motif is
140 predicted to bind ZIPIC, which is known to bind and recruit CP190, an insulator. A previous study

141 provides evidence for the co-localization of ZIPC and BEAF-32[40], which likely work together with
142 CP190 to perform insulator functions. Thus of the most enriched motifs in maternal genes (DREF, BEAF-
143 32, ZIPC), many have previously identified roles as insulators or in other ways regulating chromatin
144 state.

145 Another maternal motif identified is predicted to bind M1BP (motif-1 binding protein), which causes
146 RNA polymerase II (Pol II) to pause on the gene [41]. Pol II pausing is critical to early zygotic
147 expression[42,43] but its function in producing the maternal transcriptome is unknown. Several
148 functions have been suggested for this Pol II pausing behavior, including maximizing transcription speed
149 once certain conditions are met, synchronizing with RNA processing machinery, reacting to other
150 developmental or environmental signals, keeping chromatin accessible, and acting as an insulator. Given
151 that M1BP is both maternally deposited at high levels and has increased expression in the early embryo,
152 it is possible that M1BP has multiple functions at different time points. During oogenesis, pausing to
153 wait for external signals or RNA processing machinery seems counterproductive to maximizing
154 transcription in the ovary, but the other function of maintaining a state of open chromatin and
155 solidifying TAD boundaries may be very important. In contrast, at stage 5 it may be much more
156 important to maximize expression in response to certain signals.

157 In searching for motifs associated with zygotic expression, we recovered motifs for well-known
158 regulators of the zygotic genome (Table 1). We only identified a small number of highly enriched motifs
159 at this stage, and thus were able to predict a much smaller number of predicted factors binding to these
160 motifs, including Trl (or GAGA factor) and Zelda. Trl is a known early zygotic activator and chromatin
161 remodeler [44–46] and Zelda is known as a “master key regulator” to early developmental genes [9,47]
162 and appears to be a pioneer transcription factor that establishes the initial chromatin landscape of the
163 zygotic genome [5] . In addition to these high-quality motifs, we found a large number of motifs with
164 lower quality scores (Table S1). These motifs may regulate spatio-temporal specific genes that we

165 observe in the early embryo, and thus have a lower enrichment score due to our whole-embryo
166 approach being ill-equipped to finding such specific patterns.

167 Similar motifs appear in different species

168 To quantify the conservation of the discovered motifs across the 11 species in our study, we used
169 Tomtom [33] to measure the similarity between the sets of motifs discovered in different species. For a
170 motif to be considered conserved between two species, we required that it be discovered by HOMER in
171 both species and for Tomtom to report a statistically significant alignment score (see Methods). At the
172 maternal stage, we found that high quality (q-value < 1e-100 by HOMER, see Methods) motifs tended to
173 be well-conserved (Fig 1A) with a large percentage of the total discovered motif content shared across
174 species. We observed that sister species *D. pseudoobscura* and *D. persimilis* are unique in that they have
175 the highest number of motifs that are either species-specific or are only shared with each other, and
176 have the fewest number of motifs shared with the rest of the species. This is especially noteworthy
177 considering that this lineage is roughly in the middle of the distribution of divergence times from most
178 of the other species, and thus many more distantly related species comparisons have a higher degree of
179 motif conservation than do any comparisons with these two species. This is consistent with previous
180 results [24] that this lineage has a disproportionately high number of changes in transcript abundance
181 for its phylogenetic position, and suggests that these large number of changes in transcript abundance
182 may be due to the large scale changes in regulation in these species observed here . When comparing
183 the rest of the species, we found a relatively higher number of conserved motifs shared between pairs
184 of species within the *Drosophila melanogaster* species group (*D. melanogaster*, *D. simulans*, *D. sechellia*,
185 *D. yakuba*, *D. erecta*, *D. ananassae*), and a slightly reduced number of conserved motifs between the *D.*
186 *melanogaster* group species and the more distantly related species (*D. willistoni*, *D. mojavensis*, *D. virilis*)
187 (Fig 1B). At stage 5, we do not observe a high percentage of conserved motifs between species, rather
188 we observe many motifs that are significantly enriched in just one or two species. We also observe little

189 phylogenetic signal in the data, with the only detectable pattern being that the species with the longest
190 divergence time from the rest of the species, *D. virilis* and *D.mojavensis,* have slightly fewer shared
191 motifs (Fig 1 C,D). If the unique motifs at either stage indeed represent newly evolved regulatory
192 mechanisms, we expect that these motifs to be rare or to have a smaller frequency difference between
193 transcribed and non-transcribed genes. Either of these effects would raise the false discovery rate as
194 reported by HOMER, which makes the number of species-specific zygotic motifs identified all the more
195 remarkable.

196 Motif conservation by gene

197 While these results show that some motifs are important to regulation in the genomes of multiple
198 species, do not speak to whether orthologous genes in different species tend to contain similar motifs.
199 To investigate whether regulation was conserved at the level of individual genes, we compared the
200 motif content of each *D. melanogaster* gene (see Methods) to the motif content of each of its orthologs
201 from other species. We counted motifs as conserved between two species if the motif appeared in both
202 orthologs. For both stage 2 and stage 5, we categorized motifs based on the percent of orthologs for
203 which the motif was conserved (Fig 1E). Motifs have different levels of gene-specific conservation
204 between stages, with maternal stage motifs appearing to have lower conservation across orthologues
205 than zygotic stage motifs, where a larger proportion of orthologues possess the same motif. It is possible
206 that the genes that are missing motifs compared to their orthologues are regulated by different motifs,
207 or that the same motifs that are in radically different positions in different species. As many different
208 maternal motifs appear to be regulating transcription at the level of chromatin state, these motifs may
209 be able to function interchangeably, thus leading to more motif turnover at this stage but with higher
210 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 (Fig S2). 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 [48], there is some evidence that motifs do not behave in a strand-specific manner[49]. To evaluate the importance of the strandness 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, $p < .001$). Almost all motifs had the

234 same sign coefficients, but with different magnitudes, indicating that while motifs had the same effect
235 regardless of orientation, their efficiency could be increased if the orientation was optimal.

236 While the strandedness of motifs may play a small role in their overall effect, we want to know if
237 strandedness makes a qualitative difference to our motifs effects on transcript level, and if we can use
238 motif strand to improve our model. To determine this, we ran HOMER exclusively on the same strand
239 that the gene appeared on, rather than the default mode of scanning both strands. This resulted in the
240 same set of motifs being discovered. This is consistent with the regression results that show that the
241 two strands of the motifs generally have the same sign, indicating an effect in the same direction on
242 transcript level. To evaluate the overall effectiveness of this strategy, we constructed another regression
243 using only the data from the same-strand motifs. This regression performed less well than the
244 regression using motifs from both strands (AIC = 7915.8 for the unstranded regression, AIC = 8612.5 for
245 the stranded regression for a representative species *D. ananassae*). Overall, this suggests that motif
246 binding elements need not bind in a strand specific manner to induce their effects, though the optimal
247 orientation provides measurable increase in their effect on transcription. This result is the same at both
248 stage 2 and stage 5.

249 GO analysis

250 While we have identified a set of motifs that together seem to be responsible for early embryonic RNA
251 content, we next asked if these motifs are likely to be regulating genes with specific types of functions.
252 To this end, we performed gene ontology (GO) analysis on groups of genes, based on their motif
253 content. To simplify this analysis, we chose to focus on the top 8 motifs as reported by HOMER, and for
254 each of the 8 motifs, we compared their enrichment relative the enrichGO function in the clusterProfiler
255 R package [50,51]. 8 motifs were chosen to include representatives from both known and unknown
256 motifs. When comparing stages, we observe no overlap between GO terms. (Fig 3A), which is consistent

257 with our expectations that the genes that are activated in the zygote have different functionality to
258 those transcripts that are maternally deposited. Many of the stage 2 motifs show a similar pattern in the
259 GO categories they are associated with, with the strongest associations belonging to the DREF motif,
260 which is strongly associated with most identified categories (Fig 3B). This could be an indication that
261 there is a high degree of homogeneity in terms of the types of genes these motifs may regulate. In
262 contrast, the stage 5 motifs show more variety in the GO terms of genes they are associated with (Fig
263 3C), which could be indicative of more specific regulation at this stage.

264 Predicted maternal motif binding proteins are enriched in the ovary

265 Next, we investigated whether the potential motif binding proteins we identified were plausible
266 regulators of maternal deposition. It is unclear whether the motifs we identified as enriched in
267 maternally deposited genes are associated specifically with maternal deposition, given that chromatin
268 regulators are important at all stages in all tissues. To investigate, we used modENCODE [52] transcript
269 abundance data to compare the mRNA transcript levels for proteins predicted to bind our discovered
270 motifs, and found increased expression in ovaries (Fig 4A) as compared to other tissues sampled. This
271 pattern exists, though to a lesser extent, in the FlyAtlas 2 dataset [53] , which is a tissue-specific
272 database of transcript levels that utilizes RNA-seq data rather than microarray analysis. The discrepancy
273 between the two datasets could be due to the differences in gene expression measurement method or
274 in experimental methods. While it has been demonstrated that mRNA levels do not necessarily mirror
275 protein levels [54], the enrichment of mRNA in ovaries compared to other tissues is reasonable evidence
276 that these proteins are important in ovaries.

277 To investigate whether these proteins are acting to affect transcription in the ovaries specifically, we
278 examined the expression profiles of RNA in various tissue types (referenced in Fig 3B) from existing RNA
279 quantification datasets [52,55]. For each instance of a motif of interest, we extracted the transcript level

280 from within a 20kb window surrounding the motif and measured the normalized relative transcript level
281 for each position (an example of this is shown in Fig 3B). While the relative normalized transcript level
282 changes in each of the measured tissues, the effect is strongest in ovaries, indicating that the presence
283 of one of these binding sites is associated with a higher increase in transcript levels in the ovary
284 compared to other tissues.

285 As the motifs associated with maternal transcription also act to some degree in other tissues, we next
286 wanted to ask whether the motifs were more enriched in maternally deposited genes than in genes
287 expressed in other tissues. To determine whether regulation in different tissue types were associated
288 with different motifs, we ran HOMER in the same manner as with the maternal stage data to discover
289 enriched motifs (see Methods) in transcripts present in other tissues, as identified from ModENCODE
290 data [56]. We found that most other tissue types were also enriched in the same motifs discovered in
291 transcripts present in stage 2 embryos. However, examining the frequency of motifs in specific genes
292 revealed that the majority of those motifs were from genes that were shared between those tissue
293 types and stage 2 embryos. When we exclude genes that are expressed in stage 2 embryos, HOMER fails
294 to identify the original set of motifs as enriched in male larval gonads, male reproductive tract, adult
295 heads and adult midgut. Furthermore, HOMER detects the motifs at a lesser rate in larval ovaries, larval
296 CNS, and intestinal tract. Despite being identified in fewer tissue types and at a lesser rate in other
297 tissue types as compared to the stage 2 expression levels, the observation that these motifs may also
298 have important functions in other tissue types is consistent with the literature. For example, DREF is
299 known to be important for cell proliferation and chromatin regulation, and is active in many other
300 tissues [57,58]. These motifs are likely associated with many housekeeping genes that are vital to a
301 variety of tissue types.

302 ## Maternally deposited genes are physically clustered on the genome

303 In addition to motifs, we observed several other effects that were related to early embryonic RNA
304 content. Given that many of our discovered motifs bind architectural proteins, we hypothesize many
305 effects may be linked to the physical location of genes on the chromosome. We examined the positional
306 distribution of transcribed genes in various tissue types. As previous papers utilizing the Hi-C method
307 have shown correlation with active topologically associated domains (TADs) and gene expression
308 [59,60], we predicted that any tissue type where regulation is dominated by architectural proteins to
309 transcribe a set of genes physically clustered on the chromosome. To compare the physical gene
310 clustering of transcription at the maternal stage with that of other tissue types, we acquired several
311 RNAseq datasets from NCBI/GEO [61] and performed a Wald–Wolfowitz runs test [62] on each tissue of
312 the previously described tissue types. While all tissues examined showed a strong preference for
313 groupings of transcribed genes, ovaries were the most grouped (Fig 4B). This result was robust to
314 changes in the threshold of what is considered to be expressed (see Methods). This pattern of physical
315 co-expressed gene clustering on the chromosome is consistent with our model of regulation via
316 architectural proteins.

317 While these results speak to the pattern of clustering of expression for maternal genes in terms of
318 adjacent genes being on or off, they do not account for the distance between genes. To answer the
319 question of whether this clustering phenomenon is dependent on distance, we examined the distance to
320 adjacent genes. We observed a trend whereby proximity to an active promoter increases the odds of
321 maternal deposition (Fig 4C). This effect was slightly affected by the strandedness of the two genes
322 whereby genes that have an opposite orientation are more likely to have different expression. This is
323 consistent with observations from previous studies [34] that consecutive genes on the same strand were
324 more likely to show co-expression, while consecutive genes on opposite strands were more likely to
325 have different expression.

326 Many previous studies have observed that zygotic genes tend to be short in length [24,30,63,64]. In
327 addition to affecting transcription speed, shorter gene lengths result in a smaller distance between
328 transcriptional units along the chromosome, especially when considering which strand the gene is on. To
329 explore gene length in maternal genes and the relationship between gene length and the position on
330 the chromosome, we measured the maternal deposition rates with respect to gene length. We observed
331 a trend that in most species, shorter genes are less likely to be maternally deposited. There are
332 differences in the length of maternal genes across species, and this trend could be partly due to the bias
333 for more highly annotated genomes to be enriched in shorter genes (Fig 4D). Additionally, chromatin
334 context seems to heavily influence this effect: when the adjacent genes are off, gene length is much
335 more important (Fig 4C) and very short genes are very likely to be off. This could be because shorter
336 genes are more likely to be influenced by the regulatory machinery of a nearby gene. Alternatively,
337 longer genes might be long enough to physically isolate themselves more effectively and establish their
338 own unique regulatory environment.

339 Given that a number of motifs found in this study are bound by proteins annotated as insulators, and
340 the motifs are similar to those that are associated with TAD boundaries, we asked whether the dataset
341 shows properties consistent with these motifs acting as insulators. To determine whether these motif
342 binding sites were associated with proteins acting as insulators, in the sense that they were blocking the
343 action of a nearby enhancer or the spread of chromatin state, we assessed the transcription of nearby
344 genes relative to the transcription of a gene with these identified motifs. For each regulatory region, the
345 gene nearest to that regulatory region was examined, as well as two genes downstream and two
346 upstream. The frequency of motifs was measured based on the transcript abundance pattern of these
347 five genes. Many of the top motifs including Dref, M1BP, Zipic, and E-box, occur more frequently in the
348 center of maternally deposited gene clusters, rather than on the edge of clusters. (t-test p-values 7e-3,

349 2e-6,3e-10, and 1e-4 respectively). This is consistent with previous results that suggest architectural
350 proteins are prevalent in the centers of TADs as well as the edges[34].

351 Stage-specific genes are isolated on the genome

352 Given that maternally deposited genes are physically clustered together in the genome, we wanted to
353 examine if this pattern held with the set of genes that were stage-specific. To determine if consecutively
354 expressed cluster size is related to stage-specificity of transcript representation, we examined maternal-
355 only (transcripts present at stage 2 and entirely degraded by stage 5) and zygotic-only genes (transcripts
356 present at stage 5, not present at stage 2; for both stage-specific categories, see Methods for further
357 definitions) and their frequencies in clusters of different sizes. We determined that for most species,
358 both maternal-only and zygotic-only genes are more likely to be in smaller (1-3 consecutive maternally
359 deposited) groups than in larger groups (more than 3 consecutive genes) (Fig 5, A and B). For zygotic-
360 only genes, this could be an indication that control at the zygotic level is more specific, affecting single
361 genes rather than larger clusters. Meanwhile, maternal-only genes being in small groups could more
362 likely to have more specific regulation than that of all maternal genes as a whole

363 GC-content of upstream regions

364 In *Drosophila*, transcription start sites are frequently associated with a spike in GC content [65]. To
365 explore this phenomenon with respect to the two developmental stages we examined, we evaluated the
366 average GC content of upstream regions for genes in stage 2 and stage 5. When comparing the GC-
367 content of putative cis-regulatory sequences in maternally versus non-maternally deposited genes, we
368 observed an increase in GC-content upstream of the TSS (Fig S3), as well as a dip in GC content ~200bp
369 upstream of these genes. This modulation in GC-content has been previously observed in *Drosophila* as
370 well as humans [65]. To determine whether this modulation of GC-content was predictive of maternal
371 deposition, we constructed four generalized linear models using the GC-content, the motif data, and

372 both the motif data and GC-content as data sources (see Methods). Adding the GC-content to the model
373 that already included motif data improved the model (AIC: 185589 without GC content AIC:183079 with
374 GC content), hence increased GC content upstream of TSS is somewhat predictive of maternal
375 deposition, even when accounting for motif presence in this region.

376

377 Discussion

378 Maternally deposited gene products are responsible for the first stages of embryonic development in all
379 animals [66]. It is therefore critical that the required kind and amount of mRNAs and proteins are
380 deposited into the unfertilized egg. Later in development, the zygotic genome becomes transcriptionally
381 active and takes over control of development from maternal mRNAs. Failure in maternal mRNA
382 deposition, zygotic genome activation, or the transfer of developmental control between the two
383 genomes can lead to lethality [1,9,13], thus the gene products regulating early development are critical
384 to organismal survival.

385 Previous research has shown that the maternal and zygotic mRNA expression profiles of different
386 species of *Drosophila* are generally conserved, but with some noticeable differences [24]. To identify the
387 regulatory basis of trascription at these stages, we leveraged a large comparative dataset to identify the
388 transcription factor binding motifs found in the *cis*-regulatory sequences of these genes. We identified
389 transcription factor binding motifs that are associated with mRNA expression across species for
390 maternally deposited transcripts and zygotically expressed transcripts. We also investigated the effects
391 of other regulatory mechanisms such as chromatin state on maternal and zygotic expression of mRNAs,
392 as well as the association of transcript levels at these two stages of embryogenesis with gene length,
393 strandedness, and GC content.

394 ## Maternal Regulation

395 We found that motifs associated with putative *cis*-regulatory regions of maternally deposited genes
396 are predominantly annotated as insulator binding sites. An insulator is a type of regulatory element
397 that can block the interactions of *cis*-regulatory elements with promoters or prevent the spread of
398 chromatin state. Insulators are known to be important in creating and maintaining the gene
399 expression patterns, ubiquitous in *Drosophila*, and potentially a key factor for *Drosophila* to maintain
400 such a high gene density [40]. Here, we find that the process of maternal deposition may rely heavily
401 on insulators to express a large percentage of the genome. Because the roles and mechanisms of
402 factors annotated as insulators are not well understood, using the term "Architectural Protein"
403 instead of insulator binding protein may be more appropriate [67]. Recently, these proteins have
404 been studied using genome-wide chromatin organization methods, such as Hi-C, which detects
405 regions of interacting chromatin known as Topologically Associated Domains (TADs) and identifies
406 boundaries between them. Histone marks appear to be enriched in certain TADs but stop abruptly at
407 TAD boundaries, supporting the idea that certain TADs are entirely transcriptionally silenced while
408 others are expressed [34]. Furthermore, ChIP-seq has demonstrated that TAD boundaries in other
409 tissues are enriched in architectural protein binding sites [34] , including several those that we
410 identified in this study.

411 To understand the connection between architectural proteins and maternal deposition, we need to
412 examine where these transcripts are produced to understand the cellular context. In the ovary, nurse
413 cells are responsible for the transcription of maternally deposited genes, and there is a considerable
414 body of literature devoted to nurse cell biology. Much study has been directed towards elucidating
415 how nurse cells transport their products into the oocyte and how post translational control
416 mechanisms fine-tune protein levels of maternal transcripts [3,7,18–23,68]. However, despite this
417 wealth of knowledge, the regulatory mechanisms by which the nurse cells specify which genes to

418 transcribe are largely unknown. One unusual feature of nurse cells is that they are highly polyploid
419 [69,70]. One of the major benefits of this could be an across-the-board increase in transcription rates
420 necessary to provision the embryo with all necessary transcripts. These transcripts represent a large
421 proportion of the genome, with estimates ranging from 50-75%, depending on experimental
422 conditions [3], and necessitate large amount of transcription overall in a short period of time. We
423 extract >100ng total RNA from an embryo; this is an astonishingly large amount of RNA to be present
424 in what is essentially at the time of fertilization a single, albeit a highly specialized, cell. One point of
425 comparison is Abruzzi et al. 2015, [71] who extracted 2-5pg RNA per *Drosophila* neuron. A
426 transcriptional environment that is optimized to quickly transcribe huge numbers of genes might be
427 more amenable to control via chromatin state.

428 Given the amount of overlap between the motifs enriched in the *cis*-regulatory regions of maternally
429 deposited genes and the motifs associated with TAD boundaries, it is possible that these same
430 architectural proteins are functioning to define which genes are maternally transcribed and then
431 deposited into the embryo. We found that the maternally deposited genes are highly clustered on
432 the genome, which is indicative of control via architectural proteins. Additionally, we uncovered that
433 proximity to nearby expressed genes is highly correlated with expression. We also identified a
434 pattern whereby the relative strandedness of adjacent genes is indicative of whether they will be
435 maternally deposited, which is a pattern that has been previously observed with insulators [34]. Each
436 of these results is consistent with known behavior of architectural proteins, suggesting that
437 expression at stage 2 is controlled locally on the chromosome by activating TADs rather than specific
438 genes.

439 As architectural proteins are important in determining genome organization and regulating transcription
440 to some degree in all tissues and stages, we investigated whether the regulatory patterns we observed

441 for maternal genes were ovary-specific or shared across all stages and tissues. Many of the motif binding
442 elements discovered in this analysis appear to be enriched in ovaries, although these proteins have
443 important functions in other tissues as well. Some of the proteins predicted to bind our motifs have
444 been noted for being enriched in the regulation of housekeeping genes, and as maternally deposited
445 genes themselves are enriched in housekeeping genes, this result is perhaps unsurprising. A number of
446 studies have suggested that in addition to the common architectural proteins shared across conditions
447 and developmental stages, there may exist tissue-specific architectural proteins that integrate into the
448 canonical protein complex to produce tissue-specific TAD patterns [72–74]. Perhaps this is the case with
449 the ovary, and further study will reveal whether there are ovary-specific factors that may interact with
450 the common architectural proteins whose binding sites we find enriched here. For example, the authors
451 of Mataz et al. 2012 [75] suggest that Shep may be a tissue-specific factor interacting with architectural
452 proteins in the central nervous system. The enrichment Shep in the central nervous system is even less
453 extreme than the enrichment we observe of CP190 (a known interaction partner of ZIPC, one of our
454 maternal expression associated motifs) in ovaries, suggesting that CP190 could also qualify as tissue-
455 specific. Alternatively, the polyploid nature of nurse cells and the extensive and rapid transcription that
456 occurs in these cells may instead provide an extreme enrichment of the common architectural proteins,
457 without the need for stage or tissue specific architectural proteins.

458 Our results show that regulation in ovaries is accomplished primarily through architectural proteins that
459 establish general regions of open chromatin. This process can turn on a large percentage of the genome,
460 without the need to maintain specific motifs within specific genes. However, this leaves us with the
461 question of how the stage 2 mRNA content is so highly conserved across species overall [24], as
462 regulation at the chromatin level would appear less precise than gene-specific regulation. Perhaps
463 regulatory control primarily at the level of chromatin provides redundancy to maintain transcription
464 despite the gain or loss of individual binding sites. Alternatively, there could be other levels of regulatory

465 control that we are unable to detect, with the signal from chromatin-level control being so strong during
466 this time. The high level of conservation of maternal transcripts is also remarkable given the importance
467 of post-transcriptional regulators at this stage [3,19,23,76], as it is not clear if conservation at the
468 transcript level is necessary for conservation at the protein level.

469 Zygotic Regulation

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

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

Conclusions

In this study, we examined regulatory elements associated with maternal transcripts present at stage 2 of embryogenesis and zygotic transcripts present at 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, which speaks to a conservation of regulatory mechanisms across the genus. In general, the high degree of conservation in regulatory elements at the maternal stage and the zygotic stage, while different from one another, 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 are 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 appears contradictory. In contrast, we found zygotic regulatory elements to be considerably more gene-specific. 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.

508 # Methods

509 ## Data Acquisition

510 RNA-seq data utilized for this study was generated previously [24], and is available at NCBI/GEO at
511 accession number GSE112858. This dataset contains RNA-Seq data from single embryos. Embryos were
512 collected either at stage 2, representing a time point before zygotic genome activation, and at the end
513 of stage 5, representing a time point after widespread zygotic genome activation. Embryos were
514 collected from 14 species, however we only used the data from 11 (*D. simulans, D. sechellia*, *D.*
515 *melanogaster*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*)
516 due to annotation deficiencies in the remaining 3. GTF files and references genomes from previously
517 sequenced species [28] were downloaded from Flybase [77].

518 To determine whether a gene would be labeled as 'off' or as 'on', the overall distribution of FPKMs was
519 analyzed. For all species, for both stage 2 and stage 5, a bimodal distribution appeared, with one peak at
520 0 and another at approximately $e^{3.5}$. The commonly used cutoff of FPKM=1 [78,79] was chosen as it falls
521 between these two distributions.

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

524 ## Sequence Selection

525 Preliminary tests were performed to determine which regions were most likely to have regulatory
526 elements. For each gene, several regions were extracted: 10kb upstream,5kb upstream, 2kb upstream,
527 1kb upstream, 500bp upstream, 5' UTR, total introns, total exons, and 3' UTR. For each region,
528 boundaries were obtained from the appropriate GTF and sequences were extracted using BioPython
529 (Version 1.73, (Cock et al., 2009)). The 2kb upstream region showed the highest quality motifs (Fig S1),

530 and thus were used for matching motifs in external databases, measuring motif overlap between
531 species, analysing motif position distributions, and GO analysis. For these analysis (featured in figures 1
532 through 3 double check), UTRs were ignored as not every species had annotated UTRs.

533 Motif Discovery

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

543 To validate the HOMER output files we used MEME[33] v4.12.0 and RSAT [80]. MEME was run using-
544 mod zoops -nmotifs 2 -minw 8 -maxw 12 -revcomp. The RSAT analysis uses the purge-sequences tool,
545 followed by oligo-analysis using the following parameters: -lth occ_sig 0 -uth rank 5000 -return
546 occ,proba,rank -2str -noov -quick_if_possible -seqtype dna -l 8, followed by pattern-assembly using the
547 following parameters: -v 1 -subst 1 -toppat 5000 -2str, followed by matrix-from-patterns using the
548 following parameters: -v 1 -logo -min_weight 5 -flanks 2 -max_asmb_nb 10 -uth Pval 0.00025 -bginput -
549 markov 0 -o purged_result.

550 Zygotic-Only Motifs

551 To examine the regulatory elements associated with zygotic transcripts, we defined genes as being
552 zygotic only if they were off at stage 2 (FPKM <1) and on at stage 5 (FPKM >1). It is necessary to impose

553 such a restriction, as a large percentage (approximately 95%) of genes that are maternally deposited are
554 still expressed at stage 5, and analysis of their regulatory mechanisms would confound the signal of
555 stage 5 only genes.

556 Motif Sharing

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

564 To refine the results of shared motifs, we applied an additional quality cutoff. For stage 2, motifs were
565 first filtered for a q-value of less than 1e-100, and for stage 5, motifs were first filtered for a q-value of
566 1e-10. The difference in the cutoffs used at the two different stages was due to the differences in the
567 overall distribution of q-values for these stages due to a reduced number of zygotic-only genes (see
568 zygotic-only motifs above).

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

573 To find proteins that bind to the discovered motifs, we used Tomtom to query JASPAR and Combined
574 Drosophila Databases using the default parameters[81].

575 ## Motif Position and Count

576 Motif position was determined by using the scanMotifGenomeWide tool to in the HOMER package.
577 Queries were performed by scanning the discovered motifs against the fasta files for each gene. The 5'
578 boundary of the motif was used as the motif position. For the motif counts per gene used in many
579 downstream analyses analysing motif position distributions, GO analysis, GC content analysis, and motif
580 strand analysis. We used this output and counted the occurrence of a given motif in the target region.
581 To quantify positional distribution similarity, we used the stats.anderson_ksamp function from the scipy
582 library V1.2.1[82]. Distributions were considered to be different at α = .05 after Bonferroni correction.

583 ## Transcript Enrichment by Tissue

584 Expression data for various adult tissues was downloaded from modENCODE[56]. To compare
585 enrichment for transcripts with different magnitudes of abundance, we applied an additional
586 normalization. For each transcript, transcript levels in FPKMs were divided by a scaling factor equal to
587 the average of the expression levels in ovaries. This normalization preserves the relative abundances
588 within each transcript, but allows for visualization of transcript levels with dramatically different overall
589 expression levels.

590 ## Expression by Position

591 *D. melanogaster* expression data by position was downloaded from modENCODE [56] for several tissue
592 types. Positions for each motif was determined as previously described in the Motif Position and Count
593 section above. For each instance of the motif of interest, we determined expression values in area from
594 -10kb to +10kb. Transcript abundance in FPKMs were then normalized by the average FPKM reported on
595 the track.

596 ## GO Analysis

597 We used the R package clusterProfiler 3.10.1 [50] and the org.Dm.eg.db 3.7.0 [83] dictionary to perform
598 gene ontology (GO) analysis. For the stage 2 comparison, we generated a test set of the melanogaster
599 gene names for every gene in our dataset that was maternally deposited in at least any 7 of our species,
600 and performed an enrichment analysis using enrichGO's default parameters using a background set of
601 all *D. melanogaster* genes. For the stage 5 comparison, we generated a test set of the *D. melanogaster*
602 gene names for which at least two orthologues in our dataset showed zygotic-only expression (see
603 Zygotic-only motifs section above for definition). This threshold approximates the percent of the
604 genome that we observed to be zygotic only. We then performed an enrichment analysis using
605 enrichGO's default parameters using a background set of *D. melanogaster* genes that are not maternally
606 deposited in at least two species. This analysis therefore specifically examines the zygotically activated
607 genes in the context of genes that are "off" at stage 2 (FPKM<1 at this stage). For our analysis of stage 2
608 motifs, we generated a test set for each motif consisting of genes that contained that motif in at least
609 two species and were maternally deposited (FPKM > 1) in at least two species. We then performed an
610 enrichment analysis using enrichGO's default parameters using a background set of all *D. melanogaster*
611 genes. For our analysis of stage 5 motifs, we generated a test set for each motif using genes that were
612 represented by transcripts >1 FPKM at stage 5 in at least two species and had the motif of interest in at
613 least two species. We then performed an enrichment analysis using enrichGO's default parameters
614 using a background set of *D. melanogaster* genes that were represented by transcripts >1 FPKM at stage
615 5. To visualize our results, we employed the dotplot method for enrichGO objects, also from the
616 clusterProfiler package. For each motif, the top 3 GO terms were identified and added to the y-axis
617 labels. Whenever any GO category from another motif was identified as statistically significant (α = .05),
618 that GO category was shaded appropriately.

619 ## Model Fitting

620 Logistic regression was performed using the "glm" function in R, using the logit link function. As inputs,
621 we used the list of motifs generated from HOMER and their counts as described in the "Motif Position
622 and Count" section above. To avoid redundant motifs in our model, only motifs of size 10 were
623 considered. To evaluate the strand-specificity of motifs, we compared two generalized linear models
624 using the formulas indicated in S1_Model_Generation.pdf. To identify the most important motifs, the R
625 function stepAIC from the MASS library 7.3-51.4[84] was used to find generate an ordered list of motifs.
626 The base model used contained no additional features (chromatin state, etc). StepAIC was run 8 steps to
627 generate a short list of motifs for evaluation.

628 ## Analysis of physical clustering of co-expressed genes

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

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

635 ## Tissue-specific RNA Levels

636 modENCODE tissue profiles [52] were downloaded from flybase.org . Flyatlas2 tissue profiles were
637 downloaded from http://flyatlas.gla.ac.uk/FlyAtlas2/ [53].

638 ## Gene length

639 To determine gene length, we examined the relevant line of the appropriate .GFF file and took the
640 difference between the end and the start positions.

641 ## Distance between genes

642 To determine the distance between genes, we look at the appropriate .GFF file and took the difference
643 of positions between adjacent genes from transcription start site (TSS) to TSS.

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

645 Genes were binned by category and by either distance or length. For the top plot, 150 bins of 70bp
646 width were used. For the bottom plot, 60 bins of 70bp width were used and bins with fewer than 6
647 genes were disregarded. confidence intervals were calculated using the binomial distribution with α =
648 .05 after Bonferroni.

649 ## GC content

650 GC content levels associated with each gene were evaluated by calculating the number of GC
651 nucleotides within a sliding window of size 50bp for each of 1950 window positions to cover the
652 upstream 2kb of each gene. To evaluate the first bin of each gene, the region from -1bp to -50bp was
653 extracted, and the number of G and C nucleotides was counted. The result was divided by 50 to get the
654 %GC for this window. To calculate the GC content for the next bin, this process was repeated on the
655 region from -2bp to -51bp. Each bin had its GC content evaluated this way until the final bin of -451bp to
656 -500bp. To evaluate how closely a particular upstream region resembled a maternally deposited-like
657 distribution or a non maternally deposited-like distribution for the purposes of modeling, we calculated
658 the average GC content for each position of maternally deposited, and not maternally deposited genes.
659 Then for each gene, we measured the correlation between the GC content and that of both category
660 averages. We used the difference in these correlations as a metric to evaluate similarity in GC content
661 for each gene.

662 # Acknowledgements

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

665

666 # References

667 1. Driever W, Nüsslein-Volhard C. The bicoid protein determines position in the Drosophila embryo
668 in a concentration-dependent manner. Cell. 1988;54: 95–104.

669 2. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, et al. Global analysis of
670 mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell.
671 2007;131: 174–187.

672 3. Vastenhouw NL, Cao WX, Lipshitz HD. The maternal-to-zygotic transition revisited.
673 Development. 2019;146. doi:10.1242/dev.161471

674 4. Schulz KN, Harrison MM. Mechanisms regulating zygotic genome activation. Nat Rev Genet.
675 2019;20: 221–234.

676 5. Ventos-Alfonso A, Ylla G, Belles X. Zelda and the maternal-to-zygotic transition in cockroaches.
677 FEBS J. 2019; doi:10.1111/febs.14856

678 6. Navarro-Costa P, McCarthy A, Prudêncio P, Greer C, Guilgur LG, Becker JD, et al. Early
679 programming of the oocyte epigenome temporally controls late prophase I transcription and
680 chromatin remodelling. Nat Commun. 2016;7: 12331.

681 7. Mische S, Li M, Serr M, Hays TS. Direct observation of regulated ribonucleoprotein transport
682 across the nurse cell/oocyte boundary. Mol Biol Cell. 2007;18: 2254–2263.

683 8. Tadros W, Westwood JT, Lipshitz HD. The mother-to-child transition. Dev Cell. 2007;12: 847–
684 849.

685 9. Liang H-L, Nien C-Y, Liu H-Y, Metzstein MM, Kirov N, Rushlow C. The zinc-finger protein Zelda is a
686 key activator of the early zygotic genome in Drosophila. Nature. 2008;456: 400–403.

687 10. Harrison MM, Li X-Y, Kaplan T, Botchan MR, Eisen MB. Zelda binding in the early
688 Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic
689 transition. PLoS Genet. 2011;7: e1002266.

690 11. Akam M. The molecular basis for metameric pattern in the Drosophila embryo.
691 Development. 1987;101: 1–22.

692 12. Ingham PW. The molecular genetics of embryonic pattern formation in Drosophila.
693 Nature. 1988;335: 25–34.

694 13. Tadros W, Goldman AL, Babak T, Menzies F, Vardy L, Orr-Weaver T, et al. SMAUG is a
695 major regulator of maternal mRNA destabilization in Drosophila and its translation is activated by
696 the PAN GU kinase. Dev Cell. 2007;12: 143–155.

697 14. Benoit B, He CH, Zhang F, Votruba SM, Tadros W, Westwood JT, et al. An essential role
698 for the RNA-binding protein Smaug during the Drosophila maternal-to-zygotic transition.
699 Development. 2009;136: 923–932.

700 15. Laver JD, Li X, Ray D, Cook KB, Hahn NA, Nabeel-Shah S, et al. Brain tumor is a sequence-
701 specific RNA-binding protein that directs maternal mRNA clearance during the Drosophila

702 maternal-to-zygotic transition. Genome Biol. 2015;16: 94.

703 16. Bushati N, Stark A, Brennecke J, Cohen SM. Temporal reciprocity of miRNAs and their
704 targets during the maternal-to-zygotic transition in Drosophila. Curr Biol. 2008;18: 501–506.

705 17. Becalska AN, Gavis ER. Lighting up mRNA localization in Drosophila oogenesis.
706 Development. 2009;136: 2493–2503.

707 18. Clark A, Meignin C, Davis I. A Dynein-dependent shortcut rapidly delivers axis
708 determination transcripts into the Drosophila oocyte. Development. 2007;134: 1955–1965.

709 19. Barckmann B, Simonelig M. Control of maternal mRNA stability in germ cells and early
710 embryos. Biochim Biophys Acta. 2013;1829: 714–724.

711 20. Cui J, Sackton KL, Horner VL, Kumar KE, Wolfner MF. Wispy, the Drosophila homolog of
712 GLD-2, is required during oogenesis and egg activation. Genetics. 2008;178: 2017–2029.

713 21. Benoit P, Papin C, Kwak JE, Wickens M, Simonelig M. PAP- and GLD-2-type poly(A)
714 polymerases are required sequentially in cytoplasmic polyadenylation and oogenesis in Drosophila.
715 Development. 2008;135: 1969–1979.

716 22. Sallés FJ, Lieberfarb ME, Wreden C, Gergen JP, Strickland S. Coordinate initiation of
717 Drosophila development by regulated polyadenylation of maternal messenger RNAs. Science.
718 1994;266: 1996–1999.

719 23. Temme C, Simonelig M, Wahle E. Deadenylation of mRNA by the CCR4-NOT complex in
720 Drosophila: molecular and developmental aspects. Front Genet. 2014;5: 143.

721 24. Atallah J, Lott SE. Evolution of maternal and zygotic mRNA complements in the early

722 Drosophila embryo. PLoS Genet. 2018;14: e1007838.

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

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

727 27. Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, et al. A cis-regulatory map
728 of the Drosophila genome. Nature. 2011;471: 527–531.

729 28. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of
730 genes and genomes on the Drosophila phylogeny. Nature. 2007;450: 203–218.

731 29. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of
732 lineage-determining transcription factors prime cis-regulatory elements required for macrophage
733 and B cell identities. Mol Cell. 2010;38: 576–589.

734 30. De Renzis S, Elemento O, Tavazoie S, Wieschaus EF. Unmasking activation of the zygotic
735 genome using chromosomal deletions in the Drosophila embryo. PLoS Biol. 2007;5: e117.

736 31. Thomsen S, Anders S, Janga SC, Huber W, Alonso CR. Genome-wide analysis of mRNA
737 decay patterns during early Drosophila development. Genome Biol. 2010;11: R93.

738 32. Lott SE, Villalta JE, Zhou Q, Bachtrog D, Eisen MB. Sex-specific embryonic gene
739 expression in species with newly evolved sex chromosomes. PLoS Genet. 2014;10: e1004159.

740 33. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools
741 for motif discovery and searching. Nucleic Acids Res. 2009;37: W202–8.

742 34. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-
743 resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun.
744 2018;9: 189.

745 35. Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, et al. Enhancer-core-
746 promoter specificity separates developmental and housekeeping gene regulation. Nature.
747 2015;518: 556–559.

748 36. Liu MM, Davey JW, Jackson DJ, Blaxter ML, Davison A. A conserved set of maternal
749 genes? Insights from a molluscan transcriptome. Int J Dev Biol. 2014;58: 501–511.

750 37. Matsukage A, Hirose F, Yoo M-A, Yamaguchi M. The DRE/DREF transcriptional regulatory
751 system: a master key for cell proliferation. Biochim Biophys Acta. 2008;1779: 81–89.

752 38. Yang J, Ramos E, Corces VG. The BEAF-32 insulator coordinates genome organization
753 and function during the evolution of Drosophila species. Genome Res. 2012;22: 2199–2207.

754 39. Nègre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, et al. A
755 comprehensive map of insulator elements for the Drosophila genome. PLoS Genet. 2010;6:
756 e1000814.

757 40. Maksimenko O, Bartkuhn M, Stakhov V, Herold M, Zolotarev N, Jox T, et al. Two new
758 insulator proteins, Pita and ZIPIC, target CP190 to chromatin. Genome Res. 2015;25: 89–99.

759 41. Li J, Gilmour DS. Distinct mechanisms of transcriptional pausing orchestrated by GAGA
760 factor and M1BP, a novel transcription factor. EMBO J. 2013;32: 1829–1841.

761 42. Chen K, Johnston J, Shao W, Meier S, Staber C, Zeitlinger J. A global change in RNA
762 polymerase II pausing during the Drosophila midblastula transition. Elife. 2013;2: e00861.

763 43. Levine M. Paused RNA polymerase II as a developmental checkpoint. Cell. 2011;145:
764 502–511.

765 44. Benyajati C, Mueller L, Xu N, Pappano M, Gao J, Mosammaparast M, et al. Multiple
766 isoforms of GAGA factor, a critical component of chromatin structure. Nucleic Acids Res. 1997;25:
767 3345–3353.

768 45. Tsai S-Y, Chang Y-L, Swamy KBS, Chiang R-L, Huang D-H. GAGA factor, a positive
769 regulator of global gene expression, modulates transcriptional pausing and organization of
770 upstream nucleosomes. Epigenetics Chromatin. 2016;9: 32.

771 46. Granok H, Leibovitch BA, Shaffer CD, Elgin SC. Chromatin. Ga-ga over GAGA factor. Curr
772 Biol. 1995;5: 238–241.

773 47. Harrison MM, Botchan MR, Cline TW. Grainyhead and Zelda compete for binding to the
774 promoters of the earliest-expressed Drosophila genes. Dev Biol. 2010;345: 248–255.

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

777 49. Lis M, Walther D. The orientation of transcription factor binding site motifs in gene
778 promoter regions: does it matter? BMC Genomics. 2016;17: 185.

779 50. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological
780 themes among gene clusters. OMICS. 2012;16: 284–287.

781 51. R Core Team. R: A Language and Environment for Statistical Computing [Internet].
782 Vienna, Austria: R Foundation for Statistical Computing; 2014. Available: http://www.R-project.org/

783 52. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, et al. Diversity and
784 dynamics of the Drosophila transcriptome. Nature. 2014;512: 393–399.

785 53. Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. FlyAtlas 2: a new version of the
786 Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. Nucleic
787 Acids Res. 2018;46: D809–D815.

788 54. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic
789 and transcriptomic analyses. Nat Rev Genet. 2012;13: 227–232.

790 55. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, et al. The
791 developmental transcriptome of Drosophila melanogaster. Nature. 2011;471: 473–479.

792 56. modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al.
793 Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science.
794 2010;330: 1787–1797.

795 57. Bauke A-C, Sasse S, Matzat T, Klämbt C. A transcriptional network controlling glial
796 development in the Drosophila visual system. Development. 2015;142: 2184–2193.

797 58. Gurudatta BV, Yang J, Van Bortle K, Donlin-Asp PG, Corces VG. Dynamic changes in the
798 genomic localization of DNA replication-related element binding factor during the cell cycle. Cell
799 Cycle. 2013;12: 1605–1615.

800 59. Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, et al. Active
801 chromatin and transcription play a key role in chromosome partitioning into topologically
802 associating domains. Genome Res. 2016;26: 70–84.

803 60. Hou C, Li L, Qin ZS, Corces VG. Gene density, transcription, and insulators contribute to

804 the partition of the Drosophila genome into physical domains. Mol Cell. 2012;48: 471–484.

805 61. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO:
806 archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41: D991–5.

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

808 63. Artieri CG, Fraser HB. Transcript length mediates developmental timing of gene
809 expression across Drosophila. Mol Biol Evol. 2014;31: 2879–2889.

810 64. Heyn P, Kircher M, Dahl A, Kelso J, Tomancak P, Kalinka AT, et al. The earliest
811 transcribed zygotic genes are short, newly evolved, and different across species. Cell Rep. 2014;6:
812 285–292.

813 65. Zhang L, Kasif S, Cantor CR, Broude NE. GC/AT-content spikes as genomic punctuation
814 marks. Proc Natl Acad Sci U S A. 2004;101: 16855–16860.

815 66. Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts.
816 Development. 2009;136: 3033–3042.

817 67. Van Bortle K, Nichols MH, Li L, Ong C-T, Takenaka N, Qin ZS, et al. Insulator function and
818 topological domain border strength scale with architectural protein occupancy. Genome Biol.
819 2014;15: R82.

820 68. Jambor H, Surendranath V, Kalinka AT, Mejstrik P, Saalfeld S, Tomancak P. Systematic
821 imaging reveals features and changing localization of mRNAs in Drosophila development. Elife.
822 2015;4. doi:10.7554/eLife.05003

823 69. Dej KJ, Spradling AC. The endocycle controls nurse cell polytene chromosome structure

824 during Drosophila oogenesis. Development. 1999;126: 293–303.

825 70. Zhimulev IF, Belyaeva ES, Semeshin VF, Koryakov DE, Demakov SA, Demakova OV, et al.
826 Polytene Chromosomes: 70 Years of Genetic Research. International Review of Cytology. Academic
827 Press; 2004. pp. 203–275.

828 71. Abruzzi K, Chen X, Nagoshi E, Zadina A, Rosbash M. Chapter Seventeen - RNA-seq
829 Profiling of Small Numbers of Drosophila Neurons. In: Sehgal A, editor. Methods in Enzymology.
830 Academic Press; 2015. pp. 369–386.

831 72. Liang J, Lacroix L, Gamot A, Cuddapah S, Queille S, Lhoumaud P, et al. Chromatin
832 immunoprecipitation indirect peaks highlight long-range interactions of insulator proteins and Pol II
833 pausing. Mol Cell. 2014;53: 672–681.

834 73. Phillips-Cremins JE, Corces VG. Chromatin insulators: linking genome organization to
835 cellular function. Mol Cell. 2013;50: 461–474.

836 74. Matzat LH, Lei EP. Surviving an identity crisis: a revised view of chromatin insulators in
837 the genomics era. Biochim Biophys Acta. 2014;1839: 203–214.

838 75. Matzat LH, Dale RK, Moshkovich N, Lei EP. Tissue-specific regulation of chromatin
839 insulator function. PLoS Genet. 2012;8: e1003069.

840 76. Vardy L, Orr-Weaver TL. Regulating translation of maternal messages: multiple
841 repression mechanisms. Trends Cell Biol. 2007;17: 547–554.

842 77. Gramates LS, Marygold SJ, Santos GD, Urbano J-M, Antonazzo G, Matthews BB, et al.
843 FlyBase at 25: looking to the future. Nucleic Acids Res. 2017;45: D663–D671.

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

846 79. Tao T, Zhao L, Lv Y, Chen J, Hu Y, Zhang T, et al. Transcriptome sequencing and
847 differential gene expression analysis of delayed gland morphogenesis in Gossypium australe during
848 seed germination. PLoS One. 2013;8: e75323.

849 80. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, et
850 al. RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Res. 2015;43: W50–6.

851 81. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al.
852 JASPAR 2018: update of the open-access database of transcription factor binding profiles and its
853 web framework. Nucleic Acids Res. 2018;46: D260–D266.

854 82. Jones E, Oliphant T, Peterson P, Others. SciPy: Open source scientific tools for Python
855 [Internet]. Available: http://www.scipy.org/

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

857 84. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. Fourth. New York:
858 Springer; 2002. Available: http://www.stats.ox.ac.uk/pub/MASS4

859

Figure Captions

Table 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).

Fig 1: 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 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 one 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

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

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

900 **Fig 3:** Identified maternal regulators are ovary enriched, as is their effect on transcription (A) RNA levels
901 of putative binding proteins by tissue type. Transcript abundances within each gene have been
902 normalized such that the average abundance in ovaries is equal to 1. While identified maternal
903 regulators have regulatory functions in multiple tissue types, they are highly enriched in ovaries
904 compared to other tissues. (B) Average normalized expression levels versus proximity to motif by tissue
905 type. Normalization was performed by dividing each expression value by the average expression from

906 9.9-10kb away. While binding sites for identified maternal regulators are present in multiple tissues, the
907 effect on gene expression is stronger in ovaries compared to other tissues.

908 **Fig 4:** Chromatin stage and maternal deposition. For the analyses in A and B, genes were categorized as
909 either expressed or not expressed (see Methods) and adjacent expressed genes were considered to be
910 clustered, with a cluster size equal to the number of constituent genes. (A) Physical clustering of
911 maternally deposited genes along the chromosome, in a representative species (*D. simulans*). The
912 shaded blue region represents the observed frequency of co-expressed maternal gene clusters of
913 various sizes. The red region represents the 95% CI constructed with 10,000 bootstrap iterations.
914 Maternal genes are co-expressed in clusters along the chromosome more often than expected, given
915 the percent of the genome that is transcribed at this stage. (B) Physical clustering of co-expressed genes
916 on chromosomes in various tissue types. In order to compensate for differing proportions of the
917 genome that are expressed in each tissue type, physical clustering was measured by performing a Wald-
918 Wolfowitz runs test and taking the z-score (see Methods). Maternally expressed genes, represented by
919 stage 2 embryos, show the highest proportion of physical clustering of co-expressed genes, though
920 other tissues such as intestinal stem cells and larval CNS also have highly physically clustered co-
921 expressed genes. (C) Gene length by number of adjacent maternally expressed genes, "open" indicating
922 both adjacent genes are expressed, "border" indicating that one is expressed, and "closed" indicating
923 that neither are expressed. Genes that with more expressed neighbors are more likely to be maternally
924 deposited, regardless of length. Genes without expressed neighbors are less likely to be maternally
925 deposited, with the odds increasing as length increases. (D) Odds of maternal deposition versus distance
926 to the nearest upstream gene by upstream expression and strand. Distance is measured by from
927 transcription start site (TSS) to TSS. When the upstream gene is maternally deposited, odds of maternal
928 deposition are high, but decrease with distance regardless of strand. When the upstream gene is not

929 maternally deposited, odds of maternal deposition are low and have a strand-dependent relationship
930 with distance.

931 **Fig 5:** Stage-specific genes are more likely to be different from their chromatin neighborhood. *D.*
932 *simulans* was chosen as a representative species. (A) Cluster size distribution of maternal only genes
933 (green bars) compared with the expected frequencies based on the overall cluster size frequencies
934 observed at stage two (blue region). The expected frequencies are based on the distribution in Fig 4A
935 multiplied by a scale factor equal to the proportion of maternally deposited genes that are maternal
936 only, with the shaded region representing a 95% confidence interval. (B) Cluster size distribution for
937 zygotic only genes (green bars) compared with the expected frequencies based on the overall cluster
938 size frequencies observed at stage 5 (blue region) in a manner similar to Fig 5A. the shaded region
939 represents a 95% confidence interval. For both stages, stage-specific genes are more likely to be the
940 single gene (or one of a small number of genes) that are expressed where their neighboring genes are
941 not, representing small numbers of "on" genes in an "off" chromatin environment.

942

943 # Supporting Information Figure Captions

944 **S1 Table:** A summary of the top ranked zygotic motifs. Motifs were selected if had enrichment if they
945 were enriched in the combined upstream regions of all species with a q-value < 1e-50 and a Tomtom
946 match to any motif in an existing database with q < .1. If there were more than one, the best two
947 matches to motifs in existing databases were reported in the Best Match column. Some motifs are
948 plausible binding sites for known embryonic regulators.

949 **S1 Fig:** Distribution of motif qualities by location in a representative species in each stage. *D. ananassae*
950 was selected as a representative species. Motif qualities are given by the negative natural logarithm of

951 the q-value outputted by HOMER. High quality motifs enriched for stage 2 (A) are most likely to be
952 found in the 2kb upstream of a gene. Motifs for stage 5 are generally less high quality by this metric, and
953 while the highest quality tend to also be enriched 2kb upstream, some are enriched in 2kb upstream
954 regions of non-expressed genes or enriched in exons.

955 **S2 Fig:** Representative positional distributions of motifs. Distributions for both maternally deposited
956 genes ("on") and non- maternally deposited genes ("off") are shown. (A) The positional distribution of
957 the DREF motif, which follows the same pattern as M1BP, Zipic, Ohler-6, and E-box, and many motifs
958 without identified factors that bind them. These motifs are found upstream of maternally deposited
959 genes (red), with a higher frequency closer to the transcription start site. They are not found with any
960 frequency in non-maternally deposited genes (blue). (B,C) Positional distribution patterns of some rare,
961 undocumented motifs. In both, we see that the motif is more enriched in maternally deposited genes
962 than in non-maternally deposited genes, but that the enrichment difference is less than those motifs
963 represented by (A) above. In (B), this motif is most highly enriched upstream, less enriched around the
964 transcription start site (TSS), and more highly enriched again downstream of the TSS (though less so
965 than upstream). In (C), we see the highest enrichment downstream of the TSS, with a dip in enrichment
966 around the TSS, and less enrichment upstream of the TSS than downstream

967 **S3 Fig:** GC content of the region upstream of the TSS. GC content for each gene in a sliding window with
968 50bp width is summed for each gene in the category. (A) Maternally deposited genes. (B) Non-
969 maternally deposited genes. (C) Zygotic only genes. Note the high number of genes with higher GC
970 content immediately upstream of maternally deposited genes, and the lower GC content upstream of
971 this GC-enriched region.

972

973

	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 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

Table 1

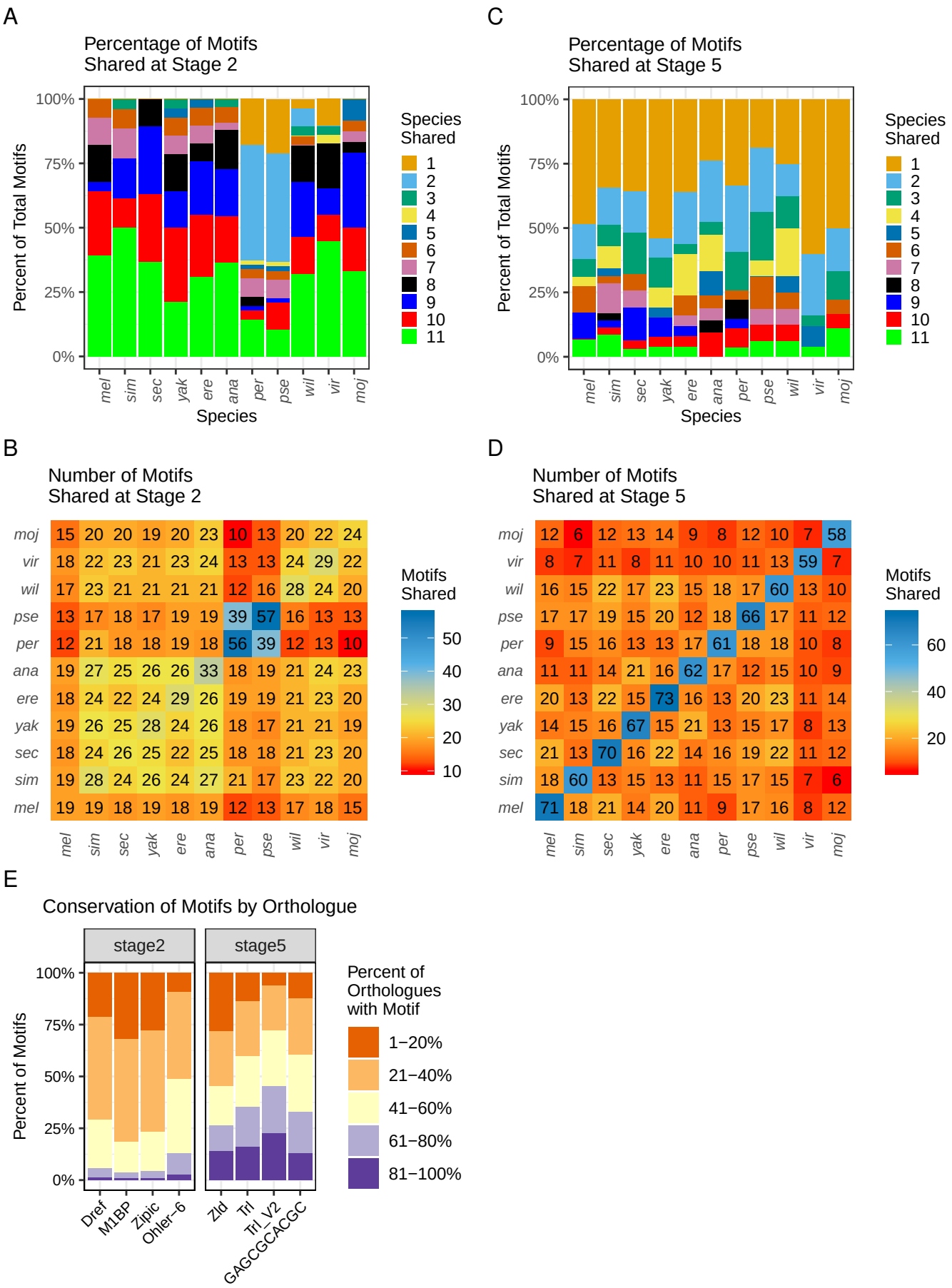
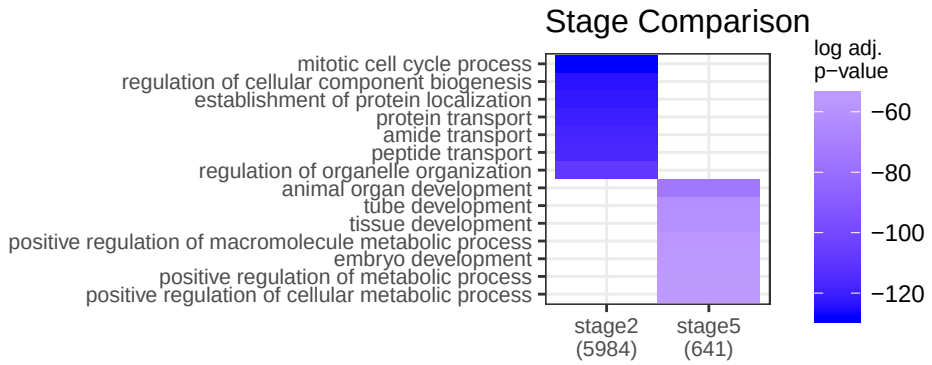
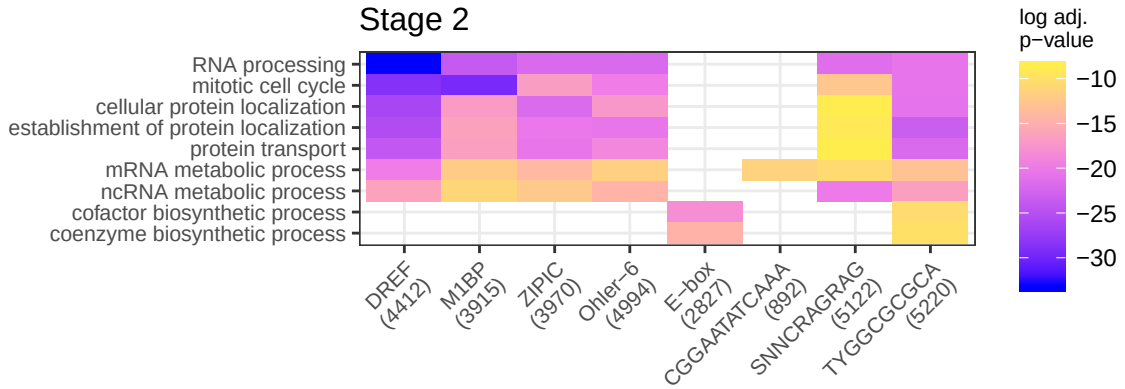

A
Percentage of Motifs
Shared at Stage 2
Percent of Total Motifs
100%
75%
50%
25%
0%
mel sim sec yak ere ana per pse wil vir moj
Species
Species
Shared
1
2
3
4
5
6
7
8
9
10
11
C
Percentage of Motifs
Shared at Stage 5
Percent of Total Motifs
100%
75%
50%
25%
0%
mel sim sec yak ere ana per pse wil vir moj
Species
Species
Shared
1
2
3
4
5
6
7
8
9
10
11
B
Number of Motifs
Shared at Stage 2
moj 15 20 20 19 20 23 10 13 20 22 24
vir 18 22 23 21 23 24 13 13 24 29 22
wil 17 23 21 21 21 21 12 16 28 24 20
pse 13 17 18 17 19 19 39 57 16 13 13
per 12 21 18 18 19 18 56 39 12 13 10
ana 19 27 25 26 26 33 18 19 21 24 23
ere 18 24 22 24 29 26 19 19 21 23 20
yak 19 26 25 28 24 26 18 17 21 21 19
sec 18 24 26 25 22 25 18 18 21 23 20
sim 19 28 24 26 24 27 21 17 23 22 20
mel 19 19 18 19 18 19 12 13 17 18 15
mel sim sec yak ere ana per pse wil vir moj
Motifs
Shared
50
40
30
20
10
D
Number of Motifs
Shared at Stage 5
moj 12 6 12 13 14 9 8 12 10 7 58
vir 8 7 11 8 11 10 10 11 13 59 7
wil 16 15 22 17 23 15 18 17 60 13 10
pse 17 17 19 15 20 12 18 66 17 11 12
per 9 15 16 13 13 17 61 18 18 10 8
ana 11 11 14 21 16 62 17 12 15 10 9
ere 20 13 22 15 73 16 13 20 23 11 14
yak 14 15 16 67 15 21 13 15 17 8 13
sec 21 13 70 16 22 14 16 19 22 11 12
sim 18 60 13 15 13 11 15 17 15 7 6
mel 71 18 21 14 20 11 9 17 16 8 12
mel sim sec yak ere ana per pse wil vir moj
Motifs
Shared
60
40
20
E
Conservation of Motifs by Orthologue
stage2
stage5
Percent of Motifs
100%
75%
50%
25%
0%
Dref
M1BP
Zipic
Ohler−6
Zld
Trl
Trl_V2
GAGCGCACGC
Percent of
Orthologues
with Motif
1−20%
21−40%
41−60%
61−80%
81−100%

Fig 1

A


Stage Comparison
mitotic cell cycle process
regulation of cellular component biogenesis
establishment of protein localization
protein transport
amide transport
peptide transport
regulation of organelle organization
animal organ development
tube development
tissue development
positive regulation of macromolecule metabolic process
embryo development
positive regulation of metabolic process
positive regulation of cellular metabolic process
stage2 (5984)
stage5 (641)
log adj. p−value
−60
−80
−100
−120

B


Stage 2
RNA processing
mitotic cell cycle
cellular protein localization
establishment of protein localization
protein transport
mRNA metabolic process
ncRNA metabolic process
cofactor biosynthetic process
coenzyme biosynthetic process
DREF (4412)
M1BP (3915)
ZIPIC (3970)
Ohler−6 (4994)
E−box (2827)
CGGAATATCAAA (892)
SNNCRAGRAG (5122)
TYGGCGCGCA (5220)
log adj. p−value
−10
−15
−20
−25
−30

C

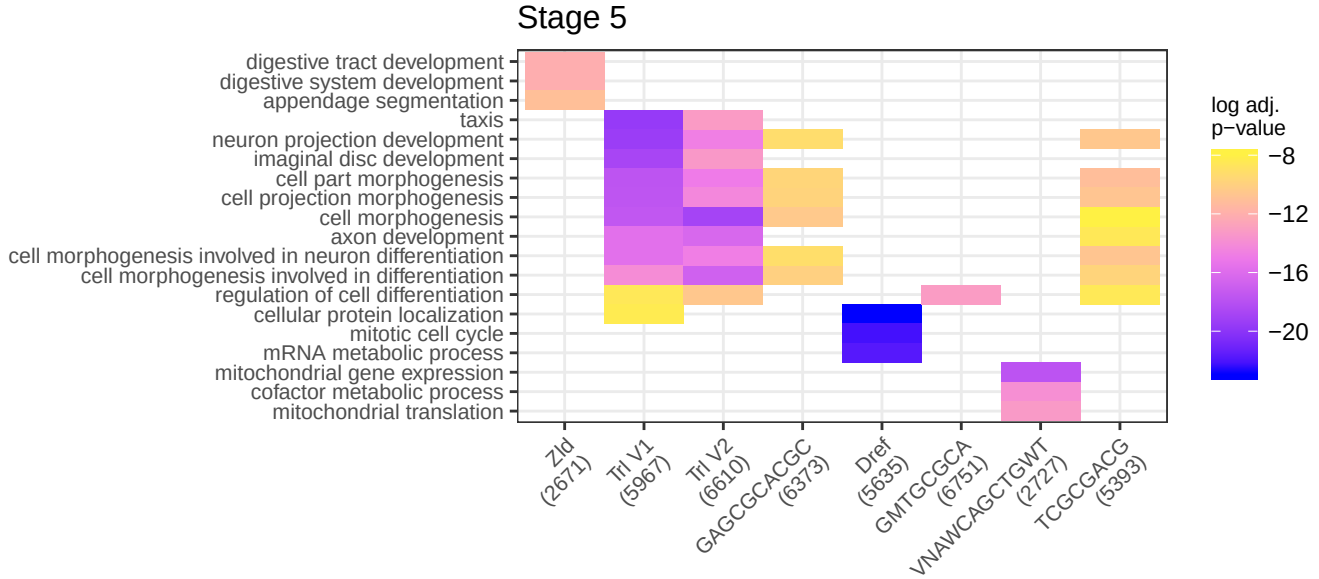
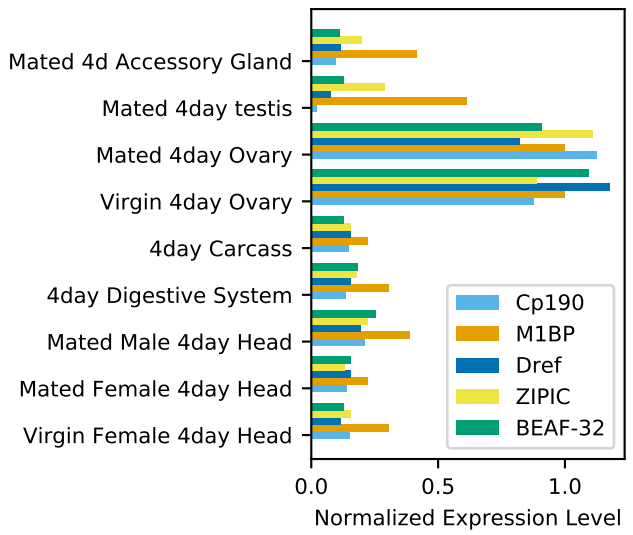

Stage 5
digestive tract development
digestive system development
appendage segmentation
taxis
neuron projection development
imaginal disc development
cell part morphogenesis
cell projection morphogenesis
cell morphogenesis
axon development
cell morphogenesis involved in neuron differentiation
cell morphogenesis involved in differentiation
regulation of cell differentiation
cellular protein localization
mitotic cell cycle
mRNA metabolic process
mitochondrial gene expression
cofactor metabolic process
mitochondrial translation
Zld (2671)
Trl V1 (5967)
Trl V2 (6610)
GAGCGCACGC (6373)
Dref (5635)
GMTGCGCA (6751)
VNAWCAGCTGWT (2727)
TCGCGACG (5393)
log adj. p−value
−8
−12
−16
−20

Fig 2

A


Mated 4d Accessory Gland
Mated 4day testis
Mated 4day Ovary
Virgin 4day Ovary
4day Carcass
4day Digestive System
Mated Male 4day Head
Mated Female 4day Head
Virgin Female 4day Head
Cp190
M1BP
Dref
ZIPIC
BEAF-32
0.0
0.5
1.0
Normalized Expression Level

B

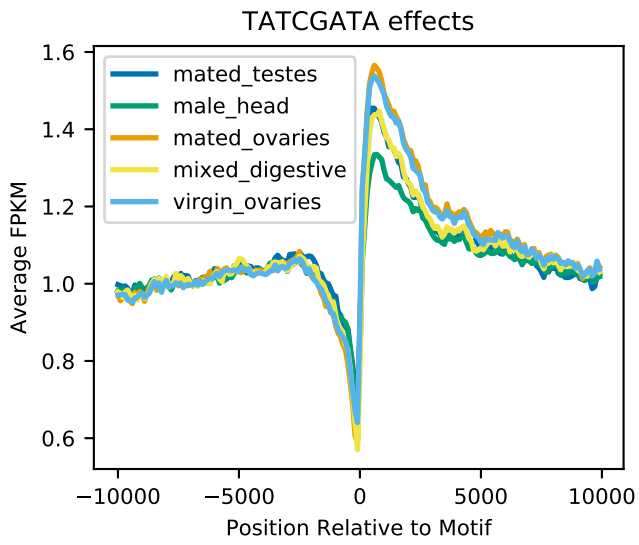

TATCGATA effects
mated_testes
male_head
mated_ovaries
mixed_digestive
virgin_ovaries
Average FPKM
1.6
1.4
1.2
1.0
0.8
0.6
−10000
−5000
0
5000
10000
Position Relative to Motif

Fig 3

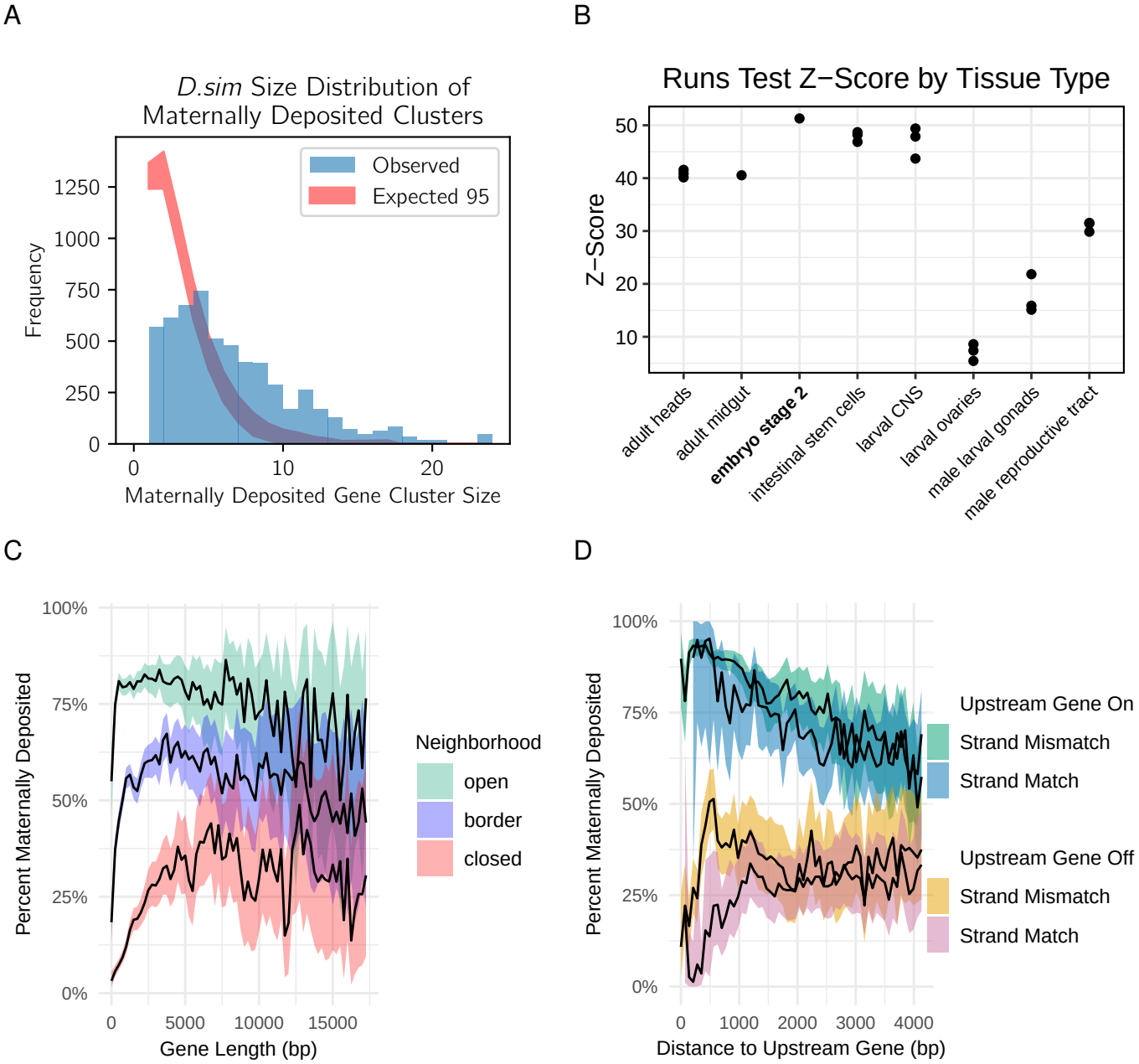
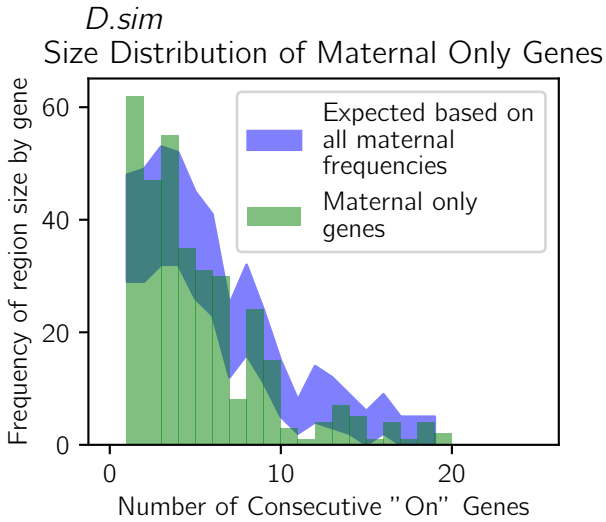

A
D.sim Size Distribution of
Maternally Deposited Clusters
Observed
Expected 95
Frequency
1250
1000
750
500
250
0
0
10
20
Maternally Deposited Gene Cluster Size
B
Runs Test Z−Score by Tissue Type
Z−Score
50
40
30
20
10
adult heads
adult midgut
embryo stage 2
intestinal stem cells
larval CNS
larval ovaries
male larval gonads
male reproductive tract
C
Percent Maternally Deposited
100%
75%
50%
25%
0%
0
5000
10000
15000
Gene Length (bp)
Neighborhood
open
border
closed
D
Percent Maternally Deposited
100%
75%
50%
25%
0%
0
1000
2000
3000
4000
Distance to Upstream Gene (bp)
Upstream Gene On
Strand Mismatch
Strand Match
Upstream Gene Off
Strand Mismatch
Strand Match

Fig 4

A


D.sim
Size Distribution of Maternal Only Genes
Frequency of region size by gene
Number of Consecutive "On" Genes
Expected based on all maternal frequencies
Maternal only genes
0
20
40
60
0
10
20

B

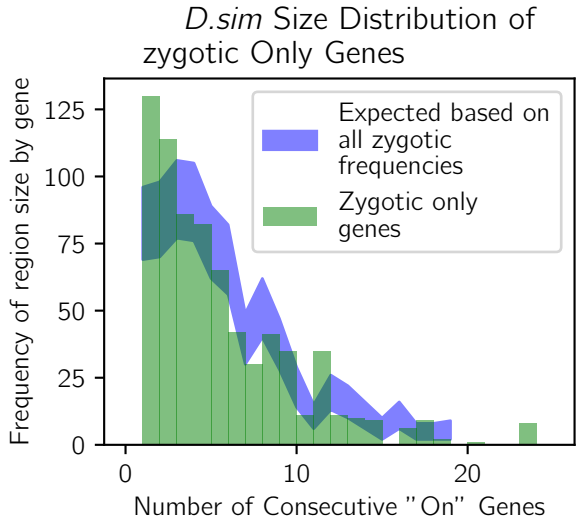

D.sim Size Distribution of zygotic Only Genes
Frequency of region size by gene
Number of Consecutive "On" Genes
Expected based on all zygotic frequencies
Zygotic only genes
0
25
50
75
100
125
0
10
20

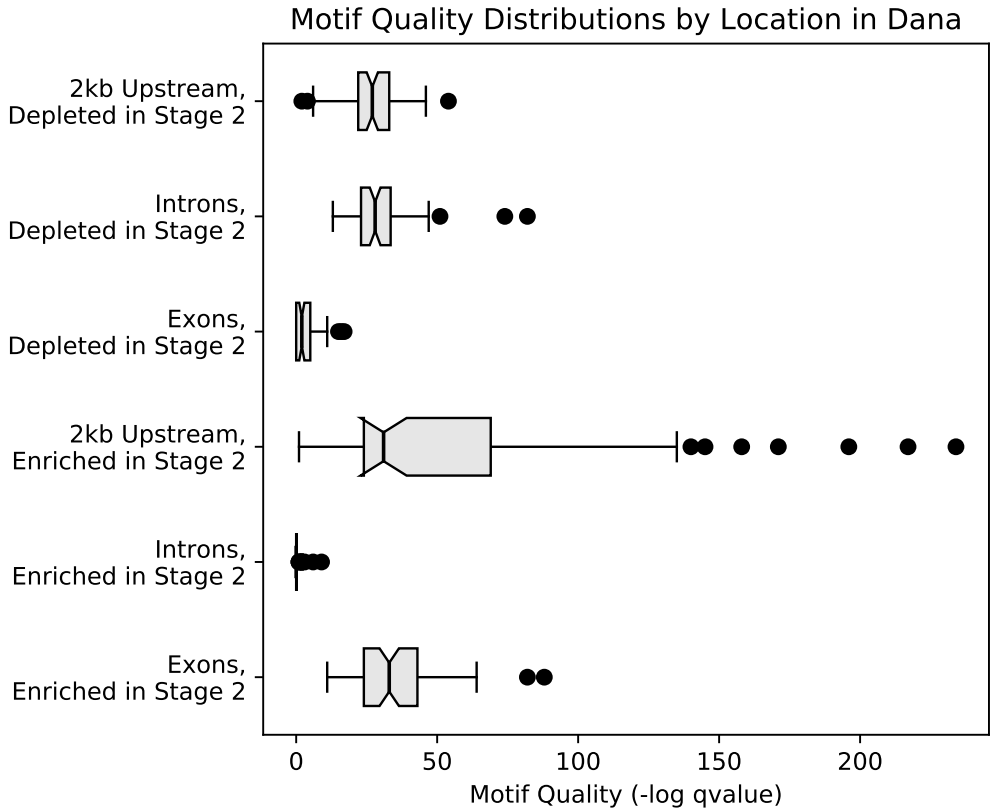
Fig 5

Supplementary

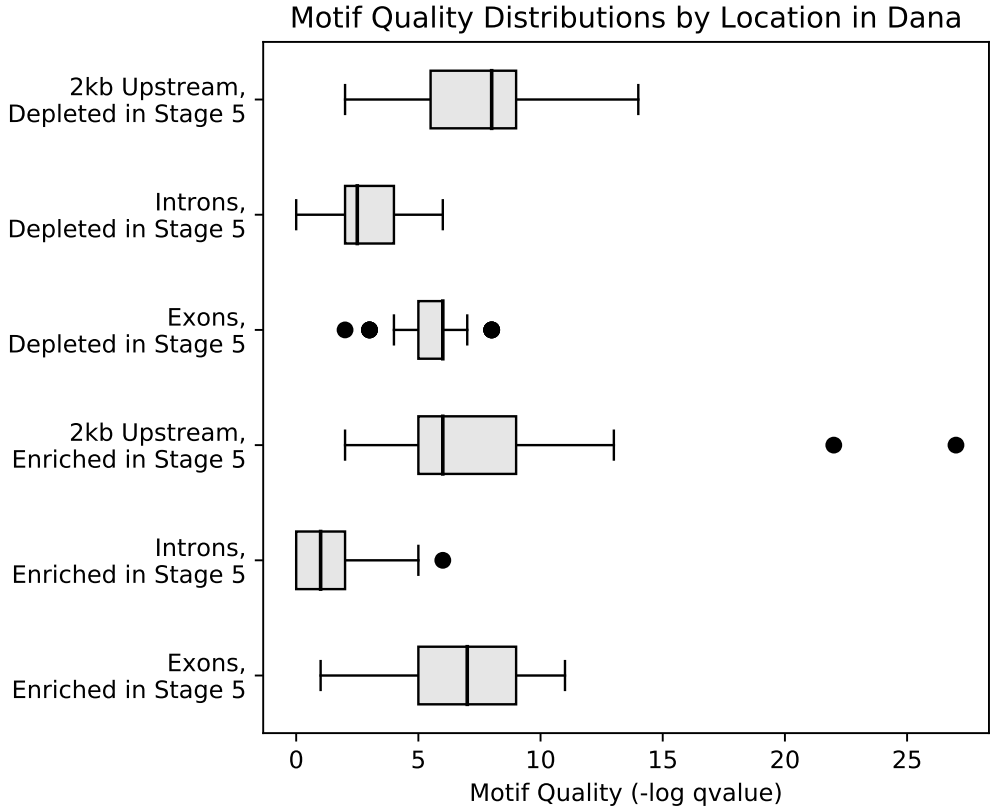
Location	Motif	Best Match	Tomtom q-value
Upstream	AHTATCGATA	Beaf-32	0.00809
Upstream	AHTATCGATA	Dref	0.03381
Upstream	BGYGYGYGCG	sr	0.01908
Upstream	BGYGYGYGCG	klu	0.04051
Upstream	CGCTCDCNCK	Trl	0.00634
Upstream	CGGCGCAGCG	Mad	0.08778
Upstream	CGTCTCTCTC	GAF	0.01615
Upstream	CGTCTCTCTC	Trl	0.01615
Upstream	GAGAGAGAGA	GAF	0.00001
Upstream	GAGAGAGAGA	Trl	0.00001
Upstream	KCGCTCGCTC	Trl	0.00059
Upstream	KCTACCTGYB	Zld	0.00002
Upstream	NNNNTACCTG	Zld	0.09762
Upstream	NNNNTACCTG	vvl	0.09762
Exon	TACATACATA	Top2	0.00277

S1 Table

A

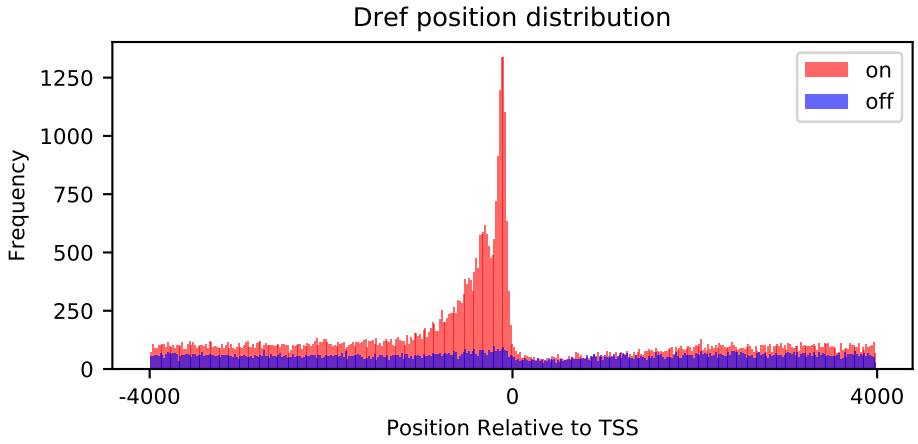

Motif Quality Distributions by Location in Dana
2kb Upstream, Depleted in Stage 2
Introns, Depleted in Stage 2
Exons, Depleted in Stage 2
2kb Upstream, Enriched in Stage 2
Introns, Enriched in Stage 2
Exons, Enriched in Stage 2
0
50
100
150
200
Motif Quality (-log qvalue)

B

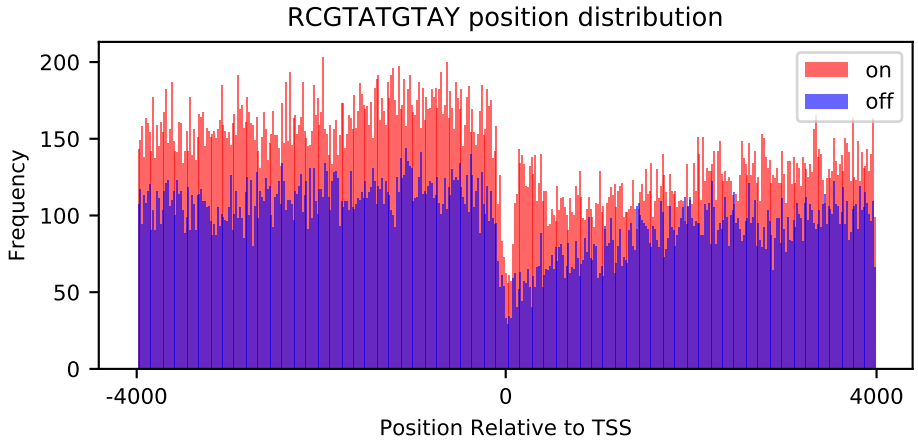

Motif Quality Distributions by Location in Dana
2kb Upstream, Depleted in Stage 5
Introns, Depleted in Stage 5
Exons, Depleted in Stage 5
2kb Upstream, Enriched in Stage 5
Introns, Enriched in Stage 5
Exons, Enriched in Stage 5
0
5
10
15
20
25
Motif Quality (-log qvalue)

S1 Fig

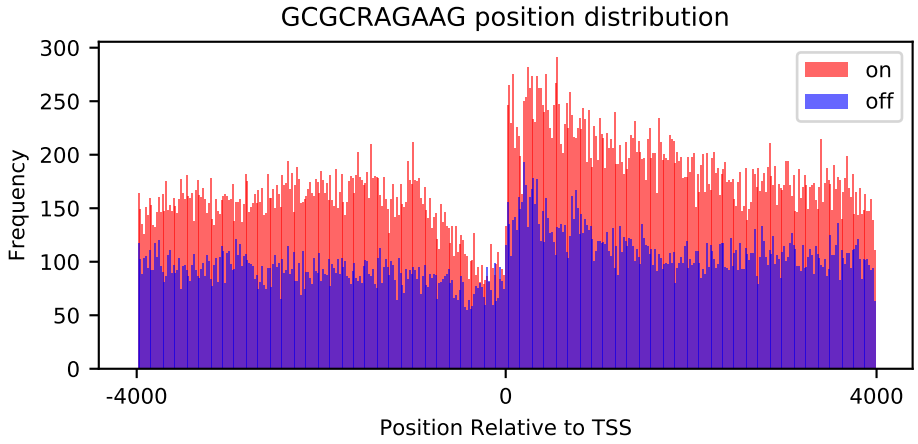
A


Dref position distribution
on
off
Frequency
1250
1000
750
500
250
0
-4000
0
4000
Position Relative to TSS

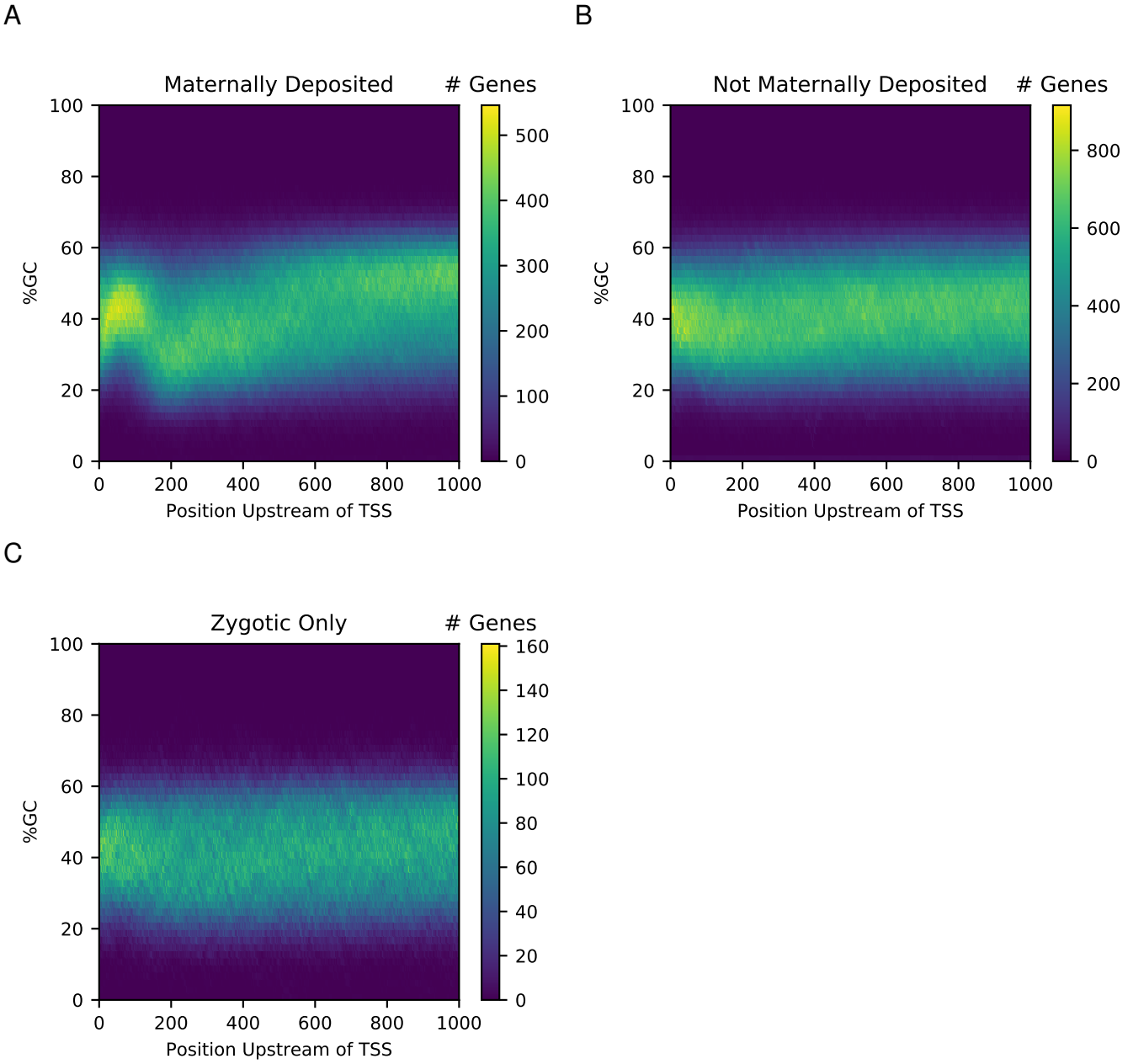
B


RCGTATGTAY position distribution
on
off
Frequency
200
150
100
50
0
-4000
0
4000
Position Relative to TSS

C


GCGCRAGAAG position distribution
on
off
Frequency
300
250
200
150
100
50
0
-4000
0
4000
Position Relative to TSS

S2 Fig


A
Maternally Deposited
Genes
%GC
Position Upstream of TSS
B
Not Maternally Deposited
Genes
%GC
Position Upstream of TSS
C
Zygotic Only
Genes
%GC
Position Upstream of TSS

S3 Fig