Transcription and splicing dynamics during early *Drosophila* development

Pedro Prudêncio [1, 2 §], Rosina Savisaar [1§], Kenny Rebelo [1§$], Rui Gonçalo Martinho[1, 2, 3*] and Maria Carmo-Fonseca[1*]

[1] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal.

[2] Algarve Biomedical Center Research Institute (ABC-RI), Universidade do Algarve, 8005-139 Faro, Portugal.

[3] Department of Medical Sciences and Institute for Biomedicine (iBiMED), Universidade de Aveiro, 3810-193 Aveiro, Portugal.

[§]These authors have contributed equally to the work.

[$]Present address: Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada

[*]Joint corresponding authors: carmo.fonseca@medicina.ulisboa.pt and rgmartinho@ua.pt

RUNNING TITLE: Transcription and splicing dynamics in *Drosophila*

KEYWORDS: splicing kinetics; Pol II pausing; transcription termination; NET-seq; *Drosophila melanogaster* embryo

ABSTRACT

Widespread co-transcriptional splicing has been demonstrated from yeast to human. However, the precise kinetics of splicing relative to transcription remains unclear, with discrepancies between studies that use different experimental approaches. Here, we adapted native elongating transcript sequencing technology (NET-seq) to measure co-transcriptional splicing dynamics during the early developmental stages of *Drosophila melanogaster* embryos. Our results reveal the position of RNA polymerase II (Pol II) when both canonical and recursive splicing occur. We found heterogeneity in splicing dynamics, with some RNAs spliced immediately after intron transcription, whereas for other transcripts no splicing was observed over the first 100 nucleotides of the downstream exon. Introns that show splicing completion before Pol II has reached the end of the downstream exon are necessarily intron-defined. We studied the splicing dynamics of both nascent pre-mRNAs transcribed in the early embryo, which have few and short introns, as well as pre-mRNAs transcribed later in embryonic development, which contain multiple long introns. As expected, we found a relationship between the proportion of spliced reads and intron size. However, intron definition was observed at all intron sizes. We further observed that transcription termination was very efficient for isolated genes but that the presence of an overlapping antisense gene was often associated with transcriptional readthrough. Taken together, our data unravels novel dynamic features of Pol II transcription and splicing in the developing *Drosophila* embryo.

52 **INTRODUCTION**

53 It is by now largely accepted that splicing can happen, and often does happen, whilst transcription is
54 still in progress (Ameur et al. 2011; Khodor et al. 2011; Beyer and Osheim 1988; Brugiolo et al. 2013;
55 Carrillo Oesterreich et al. 2010; Windhager et al. 2012; Nojima et al. 2015; Alpert et al. 2017). It is also
56 known that the processes of splicing and transcription are tightly interlinked. The RNA Polymerase II
57 (Pol II) elongation rate can affect exon inclusion (Fong et al. 2014; Maslon et al. 2019; Aslanzadeh et
58 al. 2018), and many of the proteins involved in splicing associate with the Pol II large subunit C-
59 terminal domain (CTD) (Hsin and Manley 2012; Emili et al. 2002; David et al. 2011; Görnemann et al.
60 2011; Morris and Greenleaf 2000; Nojima et al. 2018). Conversely, splicing may also affect
61 transcription, with evidence suggesting that Pol II slows down at exons (Alexander et al. 2010; Carrillo
62 Oesterreich et al. 2010; Jonkers et al. 2014; Mayer et al. 2015; Veloso et al. 2014), potentially to allow
63 for splicing to complete.

64 The exact timing of intron excision relative to Pol II elongation remains more controversial. A
65 nascent RNA sequencing study (Carrillo Oesterreich et al. 2016) showed that in *Saccharomyces*
66 *cerevisiae*, splicing could occur as soon as the 3' splice site had emerged from the transcription
67 machinery, suggesting that splicing may be completed immediately after intron transcription. Previous
68 attempts to determine the duration of splicing *in vivo* had returned estimates ranging from a few seconds
69 to several minutes (Alpert et al. 2017). The lowest of these estimates, such as the few-second estimate
70 reported in Martin et al. (2013), are consistent with splicing being completed right after the 3' splice
71 site is transcribed.

72 More recently, two long-read nascent RNA sequencing studies addressed the timing of splicing
73 in metazoa. One reported little immediate splicing in human and *Drosophila* cultured cell lines, with
74 most splicing occurring when Pol II had elongated thousands of nucleotides past the splice junction
75 (Drexler et al. 2020). Another found that splicing in mouse cells occurred immediately after
76 transcription, when Pol II was within 75-300 nucleotides of 3' splice sites (Reimer et al. 2021). These
77 discrepant results probably reflect biases introduced by the distinct methodological approaches used.
78 Indeed, one study (Drexler et al. 2020) relies on metabolic labelling of newly synthesized RNA with
79 4sU, while the other extracts RNA from chromatin (Reimer et al. 2021).

80 An alternative approach to analyse nascent pre-mRNA is Native Elongating Transcript
81 Sequencing (NET-seq), a technique that involves isolation of transcription complexes formed by Pol
82 II, the DNA template and the nascent RNA by immunoprecipitation, without crosslinking (Churchman
83 and Weissman 2011; Nojima et al. 2015). Solubilisation of Pol II complexes under native conditions is
84 typically incomplete in metazoan cells (Kimura et al. 1999), but this problem was circumvented by
85 extensive micrococcal nuclease (MNase) digestion of isolated native chromatin (Nojima et al. 2015).
86 Elongation complexes can then be immunoprecipitated using antibodies that specifically recognize the
87 different phosphorylation states of the Pol II CTD. The main advantage of this approach is that it
88 provides global, bidirectional maps of Pol II CTD phosphorylation specific nascent transcripts and
89 coupled RNA processing at single nucleotide resolution. Although initially applied to mammalian cells,
90 and thus termed mNET-seq (Nojima et al. 2015), the technology can be adjusted to any organism, as
91 recently illustrated in plants (Zhu et al. 2018). Here, we have adapted NET-seq to study the dynamic
92 properties of transcription and co-transcriptional splicing during early stages of development in
93 *Drosophila melanogaster* embryos. An advantage of this model system is that compared to human, the
94 *Drosophila* genome is more compact (Graveley et al. 2011) and thus the coverage of NET-seq reads on
95 intragenic regions is higher. Another advantage of the *Drosophila* model is that genes transcribed
96 initially in the embryo have few and short introns (like yeast genes), whereas genes transcribed later
97 contain multiple long introns (more similar to human genes). An additional feature of *Drosophila* is the
98 presence of many genes with exceptionally long introns that are subdivided by a non-canonical
99 mechanism termed recursive splicing (Duff et al. 2015; Joseph et al. 2018; Pai et al. 2018). For all these
100 reasons, *Drosophila* is an attractive model to study co-transcriptional splicing dynamics.
101 *Drosophila* early development is characterized by rapid mitotic cycles that lack cytokinesis,
102 resulting in nuclear proliferation in a syncytial cytoplasm (Campos-Ortega 1985; Laver et al. 2015).
103 During these initial mitotic divisions, which impose significant constraints on transcription and splicing
104 (Shermoen and O'Farrell 1991; Rothe et al. 1992; Guilgur et al. 2014; Martinho et al. 2015; Sandler et
105 al. 2018; Kwasnieski et al. 2019), the embryo largely relies on maternally deposited mRNAs.
106 Subsequently, the duration of the cell cycle is progressively expanded, membranes form between the
107 nuclei and segregate them into cells, and the zygotic transcriptome starts to be fully expressed (Campos-

108 Ortega 1985; Yuan et al. 2016; Blythe and Wieschaus 2015). By performing *d*NET-seq in the
109 developing *Drosophila* embryo, we have unravelled novel dynamic features of Pol II transcription and
110 splicing.
111

112 **RESULTS**

113

114 **Native elongating transcript sequencing in *Drosophila* embryos (*d*NET-seq)**

115 Our first task was to adapt the NET-seq technology for use in *Drosophila* embryos (henceforth referred
116 to as *d*NET-seq). We collected embryos at 2-3 hours after fertilization (referred to as early embryos)
117 and 4-6 hours after fertilization (referred to as late embryos) (**Fig 1A**). Analysis of embryos stained
118 with a fluorescent dye to visualize DNA (**Fig 1B**) revealed that early embryos were predominantly in
119 mitotic cycle 14 (stage 5), when membranes form between the nuclei, which are evenly distributed in a
120 monolayer under the egg surface, segregating them into separate cells (Campos-Ortega 1985; Laver et
121 al. 2015). The majority of late embryos were in the late stage of germ-band extension (stage 10; **Fig**
122 **1C**), when the embryo trunk (also known as the germ-band) elongates in the antero-posterior axis and
123 narrows in the dorso-ventral axis (Campos-Ortega 1985; Laver et al. 2015).

124 In adapting mNET-seq to *Drosophila* embryos, we optimized buffers and washing conditions
125 to purify the chromatin fraction from manually sorted embryos, solubilize the transcription complexes
126 with MNase digestion and immunoprecipitate Pol II with antibodies. We used rabbit polyclonal
127 antibodies raised against synthetic peptides of the YSPTSPS repeat of the CTD of the largest Pol II
128 subunit in *Saccharomyces cerevisiae*, phosphorylated at either S5 (ab5131 Abcam) or S2 (ab5095
129 Abcam). Both antibodies have been extensively used for chromatin immunoprecipitation experiments
130 in *Drosophila melanogaster* (Boija et al. 2017; Vizcaya-Molina et al. 2018; Akhtar et al. 2019; Dahlberg
131 et al. 2015; Lee et al. 2015; Yan et al. 2014; Arzate-Mejía et al. 2020).

132 To enable directional sequencing, the 5' hydroxyl (OH) generated by MNase digestion of RNA
133 was first converted to a 5' phosphate by T4 polynucleotide kinase (**Fig 1D**). RNA was then purified
134 from the immunoprecipitated Pol II complexes and size-selected using an RNA purification kit
135 procedure that combines a unique buffer system with a column technology (see the methods section for
136 more detail). RNAs with a size above 60 nucleotides (nt) were used for subsequent ligation of specific
137 adapters to the 5' P and 3' OH ends of each RNA fragment followed by PCR-based preparation of a
138 cDNA library for high-throughput Illumina sequencing (**Fig 1D**). After sequencing, adapter sequences
139 were trimmed and paired-end reads with sequence overlaps were merged into a single read that spans

140 the full length of the original RNA fragment (dark orange; **Fig 1E**). The resulting single reads were
141 aligned to the *Drosophila* reference genome. The nucleotide at the 3' end of each RNA fragment was
142 identified and its genomic position recorded (asterisk; **Fig 1E**).

143 Two to three NET-seq libraries were independently prepared from early and late embryos using
144 S5P antibody; three additional libraries were prepared from late embryos using S2P antibody. Each
145 library was sequenced to a high coverage with a read length of 150 bp (Supplemental **Fig S1A**; see the
146 methods section for more detail). Experimental reproducibility was demonstrated by strong agreement
147 of uniquely aligned read density between biological replicates prepared with antibodies raised against
148 the CTD phosphorylated on either serine 5 (S5P) or serine 2 (S2P) positions (Supplemental **Fig S1 B-**
149 **E**). This suggests that both S5P and S2P antibodies recognize the CTD of elongating Pol II in
150 *Drosophila* embryos.
151

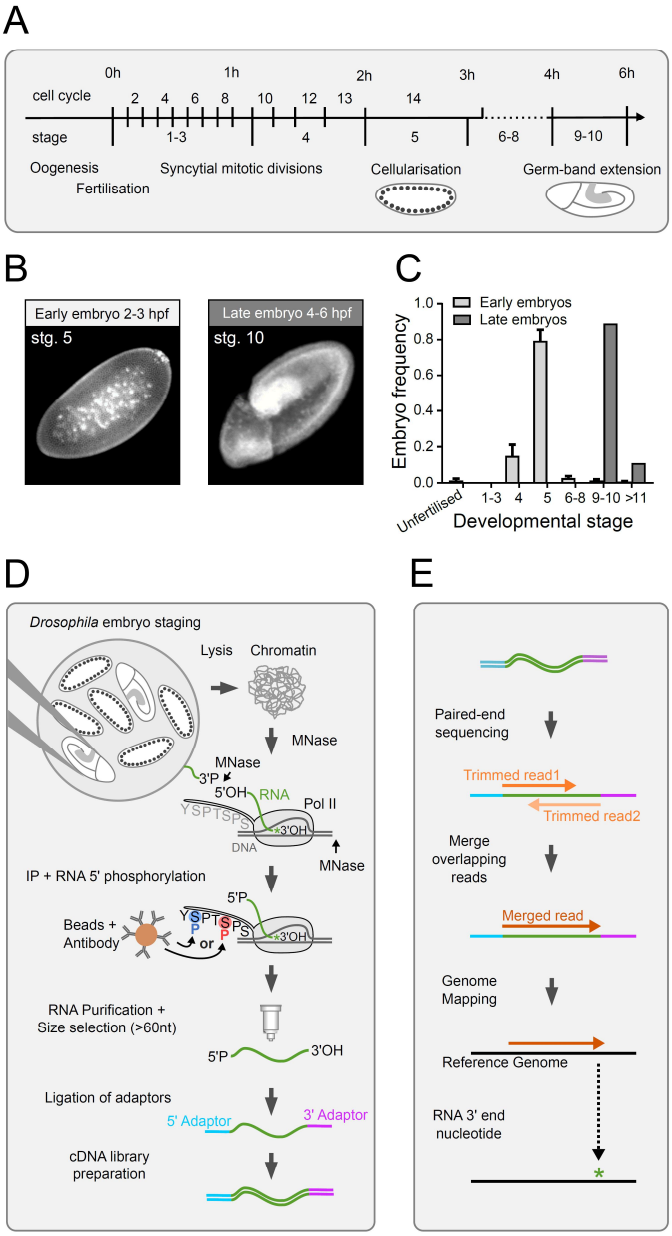


Figure 1. Native elongating transcript sequencing in *Drosophila* embryos.
(A) Timeline of *Drosophila* early embryonic development. **(B)** Representative images (stained for DNA) of embryos in mitotic cycle 14 (stage 5) and late germ-band expansion (stage 10). **(C)** The graph depicts the developmental stage of embryos sorted into the "early" and "late" groups. **(D)** Outline of the *d*NET-seq experimental protocol. **(E)** Outline of *d*NET-seq data analysis.

_d_NET-seq captures splicing intermediates and spliceosomal snRNAs.

NET-seq captures not only the final (3' OH end) nucleotide of nascent RNA but also the 3' OH end of RNAs that associate with the Pol II elongation complex (Nojima et al. 2015, 2018; Schlackow et al.

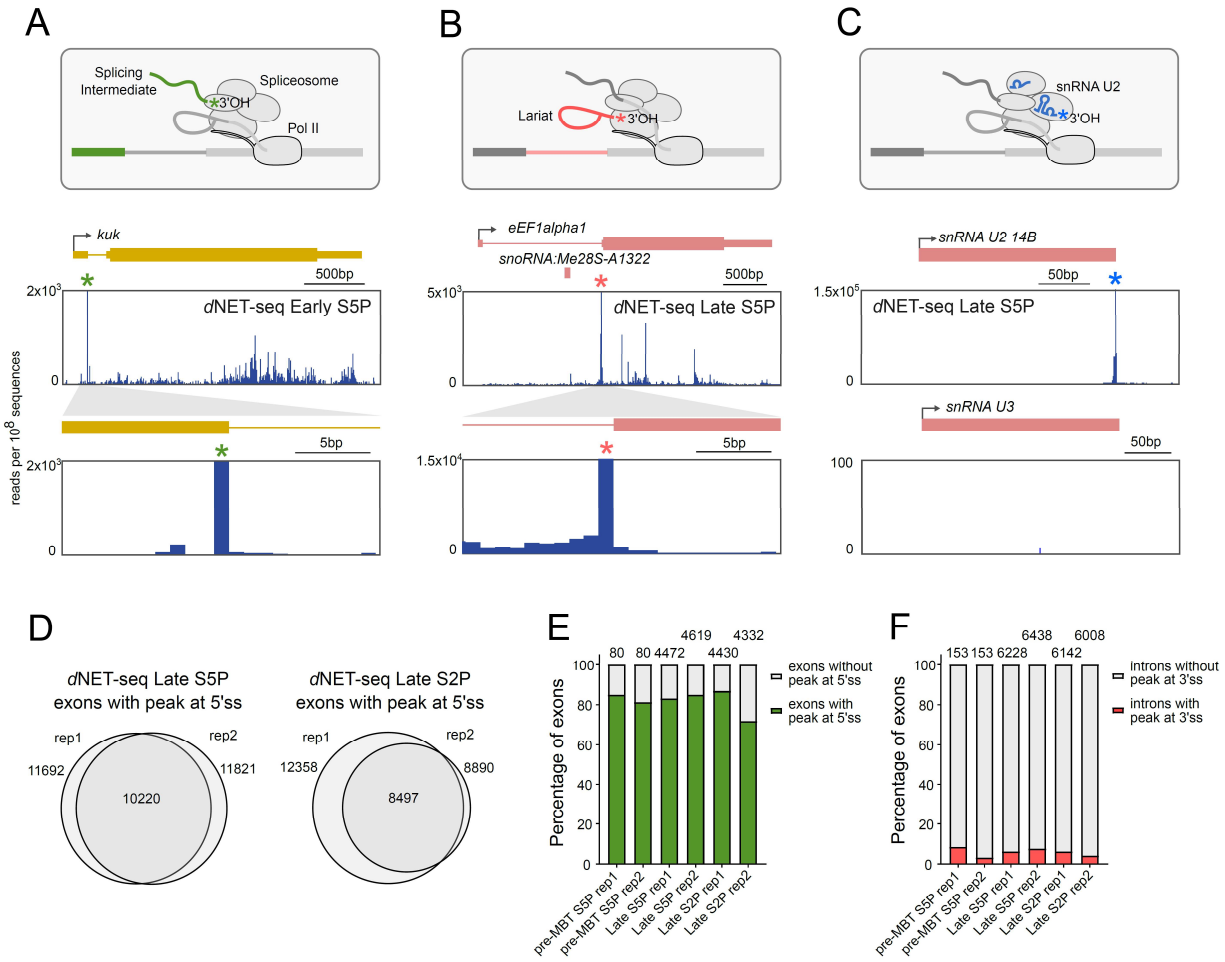
164 2017). Notably, in humans, mNET-seq of Pol II phosphorylated on CTD serine 5 detected splicing
165 intermediates formed by cleavage at the 5' splice site after the first splicing reaction. The presence of
166 such intermediates manifests as an enrichment of reads whose 3' ends map precisely to the last
167 nucleotide of an exon. In both early and late *Drosophila* embryos, we indeed observed large peaks of
168 *d*NET-seq/S5P reads mapping to the last nucleotide of spliced exons, as shown for the *kuk* gene
169 (asterisk; **Fig 2A**). We also detected *d*NET-seq/S5P peaks at the last nucleotide of introns, as shown for
170 the *eEF1alpha1* gene (asterisk; **Fig 2B**); enrichment for these reads results from co-
171 immunoprecipitation of released intron lariats after completion of the splicing reaction (**Fig 2B**). In
172 addition, we observed reads corresponding to mature snRNAs engaged in co-transcriptional
173 spliceosome assembly, suggesting that *d*NET-seq was capturing the free 3' OH ends of the snRNAs.
174 (**Fig 2C**). We found prominent peaks at the end of spliceosomal U1, U2, U4 and U5 snRNAs. As
175 expected, no peak was detected mapping to the end of the U3 snRNA, which is involved in the
176 processing of pre-rRNA synthesized by Pol I. Noteworthy, we observed an accumulation of *d*NET-seq
177 signal at the end of U6 snRNA (Supplemental **Fig S2A**), contrasting with a lack of peak observed in
178 mammalian cells (Nojima et al. 2018). This is consistent with the finding that most mammalian U6
179 snRNAs contain a 2',3'-cyclic phosphate terminal group at the 3' end, whereas U6 3' ends in
180 *Drosophila* cells consist of either a cyclic 2',3'-phosphate, a 3'-phosphate or a 2',3'-hydroxyl group
181 (Lund and Dahlberg 1992).

182 To quantify how many constitutively spliced exons have *d*NET-seq peaks at the end, we applied
183 an algorithm that finds nucleotides where the NET-seq read density is at least three standard deviations
184 above the transcript mean in a local region defined by 100 bp upstream and downstream (Churchman
185 and Weissman 2011; Prudêncio et al. 2020). Upon analysing replicates of *d*NET-seq/S5P and *d*NET-
186 seq/S2P libraries, we identified over 10,000 exons showing splicing intermediate peaks (**Fig 2D**). As
187 peaks were more frequently detected on exons of genes with higher read density (Supplemental **Fig**
188 **S2B**), we classified the exons into four groups (quartiles) based on the *d*NET-seq read density of the
189 corresponding gene and restricted the analysis to exons in the fourth quartile, i.e., from genes with the
190 highest read density. The results show that splicing intermediate peaks are detected in approximately
191 80% of all constitutively spliced exons. The proportion is similar for pre-MBT genes, i.e, zygotic genes

192 that start to be transcribed before cycle 14 and remain active thereafter (Chen et al. 2013), and for genes
193 expressed in late embryos (**Fig 2E**). Similar proportions were found in the S5P and the S2P datasets
194 (**Fig 2E**). We then used the same methodology and the same set of genes to detect peaks at the last
195 intronic nucleotide, corresponding to released intron lariats. Such peaks were detected in less than 10%
196 of introns (**Fig 2F**).

197 Taken together, these results demonstrate that *d*NET-seq is capable of detecting splicing
198 intermediates and spliceosomal snRNAs in *Drosophila* embryos, as previously reported in mammalian
199 cells (Nojima et al. 2015; Schlackow et al. 2017; Nojima et al. 2018).

200



201

202

203 **Figure 2. *d*NET-seq captures splicing intermediates and spliceosomal snRNAs. (A-C)** *d*NET-
204 seq/S5P profiles over the indicated genes in late embryos. The diagrams outline the 3' OH ends
205 generated by co-transcriptional cleavage at the 5' splice site (green, **A**), the 3' splice site (red, **B**) and
206 the free 3' OH end of spliceosomal snRNAs (blue, **C**). Arrows indicate the direction of transcription.

207 Exons are represented by boxes. Thinner boxes represent UTRs. Introns are represented by lines
208 connecting the exons. Asterisks denote the peak at the end of the exon (**A**), the end of the intron (**B**),
209 and the end of the U2 snRNA gene (**C**). (**D**) Comparison (Venn diagrams) of exons with a splicing
210 intermediate peak detected in biological replicates of *d*NET-seq/S5P and *d*NET-seq/S2P libraries. (**E,**
211 **F**) Frequency of peaks corresponding to splicing intermediates (**E**) and released intron lariats (**F**) in pre-
212 MBT genes and genes expressed in late embryos. Only genes with the highest read density (fourth
213 quartile) were considered.
214
215

216 ***d*NET-seq specifically captures nascent RNA.**

217 To validate that we were detecting nascent transcription in *Drosophila* embryos, we analysed maternal
218 mRNAs that are transcribed during oogenesis and loaded into the egg (**Fig 3A**). As expected, maternal
219 transcripts such as *bicoid* (**Fig 3B**), *nanos* (Supplemental **Fig S3A**), *gurken* (Supplemental **Fig S3B**)
220 and *Rab32* (Supplemental **Fig S3C**) were detected by RNA-seq in embryos collected 2-3 hours after
221 fertilization, but *d*NET-seq signal over these genes was negligible. However, robust *d*NET-seq signal
222 was found at the *pumilio* gene (Supplemental **Fig S3D**). Expression of this gene was considered to be
223 exclusively maternal based on RNA-seq (Lott et al. 2011) and Pol II ChIP-seq data (Chen et al. 2013),
224 but a GRO-seq study detected *pumilio* nascent transcripts in embryos collected 2-2.5 hours post-
225 fertilisation (Saunders et al. 2013). The detection of *pumilio* RNA in early embryos by GRO-seq and
226 *d*NET-seq highlights the sensitivity of these two techniques in capturing low-level nascent transcripts.
227 The stage at which the embryo switches from relying on maternally deposited mRNAs and
228 proteins to undergoing its own transcription is termed the mid-blastula transition (MBT). However, a
229 few genes (known as pre-MBT genes) become transcriptionally active before MBT (Kwasnieski et al.
230 2019). Strong *d*NET-seq signal was detected in early embryos over the bodies of pre-MBT genes such
231 as *snail* (**Fig 3C**), *fushi tarazu* (**Supplemental Fig S3E**) and *odd skipped* (**Supplemental Fig S3F**).
232 Altogether, we examined 117 previously identified pre-MBT genes and 35 maternal mRNAs (**Fig 3D**
233 **and Supplemental Fig S3G**).
234 A characteristic feature of pre-MBT genes is that they are short in length with few if any introns
235 (De Renzis et al. 2007; Ali-Murthy et al. 2013; Kwasnieski et al. 2019; Heyn et al. 2014). In contrast,
236 most genes expressed in late embryos contain multiple introns, as shown for *Akap200* (**Fig 3E**), *His3.3A*
237 (**Supplemental Fig S3H**) and *twinstar* (*tsr*) (**Supplemental Fig S3I**). Analysis of *d*NET-seq datasets

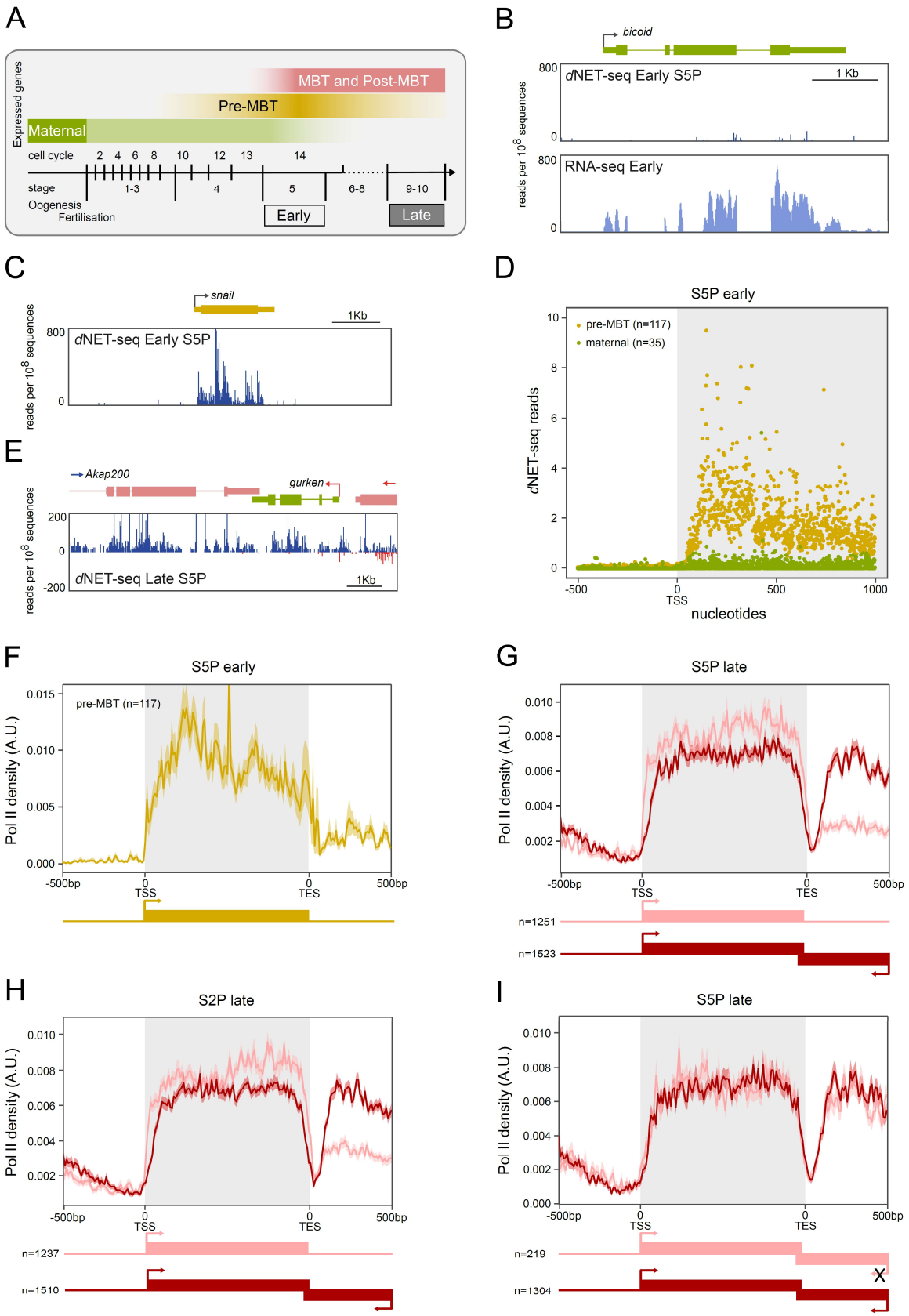
238 from late embryos revealed widespread signal on introns and regions downstream of the
239 polyadenylation site, further confirming that the technique is capturing nascent transcripts.
240 To conclude, robust *d*NET-seq signal was recovered from zygotically but not from maternally
241 expressed transcripts. In addition, we detected reads mapping to introns and downstream of the poly-
242 adenylation site. These findings suggest that *d*NET-seq is indeed specifically targeting the nascent
243 transcriptome.
244
245 ***d*NET-seq reveals transcriptional readthrough associated with the presence of overlapping**
246 **antisense genes.**
247 Having confirmed that *d*NET-seq was capturing nascent RNA, we next investigated the distribution of
248 Pol II density over transcript regions in early and late embryos. *d*NET-seq profiles (**Fig 3C-E and**
249 **Supplemental Fig S3**) do not show the characteristic higher read density near the promoter, as
250 previously described in mammalian cells (Nojima et al. 2015; Mayer et al. 2015). This is most likely
251 because Pol II typically pauses ~30-60 bp downstream of the transcription start site (TSS) (Kwak et al.
252 2013) and in our *d*NET-seq approach we enrich for RNAs longer than 60 nt (**Supplemental Fig S1F**);
253 thus, we only record the position of polymerases that have transcribed at least 60 bp past the TSS.
254 We then turned our attention to the *d*NET-seq signal around polyadenylation sites (also known
255 as the transcription end sites or TES). Many genes, such as *Akap200* (**Fig 3E**), *His3.3A* (**Supplemental**
256 **Fig S3H**) and *tsr* (**Supplemental Fig S3I**), have considerable levels of *d*NET-seq/S5P signal after the
257 TES, indicative that in these cases, Pol II continues to transcribe after cleavage and polyadenylation.
258 We noted that the 3' UTR of these three genes is overlapped by another gene (*gurken, Nepl3,* and *IntS1,*
259 respectively), which is transcribed in the opposite direction. We therefore asked whether transcriptional
260 readthrough is related to the presence of an overlapping convergent gene. In order to identify all the
261 genes that are transcriptionally active in late embryos, we used a strategy adapted from GRO-seq
262 analysis (Core et al. 2008) that relies on read density in gene desert regions as background reference for
263 absence of transcription. Very large intergenic regions (gene deserts) were divided into 50kb windows,
264 and read densities were calculated by dividing read counts in each window by the window length in bp
265 (**Supplemental Fig S4A**). Genes with *d*NET-seq signal over the gene body (in RPKM) above the 90th

266 percentile of read density for all intergenic regions analysed were considered to be transcriptionally
267 active (**Supplemental Fig S4B**). We identified approximately 7 thousand active genes, and similar
268 results were obtained from *d*NET-seq/S5P and *d*NET-seq/S2P datasets (**Supplemental Fig S4C**). This
269 set of active genes includes over 85% of the 3500 genes previously identified as actively transcribed
270 after the mid-blastula transition (MBT) based on ChIP-seq experiments (Chen et al. 2013). Next, we
271 divided the genes transcribed in late embryos into two groups, depending on whether their 3' UTR was
272 or was not overlapped by another convergent gene. The metagene analysis shows that when averaged
273 across all transcribed genes, the *d*NET-seq/S5P signal is abruptly reduced at the TES (**Fig 3G**). This is
274 expected, as when Pol II reaches this site, the nascent transcript is cleaved and polyadenylated and
275 therefore there is no RNA to be sequenced. The presence of nascent transcripts after the TES indicates
276 that Pol II fails to terminate transcription after cleavage and polyadenylation. Unexpectedly, this is
277 mainly observed on genes that have the 3' UTR overlapped by an antisense gene (**Fig 3G**). Similar
278 results were observed after metagene analysis of *d*NET-seq/S2P datasets (**Fig 3H**), confirming an
279 association between transcriptional readthrough and genomic architecture at the TES. We then asked
280 whether transcriptional readthrough depends on the transcriptional activity of the overlapping antisense
281 (convergent) gene. Using the methodology described above to identify transcribed genes, we found that
282 the vast majority (>80%) of overlapping convergent genes were transcriptionally active and only 219
283 genes were silent. The metagene analysis shown in **Fig 3I** clearly indicates that transcription of the
284 convergent overlapping gene is not required for inefficient termination at the overlapped TES.

285 The vast majority of pre-MBT genes do not show *d*NET-seq signal past the TES (**Fig 3C, 3F**
286 **and Supplemental Fig S4D, E**). Most likely, in these genes Pol II releases both the nascent RNA and
287 the template DNA within a few nucleotides from the polyadenylation site. Notably, for the majority of
288 pre-MBT genes devoid of *d*NET-seq signal past the TES (65), there is no other gene on either strand
289 within a region of 500 bp downstream of the TES (referred to henceforth as 'isolated genes'), as shown
290 for *tailless (tll)* (**Supplemental Fig S4D**). Another 26 pre-MBT genes are embedded in larger genes, as
291 shown for *nullo*, which is located within a long intron of the *CG12541* gene (**Supplemental Fig S4E**).
292 A smaller group of pre-MBT genes (21) have neighbouring genes located on either strand within a
293 region of 500 bp downstream of the TES, as shown for *Elba2* (**Supplemental Fig S4F**). We further

294 identified 5 pre-MBT genes that have the 3' UTR overlapped by an antisense convergent gene, as shown
295 for *spook (spo)* (**Supplemental Fig S4G**). Transcriptional readthrough was observed in these pre-MBT
296 genes (**Supplemental Fig S4G**). Importantly, the proportion of isolated genes is significantly higher
297 among pre-MBT genes (65/117 or ~56%) than among late genes (1251/7233 or ~17% of all
298 transcriptionally active genes for replicate 1 S5P) ($P < 2.2 * 10^{-16}$, one-tailed binomial test). This
299 suggests that pre-MBT genes may be under selection to avoid overlaps with other genes, thereby
300 minimizing transcriptional readthrough. Note, however, that we cannot exclude the possibility that in
301 some cases, a lack of NET-seq signal past the TES could be due to loss of CTD phosphorylation at this
302 position.

303 In conclusion, *d*NET-seq reveals a strong correlation between inefficient transcription
304 termination at the cleavage and polyadenylation site, and presence of an antisense convergent gene
305 overlapping the TES. The vast majority of genes that start to be transcribed before the mid-blastula
306 transition, when the mitotic cycles are still very short, do not have the TES overlapped by other genes
307 and lack *d*NET-seq signal past the TES. In contrast, many genes that start to be transcribed after the
308 mid-blastula transition, when the mitotic cycles become much longer, show transcriptional readthrough
309 and are overlapped by an antisense convergent gene.


A
Expressed genes
MBT and Post-MBT
Pre-MBT
Maternal
cell cycle
2 4 6 8 10 12 13 14
stage
1-3 4 5 6-8 9-10
Oogenesis
Fertilisation
Early
Late
B
bicoid
dNET-seq Early S5P
1 Kb
RNA-seq Early
reads per 10^8 sequences
C
snail
1Kb
dNET-seq Early S5P
D
S5P early
pre-MBT (n=117)
maternal (n=35)
dNET-seq reads
TSS
nucleotides
E
Akap200
gurken
dNET-seq Late S5P
1Kb
F
S5P early
pre-MBT (n=117)
Pol II density (A.U.)
-500bp
TSS
TES
500bp
G
S5P late
n=1251
n=1523
H
S2P late
n=1237
n=1510
I
S5P late
n=219
n=1304
X

310

311

312

313

314 **Figure 3. Transcription over gene bodies and beyond the TES.**
315 **(A)** The diagram illustrates the temporal expression of maternal, pre-MBT, MBT, and post-MBT genes
316 during *Drosophila* embryonic development. *d*NET-seq/S5P and RNA-seq profiles over the maternal
317 gene *bicoid* (**B**), the pre-MBT gene *snail* (**C**), and the post-MBT gene *Akap200* (**E**). Reads that aligned
318 to the positive strand are in blue, and reads that aligned to the negative strand are in red. (**D**) Meta-
319 analysis of mean *d*NET-seq read density around the transcription start site (TSS) in maternal and pre-
320 MBT genes (replicate 1). **(F-I)** Normalized metagene analysis in arbitrary units (A.U.). The *d*NET-seq
321 signal is depicted along the normalized gene length (grey background), as well as 500 bp upstream of
322 the transcription start site (TSS) and 500 bp downstream of the transcription end site (TES). **(F)** *d*NET-
323 seq/S5P signal on pre-MBT genes in early embryos. **(G)** *d*NET-seq/S5P signal on transcriptionally
324 active genes in late embryos; the signal over genes that have the 3' UTR overlapped by an antisense
325 gene is depicted in dark red, while the signal over genes with no other genes within 500 bp is depicted
326 in light red. **(H)** *d*NET-seq/S2P signal on transcriptionally active genes in late embryos. **(I)** *d*NET-
327 seq/S5P signal on transcriptionally active genes in late embryos; the signal over genes that have the 3'
328 UTR overlapped by a transcriptionally active antisense gene is depicted in dark red, while the signal
329 over genes with the 3' UTR overlapped by a transcriptionally inactive antisense gene is depicted in light
330 red.
331

332

333 **Analysis of *d*NET-seq read density profiles.**

334 We next focused on the distribution of *d*NET-seq reads on exons and introns. The number of nascent
335 RNA reads whose 3' ends map at a particular genomic position is proportional to the number of Pol II
336 molecules at that position. Thus, Pol II pause sites can be detected as local peaks in read density
337 (Churchman and Weissman 2011; Larson et al. 2014). To search for *d*NET-seq peaks representative of
338 paused Pol II, we excluded signal resulting from splicing intermediates (i.e., reads that map to the very
339 last nucleotide of introns and exons were discarded). We analysed both S5P and S2P data. However,
340 we were particularly interested in exploring the multiple peaks of dNET-seq/S5P signal observed on
341 both exons and introns (**Fig 4A**), because co-transcriptional splicing has been linked to S5P in humans
342 (Nojima et al. 2015).

343 A difficulty when performing the systematic identification of *d*NET-seq peaks is that transcripts
344 with higher initiation rates will contain more reads and thus peaks are more likely to be detected than
345 in more lowly transcribed genes. To control for this confound, we developed a peak calling algorithm
346 that detects regions where the local read density is significantly higher than expected by chance, given
347 the over-all read density of the transcript (**Fig 4A;** see also the methods section). We emphasize that for
348 detecting reads corresponding to splicing intermediates (**Fig 2D-F)**, we used a peak calling method that

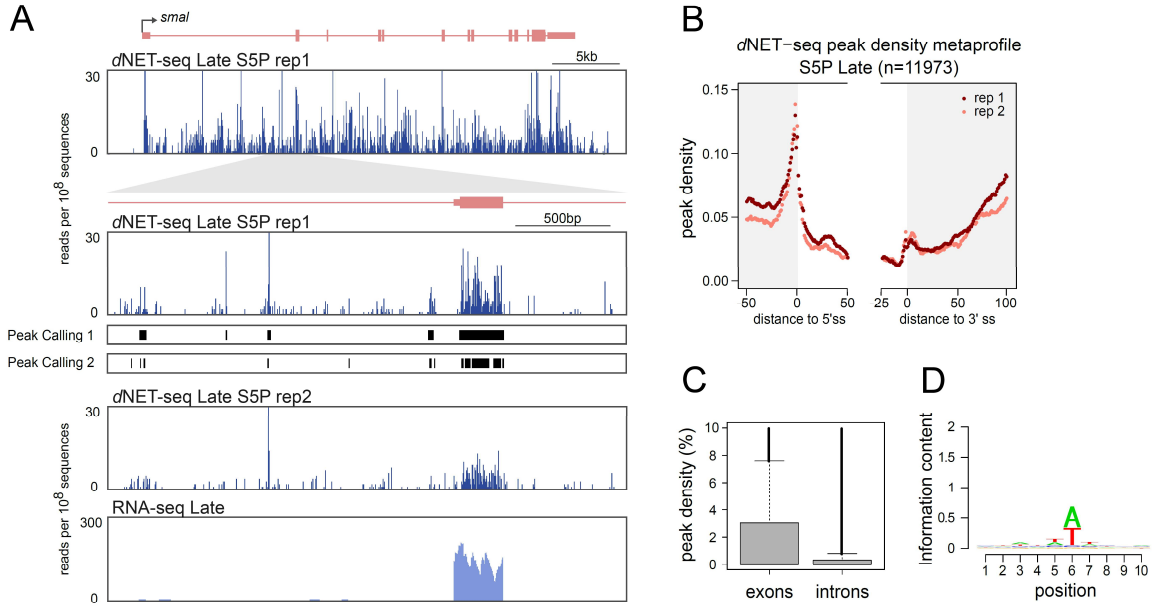
349 looks for significant single-nucleotide positions (Churchman and Weissman 2011), whereas this new
350 method detects higher read regions of variable length.

351 We then aligned exons and introns on the splice sites and calculated the average peak density
352 at each position. Exons have a higher over-all peak density than introns (mean proportion of nucleotides
353 in peaks ~0.022/~0.015 for replicate 1/replicate 2 introns and ~0.043/~0.034 for replicate 1/replicate 2
354 exons; $P < 2.2 * 10^{-16}$ for both replicates; two-tailed Mann-Whitney *U*-test with the peak densities of
355 individual introns/exons as data points), suggesting that the elongation rate is decreased over exons (**Fig**
356 **4C**). This is consistent with previous reports in mammalian (Mayer et al. 2015; Jonkers et al. 2014) and
357 *Drosophila* cells (Kwak et al. 2013). In addition, peak density is sharply increased around the 5' splice
358 site (**Fig 4B**). This could indicate Pol II pausing associated with splice site recognition. However, we
359 cannot exclude that misaligned splicing intermediate reads are contributing to the observed increase in
360 peak density around the 5' splice site.

361 The profile around the 3' splice site shows a higher peak density region just after the intron-
362 exon boundary (**Fig 4B**). A progressive increase in average exonic peak density is further observed
363 starting roughly 60 nt after the 3' splice site (**Fig 4B**), suggesting slower elongation or more frequent
364 Pol II pausing in this region. Alternatively, such increases in peak density could correspond to regions
365 where the Pol II CTD more frequently exhibits the phosphoisoform targeted by the antibodies.
366 However, note that *d*NETseq/Ser5 and *d*NETseq/Ser2 lead to very similar peak density metaprofiles
367 (**Fig 4B**; **Supplemental Fig S7I-J**).

368 A further potential caveat is that nucleotide composition varies systematically across exons and
369 introns. For example, exons tend to have a higher GC content than introns (Zhu et al. 2009). This could
370 be problematic as NET-seq relies on MNase digestion of DNA and RNA to solubilize chromatin.
371 MNase digestion of DNA is known to be sequence-biased, with most notably a preference for cleaving
372 just 5' of an adenine (Dingwall et al. 1981; Hörz and Altenburger 1981; Gaffney et al. 2012). An
373 analysis of the 5' ends of our reads revealed similar biases for MNase digestion of RNA (**Fig 4D**). This
374 sequence preference could lead to artefactual variation in read density, with more reads being sampled
375 from transcripts and transcript regions whose nucleotide composition is more similar to MNase
376 digestion biases. To verify to what extent our results were affected by this confound, we performed a

377 simulation to determine the expected distribution of reads based on the digestion bias alone
378 (Supplementary Methods). We concluded that MNase biases are unlikely to explain either the
379 enrichment of peaks in exons or the general profile of peak densities past the 3' splice site.



380

381

382 **Figure 4. Analysis of *d*NET-seq read density profiles.**
383 (**A**) *d*NET-seq/S5P and RNA-seq profiles over the post-MBT gene *smoke alarm* (*smal*). The read
384 number is depicted at two magnification levels in two biological replicates. For replicate 1, the line
385 "Peak Caller 1" shows peaks called using the "large peaks" setting, which is appropriate for detecting
386 larger regions of putative Pol II pausing. The line "Peak Caller 2" shows peaks called using the "small
387 peaks" setting, which provides higher spatial resolution and has been used for subsequent analyses.
388 RNA-seq data for the same regions is also shown. (**B**) Metagene analysis of peak density estimated
389 from *d*NET-seq/S5P data from late embryos (replicate 1, N=12799). To calculate peak density for each
390 position, we divided the number of introns that overlap with a peak at that position by the total number
391 of introns. The last 50 nucleotides of exons, the first 50 nucleotides of introns, the last 25 nucleotides
392 of introns, and the first 100 nucleotides of exons are shown. Only internal and fully coding exons from
393 transcriptionally active genes that are at least 100 nucleotides long are shown. Exons shorter than 150
394 nucleotides contribute to both the exon end and start. Only introns that were at least 50 nt long were
395 considered. **(C)** Peak density in the exons and introns of transcriptionally active genes (*d*NET-seq/S5P,
396 replicate 1). Peak density has been defined as the percentage of nucleotides within a given exon or
397 intron that overlap with a significant peak. **(D)** Sequence logo of nucleotide frequencies within a 10-nt
398 window around the 5' ends of NET-seq reads. The combined height of the bases at each position is
399 proportional to the information content. Position 6 corresponds to the 5'-most nucleotide of the read.
400 Putative internal priming reads, as well as reads mapping to the last nucleotide of exons or introns
401 (possible splice intermediate and intron lariat reads) were ignored.
402
403

404 ***d*NET-seq captures recursive splicing intermediates.**

405 Having shown that the spliceosome forms a complex with the elongating Pol II in *Drosophila* embryos,
406 we asked when splicing takes place relative to transcription. We first looked at recursive splicing of
407 long introns because this process involves the formation of inherently unstable intermediates that are
408 more likely to be formed soon after the transcription of each intronic splice site (Pai et al. 2018). In
409 recursive splicing, long introns are removed by sequential excision of adjacent sections involving
410 separate splicing reactions, each producing a distinct lariat (Hatton et al. 1998). Recursively spliced
411 intron segments are bounded at one or both ends by recursive sites or ratchet points (Burnette et al.
412 2005), which correspond to zero nucleotide exons consisting of juxtaposed 3' and 5' splice sites around
413 a central AG|GT motif, where the vertical line represents the splice junction (**Fig 5A**).

414 To capture recursive splicing intermediates using *d*NET-seq, it is essential to have a good
415 coverage of reads corresponding to nascent transcripts and spanning the splice junctions. The total
416 number of reads resulting from nascent RNA in each *d*NET-seq dataset is depicted in **Supplemental**
417 **Fig S1G**. By merging the sequencing information of overlapped paired-end reads (**Fig 1E**), we were
418 able to sequence on average ~103 nucleotides per nascent RNA (**Supplemental Fig S1F and S1H**).
419 Focusing on previously identified *Drosophila* ratchet points (Joseph et al. 2018; Duff et al. 2015), we
420 found *d*NET-seq/S5P reads that span the junction between the canonical 5' splice site at the end of the
421 exon and the first ratchet point (RP1) internal to the downstream intron, as shown for the second intron
422 of the *Megalin* gene (**Fig 5B**). Reads spanning the subsequent intronic RPs were also observed (**Fig**
423 **5B**). Overall, we detected *d*NET-seq/S5P and *d*NET-seq/S2P spliced reads supporting most of the
424 previously identified recursive splicing events (**Fig 5C and Supplemental Fig S5A**).

425 Analysis of *d*NET-seq profiles around a RP reveals an enrichment of reads in a region located
426 a few nucleotides downstream of the RP, as shown for RP2 in the first intron of the *Tenascin major*
427 gene (**Fig 5D**). Noteworthy, most of these reads are already spliced to the previous RP (**Fig 5D**). A
428 meta-analysis of *d*NET-seq/S5P reads around 137 RPs confirms that many spliced reads can be
429 observed just downstream of RPs **(Fig 5E, F)**. Taken together, these results indicate that recursive
430 splicing can occur soon after the transcription of intronic recursive sites and suggest the presence of
431 paused Pol II downstream of these sites.

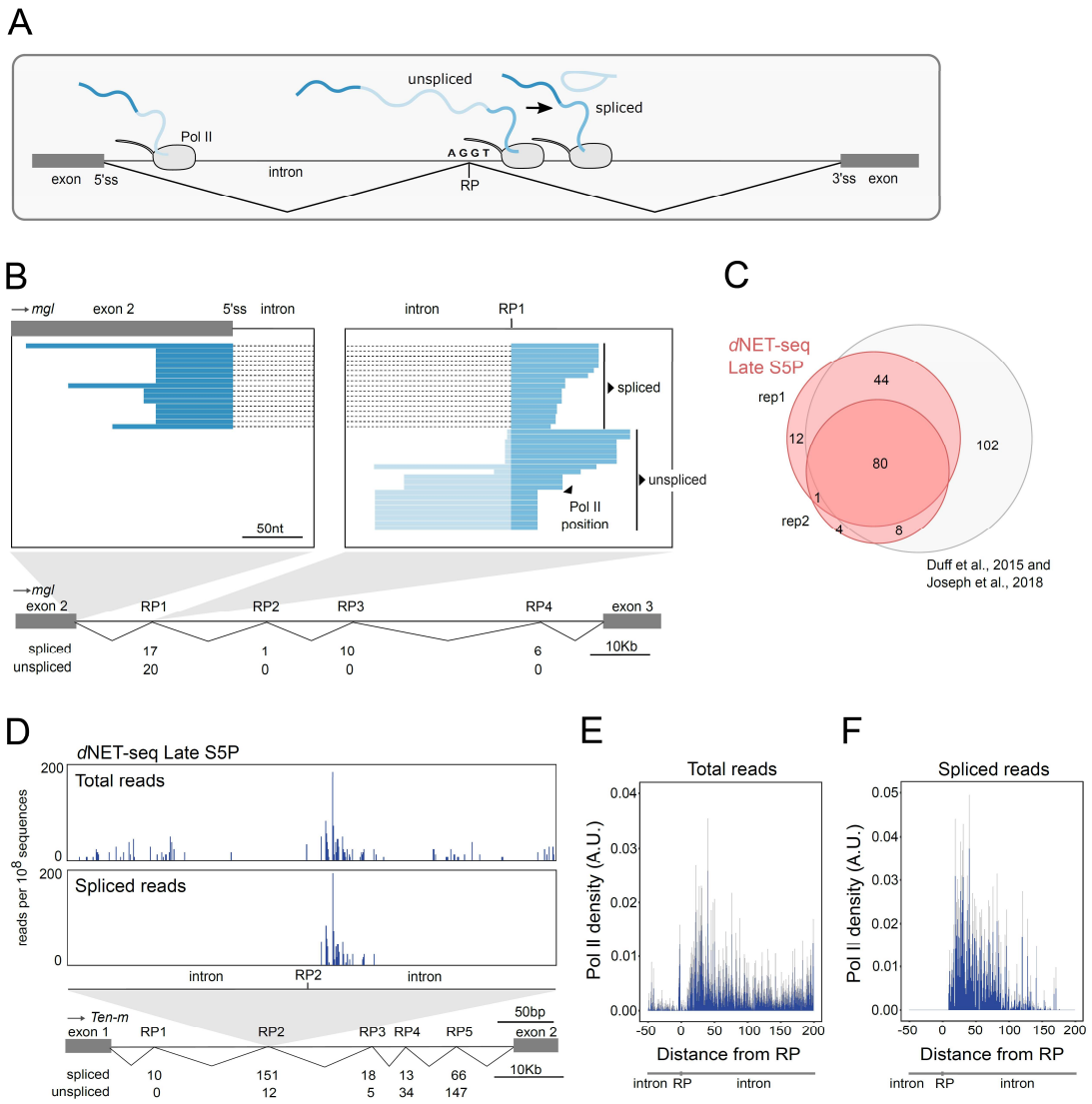


Figure 5. *d*NET-seq captures recursive splicing intermediates.
(A) Schematic illustrating recursive splicing. A ratchet point (RP) with juxtaposed acceptor and donor splice site motifs is indicated. **(B)** Visualization of *d*NET-seq/S5P reads that align to the second intron of the *Megalin* (*mgl*) gene. Recursively spliced reads align to exon 2 (green) and the intron after RP1 (dark red). The number of spliced and unspliced reads at each RP in the intron is indicated. **(C)** Venn diagram comparing RPs identified in two *d*NET-seq/S5P biological replicates and in previously reported studies (Joseph et al. 2018; Duff et al. 2015). **(D)** Number of *d*NET-seq/S5P reads that have the 3' end mapped around RP2 in the first intron of *Tenascin major* (*Ten-m*) gene. The top panel depicts all reads, and the bottom panel depicts only reads that have been spliced to RP1. **(E-F)** Meta-analysis with single nucleotide resolution of normalized *d*NET-seq/S5P reads around RPs (n=137) using all reads **(E)** or only reads spliced to the previous RP or exon **(F).**

449 **Splicing takes place as Pol II transcribes past the 3' splice site.**

450 Having established that *d*NET-seq captures recursive splicing, we then asked whether *d*NET-seq reads
451 spanning canonical exon-exon junctions were also detected **(Fig 6A)**. To identify splicing events, we
452 considered all internal and fully-coding exons that are at least 100 nt long in actively transcribed genes
453 in early and late embryos. For each splice junction, we counted how many reads had the 3' end mapped
454 to the first 100 nt of the exon. Only exons with at least 10 reads mapping to this region were considered
455 (see the methods section for justification of the threshold). Then, the *d*NET-seq splicing ratio (SR) was
456 calculated by dividing the number of spliced reads by the sum of the number of spliced and unspliced
457 reads **(Fig 6A).** Reads could be counted as spliced or unspliced if their 3' end mapped to within the first
458 100 nt of the exon and their 5' end reached upstream of the 3' splice site, allowing to check whether
459 the intron was still present. A robust agreement of estimated SR values was observed between biological
460 replicates for pre-MBT genes (**Supplemental Fig S6A**) and genes expressed in late embryos (**Fig 6B**).

461 SR values showed a bimodal distribution, with peaks at both extremes (SR = 0 and SR = 1; **Fig**
462 **6C** and **Supplemental Fig S6B**), indicating that a subset of junctions were always spliced immediately
463 after transcription (SR=1), while others remained unspliced (SR=0). Notably, differences in SR
464 distribution were found between *d*NET-seq/S5P and *d*NET-seq/S2P datasets (**Fig 6C, D and**
465 **Supplemental Fig S6B, C**). Significantly lower SR values were detected for S2P compared to S5P (P
466 $< 2.2*10^{-16}$ from binomial regression, see Fig 6 legend for statistical details). In particular, junctions
467 that were most frequently spliced (SR values close to 1) were predominantly captured by *d*NET-seq/S5P
468 (**Fig 6C and Supplemental Fig S6B**). This clearly points to a preferential association between co-
469 transcriptional splicing and S5 phosphorylation of Pol II CTD, as previously proposed in human cells
470 (Nojima et al. 2015, 2018). We therefore continued to focus solely on the S5P datasets in the remainder
471 of our splicing analysis.

472 We observed that only ~5% of splice junctions in pre-MBT genes were devoid of S5P reads
473 spanning ligated exons and thus presented an SR of 0 (~5.10% replicate 1/~4.35% replicate 2), whereas
474 in genes expressed in late embryos this proportion was ~20% (~19.72% replicate 1/~18.55% replicate
475 2) (**Fig 6E**; two-tailed binomial test for difference between late and pre-MBT, $P \sim 6.141 * 10^{-5}/9.44 *$
476 10^{-4} (replicate 1/replicate 2)). However, this difference between pre-MBT genes and genes expressed

477 in late embryos disappeared once the higher read density of pre-MBT genes had been controlled for
478 (**Fig 6E**).

479 Thus, in most cases, *Drosophila* co-transcriptional splicing can occur when Pol II is still
480 transcribing the downstream exon, implying an intron definition mechanism as previously proposed for
481 *S. cerevisiae* (Carrillo Oesterreich et al. 2016). Consistent with this view, many *Drosophila* transcripts
482 have relatively long exons separated by short introns – a gene architecture suggested to be conducive
483 to intron definition. Very long introns flanked by short exons have instead been associated with exon
484 definition (under which the downstream exon needs to be fully transcribed before splicing can take
485 place) (Keren et al. 2010). A switch to exon definition was proposed once the size of the intron surpasses
486 ~200 nt (Fox-Walsh et al. 2005). It is unclear, however, whether the choice between exon and intron
487 definition is dependent on the absolute sizes of exons and introns, or rather the ratio of intron to exon
488 size.

489 We found no relationship between the *d*NET-seq splicing ratio and exon length (**Supplemental**
490 **Fig S6D, H** for replicates 1 and 2; see **Supplementary Table 1** for details on statistical significance
491 and sample sizes for all of the gene architecture parameters discussed here). Regarding intron size, we
492 found that introns with SR = 0 (and thus no evidence for intron definition) were, on average, indeed
493 larger than other introns (**Fig 6F** and **Supplemental Fig S6G**). However, more careful examination
494 revealed a more complex picture, with a lower proportion of introns with SR = 0 both for introns of
495 intermediate size (~55-100 nt, which corresponds to ~55% of the introns studied) and for very large
496 introns (>1000 nt) (**Fig 6F**). These intron sizes may thus be optimal for fast splicing. We also uncovered
497 a lower exon to intron length ratio for introns with SR = 0 than for others (**Supplemental Fig S6E, I**;
498 note that the difference was significant for replicate 2 and near-significant for replicate 1).

499 Taken together, these results suggest that although fast splicing (implying intron definition) is
500 indeed skewed towards small introns flanked by large exons, there is also frequent and efficient intron
501 definition for large introns. Our results are inconsistent with a threshold model, where splicing would
502 systematically switch to exon definition after a given intron size is reached.

503 We also investigated the relationship between SR and several other gene architecture
504 parameters. Firstly, as the GC content in exons and introns decreases, the proportion of introns with SR

505 = 1 increases and the proportion with SR = 0 decreases, showing more efficient immediate splicing (**Fig**
506 **6G, H** and **Supplemental Fig S6K, L**). A similar effect is observed as the ratio of the downstream exon
507 GC content to intron GC content increases (**Fig 6I** and **Supplemental Fig S6M**). Thus, the most
508 efficient immediate splicing is observed when the GC content is low in both exons and introns but
509 higher in exons than introns. Secondly, as 3' splice site strength increases, the proportion of introns
510 with SR = 0 decreases (**Fig 6J** and **Supplemental Fig S6N**). We found no significant relationship
511 between SR and 5' splice site strength (**Supplemental Fig S6F, J**). However, we replicated previous
512 observations (Khodor et al. 2012) that transcripts with only a single intron tend to be spliced less
513 efficiently than multi-intron ones, with a higher proportion of SR = 0 introns (**Supplemental Fig S6O,**
514 **P**). Thirdly, similarly to previous reports (Khodor et al. 2011, 2012; Herzel et al. 2018), we uncovered
515 an effect of exon rank, whereby exons that are more central appear to be spliced more efficiently (**Fig**
516 **6K** and **Supplemental Fig S6Q**). Finally, in light of recent results in mouse linking transcriptional
517 readthrough and inefficient splicing (Reimer et al. 2021), we asked whether such a relationship could
518 be detected in our data. We found no significant difference in levels of readthrough between genes with
519 high or low mean SR (**Supplemental Fig S6R, S**).
520 In conclusion, *d*NET-seq reveals that splicing can occur immediately as Pol II transcribes past
521 the 3' splice site, yet many nascent transcripts remain unspliced. As expected, we found a relationship
522 between the proportion of spliced reads and intron size. However, immediate splicing was observed at
523 all intron sizes.

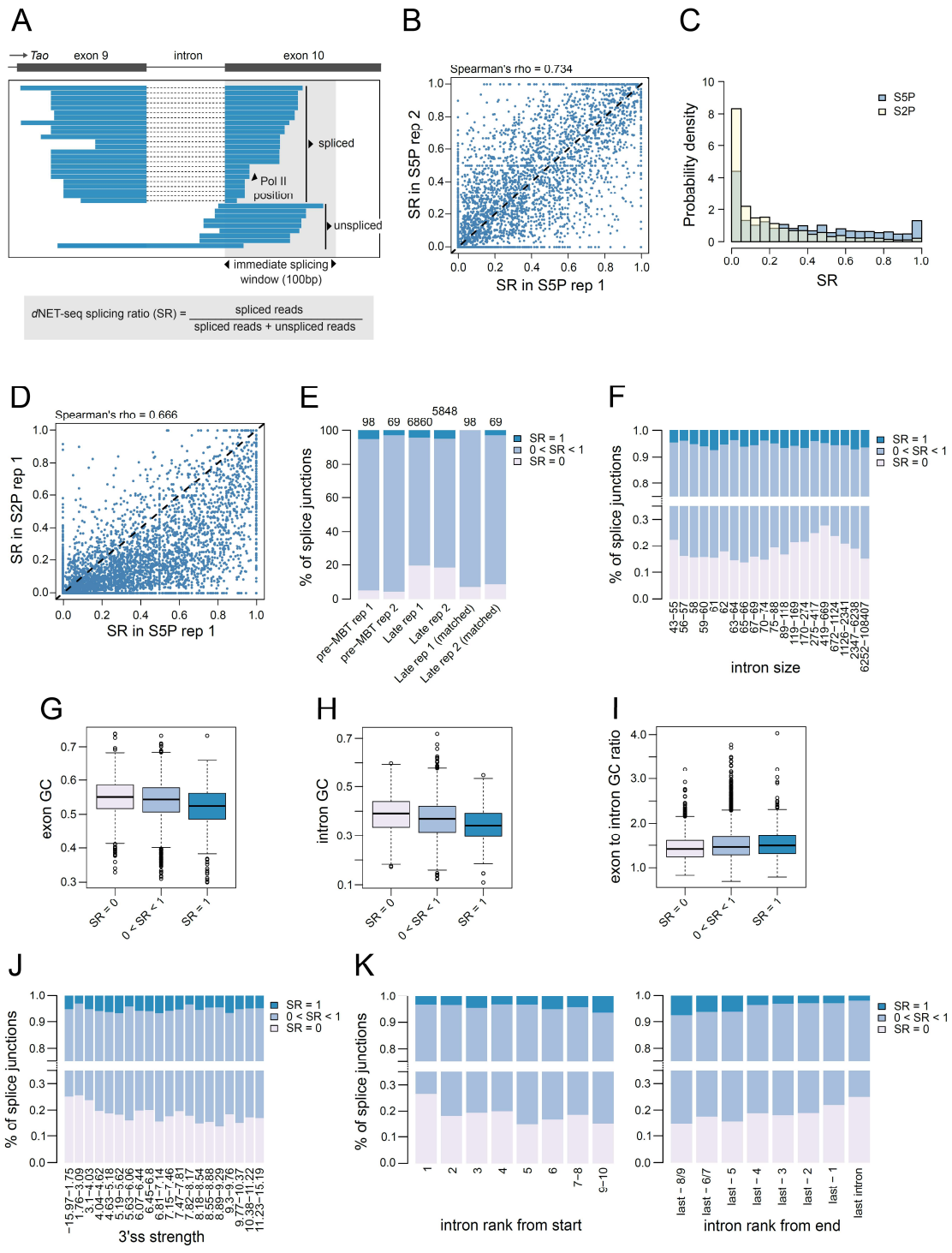


Figure 6. *d*NET-seq reveals immediate splicing at all intron sizes.
(A) Visualization of *d*NET-seq/S5P reads that align to exon 10 of the *Tao* gene. For **(B-J)**, unless otherwise specified, only introns from transcriptionally active genes where the downstream exon is a fully coding internal exon at least 100 nt long were included. In addition, enough spliced/unspliced reads had to end within the first 100 exonic nucleotides that obtaining a splicing ratio (SR) of 0 or 1 by chance alone was highly unlikely (see the methods section for details). For genes expressed in late embryos, this threshold was 10 reads for both replicates. For pre-MBT genes, it was 14 for replicate 1 and 9 for replicate 2. For S2P data, we used a threshold of 10 to enable better comparison with S5P. **(B)** SR values estimated in two biological replicates of *d*NET-seq/S5P datasets from late embryos

535 (Spearman correlation, ρ = ~0.734, $P < 2.2 * 10^{-16}$; N = 3708). **(C)** Histogram of SR values for *d*NET-
536 seq/S5P (N = 5626) and *d*NET-seq/S2P (N = 6888). To test the significance of the difference between
537 S5P and S2P, a binomial regression with a logit link was performed without filtering by read number
538 (N = 12833 for S5P; N = 13229 for S2P). The number of spliced and unspliced reads was specified as
539 the dependent variable and the status of each data point as S5P or S2P was the sole predictor. The model
540 predicted a splicing ratio of ~0.369/~0.394 for S5P replicate 1/2 and of 0.162/0.269 for S2P replicate
541 1/2. (**D**) SR values estimated in replicate 1 of *d*NET-seq/S5P and *d*NET-seq/S2P datasets from late
542 embryos (Spearman correlation, ρ = ~0.666, *P* < 2.2 * 10-16; N = 4773). **(E)** Proportion of splice
543 junctions in pre-MBT genes and genes expressed in late embryos classified according to their SR values.
544 As many pre-MBT genes are single-intron, last introns were exceptionally included in this analysis. To
545 make the two columns on the right, we used a subset of the genes expressed in late embryos (post-MBT
546 genes) that was as similar as possible to the pre-MBT set in read number. Concretely, to match each
547 pre-MBT gene, we picked the post-MBT gene that had the most similar total count of spliced and
548 unspliced reads, making sure that every post-MBT gene only appeared in the subset once. **(F-K)** Several
549 parameters of gene architecture show a relationship with SR. For the sample sizes and statistical tests
550 used, see **Supplementary Table 1**. Note that in **(F, J-K)**, the bin ranges have been set so that intron
551 numbers would be as equal as possible between bins.
552

553

554 **Immediate splicing associates with specific read density profiles.**

555 Analysis of individual *d*NET-seq profiles around constitutive splice junctions with a high splicing ratio
556 frequently revealed an enrichment of reads downstream of the 3' splice site, coincident with the
557 appearance of spliced reads, as shown for the *cno* gene (**Fig 7A**). In contrast, profiles around a
558 constitutive splice junction with SR = 0 had an accumulation of reads further along the exon, as shown
559 for the *ND-51* gene (**Fig 7B**). A clearly distinct type of profile was observed on skipped exons, on which
560 very few reads were observed, as shown for the *zip* gene (**Fig 7C**).

561 We next performed a meta-analysis of *d*NET-seq peak densities over different exonic and
562 intronic regions for introns with differing SR values. A difficulty of analysing the peak density profile
563 around the 3' splice site is that reads mapping to the final nucleotide of the intron may represent intron
564 lariats and thus be non-nascent. We removed reads mapping to this position prior to peak calling.
565 However, through misalignment, intron lariat reads may also map to the few nucleotides around the 3'
566 splice site and thus still affect the final meta-profile. In order to minimize the impact of such
567 misalignment events, we excluded from the meta-analysis all introns with a read mapping to the final
568 intronic nucleotide, as this is expected to also discard the introns most likely to contain misaligned
569 intron lariat reads (**Fig 7D, Supplemental Fig S7A**). When such filtering is not done, a sharp peak is
570 observed around the 3' splice site, notably when SR values are low (**Supplemental Fig S7 B-C**). It

571 appears that intron lariat reads are primarily captured for introns with low splicing ratios (**Fig 7E and**
572 **Supplemental Fig S7D**). This contrasts with splicing intermediate reads from the end of the upstream
573 exon, which are associated to higher splicing ratios instead (**Fig 7E and Supplemental Fig S7D**).

574 The peak density profile downstream of the 3' splice site contains two primary regions of
575 interest (highlighted in **Fig 7D and Supplemental Fig S7 A-C, G-H**). The first is a region of increased
576 peak density located roughly 10-20 nucleotides after the 3' splice site. This peak is independent of the
577 presence of putative intron lariat reads (**Fig 7D, Supplemental Fig S7 A-C**), and is thus likely to
578 correspond to a genuine region of Pol II pausing or slow-down. However, it could also represent a
579 region of preferential Ser5 phosphorylation, with no concomitant pausing. Noteworthy, pausing in this
580 region has also been observed in studies that did not target specific Pol II phosphoisoforms (e.g. (Kwak
581 et al. 2013)). Interestingly, this peak is more prominent when the SR is higher (**Fig 7D, Supplemental**
582 **Fig S7 A-C, G-H**), suggesting an association between immediate co-transcriptional splicing and
583 increased Pol II pausing at the start of the exon.

584 The second feature of interest is a rapid increase in peak density ~60 nucleotides into the exon.
585 Contrary to the first region of interest, the peak density in this region is higher when SR is *lower* (**Fig**
586 **7D, Supplemental Fig S7 A-C, G-H**). We found no evidence that this increase in peak density could
587 be a result of MNase digestion biases (**Supplementary Methods**). A further hypothesis is that this
588 pattern could be reflecting a gradual slowdown of Pol II as it moves from an AT-rich intronic region to
589 a GC-rich exonic one. The peak is indeed more prominent in introns where the ratio of exonic to intronic
590 GC content is higher (**Supplemental Fig S7 E-F**). However, when SR=1, then even exons where this
591 ratio is high do not display the peak (**Supplemental Fig S7 E-F**). Moreover, introns with higher SR
592 tend to have *higher* exonic to intronic GC ratios (**Fig 6I, Supplemental Fig S6M**). Hence, the co-
593 variation with SR cannot be explained simply through the correlation between SR and GC content
594 patterns. This suggests that there is an effect of the SR beyond any GC content effect. Potentially, this
595 region could correspond to where delayed splicing starts (because these events have low immediate
596 splicing ratios), and could represent splicing-associated Pol II pausing.

597 Note that we obtained qualitatively similar meta-profiles when considering read rather than
598 peak densities (**Supplemental Fig S7 G-H**). Our results therefore cannot be explained through any

599 biases introduced by the peak calling approach. The meta-profile obtained using *d*NET-seq/S2P data
600 was also qualitatively similar, although we did not divide the introns by SR given the rarity of reads
601 with high SR in S2P data (**Supplemental Fig S7 I-J**). We also attempted a separate meta-analysis of
602 skipped exons. These exons showed no accumulation of Pol II. However, the sample size was too small
603 for meaningful comparisons with included exons.

604 Finally, we asked how prevalent co-transcriptional splicing is in the developing *Drosophila*
605 embryo. As shown in **Fig 2E**, approximately 80% of exons in pre-MBT genes and genes expressed in
606 late embryos covered by a high density of NET-seq reads have a peak corresponding to the splicing
607 intermediate formed after cleavage at the 5' splice site but before exon-exon ligation. The majority of
608 these exons (~85%) were covered by *d*NET-seq reads that span the junction to the downstream exon
609 either directly on nascent transcripts or indirectly on splicing intermediates formed by cleavage at the
610 5' splice site of the downstream exon **(Fig 7F, H)**, confirming that they are co-transcriptionally spliced.
611 We then focused on those exons for which no splicing intermediate spike was detected by the peak
612 calling algorithm, as illustrated for the *Doc3* gene **(Fig 7G, I)**. Over 71% of these exons were also
613 covered by *d*NET-seq reads that span the junction to the downstream exon either directly on nascent
614 transcripts or indirectly on splicing intermediates formed by cleavage at the 5' splice site of the
615 downstream exon **(Fig 7I)**, arguing that even exons without a detectable splicing intermediate peak are
616 co-transcriptionally spliced. Altogether, our *d*NET-seq results support co-transcriptional splicing for
617 over 95% of the analysed exons.

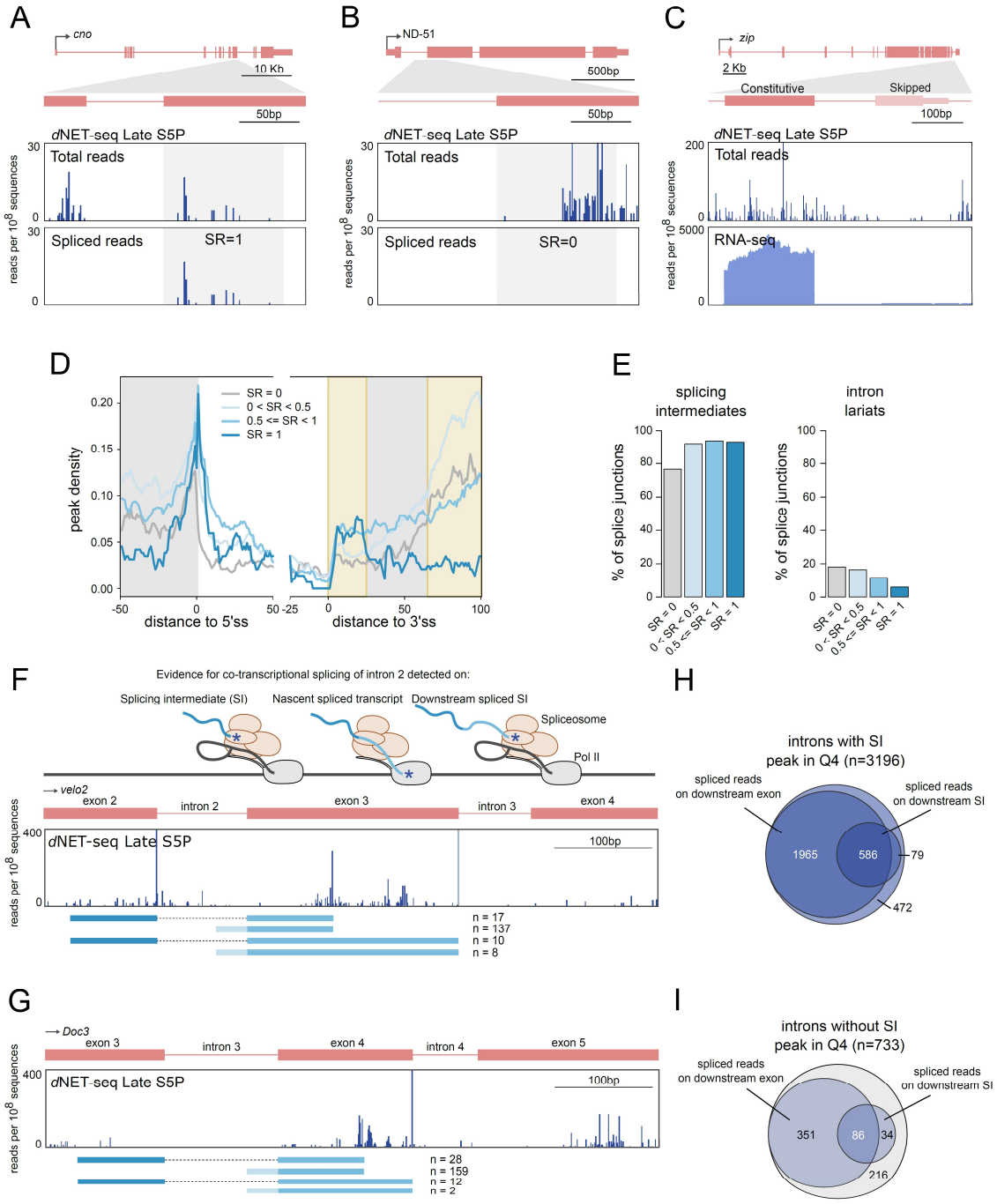


Figure 7. Immediate splicing associates with higher density of *d*NET-seq signal.
(A-C) *d*NET-seq/S5P profiles surrounding the indicated exons in the post-MBT genes *cno* (**A**), *ND-51* (**B**), and *zip* (**C**). The top panels depict all reads. The bottom panels depict either the 3' end coordinate of reads that span the splice junction **(A, B)**, or the RNA-seq profile (**C**). **(D)** Metagene analysis of peak density estimated from *d*NET-seq/S5P datasets from late embryos (replicate 1) for different ranges of SR values. Peak density has been calculated as the proportion of introns that overlap with a peak at any given position. The last 50 nucleotides of exons, the first 50 nucleotides of introns, the last 25 nucleotides of introns, and the first 100 nucleotides of exons are shown. Only internal and fully coding exons from transcriptionally active genes that are at least 100 nucleotides long are shown (N = 4783). In addition, at least 10 spliced/unspliced reads had to end within the first 100 nucleotides of the exon. (**E**) The proportion of introns with at least one read whose 3' end maps to the final position of the

upstream exon (putative splicing intermediates) or to the final position of the intron (putative intron lariats) in *d*NET-seq/S5P late replicate 1. (**F, G**) *d*NET-seq/S5P profiles on the indicated regions of the *velo2* and *Doc3* genes. Below, spliced reads are depicted. Asterisks denote the splicing intermediate peaks at the end of exons. **(H, I)** Venn diagrams showing how many junctions with or without a splicing intermediate peak are covered by spliced reads or have a downstream splicing intermediate covered by spliced reads.

DISCUSSION

In this study, we used *d*NET-seq to map Pol II over the bodies of genes that become transcriptionally active in *Drosophila* embryos during the initial stages of development. The use of embryos allowed us to perform an important test to verify that the captured RNA is truly nascent. Indeed, early *Drosophila* embryos contain abundant mRNAs that are transcribed during oogenesis and loaded into the egg. In our analysis of early embryos, these maternal transcripts were readily detected by RNA-seq, but not by *d*NET-seq.

Transcription in the *Drosophila* early embryo begins in mitotic cycle 8 for a few genes (Erickson and Cline 1993; Pritchard and Schubiger 1996) and then the number of active genes gradually increases until cycle 14 (Kwasnieski et al. 2019). Notably, the initial mitotic cycles have a duration of approximately 10 minutes, with cycle 13 taking about 21 minutes and cycle 14 lasting for at least 65 minutes (Ji et al. 2004). Thus, assuming an elongation rate of 2.4-3.0 kb per minute (Fukaya et al. 2017), there are significant constraints on the transcription of genes that are active before the end of cycle 13 (pre-MBT genes). Although pre-MBT genes are on average shorter than genes expressed at later stages of development (Artieri and Fraser 2014; Hoskins et al. 2011), longer genes are nevertheless transcribed before MBT. However, before cycle 14, transcriptional elongation of long genes is either prematurely terminated (Sandler et al. 2018) or aborted (Shermoen and O'Farrell 1991; Kwasnieski et al. 2019). It has also been reported that pre-MBT expression is associated with the generation of DNA damage due to stalling of DNA replication at transcriptionally engaged loci (Blythe and Wieschaus 2015).

Our *d*NET-seq analysis revealed a novel feature of pre-MBT genes that most likely contributes to reduce conflicts between transcription and replication during the short interphases of early embryos. We found that in the majority of pre-MBT genes, no *d*NET-seq signal is detected after the TES, whereas many of the genes that are only expressed after MBT show high *d*NET-seq signal past the TES. We

662 also observed that most pre-MBT genes are isolated from other transcriptional units, while the TES of
663 many post-MBT genes is overlapped by another gene transcribed in the opposite direction. Taken
664 together, these results suggest that *Drosophila* genome architecture may be under selection for efficient
665 transcription termination of pre-MBT genes. Our observations further reveal an unexpected link
666 between transcriptional readthrough and the presence of an overlapping convergent gene in the
667 *Drosophila* genome. The finding that readthrough is detected even when the convergent overlapping
668 gene is not transcribed is intriguing. One possibility is the induction of local conformational changes in
669 chromatin and/or Pol II, Indeed, pioneer transcription factors can potentially regulate gene expression
670 at later stages of development by inducing significant chromatin conformational changes during early
671 embryonic development (Blythe and Wieschaus 2016; Schulz et al. 2015). Although such chromatin
672 remodelling events can potentially influence transcriptional read-through, regardless of convergent
673 gene transcription, further studies are needed to understand how transcription termination is regulated
674 during early *Drosophila* development.

675 A classical view in the splicing field is that for some introns, the 5' and 3' splice sites are
676 recognized directly by 'intron definition'. For other introns, spliceosome assembly starts with the
677 recognition of the downstream exon, and only at a later stage a cross-intron complex is formed ('exon
678 definition') (Robberson et al. 1990; Berget 1995; Sterner et al. 1996; Fox-Walsh et al. 2005). These
679 alternative models are supported by *in vitro* evidence, as well as by *in vivo* reporter gene experiments
680 (Robberson et al. 1990; Berget 1995; Fox-Walsh et al. 2005; De Conti et al. 2013). There is also indirect
681 evidence, for example from analysis of exon-intron size patterns (Berget 1995). However, these models
682 have never been directly tested *in vivo* at a transcriptome-wide level. It thus remains uncertain how
683 intron definition and exon definition operate *in vivo* and what determines the choice between them.

684 Introns that show splicing completion before Pol II has reached the end of the downstream exon
685 are necessarily intron-defined, as exon definition requires the presence of the 5' splice site of the
686 downstream intron. Therefore, introns with an SR above 0 represent an experimentally-determined set
687 of introns that are spliced *via* intron definition at least part of the time, allowing us to test long-held
688 assumptions in the field. Notably, *in vitro* and reporter gene work suggested that *Drosophila* uses intron
689 definition for introns smaller than ~200-250 nt and exon definition for larger introns (Fox-Walsh et al.

690 2005). Other studies have suggested that the crucial factor is not just intron size but the relative size of
691 the intron compared to the flanking exons (Sterner et al. 1996).

692 Using *d*NET-seq, we have found that although there is a relationship between intron size and
693 the proportion of introns with SR > 0, intron definition can be observed at all intron sizes. Our results
694 are thus inconsistent with a simple "threshold" model, where splicing would systematically switch to
695 exon definition at a particular intron size. Rather, we observe an increased proportion of introns with
696 SR > 0 within an optimal intron size range of ~55-100 nt. The median intron size in our data set is 71
697 nt. It thus appears that intron definition is best optimized for introns of a "typical" size, and is less
698 efficient when introns are unusually small or large. Intriguingly, however, for very large introns (>1000
699 nt), the proportion with SR > 0 increases again, suggesting that specific mechanisms may have evolved
700 for the intron-defined splicing of very large introns. The tendency for faster splicing kinetics with
701 average intron sizes has also been reported using metabolic labelling in both *Drosophila* (Pai et al.
702 2017) and human cells (Windhager et al. 2012).

703 We propose that the tethering of the upstream exon to Pol II (de Almeida and Carmo-Fonseca
704 2008) could be crucial for the intron-defined splicing of such large introns. Indeed, introns with SR > 0
705 where the upstream exon has one or more putative splicing intermediate reads tend to be larger than
706 ones without splicing intermediate reads (one-tailed Mann-Whitney *U*-test, W = 779623, P ~ 0.003, N
707 = 4589; exon size and read number filtering has been applied like for the gene architecture analysis).

708 In addition, Pai et al. (2017) found that GC poorer introns were spliced faster. We uncovered a
709 similar relationship with GC content, with the highest proportion of introns with SR > 0 observed when
710 both the exon and the intron were AT-rich but the exon had a higher GC content than the intron. We
711 emphasize that the gene architecture parameters that we have investigated are not independent of each-
712 other. For instance, exon GC content is correlated negatively with exon rank from start (Spearman's ρ
713 ~ -0.098; P ~ $1.196 * 10^{-12}$). Further work is thus needed to distinguish between causative and merely
714 correlative factors. In addition, it must be determined how the GC content effect relates to patterns of
715 nucleosome density (Gelfman et al. 2013).

716 Our results indicate that the majority (>80%) of *Drosophila* introns expressed in late embryos
717 can be spliced through intron definition. . A caveat of our work is that introns that show no evidence

718 for splicing completion during transcription of the downstream exon could still be spliced *via* intron
719 definition, if their splicing takes too long for it to be detected during the 100 nt window studied.
720 Therefore, we do not know whether introns with a SR of 0 are exon-defined, intron-defined but spliced
721 when Pol II has elongated past the first 100 exonic nucleotides, or a mixture of both. Similarly, introns
722 with a splicing ratio between 0 and 1 could either always be intron-defined, or they could use either
723 exon or intron definition depending on the splicing event.

724 Taken together, our results show that splicing in *Drosophila* embryos can be completed very
725 rapidly after transcription of the 3' splice site, as previously reported in other cellular models (Carrillo
726 Oesterreich et al. 2016; Reimer et al. 2021; Martin et al. 2013). Furthermore, our findings are relevant
727 to controversies regarding Pol II pausing in association with splicing. Indeed, splicing-dependent Pol II
728 pausing near the 3' splice site was first reported in yeast (Alexander et al. 2010; Chathoth et al. 2014).
729 Evidence for Pol II pausing at exon boundaries was further detected in mammalian cells using NET-
730 seq (Nojima et al. 2015; Mayer et al. 2015). Notably, NET-seq density profiles suggestive of Pol II
731 pausing were observed with S5P antibodies but not with antibodies to unphosphorylated CTD (Sheridan
732 et al. 2019; Nojima et al. 2015). In contrast, a recent PRO-seq analysis found no splicing-associated Pol
733 II pausing (Reimer et al. 2021).

734 Finally, we emphasize that our analysis of nascent *Drosophila* transcripts associated with Pol
735 II S5P CTD revealed highly heterogeneous splicing dynamics. Indeed, while only a minority of splice
736 junctions were devoid of reads spanning ligated exons (corresponding to an SR of 0), many introns
737 remained unspliced in our analysis. Is it relevant for cells whether any particular intron is rapidly
738 excised after Pol II transcribes the 3' splice site or is left unspliced while the transcript elongates? Is
739 this a stochastic process? What are the consequences for gene expression of fast versus delayed
740 splicing? Addressing these questions will likely dependent on further advances in methodologies to
741 study splicing in time and space at a transcriptome-wide level.
742

743 **METHODS**

744

745 **Embryo collection**

746

747 *Drosophila melanogaster* flies (Oregon R (OrR) strain) were raised at 25°C, in polypropylene
748 vials containing standard enriched culture medium (cornmeal, molasses, yeast, soya flour, and beetroot
749 syrup). Three-day-old flies (counting from pupae eclosion) were fattened in culture medium
750 supplemented with fresh yeast for 2 days. Embryos were collected into apple juice-agar plates
751 supplemented with fresh yeast using appropriate cages containing approximately 200 flies each. To
752 avoid female retention of older embryos, three pre-collections of 30min each were made before the first
753 collection of embryos. To maximize egg laying, and avoid larvae contamination, adult flies were
754 transferred to clean embryo collection cages every day over five days. For early stage embryos (2-3
755 hours after egg-laying), adult females were allowed to lay eggs for 1 hour in apple juice-agar plates.
756 Plates were subsequently collected and embryos were aged at 25°C for 90 min. During the following
757 30 min, embryos were harvested from the plates, dechorionated in 50% bleach solution for 2 min, and
758 washed once in Phosphate Buffered Saline supplemented with 0.1% Tween-20 (PBT) and twice in
759 deionized water. In order to discard older embryos (stage 6 and older), manual staging of collected
760 embryos was performed with the help of forceps and under a magnifier scope. Embryos were then
761 resuspended in a solution containing 120mM NaCl and 0.04% Triton, and washed twice with 120mM
762 NaCl solution. At the end of the 3 hours collection, the solution was removed and embryos were frozen
763 in liquid nitrogen and stored at -80ºC. For the late stage (4-6 hours), eggs were laid for 2 hours and aged
764 at 25°C for 3.5 hours. Embryo collection and processing was similar to the early stage embryos, but in
765 this case, no manual staging was performed.

766

767 **Embryo DNA staining**

768

769 For each embryo collection, and after dechorionation, a representative embryo sample was
770 collected from the total pool and fixed in a scintillation flask, using a solution containing 1 volume of

771 4% formaldehyde in PBT and 4 volumes of heptane, for 20 min at 100rpm. The lower aqueous phase
772 solution was subsequently removed, 4ml of methanol was added and embryos were shaken vigorously
773 during 1 min. Embryos were then collected from the bottom of the scintillation flask, washed twice
774 with methanol, and frozen at -20ºC in methanol. To rehydrate the embryos, they were washed for 5
775 min each, with 3:1, 1:1, and 1:3 mix solutions of methanol:PBT. Embryos were subsequently washed
776 twice in PBT and incubated with 1:5000 Sytox green (Invitrogen), supplemented with 5µg/ml RNase
777 A (Sigma-Aldrich) in PBT for 15 min. After washing with PBT, embryos were mounted in fluorescence
778 mounting medium (Dako) and examined in a Zeiss AxioZoom V16 Fluorescence Stereo Microscope
779 for image acquisition and embryo staging. Images were processed using ImageJ software (NIH).
780
781 ***d*NET-seq and library preparation**
782
783 The *d*NET-seq protocol was adapted from mNET-seq (Nojima et al. 2016). Briefly, 300 ul of
784 frozen embryos was resuspended in 3.5ml of Buffer B1 (15mM HEPES-KOH, pH 7.6; 10mM KCl;
785 5mM MgCl2; 1mM DTT; 0.1mM EDTA; 0.35M Sucrose; 4µg/ml Pepstatin; 10mM Sodium
786 Metabisulfite; 0.5mM EGTA supplemented with complete EDTA free protease inhibitor (Roche) and
787 PhoSTOP (Roche)). Embryos were homogenised in a Dounce homogenizer with 11x strokes using a
788 tight pestle on ice. The suspension was centrifuged at 7700g for 15 min at 4ºC, the supernatant was
789 discarded and the white pellet containing the nuclei was resuspended in 500 µl of Buffer B1. The
790 suspension was again homogenised in the Dounce with 4x strokes and loaded without mixing on the
791 top of buffer B2 (15mM HEPES-KOH, pH 7.6; 10mM KCl; 5mM MgCl2; 1mM DTT; 0.1mM EDTA;
792 0.8M Sucrose; 4µg/ml Pepstatin; 10mM Sodium Metabisulfite; 0.5mM EGTA supplemented with
793 complete EDTA free protease inhibitor (Roche) and PhosSTOP (Roche)). The suspension was
794 centrifuged at 1310g during 30 min at 4ºC, and the pellet was resuspended with 125 µl of NUN1 buffer
795 (20mM Tris-HCl (pH 7.9); 75 mM NaCl; 0.5mM EDTA and 50% Glycerol). 1.2ml of Buffer NUN2
796 (300mM NaCl, 7.5mM MgCl2, 1% NP-40, 1M Urea supplemented with complete EDTA free protease
797 inhibitor (Roche) and PhoSTOP (Roche) was mixed with the nuclei and incubated on ice during 15 min
798 preforming short vortex every 3 min. Chromatin was then centrifuged at 10000g for 10 min at 4ºC,

799 washed with 100 µl of 1x MNase Buffer and incubated in 100µl of MNase reaction mix (1x MNase
800 buffer and 30 gel unit/µl MNase (New England Biolabs)) during 3 min at 37ºC with mix at 1400rpm.
801 The reaction was stopped with 10 µl of 250mM EGTA, centrifuged at 10000g for 5 min at 4ºC, and the
802 supernatant containing the solubilized chromatin was recovered. For the early embryos sample, 2x
803 300µl of embryos were prepared in parallel and pooled together after the chromatin solubilization.
804 Immunoprecipitation of Pol II-RNA complexes was performed using 50µl of Protein G Dynabeads
805 (Thermo Fisher Scientific), pre-incubated over night with 5µg of the correspondent antibody: anti-Pol
806 II CTD S5P (ab5131 Abcam) or anti-Pol II CTD S2P (ab5095 Abcam) in 100µl NET2 (50mM Tris-
807 HCl pH 7.4; 150mM NaCl and 0.05% NP-40) and washed 3 times with NET2. Beads were incubated
808 with the solubilized chromatin in 1ml total volume of NET2 during 1 hour at 4ºC, washed 7 times with
809 500µl of NET2 and once with 100µl of PNKT (1x PNK buffer and 0.1% tween) before incubation
810 during 6 min in 50µl of PNK reaction mix (1x PNKT, 1 mM ATP and 0.05 U/ml T4 PNK 3′phosphatase
811 minus (NEB) in a thermomixer at 37ºC and 1400rpm. After washing the beads with NET2, long RNA
812 fragments were isolated using Quick-RNA MicroPrep (Zymo research): 300µl of RNA Lysis Buffer in
813 33% EtOH was mixed to the beads by pipetting. Beads were discarded and the suspension was loaded
814 into a Zymo spin column that was centrifuged at 10000g during 30sec. The column was washed once
815 with 400µl RNA prep buffer and twice with 700µl and 400µl RNA wash buffer respectively. RNA was
816 then eluted in 15µl of DNase/RNase-Free water (zymo) and stored at -80ºC. 100ng of RNA was used
817 to prepare each library, following the standard protocol of the Truseq small RNA library prep kit
818 (Illumina). After adapter ligation and reverse transcription, the libraries were PCR amplified using 16
819 PCR cycles and cDNA libraries were fractionated in the gel between 130 to 300bp. The libraries were
820 sequenced using PE-150 on the Illumina HiSeq X platform by Novogene Co., Ltd.
821
822
823
824
825
826

827 **QUANTIFICATION AND STATISTICAL ANALYSIS**

828 ***d*NET-seq data processing**

829

830 Adapter sequences were removed from all *d*NET-seq paired-end samples using Cutadapt
831 (version 1.18) (Martin 2011) with the following parameters: -a TGGAATTCTCGGGTGCCAAGG -A
832 GATCGTCGGACTGTAGAACTCTGAAC -m 10 -e 0.05 --match-read-wildcards -n 1. Paired-end
833 read merging was performed using bbmerge.sh from BBMap (Bushnell et al. 2017) with the 'xloose'
834 parameter. Merged reads were then aligned to the *Drosophila* reference genome (*dm6*; Ensembl release
835 95) (Cunningham et al. 2019) using STAR (version 2.6.0b) (Dobin et al. 2013) with –chimSegmentMin
836 set to 20. Only uniquely mapped reads were considered, extracted using SAMtools (version 1.7) (Li et
837 al. 2009) with -q set to 255. Exceptionally, in **Supplemental Fig S2A**, HiSat2 was used, using the same
838 dm6 genome annotation file for genome indexing and using default parameters, with the --no-
839 discordant –no-mixed flags set. PCR internal priming events generated during library preparation were
840 removed using a custom Python script (Prudêncio et al. 2020) with the following parameters: -a TGG..
841 -s paired. To obtain single-nucleotide resolution, a custom Python script (Prudêncio et al. 2020) was
842 used to extract the 5' end nucleotide of read 2 (after trimming) in each sequencing pair, with the
843 directionality indicated by read 1 (**Fig 1E**).

844

845 **Publicly available RNA-seq datasets used**

846

847 Publicly available *Drosophila* embryonic transcriptome sequencing data (Poly(A) RNA-seq),
848 performed in developmental stages similar to the dNET-seq early and late samples, were used in this
849 study. RNA-seq datasets corresponding to cycle 14B (Lott et al. 2011) were obtained from the Gene
850 Expression Omnibus (GEO) (samples GSM618409, GSM618410, GSM618421 and GSM618422 from
851 dataset GSE25180). RNA-seq datasets from 4-6h old embryos were obtained from modENCODE
852 project PRJNA75285 (Graveley et al. 2011) (accessions SRR023696, SRR023746, SRR023836,
853 SRR035220, SRR023669, SRR035405, SRR035406, SRR024014 and SRR023539).

854

855 **RNA-seq data processing**

856

857 Adapters were removed from all datasets using Trim Galore (version 0.4.4)
858 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; last accessed 26 April 2020).
859 Datasets from modENCODE and GEO were aligned to the *dm6 Drosophila* reference genome (Ensembl
860 release 95) (Cunningham et al. 2019) using STAR (version 2.6.0b) (Dobin et al. 2013) with --
861 chimSegmentMin set to 20. Stringtie (version 1.3.3b) (Pertea et al. 2015) was used to quantify
862 normalized gene expression as Transcripts Per Kilobase Million (TPM) values with the following
863 parameters: -a 5 -e. In addition, the isoform list was provided together with the -G parameter
864 corresponding to the *dm6* (Ensembl release 95) GTF file. Genes with TPM values above 2 were
865 considered to be expressed.

866

867 **Selection of genes and isoforms for analysis**

868

869 Similar to a previous GRO-seq analysis (Core et al. 2008), we used the read density of very
870 large intergenic regions (gene desert regions) to define the reference for absence of transcription. Gene
871 deserts were divided into 50kb windows, and *d*NET-seq read densities were calculated by dividing the
872 read counts in each window by the window length (in bp). Read counts per window were obtained with
873 bedtools genome coverage (version 2.27.1-1-gb87c465) (Quinlan and Hall 2010), and an arbitrary
874 density threshold was defined as the 90th percentile of the read density distribution (**Supplemental Fig**
875 **S4A**). Transcripts whose gene body *d*NETseq read density exceeded this threshold were considered to
876 be transcriptionally active (**Supplemental Fig S4B**). For all of the analyses performed on late genes,
877 only transcriptionally active genes were considered. In addition, only one isoform per gene was
878 considered for all analysis, selected as the isoform with the highest RPKM value in the RNA-seq dataset
879 for the corresponding developmental stage.

880 The coordinates of previously identified pre-MBT genes were obtained from Chen et al. (2013)
881 and converted to *dm6* coordinates using the *liftOver* tool (https://genome.ucsc.edu/cgi-

882 bin/hgLiftOver). The most representative isoform for each gene was manually selected through
883 visualization of individual profiles.
884
885 **Splicing intermediate and lariat detection**
886
887 Exons containing splicing intermediates or introns containing lariats were identified using a
888 peak finder algorithm (NET_snrPeakFinder) (Churchman and Weissman 2011; Prudêncio et al. 2020)
889 that detects the presence of a peak in the last nucleotide of an exon (splicing intermediate) or an intron
890 (lariat), by comparing the accumulation of 3' end reads mapping at that position with the mean read
891 density of the flanking 200 nucleotides. A peak is called when the read density at the peak is superior
892 to the mean of this surrounding region plus 3 standard deviations (Churchman and Weissman 2011).
893 Since gene read density influences peak detection, the exons were divided into quartiles based on the
894 *d*NET-seq read density of the corresponding gene. Only exons from the highest quartile (i.e. from genes
895 with the highest read density) were considered in **Fig 2** for *d*NET-seq Late analysis.
896
897 **Analysis of read density and peak calling**
898
899 RPKM values for the merged *d*NET-seq datasets were calculated in the following manner:
900 $$RPKM\ (transcript) = \frac{\text{reads} * \ 10^4 * \ 10^6}{(\text{total uniquely mapped reads}) * (gene\ length\ in\ bp)}$$
901 Where 10^4 normalizes for gene length and 10^6 normalizes for sequencing depth. Queries of gene
902 3'UTR overlaps (**Figure 3**) between genes were performed with bedtools intersect (version 2.27.1-1-
903 gb87c465) (Quinlan and Hall 2010).
904 To detect splicing intermediate and intron lariat peaks, the algorithm from Churchman and
905 Weissman (2011) was adapted as described in Prudêncio et al. (2020). To be able to call larger regions
906 as peaks, a custom algorithm was developed and implemented in Python 3.7. Note that like for all of
907 the custom Python code reported here, there is heavy reliance on *bedtools* v2.29.0 (Quinlan and Hall
908 2010) for operations on coordinate intervals. In addition, NumPy v1.17.2 (Harris et al. 2020) and SciPy

1.4.0 (SciPy 1.0 Contributors et al. 2020) were used. Our peak caller detects regions where the local read density is significantly higher than expected by chance given the overall read density of the transcript. Almost all of the numerical parameters used by the peak caller can be adjusted. In **Fig 4A**, results obtained with two particular parameterizations are shown. Peak Caller 1 ("large peaks") is adapted to detecting larger peaks and provides results that are more intuitive to a human observer. Peak Caller 2 ("small peaks") provides a finer spatial resolution, and corresponds to the settings used in all of the analyses in this study. The peak caller takes as input a BED file with the 3' ends of reads (single-nucleotide resolution), as well as a GTF file with transcript and exon annotations and a list of transcripts to analyse. It functions by calculating a sliding average of read density within each transcript (window size 5/21 for small/large; only reads mapping to the same strand as the annotated transcript are considered). It then randomly shuffles the positions of the reads within the transcript and recalculates the sliding averages to determine the random expectation. This can be repeated several times (5 in this study) for more robustness. Windows obtained with the true read distribution are called as significant if their read density is higher than the 99th percentile of the simulated windows. Note that in this study, we used a setting whereby this threshold is calculated separately for each exon (and its upstream intron, for exons other than the first), by excluding the intron-exon pair of interest and reads overlapping it during the simulation step. This is necessary so that when calling peaks within a given exon (and its upstream intron), the threshold set would not be affected by the reads within that particular exon and its upstream intron. This way, for instance, the calling of a peak in the beginning of the exon is not affected by the calling of a peak in the middle of the same exon (except through potential merging, see below). After the initial peaks are called, they are filtered to remove peaks where more than 90% of the reads come from a single nucleotide (probable PCR duplicates), that are shorter than 5 nucleotides, or that overlap with fewer reads than a specified threshold (10/5 for large/small). Finally, peaks that are within a specified distance of each-other (21/5 nucleotides for large/small) are merged together.

937 **Individual gene profiles**

938

939 Individual dNET-seq gene profiles were generated by separating reads by strand using
940 SAMtools (version 1.7) (Li et al. 2009). Strand-separated read data was converted to bedGraph format
941 using bedtools genomecov with the -bg flag (version 2.27.1-1-gb87c465) (Quinlan and Hall 2010).
942 Coverage values were normalized per nucleotide accounting for the total number of uniquely aligned
943 reads and with the scale set to reads per 10^8 sequences. The outcome was converted to bigwig files
944 through the bedGraphToBigWig tool (Kent et al. 2010) and uploaded to the UCSC genome browser
945 (James Kent et al. 2002).

946

947 **Metagene analysis**

948

949 The read density metagene plots in **Fig 3F-I** and **Fig 5E,F** were created with deepTools (version
950 3.0.2) (Ramírez et al. 2016). Metagenes with normalized gene size (Figure 3) have bins of 10 bp while
951 all other metagene plots in this study have single nucleotide resolution. Normalized gene and intron
952 lengths (**Fig 3F-I**) were obtained through the scale-regions option. Exon-intron junctions without
953 normalized lengths were obtained using the reference-option set to the 3'SS or the 5'SS. For
954 normalization, we divided the number of reads at each nucleotide (or bin) by the total number of reads
955 in the entire genomic region under analysis. These values were then used to calculate the mean for each
956 nucleotide, and the results were plotted in an arbitrary units (A.U.) ranging from 0 to 1.

957 The peak or read density metagene plots in **Fig 3D**, **4B** and **7D**, and **Supplemental Fig S3G,**
958 **S7A-G** and **S7E-J** were prepared using custom Python and R scripts
959 (https://github.com/rosinaSav/dNETseq_code). The peak density value represents the proportion of
960 introns/exons that overlap with a peak at that position. Only internal fully coding exons that were at
961 least 100 nucleotides long were included. In addition, the intron just upstream of the exon had to be at
962 least 50 nt long. For **Fig 7D**, **S7A-G** and **S7E-H**, further filtering based on read coverage was performed
963 (see below). Note that exons shorter than 150 nucleotides contribute both to the upstream and
964 downstream exonic proportion of the plot.

965 **Immediate splicing analysis**

966

967 The immediate splicing analysis was performed solely on the S5P data sets. Only the 117
968 previously annotated pre-MBT genes (Chen et al. 2013) were analysed for early data sets. Reads were
969 considered as spliced if they contained 'N's in the CIGAR string, in a position corresponding to an
970 annotated intron. Reads that overlapped both the (unspliced) intron and the downstream exon were
971 considered as unspliced. In both cases, only reads where the 3' end was located at least 5 nt downstream
972 of the 3' ss were included, to avoid analysing misaligned reads whose 3' end should have mapped to
973 the end of the upstream exon instead. Spliced reads that had the 5' end mapped to the upstream exon
974 and the 3' end mapped to the intron were considered indicative of recursive splicing if the first
975 nucleotide of the downstream end (indicative of the ratchet point position) matched the second G in the
976 AGGT canonical splicing motif. If the last nucleotide of a read matched the last nucleotide of an exon,
977 it was considered a splicing intermediate read and not representative of nascent RNA. A splicing ratio
978 was calculated by dividing the number of nascent RNA spliced reads by the sum of the number of
979 spliced and unspliced nascent RNA reads, only including reads whose 3' ends mapped to the first 100
980 nt of the downstream exon. Only fully coding internal exons at least 100 nt long were considered
981 (exceptionally, in **Fig 6E** and **K** and **Supplemental Fig S6O-Q**, the 3' most coding exon was also
982 analysed). Finally, we performed filtering to remove exons where the read coverage was too low to
983 allow for robust estimation of the splicing ratio. The relevant threshold was calculated for each dataset
984 separately. We calculated the total proportion of spliced reads out of all spliced/unspliced reads for the
985 dataset to obtain the expected splicing ratio. We then performed a binomial test to know the probability
986 of sampling only spliced/unspliced reads by chance under the null that the true splicing ratio equalled
987 this expectation. We set the threshold as the lowest number of reads that had to be sampled for the
988 probability to be below 0.01. Through his procedure, the threshold was set at >=10 reads for replicate
989 1 and 2 of the late data set (no terminal coding exons), at >=11/10 reads for replicate 1/2 of the late data
990 set including terminal coding exons, and at >=14/9 reads for replicate 1/2 of the early data set (terminal
991 coding exons always included).

992

993 **Gene architecture and nucleotide composition analysis**

994

995 Gene architecture and nucleotide composition parameters were calculated using custom Python
996 and R scripts based on Ensembl annotations for dm6.18 (Cunningham et al. 2019). Splice site strength
997 scores were calculated using MaxEntScan (Yeo and Burge 2004) with default parameters. Significance
998 testing was performed using Kruskal-Wallis tests with a Bonferroni correction for multiple
999 comparisons, with the correction applied separately for either replicate. For predictors where the
1000 corrected *P*-value was <0.05, Dunn's test was performed on the pairwise comparisons using *R* package
1001 *dunn.test* (Dinno 2017). The sequence logo in **Fig 4D** was generated using the *seqLogo* package version
1002 1.52.0 in R (Bembom 2019).

1003

1004 **Data access**

1005

1006 All raw and processed sequencing data generated in this study have been submitted to the
1007 NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number
1008 GSE152585. The Python code used is available at https://github.com/rosinaSav/dNETseq_code and
1009 https://github.com/kennyrebelo. The Python code can also be found in the **Supplemental Code** file.

1010

1011 **Competing interests**

1012 The authors declare that they have no competing interests.

1013

1014 **Acknowledgements**

1015

1016 We thank Takayuki Nojima and Nicholas Proudfoot (University of Oxford, UK) for critical discussion.
1017 This work was supported by funding to M.C.-F. (Fundação para a Ciência e Tecnologia, FCT/
1018 Ministério da Ciência, Tecnologia e Ensino Superior - Fundos do Orçamento de Estado
1019 (UIDB/50005/2020), and FCT/FEDER/POR Lisboa 2020, Programa Operacional Regional de Lisboa
1020 PORTUGAL 2020, grant LISBOA-01-0145-FEDER-016394) and to R.G.M (FCT grant PTDC/BIA-

BID/28441/2017). P.P. was a recipient of a FCT fellowship (SFRH/BD/109689/2015). R.S. was a recipient of an EMBO Long-Term Fellowship (EMBO ALTF 101-2019). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 842695 (Marie Skłodowska-Curie Actions) and No 857119 (RiboMed).

Author Contributions

P.P. performed all molecular biology experiments. R.S. performed the majority of splicing kinetics analyses and developed the peak caller for detecting putative pause sites. K.R. performed additional bioinformatics analyses. P.P., R.S., R.G.M. and M.C.-F conceived the study and wrote the paper.

1043 **REFERENCES**

1044

1045 Akhtar J, Kreim N, Marini F, Mohana G, Brüne D, Binder H, Roignant J-Y. 2019. Promoter-
1046 proximal pausing mediated by the exon junction complex regulates splicing. *Nat*
1047 *Commun* **10**: 521.

1048 Alexander RD, Innocente SA, Barrass JD, Beggs JD. 2010. Splicing-dependent RNA
1049 polymerase pausing in yeast. *Molecular Cell* **40**: 582–593.

1050 Ali-Murthy Z, Lott SE, Eisen MB, Kornberg TB. 2013. An Essential Role for Zygotic
1051 Expression in the Pre-Cellular Drosophila Embryo. *PLoS Genetics* **9**.

1052 Alpert T, Herzel L, Neugebauer KM. 2017. Perfect timing: splicing and transcription rates in
1053 living cells. *WIREs RNA* **8**.

1054 Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L. 2011.
1055 Total RNA sequencing reveals nascent transcription and widespread co-
1056 transcriptional splicing in the human brain. *Nature Structural and Molecular Biology*
1057 **18**: 1435–1440.

1058 Artieri CG, Fraser HB. 2014. Transcript length mediates developmental timing of gene
1059 expression across drosophila. *Molecular Biology and Evolution* **31**: 2879–2889.

1060 Arzate-Mejía RG, Josué Cerecedo-Castillo A, Guerrero G, Furlan-Magaril M, Recillas-Targa
1061 F. 2020. In situ dissection of domain boundaries affect genome topology and gene
1062 transcription in Drosophila. *Nat Commun* **11**: 894.

1063 Aslanzadeh V, Huang Y, Sanguinetti G, Beggs JD. 2018. Transcription rate strongly affects
1064 splicing fidelity and cotranscriptionality in budding yeast. *Genome Research* **28**: 203–
1065 213.

1066 Bembom O. 2019. seqLogo: Sequence logos for DNA sequence alignments. R package
1067 version 1.52.0.

1068 Berget SM. 1995. Exon Recognition in Vertebrate Splicing. *The Journal of Biological*
1069 *Chemistry* **270**: 2411–2414.

1070 Beyer AL, Osheim YN. 1988. Splice site selection, rate of splicing, and alternative splicing
1071 on nascent transcripts. *Genes & Development* **2**: 754–765.

1072 Blythe SA, Wieschaus EF. 2015. *Coordinating Cell Cycle Remodeling with Transcriptional*
1073 *Activation at the Drosophila MBT*. 1st ed. Elsevier Inc.

1074 Blythe SA, Wieschaus EF. 2016. Establishment and maintenance of heritable chromatin
1075 structure during early Drosophila embryogenesis. *Elife* **5**.

1076 Boija A, Mahat DB, Zare A, Holmqvist P-H, Philip P, Meyers DJ, Cole PA, Lis JT, Stenberg
1077 P, Mannervik M. 2017. CBP Regulates Recruitment and Release of Promoter-
1078 Proximal RNA Polymerase II. *Molecular Cell* **68**: 491-503.e5.

1079 Brugiolo M, Herzel L, Neugebauer KM. 2013. Counting on co-transcriptional splicing.
1080 *F1000Prime Reports* **5**.

1081 Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J, Lopez AJ. 2005. Subdivision of
1082 large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics* **170**:
1083 661–674.

1084 Bushnell B, Rood J, Singer E. 2017. BBMerge – Accurate paired shotgun read merging via
1085 overlap. *PLoS ONE* **12**.

1086 Campos-Ortega JA. 1985. Genetics of early neurogenesis in Drosophila melanogaster. *Trends*
1087 *in Neurosciences* **8**: 245–250.

1088 Carrillo Oesterreich F, Herzel L, Straube K, Hujer K, Howard J, Neugebauer KM. 2016.
1089 Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell*
1090 **165**: 372–381.

1091 Carrillo Oesterreich F, Preibisch S, Neugebauer KM. 2010. Global analysis of nascent RNA
1092 reveals transcriptional pausing in terminal exons. *Molecular Cell* **40**: 571–581.

1093 Chathoth KT, Barrass JD, Webb S, Beggs JD. 2014. A Splicing-Dependent Transcriptional
1094 Checkpoint Associated with Prespliceosome Formation. *Molecular Cell* **53**: 779–790.

1095 Chen K, Johnston J, Shao W, Meier S, Staber C, Zeitlinger J. 2013. A global change in RNA
1096 polymerase II pausing during the Drosophila midblastula transition. *eLife* **2013**.

1097 Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at
1098 nucleotide resolution. *Nature* **469**: 368–373.

1099 Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA Sequencing Reveals Widespread Pausing
1100 and Divergent Initiation. *Science* **322**: 1845–1849.

1101 Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J,
1102 Billis K, Boddu S, et al. 2019. Ensembl 2019. *Nucleic Acids Research* **47**: D745–
1103 D751.

1104 Dahlberg O, Shilkova O, Tang M, Holmqvist P-H, Mannervik M. 2015. P-TEFb, the Super
1105 Elongation Complex and Mediator Regulate a Subset of Non-paused Genes during
1106 Early Drosophila Embryo Development ed. G.P. Copenhaver. *PLOS Genetics* **11**:
1107 e1004971.

1108 David CJ, Boyne AR, Millhouse SR, Manley JL. 2011. The RNA polymerase II C-terminal
1109 domain promotes splicing activation through recruitment of a U2AF65-Prp19
1110 complex. *Genes and Development* **25**: 972–982.

1111 de Almeida SF, Carmo-Fonseca M. 2008. The CTD role in cotranscriptional RNA processing
1112 and surveillance. *FEBS Letters* **582**: 1971–1976.

1113 De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing.
1114 *Wiley interdisciplinary reviews RNA* **4**: 49–60.

1115 De Renzis S, Elemento O, Tavazoie S, Wieschaus EF. 2007. Unmasking activation of the
1116 zygotic genome using chromosomal deletions in the Drosophila embryo. *PLoS*
1117 *Biology* **5**: 1036–1051.

1118 Dingwall C, Lomonossoff GP, Laskey RA. 1981. High sequence specificity of micrococcal
1119 nuclease. *Nucleic Acids Research* **9**: 2659–2674.

1120 Dinno A. 2017. dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums. R
1121 package version 1.3.5.

1122 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
1123 Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:
1124 15–21.

1125 Drexler HL, Choquet K, Churchman LS. 2020. Splicing Kinetics and Coordination Revealed
1126 by Direct Nascent RNA Sequencing through Nanopores. *Molecular Cell* **77**: 985-
1127 998.e8.

1128 Duff MO, Olson S, Wei X, Garrett SC, Osman A, Bolisetty M, Plocik A, Celniker SE,
1129 Graveley BR. 2015. Genome-wide identification of zero nucleotide recursive splicing
1130 in Drosophila. *Nature* **521**: 376–379.

1131 Emili A, Shales M, McCracken S, Xie W, Tucker PW, Kobayashi R, Blencowe BJ, Ingles
1132 CJ. 2002. Splicing and transcription-associated proteins PSF and p54nrb/NonO bind
1133 to the RNA polymerase II CTD. *Rna* **8**: 1102–1111.

1134 Erickson JW, Cline TW. 1993. A bZIP protein, sisterless-a, collaborates with bHLH
1135 transcription factors early in Drosophila development to determine sex. *Genes and*
1136 *Development* **7**: 1688–1702.

1137 Fong N, Kim H, Zhou Y, Ji X, Qiu J, Saldi T, Diener K, Jones K, Fu XD, Bentley DL. 2014.
1138 Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate.
1139 *Genes and Development* **28**: 2663–2676.

1140 Fox-Walsh KL, Dou Y, Lam BJ, Hung S -p., Baldi PF, Hertel KJ. 2005. The architecture of
1141 pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National*
1142 *Academy of Sciences* **102**: 16176–16181.

1143 Fukaya T, Lim B, Levine M. 2017. Rapid Rates of Pol II Elongation in the Drosophila
1144 Embryo. *Current Biology* **27**: 1387–1391.

1145 Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K,
1146 Widom J, Gilad Y, Pritchard JK. 2012. Controls of Nucleosome Positioning in the
1147 Human Genome. *PLoS Genetics* **8**.

1148 Gelfman S, Cohen N, Yearim A, Ast G. 2013. DNA-methylation effect on cotranscriptional
1149 splicing is dependent on GC architecture of the exon-intron structure. *Genome*
1150 *Research* **23**: 789–799.

1151 Görnemann J, Barrandon C, Hujer K, Rutz B, Rigaut G, Kotovic KM, Faux C, Neugebauer
1152 KM, Séraphin B. 2011. Cotranscriptional spliceosome assembly and splicing are
1153 independent of the Prp40p WW domain. *Rna* **17**: 2119–2129.

1154 Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, Baren
1155 MJ Van, Boley N, Booth BW, et al. 2011. The Developmental Transcriptome of
1156 Drosophila melanogaster. *Nature* **471**: 473–479.

1157 Guilgur LG, Prudêncio P, Sobral D, Liszekova D, Rosa A, Martinho RG. 2014. Requirement
1158 for highly efficient pre-mRNA splicing during Drosophila early embryonic
1159 development. *eLife* **3**.

1160 Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E,
1161 Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature*
1162 **585**: 357–362.

1163 Hatton AR, Subramaniam V, Lopez AJ. 1998. Generation of alternative Ultrabithorax
1164 isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions.
1165 *Molecular Cell* **2**: 787–796.

1166 Herzel L, Straube K, Neugebauer KM. 2018. Long-read sequencing of nascent RNA reveals
1167 coupling among RNA processing events. 1008–1019.

1168 Heyn P, Kircher M, Dahl A, Kelso J, Tomancak P, Kalinka AT, Neugebauer KM. 2014. The
1169 earliest transcribed zygotic genes are short, newly evolved, and different across
1170 species. *Cell Reports* **6**: 285–292.

1171 Hörz W, Altenburger W. 1981. Sequence specific cleavage of DNA by micrococcal nuclease.
1172 *Nucleic Acids Research* **9**: 2643–2658.

1173 Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Booth BW,
1174 Zhang D, Wan KH, Yang L, et al. 2011. Genome-wide analysis of promoter
1175 architecture in Drosophila melanogaster. *Genome Research* **21**: 182–192.

1176 Hsin JP, Manley JL. 2012. The RNA polymerase II CTD coordinates transcription and RNA
1177 processing. *Genes and Development* **26**: 2119–2137.

1178 James Kent W, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D.
1179 2002. The human genome browser at UCSC. *Genome Research* **12**: 996–1006.

1180 Ji J-Y, Squirrell JM, Schubiger G. 2004. Both Cyclin B levels and DNA-replication
1181 checkpoint control the early embryonic mitoses in Drosophila Jun-Yuan.
1182 *Development* **131**: 401–411.

1183 Jonkers I, Kwak H, Lis JT. 2014. Genome-wide dynamics of Pol II elongation and its
1184 interplay with promoter proximal pausing, chromatin, and exons. *eLife* **2014**.

1185 Joseph B, Kondo S, Lai EC. 2018. Short cryptic exons mediate recursive splicing in
1186 Drosophila. *Nature Structural and Molecular Biology* **25**: 365–371.

1187 Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed:
1188 Enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204–2207.

1189 Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: Diversification, exon
1190 definition and function. *Nature Reviews Genetics* **11**: 345–355.

1191 Khodor YL, Menet JS, Tolan M, Rosbash M. 2012. Cotranscriptional splicing efficiency
1192 differs dramatically between Drosophila and mouse. *RNA* **18**: 2174–2186.

1193 Khodor YL, Rodriguez J, Abruzzi KC, Tang CHA, Marr MT, Rosbash M. 2011. Nascent-seq
1194 indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes and*
1195 *Development* **25**: 2502–2512.

1196 Kimura H, Tao Y, Roeder RG, Cook PR. 1999. Quantitation of RNA Polymerase II and Its
1197 Transcription Factors in an HeLa Cell: Little Soluble Holoenzyme but Significant
1198 Amounts of Polymerases Attached to the Nuclear Substructure. *Molecular and*
1199 *Cellular Biology* **19**: 5383–5392.

1200 Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how
1201 promoters direct initiation and pausing. *Science* **339**: 950–953.

1202 Kwasnieski JC, Orr-Weaver TL, Bartel DP. 2019. Early genome activation in Drosophila is
1203 extensive with an initial tendency for aborted transcripts and retained introns. *Genome*
1204 *Research* **29**: 1188–1197.

1205 Larson MH, Mooney RA, Peters JM, Windgassen T, Nayak D, Gross CA, Block SM,
1206 Greenleaf WJ, Landick R, Weissman JS. 2014. A pause sequence enriched at
1207 translation start sites drives transcription dynamics in vivo. *Science* **344**: 1042–1047.

1208 Laver JD, Marsolais AJ, Smibert CA, Lipshitz HD. 2015. *Regulation and Function of*
1209 *Maternal Gene Products During the Maternal-to-Zygotic Transition in Drosophila*.
1210 1st ed. Elsevier Inc.

1211 Lee JEA, Mitchell NC, Zaytseva O, Chahal A, Mendis P, Cartier-Michaud A, Parsons LM,
1212 Poortinga G, Levens DL, Hannan RD, et al. 2015. Defective Hfp-dependent
1213 transcriptional repression of dMYC is fundamental to tissue overgrowth in Drosophila
1214 XPB models. *Nat Commun* **6**: 7404.

1215 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
1216 R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:
1217 2078–2079.

1218 Lott SE, Villalta JE, Schroth GP, Luo S, Tonkin LA, Eisen MB. 2011. Noncanonical
1219 compensation of zygotic X transcription in early Drosophila melanogaster
1220 development revealed through single-embryo RNA-Seq. *PLoS Biology* **9**.

1221 Lund E, Dahlberg JE. 1992. Cyclic 2 ', 3 "-Phosphates and Nontemplated Nucleotides at the
1222 3" End of Spliceosomal U6 Small Nuclear RNAs. *Science* **255**: 327–330.

1223 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing
1224 reads. *EMBnet.journal* **17**: 10–12.

1225 Martin RM, Rino J, Carvalho C, Kirchhausen T, Carmo-Fonseca M. 2013. Live-Cell
1226 Visualization of Pre-mRNA Splicing with Single-Molecule Sensitivity. *Cell Reports*
1227 **4**: 1144–1155.

1228 Martinho RG, Guilgur LG, Prudêncio P. 2015. How gene expression in fast-proliferating
1229 cells keeps pace. *BioEssays* **37**: 514–524.

1230 Maslon MM, Braunschweig U, Aitken S, Mann AR, Kilanowski F, Hunter CJ, Blencowe BJ,
1231 Kornblihtt AR, Adams IR, Cáceres JF. 2019. A slow transcription rate causes
1232 embryonic lethality and perturbs kinetic coupling of neuronal genes. *The EMBO*
1233 *Journal* **38**.

1234 Mayer A, Di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R,
1235 Stamatoyannopoulos JA, Churchman LS. 2015. Native elongating transcript
1236 sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**:
1237 541–554.

1238 Morris DP, Greenleaf AL. 2000. The splicing factor, Prp40, binds the phosphorylated
1239 carboxyl-terminal domain of RNA Polymerase II. *Journal of Biological Chemistry*
1240 **275**: 39935–39943.

1241 Nojima T, Gomes T, Carmo-Fonseca M, Proudfoot NJ. 2016. Mammalian NET-seq analysis
1242 defines nascent RNA profiles and associated RNA processing genome-wide. *Nature*
1243 *Protocols* **11**: 413–428.

1244 Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot
1245 NJ. 2015. Mammalian NET-seq reveals genome-wide nascent transcription coupled to
1246 RNA processing. *Cell* **161**: 526–540.

1247 Nojima T, Rebelo K, Gomes T, Grosso AR, Proudfoot NJ, Carmo-Fonseca M. 2018. RNA
1248 Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during
1249 Co-transcriptional Splicing. *Molecular Cell* **72**: 369-379.e4.

1250 Pai AA, Henriques T, McCue K, Burkholder A, Adelman K, Burge CB. 2017. The kinetics of
1251 pre-mRNA splicing in the Drosophila genome and the influence of gene architecture.
1252 *eLife* **6**.

1253 Pai AA, Paggi JM, Yan P, Adelman K, Burge CB. 2018. Numerous recursive sites contribute
1254 to accuracy of splicing in long introns in flies. *PLoS Genetics* **14**.

1255 Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie
1256 enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature*
1257 *Biotechnology* **33**: 290–295.

1258 Pritchard DK, Schubiger G. 1996. Activation of transcription in Drosophila embryos is a
1259 gradual process mediated by the nucleocytoplasmic ratio. *Genes and Development* **10**:
1260 1131–1142.

1261 Prudêncio P, Rebelo K, Grosso AR, Martinho RG, Carmo-fonseca M. 2020. Analysis of
1262 Mammalian Native Elongating Transcript sequencing (mNET-seq)high-throughput
1263 data. *Methods* **178**: 89–95.

1264 Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic
1265 features. *Bioinformatics* **26**: 841–842.

1266 Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F,
1267 Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data
1268 analysis. *Nucleic acids research* **44**: W160–W165.

1269 Reimer KA, Mimoso CA, Adelman K, Neugebauer KM. 2021. Co-transcriptional splicing
1270 regulates 3′ end cleavage during mammalian erythropoiesis. *Molecular Cell* **81**: 998-
1271 1012.e7.

1272 Robberson BL, Cote GJ, Berget SM. 1990. Exon Definition May Facilitate Splice Site
1273 Selection in RNAs with Multiple Exons. *Molecular and Cellular Biology* **10**: 84–94.

1274 Rothe M, Pehl M, Taubert H, Jäckle H. 1992. Loss of gene function through rapid mitotic
1275 cycles in the Drosophila embryp. *Nature* **359**: 156–159.

1276 Sandler JE, Irizarry J, Stepanik V, Dunipace L, Amrhein H, Stathopoulos A. 2018. A
1277 Developmental Program Truncates Long Transcripts to Temporally Regulate Cell
1278 Signaling. *Developmental Cell* **47**: 773-784.e6.

1279 Saunders A, Core LJ, Sutcliffe C, Lis JT, Ashe HL. 2013. Extensive polymerase pausing
1280 during Drosophila axis patterning enables high-level and pliable transcription. *Genes*
1281 *and Development* **27**: 1146–1158.

1282 Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ. 2017.
1283 Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs.
1284 *Molecular Cell* **65**: 25–38.

1285 Schulz KN, Bondra ER, Moshe A, Villalta JE, Lieb JD, Kaplan T, McKay DJ, Harrison MM.
1286 2015. Zelda is differentially required for chromatin accessibility, transcription factor
1287 binding, and gene expression in the early Drosophila embryo. *Genome Res* **25**: 1715–
1288 1726.

1289 SciPy 1.0 Contributors, Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T,
1290 Cournapeau D, Burovski E, Peterson P, Weckesser W, et al. 2020. SciPy 1.0:
1291 fundamental algorithms for scientific computing in Python. *Nat Methods* **17**: 261–
1292 272.

1293 Sheridan RM, Fong N, D'Alessandro A, Bentley DL. 2019. Widespread Backtracking by
1294 RNA Pol II Is a Major Effector of Gene Activation, 5′ Pause Release, Termination,
1295 and Transcription Elongation Rate. *Molecular Cell* **73**: 107-118.e4.

1296 Shermoen AW, O'Farrell PH. 1991. Progression of the Cell Cycle through Mitosis Leads to
1297 Abortion of Nascent Transcripts. *Cell* **67**: 303–310.

1298 Sterner DA, Carlo T, Berget SM. 1996. Architectural limits on split genes. *Proc Natl Acad*
1299 *Sci USA* **93**: 15081–15085.

1300 Veloso A, Kirkconnell KS, Magnuson B, Biewen B, Paulsen MT, Wilson TE, Ljungman M.
1301 2014. Rate of elongation by RNA polymerase II is associated with specific gene
1302 features and epigenetic modifications. *Genome Research* **24**: 896–905.

1303 Vizcaya-Molina E, Klein CC, Serras F, Mishra RK, Guigó R, Corominas M. 2018. Damage-
1304 responsive elements in *Drosophila* regeneration. *Genome Res* **28**: 1852–1866.

1305 Windhager L, Bonfert T, Burger K, Ruzsics Z, Krebs S, Kaufmann S, Malterer G, L'Hernault
1306 A, Schilhabel M, Schreiber S, et al. 2012. Ultrashort and progressive 4sU-tagging

1307 reveals key characteristics of RNA processing at nucleotide resolution. *Genome*
1308 *Research* **22**: 2031–2042.

1309 Yan D, Neumüller RA, Buckner M, Ayers K, Li H, Hu Y, Yang-Zhou D, Pan L, Wang X,
1310 Kelley C, et al. 2014. A Regulatory Network of Drosophila Germline Stem Cell Self-
1311 Renewal. *Developmental Cell* **28**: 459–473.

1312 Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with
1313 applications to RNA splicing signals. *Journal of Computational Biology* **11**: 377–394.

1314 Yuan K, Seller CA, Shermoen AW, Farrell PHO. 2016. Timing the Drosophila Mid-Blastula
1315 Transition: A Cell Cycle-Centered View. *Trends in Genetics* **32**: 496–507.

1316 Zhu J, Liu M, Liu X, Dong Z. 2018. RNA polymerase II activity revealed by GRO-seq and
1317 pNET-seq in Arabidopsis. *Nature Plants* **4**: 1112–1123.

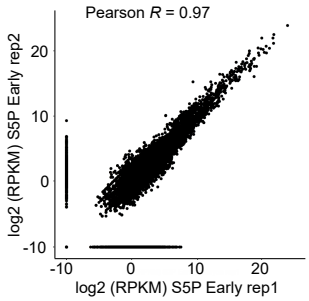
1318 Zhu L, Zhang Y, Zhang W, Yang S, Chen J, Tian D. 2009. Patterns of exon-intron
1319 architecture variation of genes in eukaryotic genomes. *BMC Genomics* **10**.

SUPPLEMENTAL FIGURES

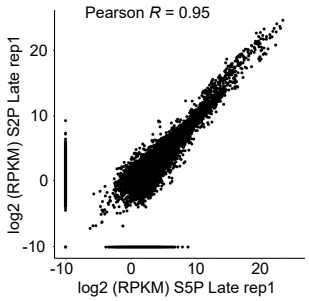
A

Dataset	Raw reads	Uniquely Mapped after Merge
Early_2-3h_S5P_rep1	62,891,575	12,959,070
Early_2-3h_S5P_rep2	56,917,697	3,610,607
Late_4-6h_S5P_rep1	65,865,258	21,564,673
Late_4-6h_S5P_rep2	75,909,131	27,375,998
Late_4-6h_S5P_rep3	71,923,876	11,987,034
Late_4-6h_S2P_rep1	77,838,463	38,595,780
Late_4-6h_S2P_rep2	64,862,996	23,137,529
Late_4-6h_S2P_rep3	76,719,886	17,912,504

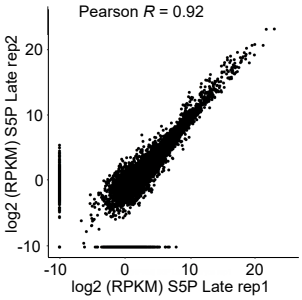
B



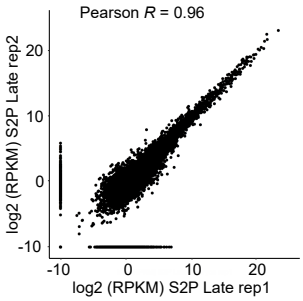
C



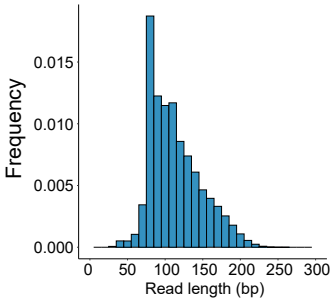
D



E



F



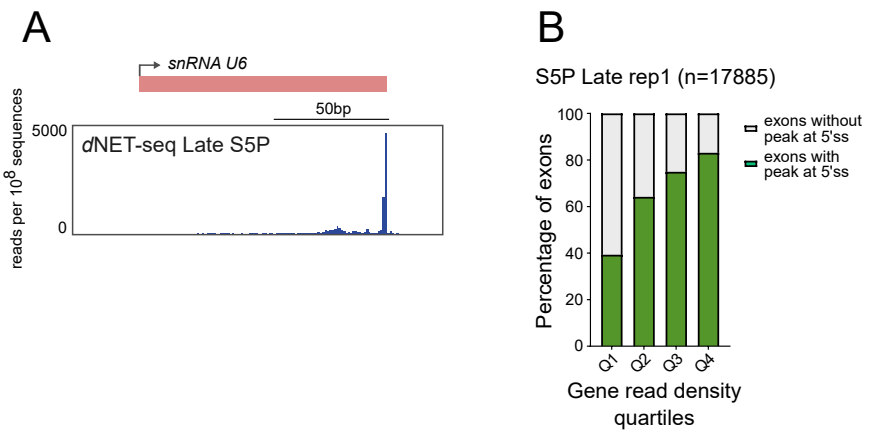
G

Dataset	Number of nascent reads	Nascent reads / Uniquely mapped reads (%)
Early_2-3h_S5P_rep1	7,984,411	61.6
Early_2-3h_S5P_rep2	2,430,876	67.3
Late_4-6h_S5P_rep1	13,413,045	62.2
Late_4-6h_S5P_rep2	13,777,880	50.3
Late_4-6h_S5P_rep3	8,505,056	71.0
Late_4-6h_S2P_rep1	23,153,659	60.0
Late_4-6h_S2P_rep2	12,926,707	55.9
Late_4-6h_S2P_rep3	10,356,743	57.8

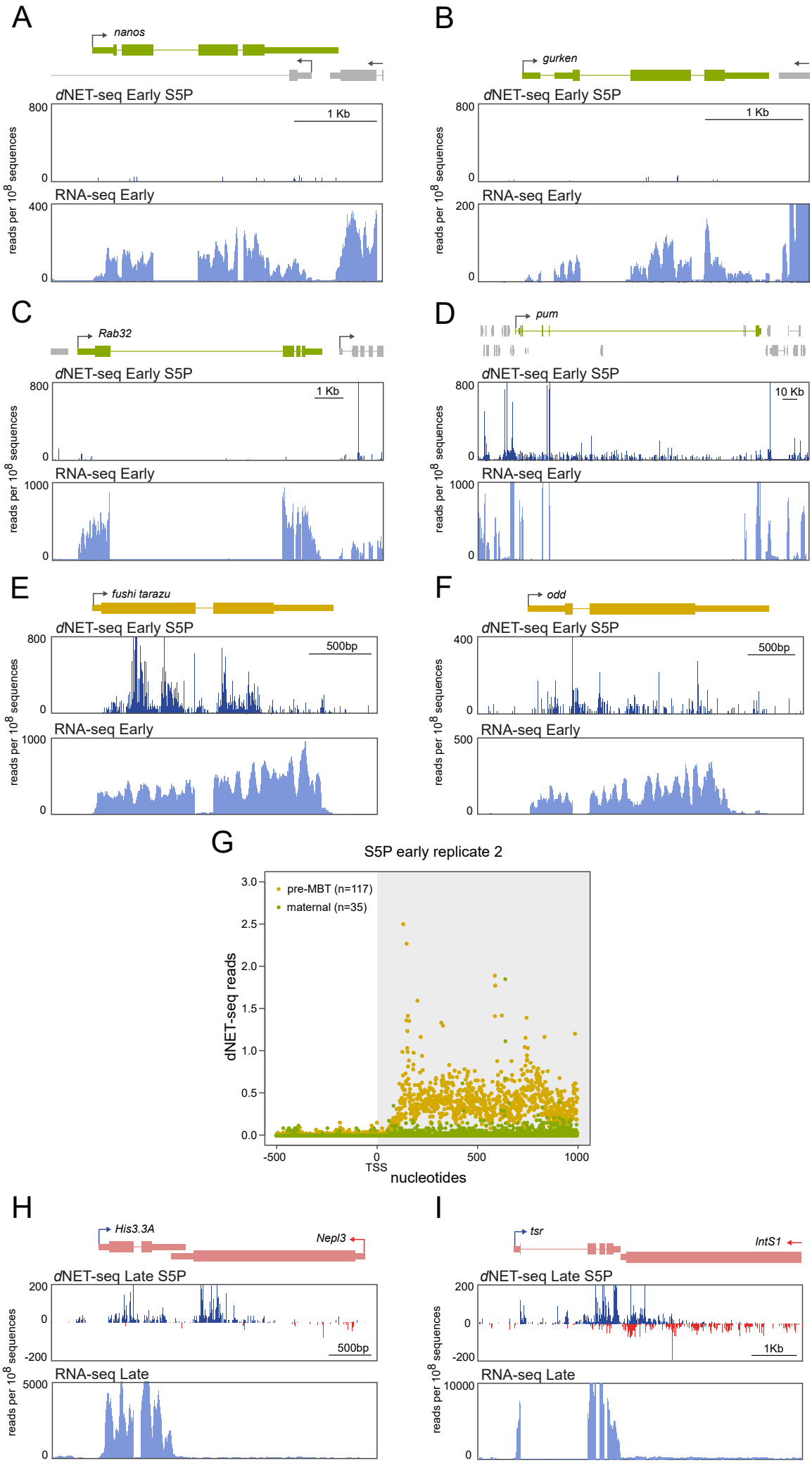
H

Dataset	min	mean	max	25th percentile	50th percentile	75th percentile
Early_2-3h_S5P_rep1	35	99.6	293	74	81	119
Early_2-3h_S5P_rep2	35	94.3	293	72	81	110
Late_4-6h_S5P_rep1	35	114.2	293	86	108	135
Late_4-6h_S5P_rep2	35	102.0	294	80	87	119
Late_4-6h_S5P_rep3	35	88.4	294	76	81	89
Late_4-6h_S2P_rep1	35	107.0	294	81	96	128
Late_4-6h_S2P_rep2	35	111.1	294	81	102	132
Late_4-6h_S2P_rep3	35	109.0	293	83	105	130

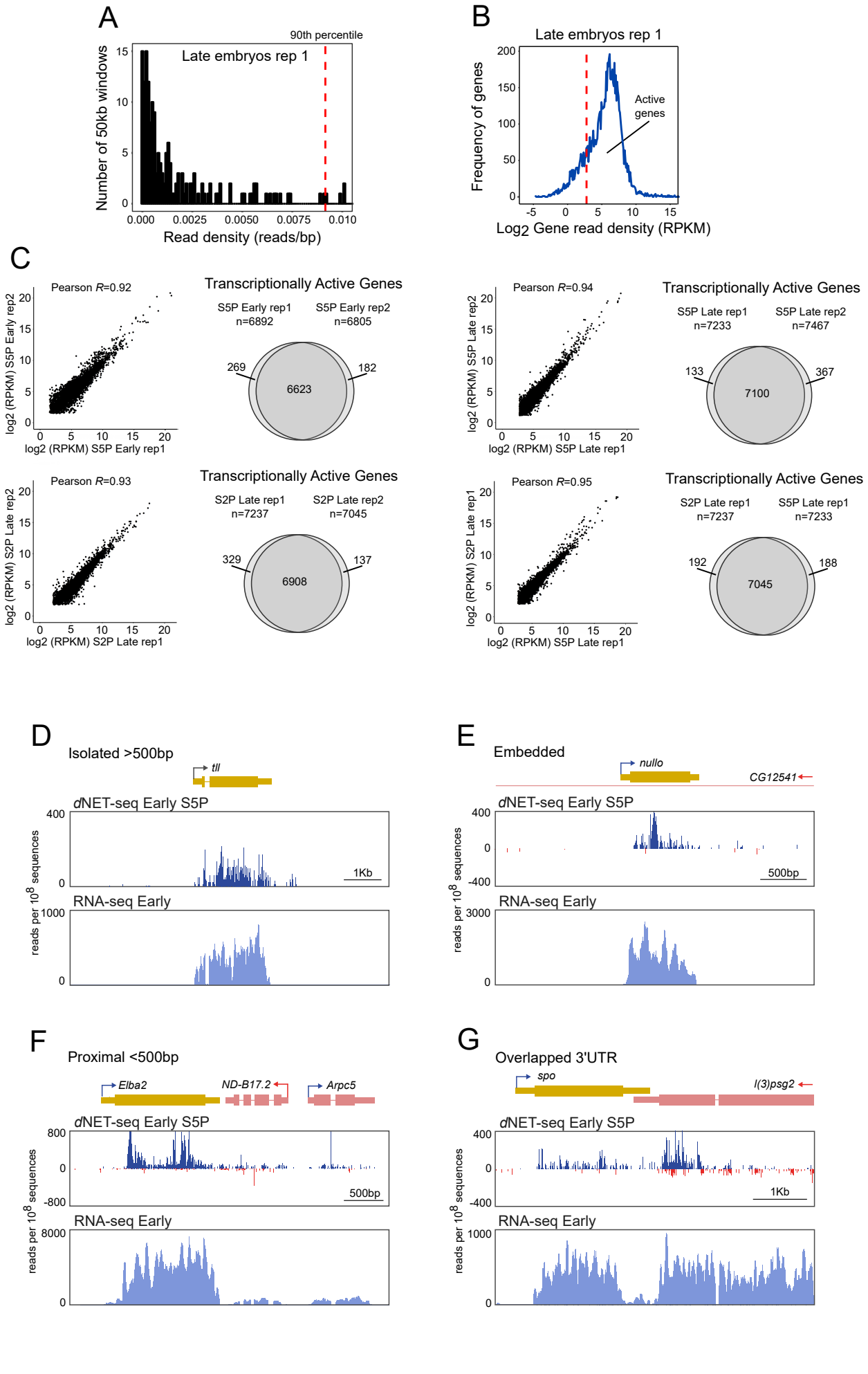
Supplemental Fig. S1. (A) For each *d*NET-seq library prepared, the number of total reads and of uniquely aligned reads is indicated. **(B)** Density of uniquely aligned reads per gene (RPKM in log2 scale) for two *d*NET-seq/S5P biological replicates from early embryos (Pearson's correlation, R = 0.97). (**C**) Density of uniquely aligned reads per gene (RPKM in log2 scale) from late embryos analysed by *d*NET-seq/S5P and *d*NET-seq/S2P (Pearson's correlation, R = 0.95). (**D, E**) Density of uniquely aligned reads per gene (RPKM in log2 scale) for two *d*NET-seq/S5P and *d*NET-seq/S2P biological replicates from late embryos (Pearson's correlation, R = 0.92; R = 0.96). **(F)** Histogram of the read lengths in *d*NET-seq/S5P data from late embryos. (**G**) For each *d*NET-seq library generated, the number of uniquely aligned reads corresponding to nascent transcripts is indicated. **(H)** Length of sequenced nascent RNA (in nucleotides).



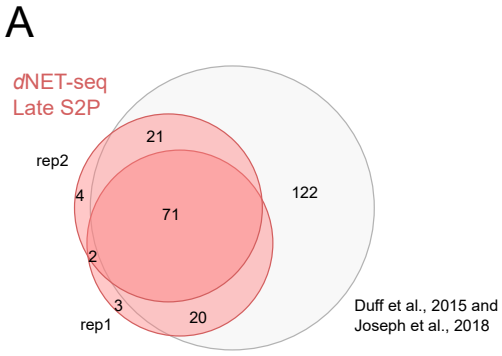
Supplemental Fig. S2. (**A**) *d*NET-seq/S5P profile over the U6 snRNA gene in the late dataset (replicate 1). (**B**) Frequency of peaks corresponding to splicing intermediates (green) detected by *d*NET-seq/S5Pon exons of genes expressed in late embryos. Genes were grouped into quartiles (Q) based on their *d*NET-seqread density.


A
nanos
dNET-seq Early S5P
1 Kb
RNA-seq Early
reads per 10^8 sequences
B
gurken
dNET-seq Early S5P
1 Kb
RNA-seq Early
C
Rab32
dNET-seq Early S5P
1 Kb
RNA-seq Early
D
pum
dNET-seq Early S5P
10 Kb
RNA-seq Early
E
fushi tarazu
dNET-seq Early S5P
500bp
RNA-seq Early
F
odd
dNET-seq Early S5P
500bp
RNA-seq Early
G
S5P early replicate 2
pre-MBT (n=117)
maternal (n=35)
dNET-seq reads
TSS
nucleotides
H
His3.3A
Nepl3
dNET-seq Late S5P
500bp
RNA-seq Late
I
tsr
IntS1
dNET-seq Late S5P
1Kb
RNA-seq Late

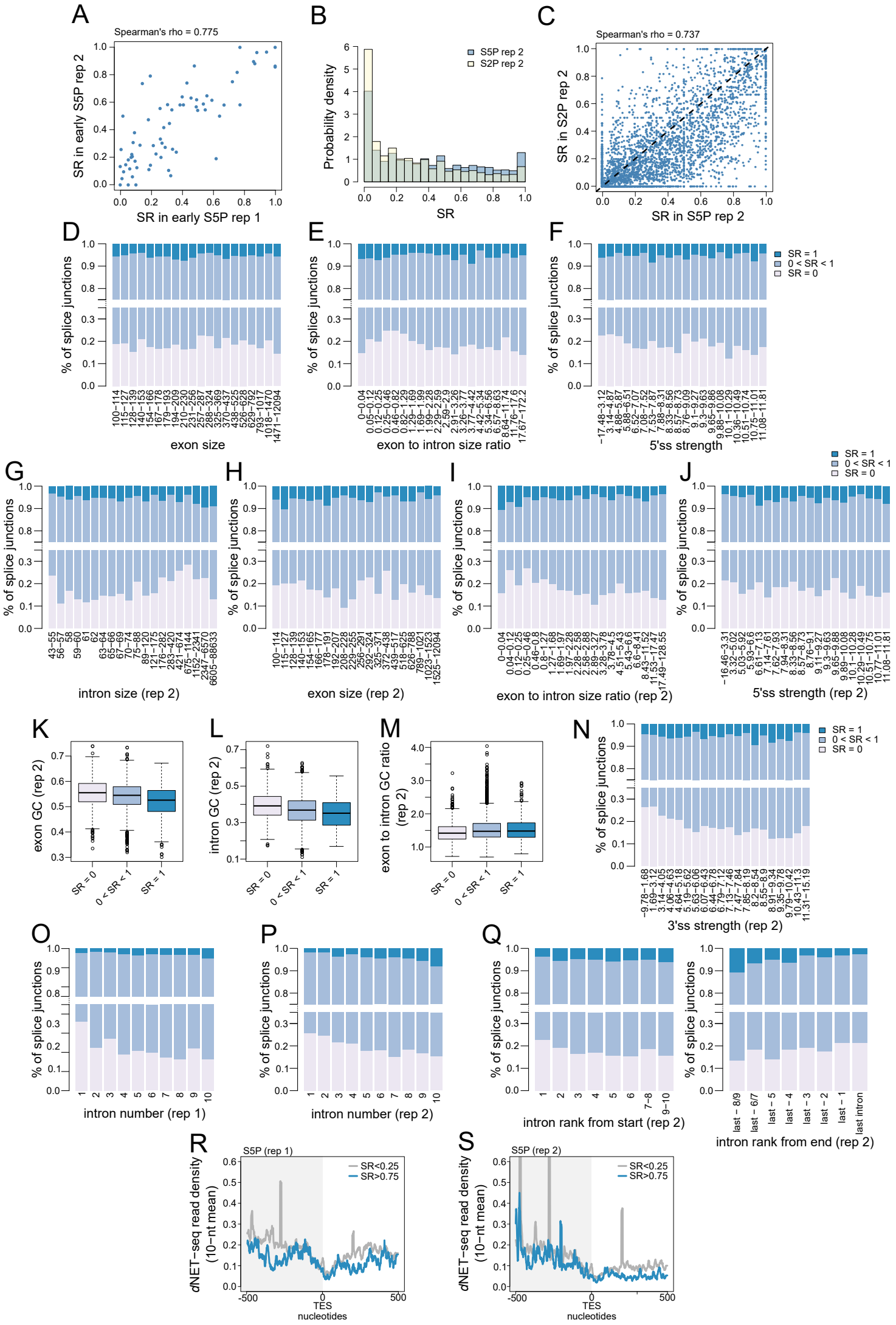
Supplemental Fig. S3. **(A-D)** *d*NET-seq/S5P and RNA-seq profiles over maternal genes in early embryos: *nanos* (**A**), *gurken* (**B**), *Rab32* (**C**) and *pumilio (pum)* (**D**). **(E-F)** *d*NET-seq/S5P and RNA-seq profiles over pre-MBT genes in early embryos: *fushi tarazu* (**E**) and *odd skipped (odd)* (**F**). (**G**) Meta-analysis of *d*NET-seq/S5P mean read density around the transcription start site (TSS) in maternal and pre-MBT genes (replicate 2). **(H-I)** *d*NET-seq/S5P and RNA-seq profiles over MBT genes in late embryos: *His3.3A* (**H**) and *twinstar (tsr)* (**I**). Reads that aligned to the positive strand are in blue, and reads that aligned to the negative strand are in red. The direction of transcription is indicated by an arrow.


A
90th percentile
Late embryos rep 1
Number of 50kb windows
Read density (reads/bp)
B
Late embryos rep 1
Active genes
Frequency of genes
Log2 Gene read density (RPKM)
C
Pearson R=0.92
log2 (RPKM) S5P Early rep2
log2 (RPKM) S5P Early rep1
Transcriptionally Active Genes
S5P Early rep1 n=6892
S5P Early rep2 n=6805
269
6623
182
Pearson R=0.94
log2 (RPKM) S5P Late rep2
log2 (RPKM) S5P Late rep1
Transcriptionally Active Genes
S5P Late rep1 n=7233
S5P Late rep2 n=7467
133
7100
367
Pearson R=0.93
log2 (RPKM) S2P Late rep2
log2 (RPKM) S2P Late rep1
Transcriptionally Active Genes
S2P Late rep1 n=7237
S2P Late rep2 n=7045
329
6908
137
Pearson R=0.95
log2 (RPKM) S2P Late rep1
log2 (RPKM) S5P Late rep1
Transcriptionally Active Genes
S2P Late rep1 n=7237
S5P Late rep1 n=7233
192
7045
188
D
Isolated >500bp
tll
dNET-seq Early S5P
1Kb
RNA-seq Early
reads per 10^8 sequences
E
Embedded
nullo
CG12541
dNET-seq Early S5P
500bp
RNA-seq Early
reads per 10^8 sequences
F
Proximal <500bp
Elba2
ND-B17.2
Arpc5
dNET-seq Early S5P
500bp
RNA-seq Early
reads per 10^8 sequences
G
Overlapped 3'UTR
spo
l(3)psg2
dNET-seq Early S5P
1Kb
RNA-seq Early
reads per 10^8 sequences

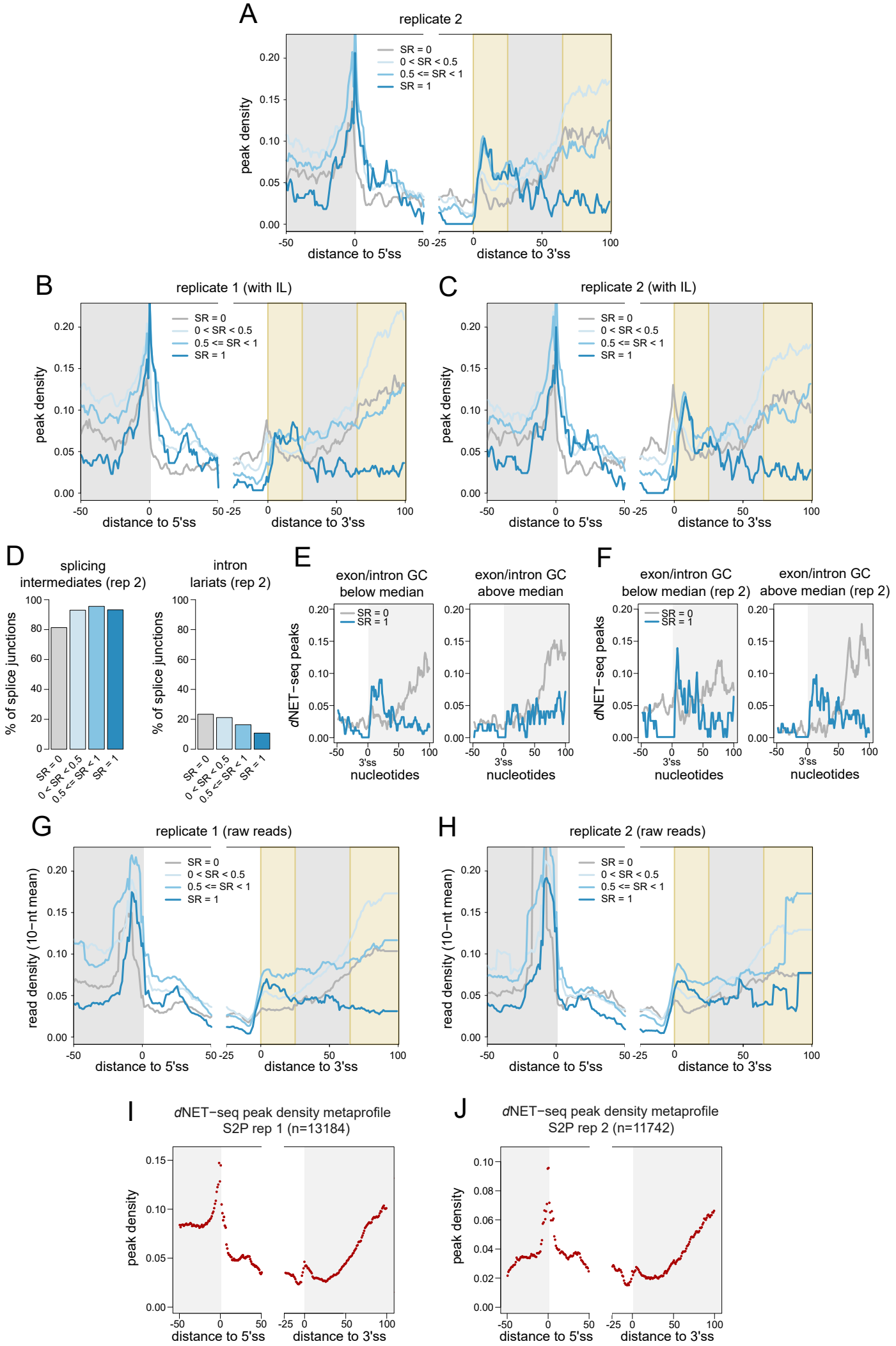
Supplemental Fig. S4. **(A-C)** Identification of transcriptionally active genes based on *d*NET-seq signal along the gene body in late embryos. **(A)** Read density distribution in intergenic regions (gene deserts). The red dashed line represents the 90th percentile of read density in all intergenic regions analysed. **(B)** Read density distribution per gene (RPKM) represented in Log2 scale. The 90th percentile of read density over gene deserts is set as threshold (red dashed line). **(C)** Comparison of read density per active gene (RPKM in log2 scale) between the indicated datasets. Total number of active genes identified in each dataset is indicated (*n*). Venn diagrams show common active genes between datasets. **(D-G)** *d*NET-seq/S5P and RNA-seq profiles over the indicated pre-MBT genes. Reads that aligned to the positive strand are in blue, and reads that aligned to the negative strand are in red. The direction of transcription is indicated by an arrow.



Supplemental Fig . S5. **(A)** Venn diagram comparing RPs identified in two *d*NET-seq/S2P biological replicates and in previously reported studies (Joseph et al. 2018; Duff et al. 2015).


A
Spearman's rho = 0.775
SR in early S5P rep 2
SR in early S5P rep 1
B
S5P rep 2
S2P rep 2
Probability density
SR
C
Spearman's rho = 0.737
SR in S2P rep 2
SR in S5P rep 2
D
% of splice junctions
exon size
E
exon to intron size ratio
F
5'ss strength
SR = 1
0 < SR < 1
SR = 0
G
intron size (rep 2)
H
exon size (rep 2)
I
exon to intron size ratio (rep 2)
J
5'ss strength (rep 2)
K
exon GC (rep 2)
L
intron GC (rep 2)
M
exon to intron GC ratio (rep 2)
N
3'ss strength (rep 2)
O
intron number (rep 1)
P
intron number (rep 2)
Q
intron rank from start (rep 2)
intron rank from end (rep 2)
R
S5P (rep 1)
dNET−seq read density (10−nt mean)
SR<0.25
SR>0.75
TES
nucleotides
S
S5P (rep 2)

Supplemental Fig. S6. In all panels, unless otherwise specified, only introns from transcriptionally active genes where the downstream exon is a fully coding internal exon at least 100 nt long were included. In addition, enough spliced/unspliced reads had to end within the first 100 exonic nucleotides that obtaining a splicing ratio (SR) of 0 or 1 by chance alone was highly unlikely (see the Methods section for details). For genes expressed in late embryos, this threshold was 10 reads for both replicates. For pre-MBT genes, it was 14 for replicate 1 and 9 for replicate 2. For S2P, a threshold of 10 was used for better comparability with S5P data. (**A**) SR values estimated in two biological replicates of *d*NET-seq/S5P datasets for pre-MBT genes in early embryos (Spearman correlation, $\rho = \sim 0.775$, $P \sim 1.348 * 10^{-14}$). **(B)** Distribution of SR values for *d*NET-seq/S5P (N = 4510) and *d*NET-seq/S2P (N = 4373). (**C**) SR values estimated in replicate 2 of *d*NET-seq/S5P and *d*NET-seq/S2P from late embryos (Spearman correlation, $\rho = \sim 0.737$, $P < 2.2 * 10^{-16}$; N = 3295). **(D-Q)** SR values for *d*NET-seq/S5P data (replicate 1 and 2) plotted out against several gene architecture parameters. For sample sizes and statistical tests, see **Supplementary Table 1**. In **(D-J, N-Q)**, the bin ranges have been set so that intron numbers would be as equal possible between bins. (**R-S**) *d*NET-seq/S5P read density around the transcription end site (TES) (averaged over a 10-nucleotide sliding window) for genes with different mean splicing ratios in replicate 1 (**R**) and 2 (**S**). There is no significant correlation between SR and the count of reads 101-300 nucleotides past the TES (partial Spearman correlation controlling for the count of reads in the region 500 to 301 nucleotides before the TES as a proxy for gene body rad density; replicate 1: $\rho = \sim 0.013$, $P = \sim 0.519$; N = 2450; replicate 2: $\rho = \sim -0.034$, P = ~0.126; N = 2052).


A
replicate 2
SR = 0
0 < SR < 0.5
0.5 <= SR < 1
SR = 1
peak density
distance to 5'ss
distance to 3'ss
B
replicate 1 (with IL)
C
replicate 2 (with IL)
D
splicing intermediates (rep 2)
intron lariats (rep 2)
% of splice junctions
E
exon/intron GC below median
exon/intron GC above median
dNET−seq peaks
3'ss
nucleotides
F
exon/intron GC below median (rep 2)
exon/intron GC above median (rep 2)
G
replicate 1 (raw reads)
read density (10−nt mean)
H
replicate 2 (raw reads)
I
dNET−seq peak density metaprofile
S2P rep 1 (n=13184)
J
dNET−seq peak density metaprofile
S2P rep 2 (n=11742)

Supplemental Fig. S7. (**A-C**) Metagene analyses of peak density estimated from *d*NET-seq/S5P datasets from late embryos for different SR ranges. (**A**) shows replicate 2, with introns with putative intron lariat reads removed (N = 3623). (**B**) and (**C**) show replicates 1 and 2, respectively, with introns with putative intron lariat reads included (N = 5614 for replicate 1 and N = 4503 for replicate 2). Peak density has been calculated as the proportion of introns that overlap with a peak at any given position. The last 50 nucleotides of exons, the first 50 nucleotides of introns, the last 25 nucleotides of introns, and the first 100 nucleotides of exons are shown. Only internal and fully coding exons from transcriptionally active genes that are at least 100 nucleotides long are shown. In addition, at least 10 spliced/unspliced reads had to end within the first 100 nucleotides of the exon. (**D**) The proportion of introns with at least one read whose 3' end maps to the final position of the upstream exon (putative splicing intermediates) or to the final position of the intron (putative intron lariats) in *d*NET-seq/S5P late replicate 2. (**E-F**) *d*NET-seq/S5P peak density profiles for late embryos, replicate 1 (**E**) and replicate 2 (**F**). The final 50 nucleotides of the intron and the first 100 nucleotides of the downstream exon are shown. For each data point, the GC content over the first 100 exonic nucleotides was divided by the GC content over the last 50 intronic nucleotides, generating the exonic to intronic GC ratio. Introns were split into two classes along the median of this metric. Only introns with an SR of 0 or 1 were plotted. (**G-H**) Similar to 7D and S7A, respectively, but showing read densities (averaged over a 10-nucleotide sliding window) rather than peak densities. (**I-J**) Metagene analysis of peak density estimated from *d*NET-seq/S2P datasets from late embryos, in replicate 1(**I**) and replicate 2 (**J**). The last 50 nucleotides of exons, the first 50 nucleotides of introns, the last 25 nucleotides of introns, and the first 100 nucleotides of exons are shown. Only internal and fully coding exons from transcriptionally active genes that are at least 100 nucleotides long are shown (N = 13229 for replicate 1, N = 11831 for replicate 2). Exons shorter than 150 nucleotides contribute to both the exon end and start. Only introns that were at least 50 nt long were considered.

SUPPLEMENTARY METHODS

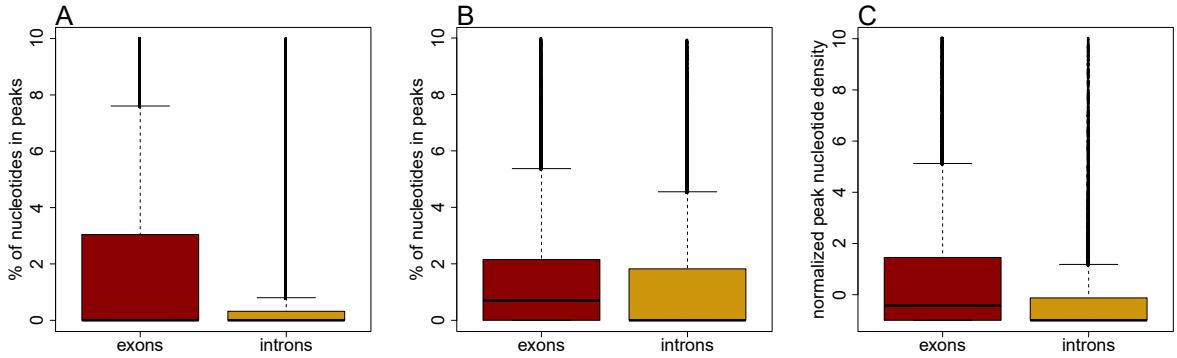
*d*NET-seq sequence biases

Previous work has shown that MNase cleaves DNA more frequently at AT than at GC base pairs (Dingwall *et al.*, 1981). More specifically, Hörz and Altenburger (1981) found a preference for sites where a few alternating AT-TA base pairs were surrounded by a more GC-rich region, rather than for sites within longer AT-rich stretches. More recent genome-wide work in humans has come to similar conclusions (Gaffney *et al.*, 2012). *d*NET-seq relies on MNase cleavage of both DNA and RNA to solubilize chromatin. Any nucleotide biases in cleavage may be problematic, as they could mean that the solubilization is more efficient in regions where the nucleotide composition is more in line with such cleavage biases. This could lead to an artefactually increased read density in such regions.

To verify whether there was any evidence for RNA cleavage biases, we determined the nucleotide composition around the locations of the 5' ends of NET-seq reads (**Figure 4D** in main text). A clear preference for cleavage just 5' of an adenine was observed, with additional weaker biases at surrounding sites. Our results are therefore broadly in line with those obtained previously for DNA.

To what extent could these biases affect our results? We generated a simulated version of a *d*NET-seq dataset (S5P late, replicate 1) where the read positions were randomized in such a way as to preserve the nucleotide biases at the 5' ends of reads. Only reads that overlapped transcriptionally active genes were considered. More specifically, for each read, we defined the "starting hexamer" as the hexamer centred on the 5' end of the read (the three nucleotides just 5' of the read and the first three nucleotides of the read). We then randomly picked another occurrence of the same hexamers from within transcriptionally active genes and defined a "simulated read" at that position, preserving the length of the initial read. We then called peaks on the simulated reads similarly to true reads.

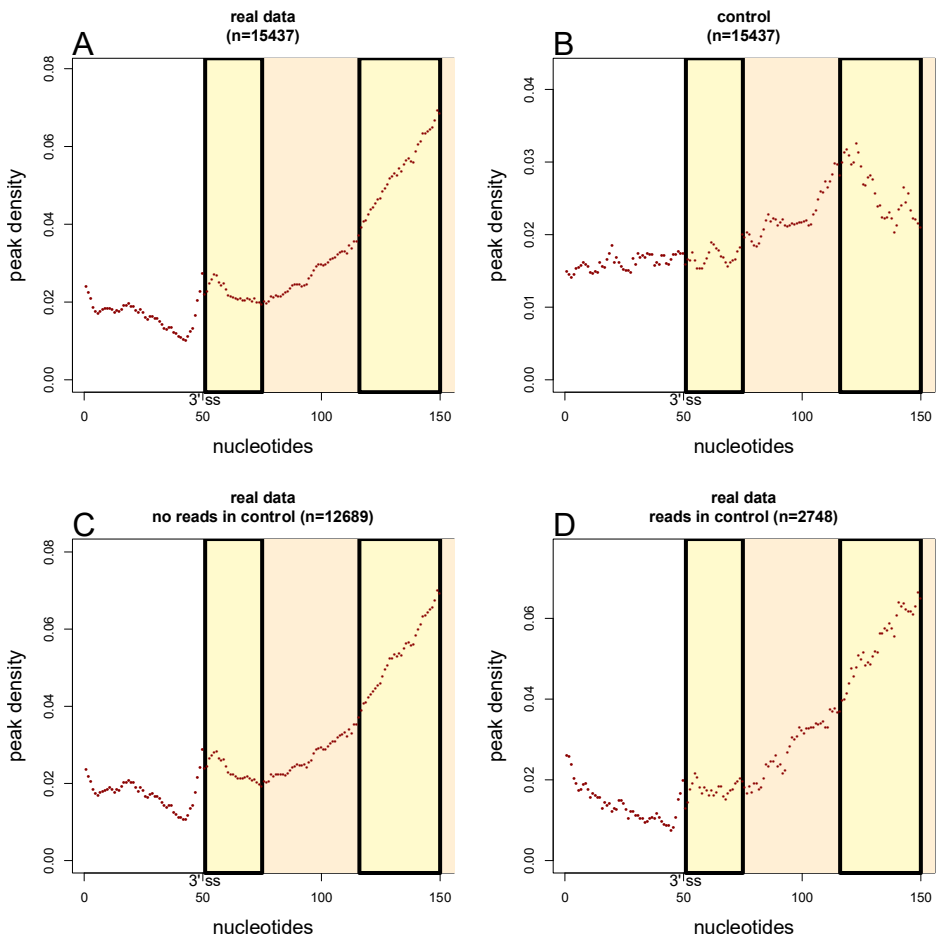
The mean proportion of nucleotides within simulated peaks was similar for both exons and introns (~0.017). However, this belies a difference in medians, with a higher peak density in exons than in introns (exon median: ~0.008; intron median: 0; $P < 2.2 * 10^{-16}$; two-tailed Mann-Whitney *U*-test; **Supplementary Methods Figure 1**). To know if the 5' end nucleotide bias effect could explain the bias towards exons observed in real data, we calculated a normalized peak density for each exon and intron, as $\frac{true\ peak\ density - simulated\ peak\ density}{simulated\ peak\ density}$. Exons and introns with a simulated peak density of 0 were discarded. Normalized peak densities were still higher for exons than for introns ($P < 2.2 * 10^{-16}$; two-tailed Mann-Whitney *U*-test). Hence, 5' end nucleotide biases may exaggerate but do not explain the bias towards exons in real data.



Supplementary Methods Figure 1: distribution of peak densities for exons and introns. A: true peaks. B: simulated peaks. C: normalized peaks, obtained as $\frac{\textit{true peak density} - \textit{simulated peak density}}{\textit{simulated peak density}}$*. For all three panels, the y axis has been limited to values below 10 for visualization reasons.*

It may seem surprising that simulated read density would be higher in exons, given that *Drosophila* exons tend to have a higher GC content than introns (Zhu *et al.*, 2009). Indeed, one would, at first sight, expect the bias to exaggerate intronic peak density, rather than exonic peak density. However, peaks are called based on the locations of the 3' ends of reads, whereas the bias affects the locations of the 5' ends of reads. Therefore, many reads whose 3' ends are in GC-rich exonic sequence are expected to have their 5' ends in AT-rich intronic sequence.

Further examination of the meta-profiles reveals that the simulated peaks do not display the increased peak density in the beginning of the exons (first highlighted region in **Supplementary Methods Figure 2B**), as observed for true data (**Supplementary Methods Figure 2A**). It thus appears that this peak cannot be explained by read 5' end biases. However, the second highlighted region shows increased peak densities both in true and simulated data (**Supplementary Methods Figure 2A-B**). To verify if the increase observed in this region with true data could be explained by read 5' end nucleotide composition biases, we divided the true data into two classes based on whether or not any simulated peaks mapped to within the second highlighted region (**Supplementary Methods Figure 2C-D**). There was no visible difference in true mean peak densities between the two groups within the second highlighted region. This suggests that any nucleotide biases acting at the 5' ends of reads are unlikely to explain this increase in peak densities.



Supplementary Methods Figure 2: average peak density for the last 50 nucleotides of introns and the first 100 nucleotides of exons (concatenated). The orange boxes show the positions of exons. Only exons at least 100 nt long have been included.

Dingwall, C, Lomonossoff, GP, and Laskey, RA (1981). High sequence specificity of micrococcal nuclease. Nucleic Acids Res 9, 2659–2674.
Gaffney, DJ, McVicker, G, Pai, AA, Fondufe-Mittendorf, YN, Lewellen, N, Michelini, K, Widom, J, Gilad, Y, and Pritchard, JK (2012). Controls of Nucleosome Positioning in the Human Genome. PLoS Genet 8, 1–13.
Hörz, W, and Altenburger, W (1981). Sequence specific cleavage of DNA by micrococcal nuclease. Nucleic Acids Res 9, 2643–2658.
Zhu, L, Zhang, Y, Zhang, W, Yang, S, Chen, J, and Tian, D (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. 12, 1–12.