RNA Polymerase Stalling at Developmental Control Genes in the Drosophila Embryo – Supplementary Information

Julia Zeitlinger, Alexander Stark, Manolis Kellis, Joung-Woo Hong, Sergei Nechaev, Karen Adelman, Michael Levine and Richard A. Young

Table of contents

Experimental protocol

ChIP-chip assays. The chromatin immunoprecipitation (ChIP) experiments coupled to micrroarrays (chip) were performed as described in Zeitlinger et al. 2007 on protein A-coupled Dynabeads but with different antibodies (see below).

Antibodies. In order to maximize Pol II enrichment in the chromatin immunoprecipitation and to detect Pol II in different states, we used a mixture of two monoclonal antibodies (8WG16 and H14). Pooled monoclonal antibodies in immunoprecipitations enable the formation of multimeric complexes, like polyclonal antibodies, but are more specific than polyclonal antibodies ("Using Antibodies" by Ed Harlow and David Lane"). The antibody 8WG16 recognizes the C-terminal heptapeptide repeat present on the largest subunit of Pol II, while H14 recognizes the phosphoserine 5 version of the heptapeptide repeat. Both antibodies work very efficiently in chromatin immunoprecipitations by themselves and have been used in many studies. Although some studies report that the 8WG16 antibody specifically recognizes the hypophosphorylated and therefore initiating form of Pol II (e.g. Kim et al. 2005; Lee et al. 2006), other studies, especially those in yeast and *Drosophila,* clearly also detect the elongating form of Pol II in chromatin immunoprecipitations using 8WG16 (e.g. Boehm et al. 2003). It is unclear at this point whether the differences in the reported specificity of 8WG16 depend on species, experimental conditions, controls, biological expectation or a combination of factors. Similar uncertainties exist with the specificity of antibodies raised against phosphorylated forms of Pol II. To avoid arguing about the specificity of the antibodies, we decided to use an approach that would maximize the detection of any form of Pol II. The results argue that we immunoprecipitated both the initiating and elongating form of Pol II with high efficiency.

Arrays. We used *Drosophila* whole-genome tiling arrays printed by Agilent as 11-array set (44k each) as described in Zeitlinger et al. 2007 or as 2-array set (244k each). Probes of 60mers span the entire eukaryotic portion of the *Drosophila melanogaster* genome. While the spacing of these probes is ~280 bp on average, an additional probe is present between the two probes that flank each known TSS. Thus, the resolution around transcriptional start sites is ~ 140 bp.

Data. The data can be downloaded from ArrayExpress (E-TABM-322) or our web site at http://web.wi.mit.edu/young/pol2/

Error model. We used the Rosetta error model to control for noise at probes, thus a probe required a p-value < 0.001. We did not use our previous algorithms for detecting bound probes and then assigning genes. Rather, we calculated parameters indicating Pol II enrichment directly for each gene (see below). A combination of Pol II enrichment at the start site and median enrichment across the gene were used to classify Pol II as either absent, stalled or active.

Permanganate footprint assays. Transcription bubble assays with $KMnO_4$ were performed as described previously (Gilmour and Lis 1986; Wang et al. 2007). Embryos were collected 2-4 hours after egg deposition, dechorionated and partially homogenized before treatment with $KMnO_4$. Embryos were treated with 20 mM $KMnO_4$ (Adelman lab, Fig. 6) or 40 mM $KMnO_4$ (Levine lab, Fig. 3D) for 60s on ice. The transcription start sites of the examined genes were identified and confirmed using ESTs in Flybase and previous expression analysis using tiling arrays (Biemar et al. 2006). The linker primers and gene-specific primers used for ligation-mediated PCR (LM-PCR) are shown in Table S1.

Table S1. Primers used for permanganate footprint assays

linker A	5'-GCGGTGATTTAAAAGATCTGAATTC-3'
linker B	5'-GAATTCAGATC-3'
rho-LMPCR-1	5'-CATTGGTAACTTAGTTTTGC-3'
rho-LMPCR-2	5'-AACTTAGTTTTGCTGCTCGT-3'
rho-LMPCR-3	5'-TTTTGCTGCTCGTAAATCCAG-3'
Dr-LMPCR-1	5'-GATCGTTTGTGTAACTGTGG-3'
Dr-LMPCR-2	5'-TTGTGTAACTGTGGCTCGTT-3'
Dr-LMPCR-3	5'-CTGTGGctcgttAATACTGTGCT-3'
Lbe-LMPCR-1	5'-AGAGTTTCGTTTCAATTCGT-3'
Lbe-LMPCR-2	5'-TCAATTCGTTTGGTTTAGCA-3'
Lbe-LMPCR-3	5'-TCGTTTGGTTTAGCACTTAACTGT-3'
Tup-LMPCR-1	5'-GGATTTGGATCTATGGTGAG-3'
Tup-LMPCR-2	5'-TGGATCTATGGTGAGGGATT-3'
Tup-LMPCR-3	5'-GGTGAGGGATTTAAGAGTCTCTCGC-3'

Classification of genes based on Pol II profile

Calculation of the Stalling Index from whole-genome tiling data

When we started analyzing the Pol II ChIP-chip data, there was no precedent for the classification of stalled Pol II profiles. We therefore developed our own method. Recently, another study had independently developed a similar method for analyzing RNA polymerase binding profiles in *E.coli* (Reppas et al. 2006).

The basic idea is to systematically calculate the ratio between the enrichment at the transcriptional start site (TSS) versus the enrichment found at the transcription unit (TU). We termed this ratio "Stalling Index". Reppas et al. defined the inverse ratio and termed it "Traveling Ratio". The exact details are also slightly different presumably due to different data, array design, genome structure and biological question, but the idea and findings remain similar.

ASSIGNMENT OF PROBES TO TSS AND TU

The 300-600 rule. To determine which probes should enter the TSS and TU for each gene, two issues were encountered. First, we did not want to make a prior assumption of whether the maximum peak was found upstream or downstream of transcription. We therefore searched 300 bp upstream and downstream of the TSS (200 bp was found to miss the max signal of a few genes). Second, high signal from the TSS can significantly contribute to the median(TU) value. This is because the TSS peaks are often located significantly downstream of the TSS, and the tail (or shoulder) of the peak can be detected a few hundred bp further into the gene. (The shorter the gene, the greater the distortion can be.) We therefore excluded the first 600 bp from the calculation of the median(TU) value. This means that the Stalling Index cannot be calculated for genes that are less than 600 bp long or that have no probe on the array after 600 bp. (This was true for 4.6% of all genes.)

CALCULATION OF STALLING INDEX

Max and Median. Across the TU, the signal from different probes varies slightly but is overall constant. Therefore, the more probes are chosen, the more robust the result. Furthermore, because of outliers (e.g. additional TSSs among the TU), the median is a more robust way of calculating the overall signal than the mean. (Calculating the maximum would be particularly unreliable since it selects for outliers). For the TSS, the Pol II signal, if present, is not constant across a certain distance but has a clear peak near the TSS and then tails off at both sides (unless it is also found at the TU). Because of this different signal profile, calculating the median does not work very well because the result highly depends on how the probes are chosen, whether the window of the peak was chosen correctly and what the experimental conditions were (the tails depend on the DNA fragment size distribution). We therefore identified the maximum signal for TSS probes. The maximum enrichment is also commonly displayed in conventional representations of ChIP enrichment ratios and seems to be biologically meaningful[*]. In summary, the stalling index for each gene was calculated as the maximum enrichment at the TSS divided by the median enrichment across the TU. This means that on average the maximum(TSS) is slightly higher than the median(TU) at genes where the Pol II signal is uniformly distributed across the gene. We controlled for this difference in later steps and did not find that it influences the results in significant ways.

Low signal. If there is no Pol II present at a gene, the Stalling Index could become sensitive to noise. For example, the max(TSS) might be 1.2, whereas the median(TU) might be 0.3, producing an artificial Stalling Index of 4. Fortunately, the ChIP-chip data from this study did not display this level of noise. However, to prevent future problems (and criticism), we only calculated the Stalling Ratio for genes with a max(TSS) enrichment significantly above noise (Rosetta error model p value < 0.001).

Handling of alternative transcripts. At the ~16% of genes with several transcripts (because of alternative start sites or alternative splicing), a method was developed to automatically determine the transcript that is likely to be dominantly produced. First, the transcript with the highest maximum(TSS) was chosen. If equal, the transcript with the highest median(TU) was chosen. The maximum(TSS) was found to be a better indicator for dominant transcripts than the median(TU) because in cases were additional TSSs are found in the TU, the median(TU) is artificially inflated. If different transcripts were identical, both in terms of maximum(TSS) and median(TU) (because the transcript only differed in splicing or small bp differences that did not alter the probe choice), the transcripts were ranked alphabetically, e.g. CG2671-RA was chosen before CG2671-RC.

Control calculations and determination of cutoff values.

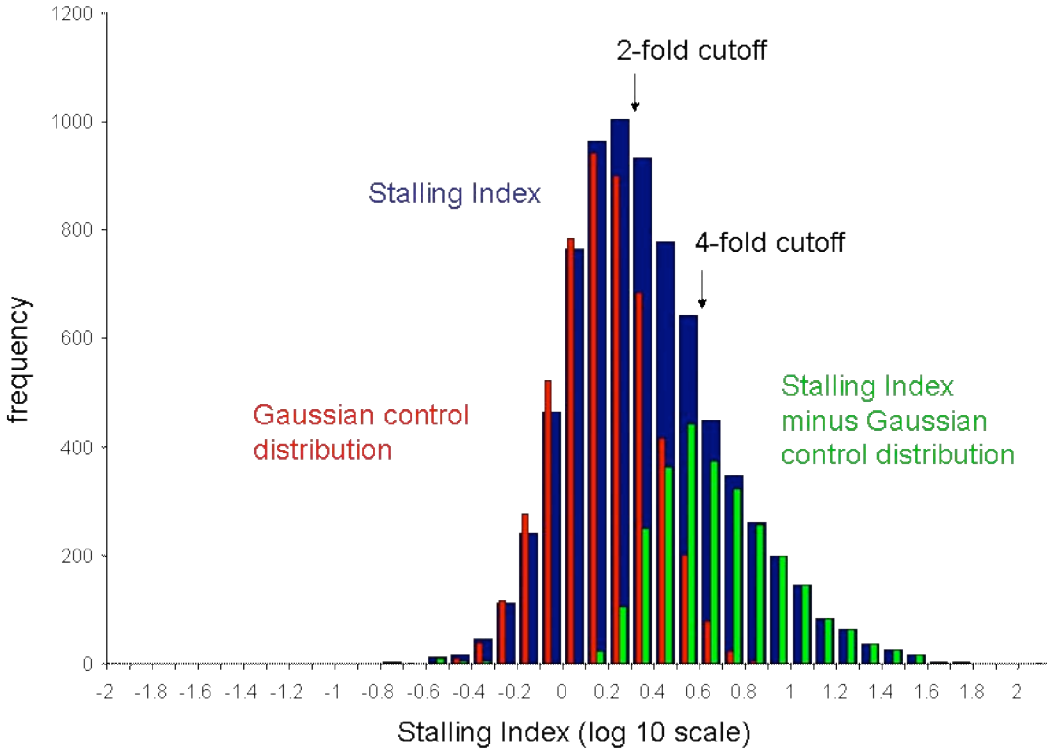
Once we felt that the Stalling Index calculations were optimized for robustness, the biggest challenge was to determine meaningful cut-off values that would classify a Pol II profile as active or stalled. The goal was to minimize false positives (due to random fluctuations in the data) without loosing too much sensitivity (too many genes classified as uncertain). To do this, we examined the distribution of all Stalling Indexes for all genes and performed several control calculations.

[*] It should be noted that the peak shape and the robustness of the maximum value may depend on the array type and design.

To separate (stalling) signal from noise, one needs an assumption about the distribution of noise. This can either be a theoretical assumption or one that is empirically derived. We tried both approaches and arrived at similar conclusions.

Gaussian distribution as control. If Pol II stalling did not exist, the assumption would be that the spread of ratios between max(TSS) and median(TU) is due to noise. We can approximate such noise using a Gaussian distribution. The mean of this Gaussian distribution, we expect to be slightly above 1 because the maximum is higher than the median on average. The standard deviation is fitted based on the distribution of the Stalling Indexes below 1 (left side of distribution), which is not affected by Pol II stalling (see Fig. S1).

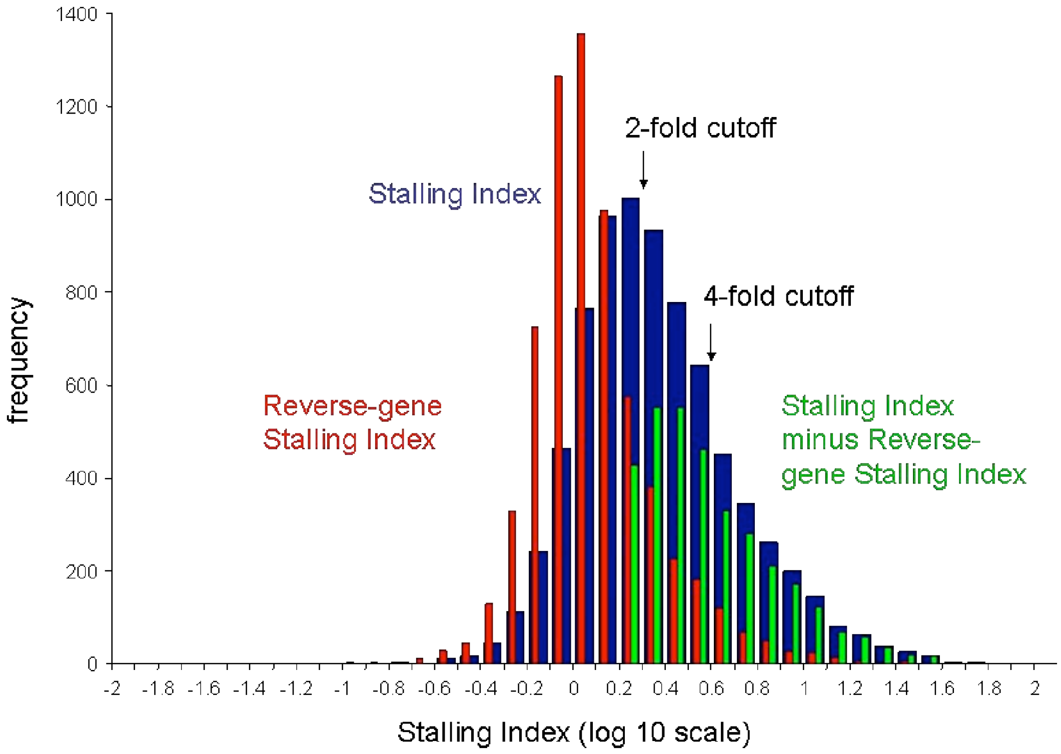
Fig. S1. Pol II Stalling Index distribution with Gaussian control



The histogram shows the frequencies of the Pol II Stalling Indexes across all genes in the genome (blue, n=13448). The distribution is asymmetric with higher Stalling Indexes disproportionately more frequent. Assuming a Gaussian control distribution for genes that are not stalled (red), a Pol II index higher than 4 is significantly above of what would be expected from noise alone. Furthermore, a Pol II index smaller than 2 is likely to occur through noise alone.

Reverse-gene control distribution. Another control distribution is based on the assumption that Pol II stalling does not occur at the end of genes. Thus, we calculated the Pol II index using the maximum ±300 bp from the end of the gene rather than from the TSS (effectively, the direction of the gene is reversed in the analysis). As expected, the distribution is more symmetric, with most values slightly above 1. However, there are significant occurrences of very high and very low values. A closer inspection of the data at these genes reveals that this is because signal from neighboring genes is detected. This occurs more often in the reverse-gene distribution because the intergenic space is sometimes small downstream of a gene but not upstream of a gene, where a promoter must be present. Nevertheless, the reverse-gene control distribution led us to the same conclusion as the Gaussian control distribution, namely that a Pol II index greater than 4 is likely to represent stalled Pol II, whereas a Pol II index smaller than 2 is not.

Fig. S2. Pol II Stalling Index distribution with reverse-gene control

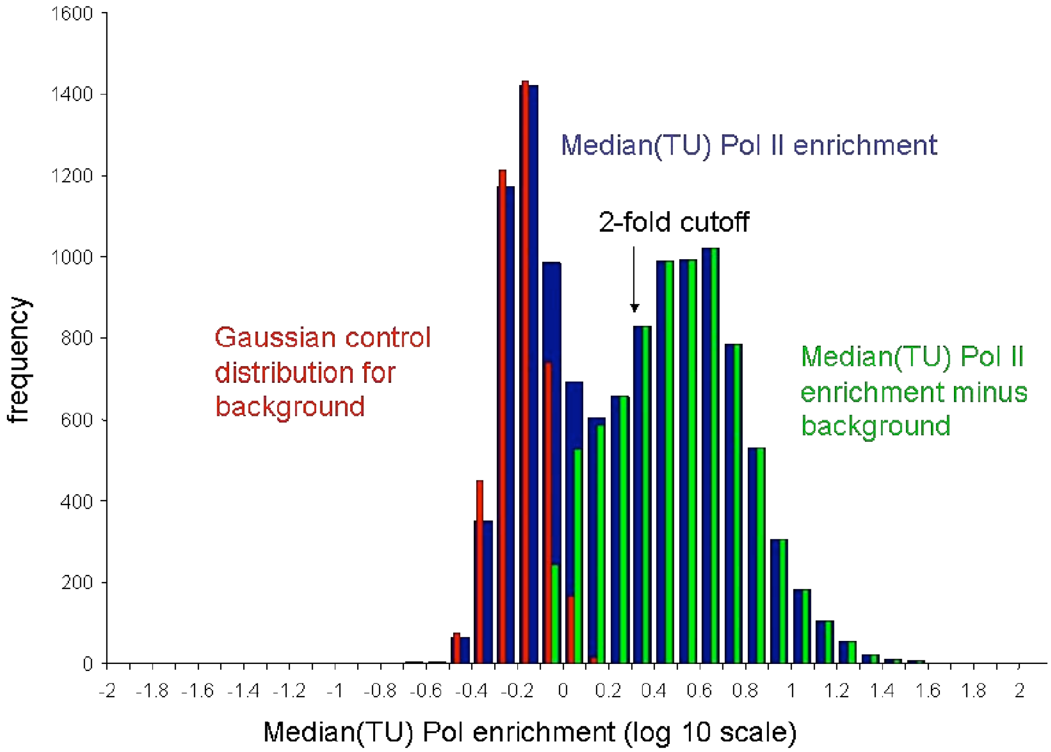


The histogram shows the frequencies of the Pol II Stalling Indexes across all genes in the genome (blue, n=13448). The distribution is asymmetric with higher Stalling Indexes disproportionately more frequent. Based on control distribution derived from reversing the orientation of all genes (red), a Pol II index higher than 4 is significantly above of what that in the control distribution. Furthermore, a Pol II index smaller than 2 is likely to occur through noise alone.

We also examined the distribution of the median(TU) across all genes in order to determine the value at which it is significantly above noise. It turns out that the distribution of these values shows a clear bimodal distribution (Fig. S3 and S4). (In fact, this distribution is commonly found for ChIP-chip data of Pol II but is rarely found for those of transcription factors. The latter look more like the distribution of the Stalling Indexes.) Due to the bimodal distribution, cutoffs can be determined fairly straight forward. It is safe to assume that Pol II is present at most genes with median Pol II values above 2-fold (Fig. S3).

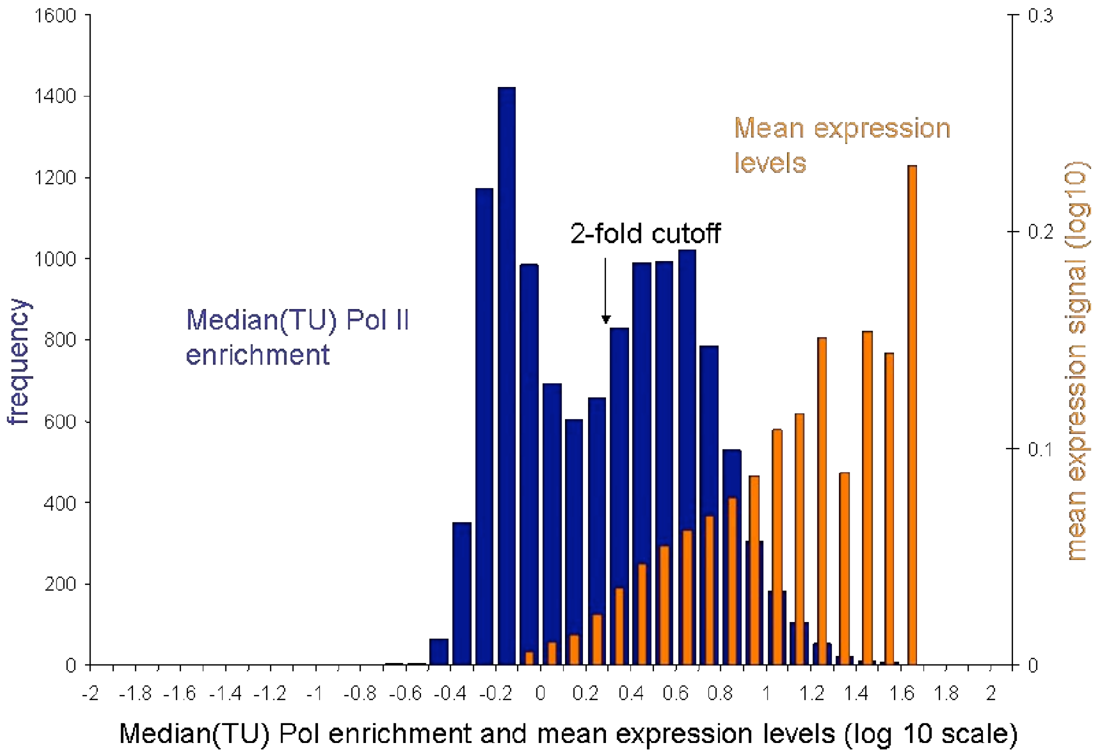
In addition, we tested the validity of the 2-fold cutoff by analyzing the relationship between the median(TU) Pol II data and the corresponding genome-wide transcripts levels of genes (Fig. S4).

Fig. S3. Median Pol II enrichment across genes and background noise.



The histogram shows the frequencies of Pol II Median(TU) values across all genes in the genome (blue, n=13448). The distribution is bimodal with the lower values arising through background noise. Assuming a Gaussian control distribution for this noise (red), the signal from Pol II occupancy can be estimated (green). A Pol II median(TU) of greater than 2 is likely to represent signal.

Fig. S4. Median Pol II enrichment across genes and expression levels.



The histogram shows the frequencies of Pol II Median(TU) values across all genes in the genome (blue, n=13448). For comparison the average expression levels (fold ratio signal to background from introns) is displayed. A Pol II median(TU) of greater than 2 is likely to reflect active Pol II with significant transcript production.

Gene classification

In accordance with the analysis described above, we classified genes the following way:

Stalled category: max(TSS) above noise and Pol II index > 4

Active gene category: max(TSS) above noise and Pol II index < 2

No Pol II category: max(TSS) below noise and median(TU) < 2 or absent

We were able to assign 76% of genes in this way. Of the genes that were not assigned, 18% had a Pol II index between 2 and 4, and 6 % were ambiguous because the gene was either too short or the known gene model was inconsistent with the Pol II binding data (e.g. median(TU)>2 but no enrichment above noise at TSS).

Comparison with other methods

While we developed this method with the best of our knowledge, we are aware that there is no perfect way to classify genes into the three groups. In fact, other groups (Church/Struhl or Adelman group) have developed slightly different methods, although the overall principle and conclusions remain the same. The main differences are the region of the gene at which data are collected (e.g. distance from TSS) and the method of calculation (e.g. maximum, median or mean). When we compared the effect of these parameters on the classification of genes, we found two major effects:

1) Since we are using a method that depends on a threshold for classification, any slight change of method will cause some genes to just make it above or below the threshold. Therefore, results from similar methods can easily differ by 10-20% when analyzing genomic microarray data.
2) The gene region used for the analysis significantly affects the results for genes that have several start sites or alternative transcripts, as well as for genes that are short or (presumably) mis-annotated. At these genes, the data can be conflicting and a simple classification into one of the groups seems inappropriate.

Analysis of transcript levels in Toll^{10b} embryos

To determine transcript levels in *Toll^{10b}* embryos, we calculated the median transcript levels of all known genes from whole-genome tiling arrays (Biemar et al. 2006). For this, we first determined the median value for each probe from triplicate experiments, collected all values from each exon of a gene and then determined the median value for each gene. To determine the background signal, we calculated the median of all gene introns. The fold-ratio transcript levels as displayed in Fig. 2 are calculated as the ratio between the median exon signal of each gene and the median intron signal of all genes (background).

We defined genes with stalled Pol II that are tightly restricted to the TSS as those stalled genes that have a median enrichment of < 2 (n= 996). To obtain active genes with similar levels of Pol II enrichment near the TSS, we searched with each maximum(TSS) of the stalled genes and identified the active gene with the closest maximum(TSS).

Metagene analysis

To determine the metagene profile, we aligned all genes in each class at their transcriptional start site and calculated the average enrichment for each position. The algorithm is similar to the one used by Pokholok et al. 2005.

GO and IMAGO analysis

We analyzed the functions of the genes found in the three classes (active, stalled and no Pol II) using the gene-ontology (GO) annotation (Ashburner et al. 2000) or embryonic RNA *in situ* hybridization patterns (ImaGO) annotation (Tomancak et al. 2002). For this, we determined the number of genes in each GO category (n= 4693) and ImaGO category (n=345) in each of the three classes and assessed whether the category was significantly over-represented and under-represented. The p-value was calculated based on the hypergeometric distribution (i.e. drawing without replacement) with and without Bonferroni correction.

We then selected the major over-represented categories at the top of each list and displayed them graphically in Fig. 3. The statistics for these data is shown in Table S2 and Table S3. Only categories with p-value < 10^{-10} were considered. Due to the hierarchical structure of both databases where many genes are shared between categories, we displayed the p-values without Bonferroni correction, which would be an overly conservative correction (yet the reported p-values would remain highly significant).

Table S2. The three Pol II classes and ImaGO categories

	Embryo stage	IMAGO Description	Gene overlap	Total genes in IMAGO category	Total genes in Pol II category	Total genes in any IMAGO category	P-value over-representation	P-value under-representation
No Pol II	stage 1-3	no staining (1-3)	928	1883	1728	6315	4.27E-136	1
	stage 7-8	ubiquitous (7-8)	66	1280	1728	6315	1	1.28E-111
	stage 4-6	subset (4-6)	39	236	1728	6315	0.999977	4.72E-05
	stage 4-6	ectoderm AISN (4-6)	25	195	1728	6315	1.000000	5.62E-07
Active Pol II	stage 1-3	no staining (1-3)	464	1883	2435	6315	1	1.12E-51
	stage 7-8	ubiquitous (7-8)	798	1280	2435	6315	1.93E-83	1
	stage 4-6	subset (4-6)	75	236	2435	6315	0.988469	0.016485
	stage 4-6	ectoderm AISN (4-6)	68	195	2435	6315	0.8751811	0.158687
Stalled Pol II	stage 1-3	no staining (1-3)	206	1883	946	6315	1.000000	1.19E-09
	stage 7-8	ubiquitous (7-8)	92	1280	946	6315	1	5.10E-21
	stage 4-6	subset (4-6)	100	236	946	6315	2.91E-25	1
	stage 4-6	ectoderm AISN (4-6)	83	195	946	6315	3.17E-21	1

Table S3. The three Pol II classes and GO categories

	GO term	Description	Gene overlap	Total genes in GO category	Total genes in Pol II category	Total genes in any GO category	P-value over-representation	P-value under-representation
No Pol II	GO:0001584	rhodopsin-like receptor activity	133	170	2410	6911	1.28E-31	1
	GO:0042302	structural constituent of cuticle	75	79	2410	6911	6.58E-30	1
	GO:0008152	metabolism	1073	3820	2410	6911	1	1.25E-39
	GO:0008283	cell proliferation	106	685	2410	6911	1	9.21E-33
	GO:0007275	development	318	1236	2410	6911	1	2.10E-14
	GO:0007399	neurogenesis	94	381	2410	6911	0.999996	6.84E-06
	GO:0007398	ectoderm development	66	194	2410	6911	0.626329	0.432966
	GO:0007517	muscle development	18	75	2410	6911	0.984721	0.028541
Active Pol II	GO:0001584	rhodopsin-like receptor activity	15	170	2248	6911	1	1.32E-13
	GO:0042302	structural constituent of cuticle	1	79	2248	6911	1	1.01E-12
	GO:0008152	metabolism	1455	3820	2248	6911	1.73E-28	1
	GO:0008283	cell proliferation	338	685	2248	6911	4.70E-22	1
	GO:0007275	development	408	1236	2248	6911	0.356549	0.668061
	GO:0007399	neurogenesis	114	381	2248	6911	0.880185	0.144153
	GO:0007398	ectoderm development	45	194	2248	6911	0.998512	0.002519
	GO:0007517	muscle development	19	75	2248	6911	0.930645	0.110980
Stalled Pol II	GO:0001584	rhodopsin-like receptor activity	12	170	1002	6911	0.999189	0.001964
	GO:0042302	structural constituent of cuticle	1	79	1002	6911	0.999996	5.71E-05
	GO:0008152	metabolism	555	3820	1002	6911	0.482461	0.544864
	GO:0008283	cell proliferation	112	685	1002	6911	0.083232	0.932498
	GO:0007275	development	344	1236	1002	6911	1.03E-42	1
	GO:0007399	neurogenesis	137	381	1002	6911	3.45E-27	1
	GO:0007398	ectoderm development	79	194	1002	6911	1.18E-19	1
	GO:0007517	muscle development	38	75	1002	6911	1.10E-13	1

Analysis of signal transduction pathways components

To determine which signal transduction pathway genes contain stalled Pol II, we tested for enrichment of KEGG categories among the three Pol II groups (Fig. S6). The p-value was calculated based on the hypergeometric distribution. As expected, stalled Pol II genes are enriched for signal transduction pathways, in particular the JAK-STAT, Notch, MAPK, Wnt and TGF-beta pathways. The p-values are not as significant as with GO and IMAGO analysis because the KEGG pathway members are not well documented for *Drosophila* and the numbers in each group are small. We therefore annotated the signal transduction genes with stalled Pol II by hand (Table S4).

Fig. S5. KEGG pathways enrichment

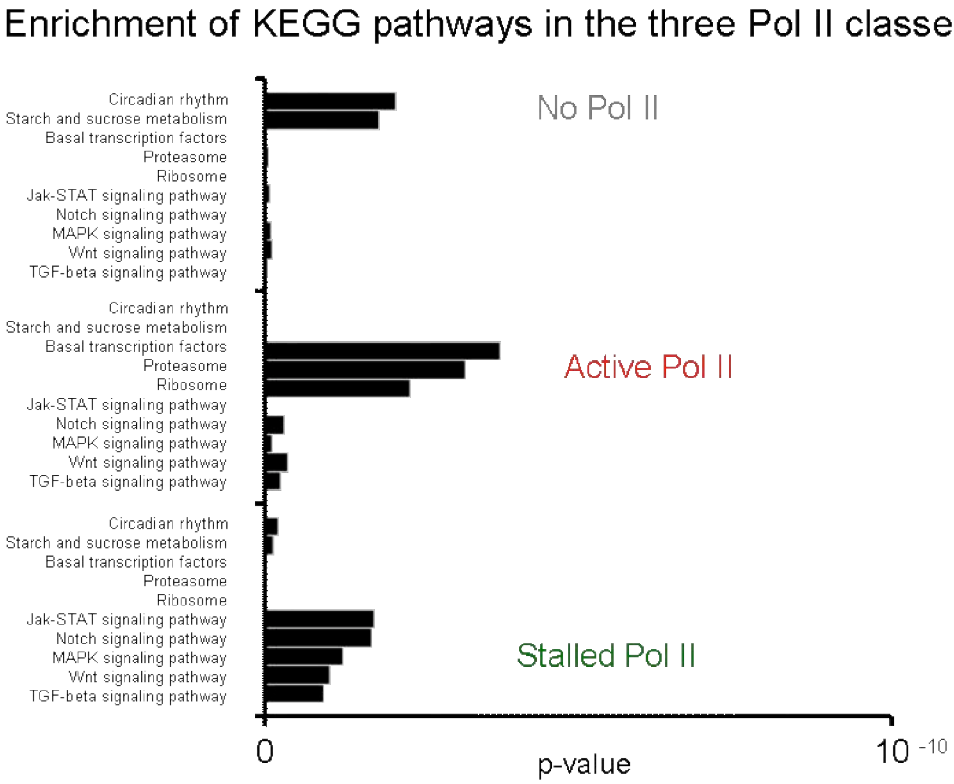


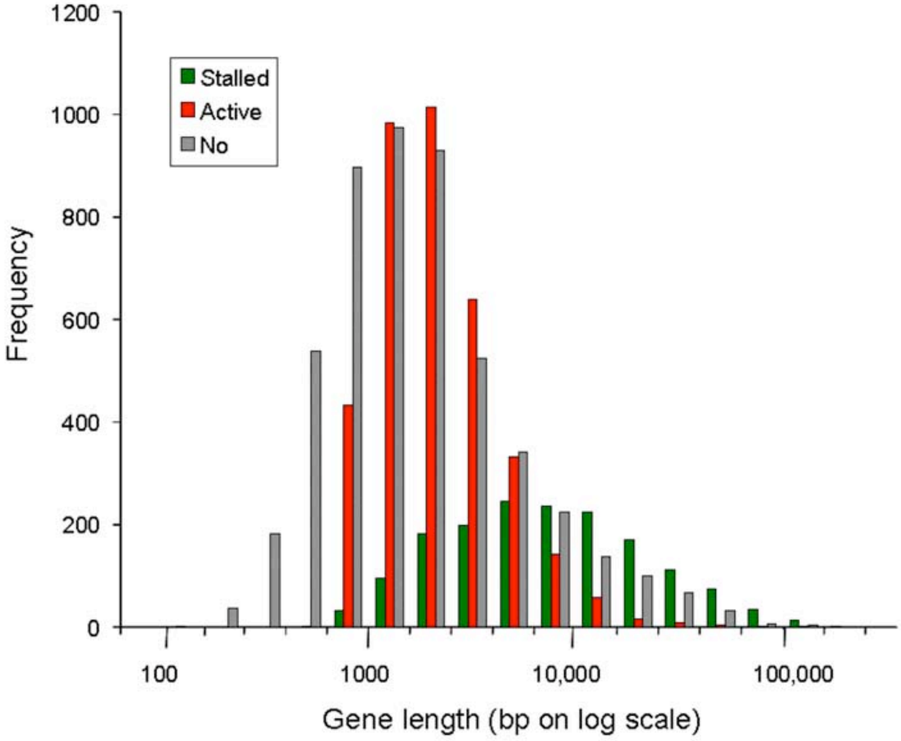
Table S4. Signal transduction genes with stalled Pol II

Signaling Pathway	Gene	Name
Jak-STAT	CG10155	CG10155
	CG6033	drk
	CG1594	hop
	CG2699	Pi3K21B
	CG1921	sty
Notch	CG7147	kuz
	CG8118	mam
	CG3936	N
	CG3779	numb
	CG3497	Su(H)
TGF-beta	CG5201	Dad
	CG9885	dpp
	CG4943	lack
	CG7904	put
	CG8416	Rho1
	CG9224	sog
Wnt	CG14622	CG14622
	CG2185	CG2185
	CG4974	dally
	CG4974	dally
	CG32146	dlp
	CG17348	drl
	CG17697	fz
	CG9739	fz2
	CG9739	fz2
	CG16785	fz3
	CG4379	Pka-C1
	CG8416	Rho1
	CG3135	shf
	CG4889	wg
	CG1916	Wnt2
	CG1916	Wnt2
	CG4698	Wnt4
	CG6407	Wnt5
MAPK		
EGF	CG4531	argos
	CG4426	ast
	CG12283	kek1
	CG17077	pnt
	CG1004	rho
	CG4385	S
TNF	CG12919	egr
	CG6531	wgn
MAPK -cont.		
FGF	CG4608	bnl
	CG32134	btl
	CG1921	sty
PVR	CG7103	Pvf1
	CG13780	Pvf2
	CG8222	Pvr
ERK	CG3166	aop
	CG6033	drk
	CG6721	Gap1
	CG9768	hkb
JNK	CG15509	kay
	CG7850	puc
other	CG3954	csw
	CG12244	lic
	CG10379	mbc
	CG2049	Pkn
	CG1697	rho-4
	CG33304	rho-5
	CG17212	rho-6
	CG8972	rho-7
	CG5701	RhoBTB
	CG1976	RhoGAP100F
	CG1748	RhoGAP102A
	CG4937	RhoGAP15B
	CG7122	RhoGAP16F
	CG7481	RhoGAP18B
	CG1412	RhoGAP19D
	CG6477	RhoGAP54D
	CG3208	RhoGAP5A
	CG6811	RhoGAP68F
	CG32149	RhoGAP71E
	CG31319	RhoGAP88C
	CG4755	RhoGAP92B
	CG3421	RhoGAP93B
	CG32555	RhoGAPp190
	CG7823	RhoGDI
	CG9635	RhoGEF2
	CG1225	RhoGEF3
	CG8606	RhoGEF4
	CG9366	RhoL
	CG18497	spen

Other analyses

Because of the functional differences between the stalled, active and no Pol II categories, there are other gene characteristics associated with each group. For example, genes that are stalled tend to be longer (Fig. S6) and have more introns and regulatory regions. These properties are typical for developmental control genes.

Fig. S6. Pol II categories and gene length



Association of Snail-repressed genes with stalled Pol II

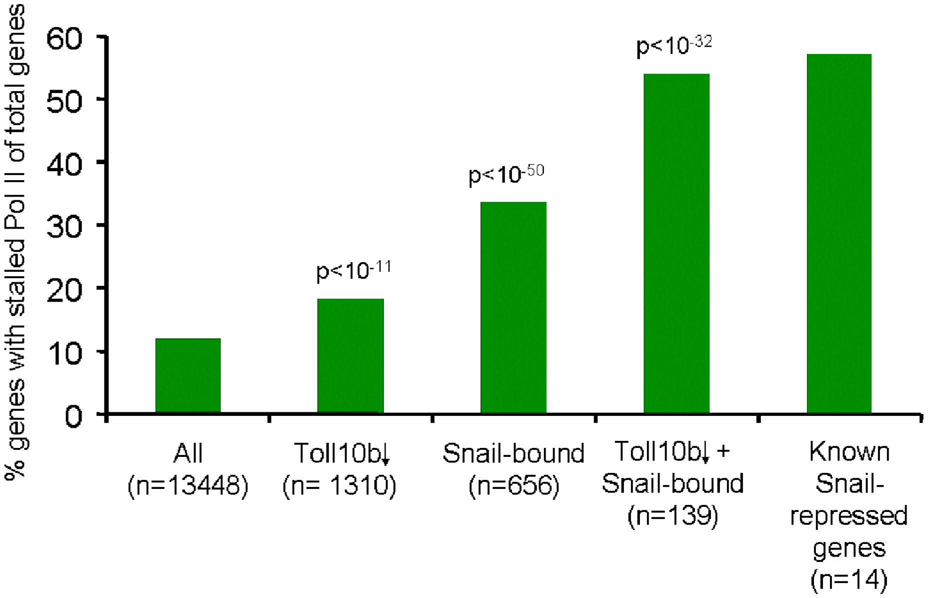
We found that genes repressed by Snail-are highly enriched in genes with stalled Pol II. Testing this hypothesis was not trivial for two reasons:

First, Snail binding to genes is not sufficient by itself to mediate repression. For example, we have previously shown that Snail also binds to genes that are specifically activated in *Toll*^{10b} mutants (Zeitlinger et al. 2007).

Second, when selecting genes based on lower expression in *Toll*^{10b} mutants versus mutants where Snail is not present, we select for genes that are expressed at low levels and thus are more likely to be stalled.

We therefore combined the two criteria – binding by Snail and lower expression – in a pair-wise fashion (Fig. S5). The significance was calculated using the hypergeometric distribution. Genes that fulfilled our expression criteria were those genes with 3-fold lower transcript levels in *Toll*^{10b} mutants versus either *Toll*$^{rm9/rm10}$ or *pipe* mutants (Stathopoulos et al. 2002). In *Toll*$^{rm9/rm10}$ mutants, cells develop into neurectodermal precursors, whereas in *pipe* mutants, cells acquire dorsal ectodermal fate. The 3-fold cutoff was chosen because this was the cutoff that the authors of the original expression study used.

Fig. S7. Preferential Pol II stalling among Snail-repressed genes

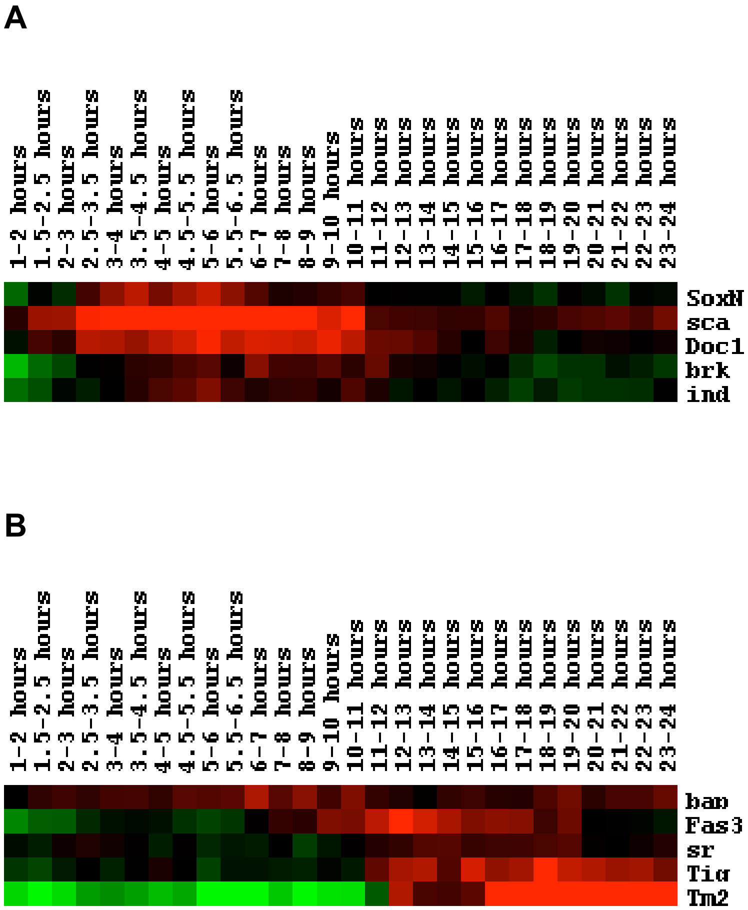


Among all genes, 12 % of genes show stalled Pol II. 18% of repressed genes show Pol II stalling ($p<10^{-11}$). Among genes that are bound by Snail, 34% show stalled Pol II ($p<10^{-50}$). Among all repressed genes that are also bound by Snail (Zeitlinger et al. 2007), 54% of genes show Pol II stalling ($p<10^{-32}$ compared to all genes or $p<10^{-23}$ compared to repressed genes). This percentage is similar to the one found at all well characterized Snail target genes.

Analysis of genes that are rapidly induced during embryogenesis

We used the data by Hooper et al. 2007, which analyzed the expression experiments by Arbeitman et al. 2002 and identified sharp transcript changes during *Drosophila* embryogenesis. We used all genes that show a sharp increase of expression with a maximum between 4-16 hours of development. Many of these also show a later decrease in expression (class II genes), whereas some genes only increase in expression (class III genes). We found that more than 24% of these genes (229 out of 937) display stalled Pol II, as compared to 12% that are expected by chance.

Fig. S8. Stalled Pol II genes and expression time-course



Examples of genes that have stalled Pol II and that are rapidly induced (red versus green) during embryogenesis. (A) Genes that are also repressed by Snail, (B) genes that are expressed during muscle development.

Analysis of genes that regulate muscle development

To identify genes that are not only induced after the time-frame of the analysis but are also expressed in the cell type that *Toll*^{10b} embryos could give rise to, we analyzed the Pol II binding pattern at a more comprehensive, experimentally derived list of genes expressed during muscle development (Sandmann et al. 2006). We selected all 260 genes that were either mentioned by name in the manuscript or were listed in Supplementary Materials. To identify the genes that are not yet expressed in muscle at the time frame of our analysis, we selected the subset of muscle genes with no detectable transcript in *Toll*^{10b} mutant embryos. We found that among these genes, 35% (44 out of 102) show stalled Pol II, although only 8% of all silent genes show stalled Pol II.

Table S5. Stalled Pol II at muscle genes

Gene	Flybase ID	Symbol	Gene	Flybase ID	Symbol
CG4807	FBgn0000011	ab	CG2679	FBgn0004919	gol
CG8376	FBgn0000099	ap	CG9042	FBgn0001128	Gpdh
CG4531	FBgn0004569	argos	CG8346	FBgn0002609	HLHm3
CG7902	FBgn0004862	bap	CG14548	FBgn0002733	HLHmbeta
CG10021	FBgn0004893	bowl	CG11312	FBgn0011674	insc
CG17124	FBgn0032297	CG17124	CG10197	FBgn0001319	kn
CG17181	FBgn0035144	CG17181	CG6545	FBgn0011278	lbe
CG18854	FBgn0042174	CG18854	CG33197	FBgn0053197	mbl
CG2330	FBgn0037447	CG2330	CG10145	FBgn0020269	mspo
CG30460	FBgn0050460	CG30460	CG8967	FBgn0004839	otk
CG31038	FBgn0051038	CG31038	CG9811	FBgn0034434	Rgk1
CG31781	FBgn0051781	CG31781	CG8643	FBgn0033310	rgr
CG3624	FBgn0034724	CG3624	CG6534	FBgn0002941	slou
CG6330	FBgn0039464	CG6330	CG11121	FBgn0003460	so
CG8547	FBgn0033919	CG8547	CG5557	FBgn0010768	sqz
CG8713	FBgn0033257	CG8713	CG7847	FBgn0003499	sr
CG9416	FBgn0034438	CG9416	CG11502	FBgn0003651	svp
CG5441	FBgn0008649	dei	CG11527	FBgn0011722	Tig
CG9885	FBgn0000490	dpp	CG4843	FBgn0004117	Tm2
CG17348	FBgn0015380	drl	CG6863	FBgn0004885	tok
CG9554	FBgn0000320	eya	CG10388	FBgn0003944	Ubx
CG5803	FBgn0000636	Fas3	CG7178	FBgn0004028	wupA
CG6992	FBgn0004620	GluRIIA			

The table shows genes involved in muscle development that are not expressed in *Toll*^{10b} embryos. Pol II stalling at these genes in mesodermal precursors may prepare them for later activation.

Genes with stalled PolII important for muscle development were also identified using the IMAGO and GO categories (see above).

Permanganate footprints of muscle genes prior to activation

For the permanganate footprint assays of muscle genes, we used wild-type embryos to ensure that the observed footprint is not an artifact of the *Toll*^{10b} embryos but occurs during wild-type development.

The use of wild-type embryos raises the question of whether PolII stalling at muscle genes is specific to mesodermal cells. The Pol II profile of muscle genes in *Toll*^{rm9}/*Toll*^{rm9} and *gd*^{7} embryos suggests that at some genes, stalled PolII is mesoderm-specific (e.g. *bap* and *tin*), while at others (e.g. *Dr* and *lbe*), stalled Pol II is found throughout the embryo (data not shown).

References

Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., and White, K.P. 2002. Gene expression during the life cycle of Drosophila melanogaster. Science 297: 2270-5.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-9.

Biemar, F., Nix, D.A., Piel, J., Peterson, B., Ronshaugen, M., Sementchenko, V., Bell, I., Manak, J.R., and Levine, M.S. 2006. Comprehensive identification of Drosophila dorsal-ventral patterning genes using a whole-genome tiling array. Proc Natl Acad Sci U S A 103: 12763-8.

Boehm, A.K., Saunders, A., Werner, J., and Lis, J.T. 2003. Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock. Mol Cell Biol 23: 7628-37.

Gilmour, D.S. and Lis, J.T. 1986. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. Mol Cell Biol 6: 3984-9.

Hooper, S.D., Boue, S., Krause, R., Jensen, L.J., Mason, C.E., Ghanim, M., White, K.P., Furlong, E.E., and Bork, P. 2007. Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. Mol Syst Biol 3: 72.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005. A high-resolution map of active promoters in the human genome. Nature 436: 876-80.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., Koseki, H., Fuchikami, T., Abe, K., Murray, H.L., Zucker, J.P., Yuan, B., Bell, G.W., Herbolsheimer, E., Hannett, N.M., Sun, K., Odom, D.T., Otte, A.P., Volkert, T.L., Bartel, D.P., Melton, D.A., Gifford, D.K., Jaenisch, R., and Young, R.A. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125: 301-13.

Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D.K., and Young, R.A. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517-27.

Reppas, N.B., Wade, J.T., Church, G.M., and Struhl, K. 2006. The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting. Mol Cell 24: 747-57.

Sandmann, T., Jensen, L.J., Jakobsen, J.S., Karzynski, M.M., Eichenlaub, M.P., Bork, P., and Furlong, E.E. 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. Dev Cell 10: 797-807.

Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M., and Levine, M. 2002. Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo. Cell 111: 687-701.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., and Rubin, G.M. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3: research0088.1-0088.14.

Wang, X., Lee, C., Gilmour, D.S., and Gergen, J.P. 2007. Transcription elongation controls cell fate specification in the Drosophila embryo. Genes Dev 21: 1031-6.

Zeitlinger, J., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., and Levine, M. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev 21: 385-90.