Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level

Elizabeth A. Rach[1]◔, Deborah R. Winter[1]◔, Ashlee M. Benjamin[1], David L. Corcoran[2], Ting Ni[2,3]¤, Jun Zhu[2,3]¤, Uwe Ohler[2,4,5]*

1 Program in Computational Biology and Bioinformatics, Duke University, Durham, North Carolina, United States of America, 2 Institute for Genome Sciences and Policy, Duke University Medical Center, Durham, North Carolina, United States of America, 3 Department of Cell Biology, Duke University, Durham, North Carolina, United States of America, 4 Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, United States of America, 5 Department of Computer Science, Duke University, Durham, North Carolina, United States of America

Abstract

The application of deep sequencing to map 5′ capped transcripts has confirmed the existence of at least two distinct promoter classes in metazoans: "focused" promoters with transcription start sites (TSSs) that occur in a narrowly defined genomic span and "dispersed" promoters with TSSs that are spread over a larger window. Previous studies have explored the presence of genomic features, such as CpG islands and sequence motifs, in these promoter classes, but virtually no studies have directly investigated the relationship with chromatin features. Here, we show that promoter classes are significantly differentiated by nucleosome organization and chromatin structure. Dispersed promoters display higher associations with well-positioned nucleosomes downstream of the TSS and a more clearly defined nucleosome free region upstream, while focused promoters have a less organized nucleosome structure, yet higher presence of RNA polymerase II. These differences extend to histone variants (H2A.Z) and marks (H3K4 methylation), as well as insulator binding (such as CTCF), independent of the expression levels of affected genes. Notably, differences are conserved across mammals and flies, and they provide for a clearer separation of promoter architectures than the presence and absence of CpG islands or the occurrence of stalled RNA polymerase. Computational models support the stronger contribution of chromatin features to the definition of dispersed promoters compared to focused start sites. Our results show that promoter classes defined from 5′ capped transcripts not only reflect differences in the initiation process at the core promoter but also are indicative of divergent transcriptional programs established within gene-proximal nucleosome organization.

Citation: Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, et al. (2011) Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level. PLoS Genet 7(1): e1001274. doi:10.1371/journal.pgen.1001274

Editor: Jason D. Lieb, The University of North Carolina at Chapel Hill, United States of America

Received June 17, 2010; **Accepted** December 13, 2010; **Published** January 13, 2011

Funding: Funding for this research was provided by NIH R01 HG 004065 to UO and JZ. DRW was supported by an NSERC award from the Canadian government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: Uwe.Ohler@duke.edu

◔ These authors contributed equally to this work.

¤ Current address: Genetics and Development Biology Center, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, United States of America

Introduction

The development of high-throughput sequencing strategies, which generate millions of 5′ sequence tags from capped RNAs transcribed by RNA polymerase II (pol II), has enabled obtaining fine-grained pictures of transcription initiation. Each of the tags originates from a transcription start site (TSSs), and mapping the tags to the genome identifies tag clusters for individual genes. In particular, the application of Cap Analysis of Gene Expression (CAGE) produced comprehensive data sets for mammalian promoters [1], and an extension of this methodology to Paired End Analysis of Transcription Start Sites (PEAT) was used to map and cluster millions of paired reads from *Drosophila melanogaster* embryos [2]. Tag clusters exhibit different initiation patterns, i.e. distributions of tags within a cluster, and have been used to define distinct promoter classes, generally falling into two basic groups: Both flies and mammals have focused promoters in which transcription occurs within a narrow genomic window of a few nucleotides, and dispersed promoters in which TSSs spread out over a larger genomic region on the order of a hundred nucleotides. Promoter classes have distinct associations to core promoter motifs and functional roles [3,4], and evidence has pointed towards enriched pausing, or stalling, of *Drosophila* pol II at focused promoters [5].

Many studies have shown a generic pattern of chromatin organization in promoters, in which a nucleosome free region (NFR) upstream of the TSS is surrounded by periodic arrangements of nucleosomes within the transcript and further upstream [6,7], illustrating the connection between chromatin features and the accessibility of the DNA to transcription factors (TFs). Nucleosomes containing H2 and H3 histone variants provide particularly strong signals for the beginnings of genes in eukaryotes [6,8,9], as they are preferentially incorporated in or near areas of active transcription. Data on frequent modifications to the N-terminal histone tails have furthermore supported a histone code specifying functional domains in the genome; for instance, the tri-

Author Summary

How are genes transcribed at the right levels and under the right conditions? Transcription regulation in eukaryotes has long been proposed to work by a division of labor: ubiquitous DNA sequence features in the core promoter region, close to the transcription start site (TSS) of genes, were thought to generically encode information to recruit RNA polymerase to initiate transcription, while specific sequence features, often distal from the genes, were thought to boost expression under the right conditions. Supporting the generic function of core promoters, genome-wide chromatin maps showed a stereotypical arrangement of well-spaced nucleosomes providing access to the TSS. High-throughput sequencing has generated genome-wide TSS maps at high resolution, which show that promoters exhibit different initiation patterns, ranging from focused start sites to dispersed regions. Linking these patterns to chromatin maps, we now find distinct core promoter classes, those in which the TSS location is defined broadly on the chromatin level and those in which the TSS is defined by precisely positioned sequence features. Notably, these architectures are conserved deeply across eukaryotes and are used for different functional classes of genes. Our work adds to the increasing understanding that core promoters contribute significantly to the complexity of eukaryotic gene expression.

methylation of H3K4 has been shown to mark the promoter regions surrounding TSSs [10]. In addition, individual instances of insulator elements have been shown or suggested to play a role in chromatin remodeling near promoter regions [11,12].

Given that the distinct promoter classes are widely conserved throughout metazoans, and nucleosomes are correlated with the accessibility of the DNA, it may be surprising that virtually no analysis has so far has directly examined whether focused or dispersed promoters are associated with different nucleosome organization and chromatin structure. Instead, the majority of reports have taken the approach of dividing genes according to chromatin or insulator patterns, and then associating the promoters in each group with sequence features [6,13] or function [14,15]. One of the main limitations of this approach has been that these characteristics are present in only a fraction of promoters. For instance, the TATA box motif is present in only ~10–20% of all eukaryotic promoters, and ~35% of focused promoters [16]. On the other hand, CpG islands are a very frequent sequence feature of mammalian regulatory regions [17,18] and have been repeatedly associated with dispersed promoters. Yet, this property is by far not unique to one initiation pattern: depending on the definition, ~70–80% of dispersed promoters coincide with the presence of a CpG island, but ~50–60% of focused promoters do so as well (Table 1). Furthermore, while chromatin features and initiation patterns are conserved at least in metazoans, CpG islands do not exist in the fruit fly genome [19], suggesting that specific sequence features may lead to enrichments but not be the sole or primary indicators of the underlying process.

In this work, we show that promoter classes defined on patterns of transcription initiation are mirrored by significant differences in nucleosome organization and histone modifications, confirming the presence of divergent strategies of transcription, as recently proposed for yeast and for special functional classes of mammalian genes [20,21]. These differences are further supported by distinct associations to recently defined *Drosophila* insulator classes [22], and are consistently present across changing expression levels, polymerase stalling, and promoters with or without CpG islands. Furthermore, computational models based on chromatin features show strong differences in their ability to identify initiation sites from the different promoter classes. Our findings are conserved between humans and flies and thus show that the initiation patterns are signatures of fundamental and divergent strategies of gene regulation across eukaryotes.

Table 1. Distribution of Promoters in the Human and Fly Datasets Used in This Study.

	HUMAN		FLY
Class	CpG/total (Frommer)	CpG/total (Jones)	TATA/total
NP	827/1409 (58.7%)	689/1409 (48.9%)	179/517 (34.6%)
BP	1375/1759 (78.2%)	1130/1759 (64.2%)	51/406 (12.6%)
WP	6510/7656 (85.0%)	5244/7656 (68.5%)	74/1054 (7.1%)

The table lists the number of promoters in each class, and indicates the presence of CpG islands (human) or TATA boxes (fly) within classes. As the table shows, individual sequence features are enriched in certain promoter classes, but any single feature does not cover any of the classes completely. CpG islands were defined using two sets of criteria: the classic definition of Gardiner-Garden & Frommer [57], and the more stringent definition of Takai & Jones [30] which aims at a better separation from Alu-repetitive elements. TATA-containing promoters were taken from [2].

doi:10.1371/journal.pgen.1001274.t001

Results

Promoter Classes Exhibit Significant Differences in Nucleosome Organization

Studies in different metazoans have identified several promoter classes based on the size of the initiation region and the distribution of initiation events within each region [1]. In our previous work in *Drosophila* [2], we defined three specific classes: Narrow Peak (NP) promoters are typical focused promoters with high occurrences of initiation at one location. They typically contain one or more canonical position-specific core promoter motifs such as the TATA box, which have been found in genes with developmental regulation and tissue-specific functions. Conversely, Weak Peak (WP) promoters are dispersed promoters, in which transcription is distributed over a larger genomic span and lacks a clear preference for a single start site. In flies, WP promoters are associated with distinct core promoter sequence elements but largely lack the canonical eukaryotic-wide core promoter motifs, and are frequently associated with housekeeping genes [14,23]. CpG islands, long stretches of CpG dinucleotides that play a role in chromatin packing and nucleosome organization [24,25], are a feature of most mammalian promoters and are more frequently present in WP promoters [1] (Table 1). Finally, an intermediate class, Broad with Peak (BP) promoters, displays both a preference for a narrow location as in NP promoters, yet with tags covering a larger genomic span as in WP promoters.

We determined TSS clusters from available human CAGE tags in the FANTOM4 database [26] (see Methods). 13% of promoter clusters fell into the NP class, 16% into the BP class, and 71% were classified as WP. We evaluated the chromatin structure within each of these promoter classes using several genome-wide datasets reflecting the positions of bulk nucleosomes, histone variants, and histone marks. We first examined H2A.Z profiles in human CD4+ T cells [10], as this histone variant has been associated with clearer signals in promoters compared to bulk nucleosomes [6]. Both BP and WP promoters showed the stereotypic confirmation of well-spaced nucleosomes upstream and downstream of the TSS, divided by a nucleosome free region. The relative locations of H2A.Z nucleosomes, and the 185 bp spacing between them, agreed with previous estimates [12,27]. However, NP promoters clearly did not fit this picture, as BP and WP promoters had a consistently higher association with H2A.Z nucleosome organization than NP (Figure 1A), with the strongest divergence observed at the +1 nucleosome ($p<10E-90$). Examining bulk nucleosome locations [7] confirmed these differences: BP and WP promoters showed defined nucleosome positions and spacing and thus a distinctly higher association with bulk nucleosome organization than NP promoters (Figure 1B). At the +1 position, WP and BP promoters showed significantly higher levels compared to a baseline calculated from random genomic locations.

To test whether these observations were reflected in DNase Hypersensitivity Sites (DHS) which reflect the accessibility of DNA by DNaseI digestion, we evaluated DHS profiles from the same human cell line. Previous studies reported that most promoters were accompanied by a DHS site [28]. However, in agreement with the NFR differences we observed between bulk nucleosome profiles, WP and BP promoters demonstrated a significantly higher peak at the NFR (~100 bp upstream), appearing at least twice as sensitive to DNase when compared with NP promoters (Figure 1C, $p<10E-56$). Notably, the increase in accessibility was not accompanied by higher levels of pol II; rather, NP and BP promoters had elevated amounts of pol II on average compared to WP promoters (Figure 1D).

The above analyses uncovered a clear division of promoters by nucleosome organization, quantified by different genome wide assays: dispersed promoters exhibited a clearly defined periodic nucleosome organization, whereas focused promoters were less organized at the chromatin level, ruling out the possibility that narrow initiation events were defined by tight nucleosome locations. To illustrate this in more detail, we plotted the distribution of H2A.Z nucleosomes within each promoter as a heatmap (Figure 2). Individual WP and BP promoters had more clearly defined nucleosome positions, and NP promoters displayed less organization and lower concentrations around specific locations. An unsupervised clustering of all promoters, based on bulk and H2A.Z nucleosomes, recovered these distinct nucleosome profiles, with clear enrichments for specific initiation patterns (Figure S1).

The Presence of CpG Islands Alone Does Not Explain Differences in Nucleosome Organization

CpG islands have frequently been used to split mammalian promoters into two distinct classes for TSS modeling or promoter analysis [17,20], and CpG island-containing promoters have been reported to show stronger nucleosome associations [20,29]. Thus, we examined whether the presence of CpG islands would recapitulate the divergent chromatin modes we observed for different initiation patterns. We extracted annotated CpG islands from the UCSC genome browser and determined the overlap of CpG islands as defined by Takai & Jones [30] with the promoters in our three classes. As previously reported [1], there were higher percentages of CpG islands at WP (69%) and BP (64%) promoters, compared to NP (49%) promoters (cf. Table 1). However, regardless of the presence of CpG islands, BP and WP promoters had significantly higher associations to nucleosomes than NP promoters. Likewise, promoters within the same class maintained

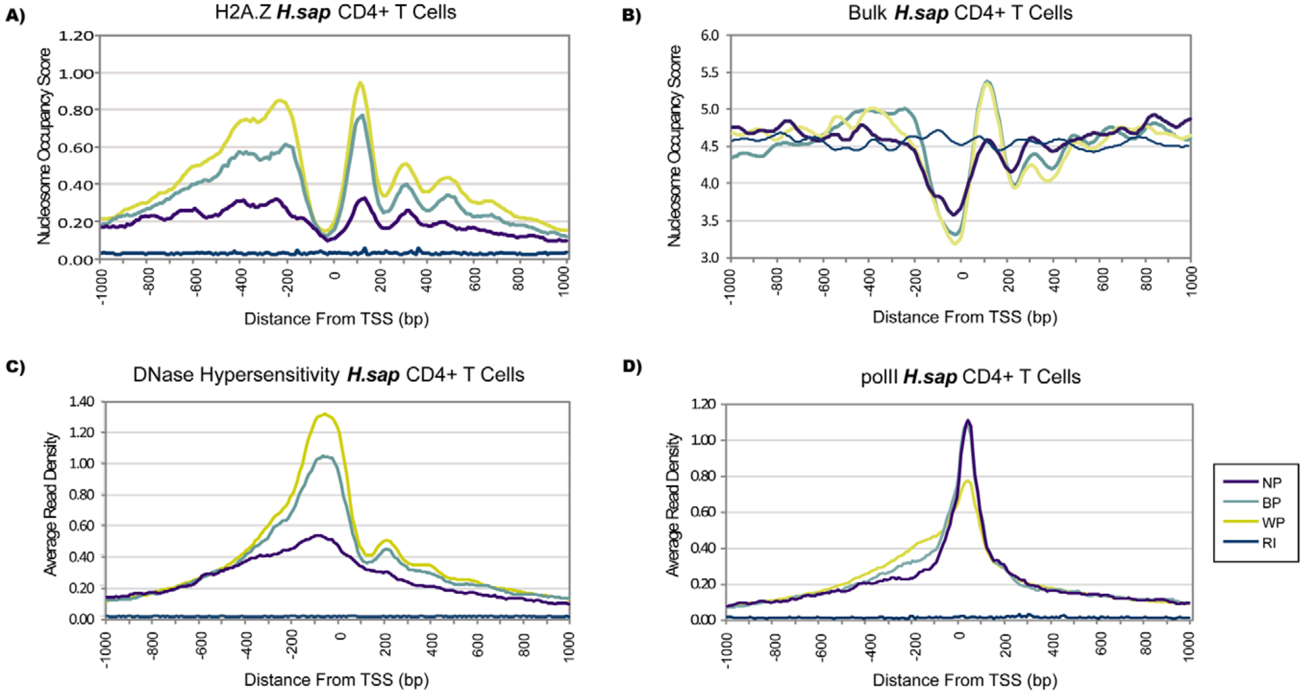

Figure 1. Promoter Classes Reflect Distinct Profiles of Nucleosome Organization. Profiles are based on promoters classified as Narrow Peak (NP), Broad with Peak (BP), and Weak Peak (WP), and show the region of –1 kb to +1 kb around the designated TSS. RI refers to average levels at random intergenic sites, which is used as a baseline. (A) Increased H2A.Z levels ($p<10E-36$), (B) increased bulk levels, and consistent spacing were observed for human BP and WP promoters compared to NP. DNase hypersensitive sites revealed a more accessible nucleosome-free region at BP and WP but not at NP promoters (C), yet pol II levels were higher at NP promoters (D).
doi:10.1371/journal.pgen.1001274.g001

qualitatively similar profiles (Figure 3, Figure S2). Specifically, H2A.Z levels between NP and WP promoters differed at highly significant levels regardless of CpG island presence ($p<10E-84$ and $p<10E-40$ for promoters with and without CpG islands, respectively), whereas H2A.Z differences between promoters with and without a CpG island within the same class were notably less pronounced (WP promoters $p<10E-07$; NP promoters $p<10E-08$; no significance for BP promoters). Due to the much smaller number of focused promoters in the genome and the larger fraction of dispersed promoters containing CpG islands, splitting all promoters in two groups based on the presence of CpG islands as in previous reports will, indeed, lead to different profiles. Regardless, these differences can be explained away by accounting for initiation patterns.

Previous studies had generally observed a stronger correlation of periodic nucleosome organization with more highly expressed genes [7,28]. To rule out the possibility that the observations above could be explained by an overall lower activity of specific promoter classes, we divided the human CD4+ T cell data into four groups based on expression levels (Figure S3). The class-

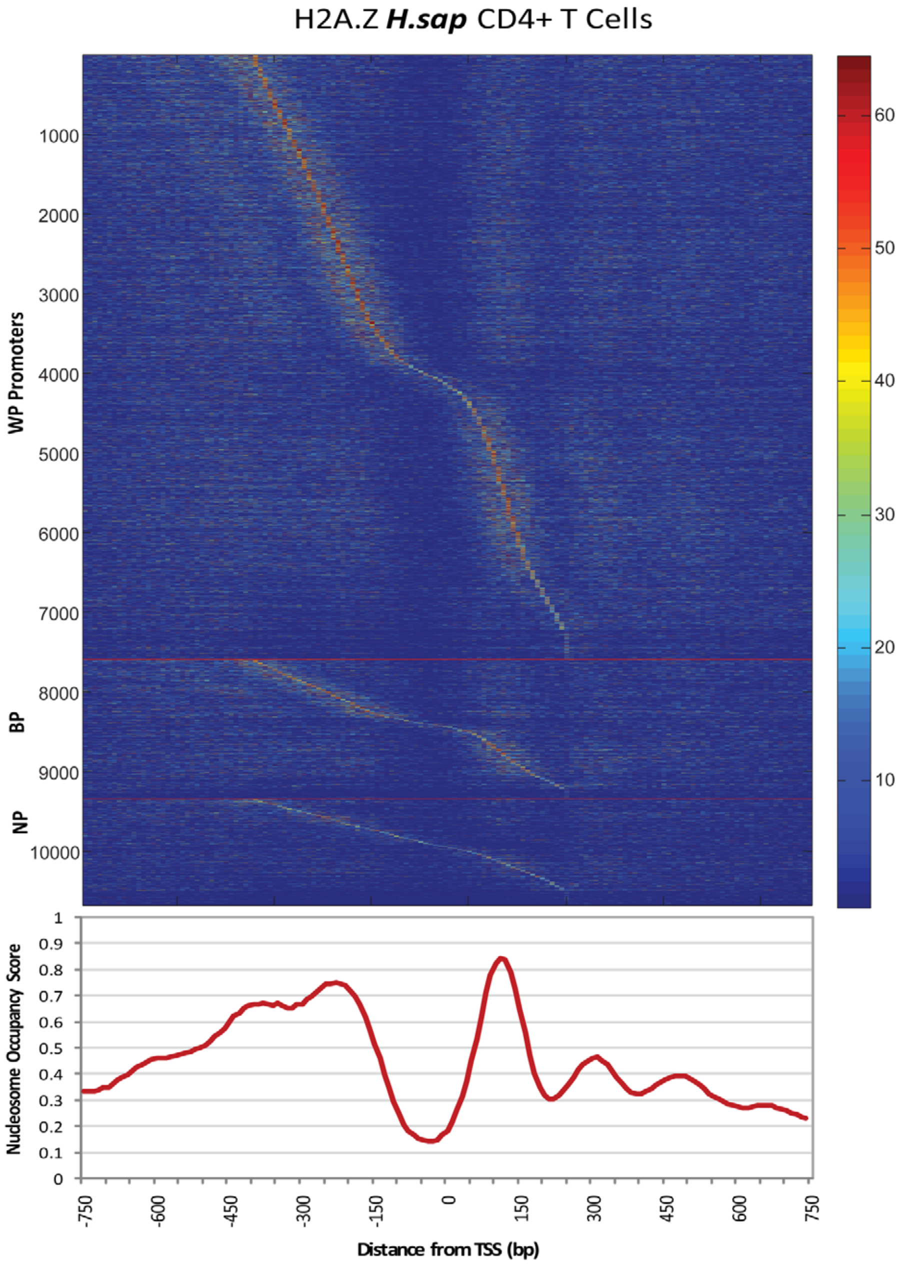

Figure 2. Heatmap of Nucleosome Occupancy within Individual Promoters. Raw H2A.Z nucleosome occupancy values for each human promoter were partitioned into the three classes. The lower panel shows the average occupancy profile across all three classes. Within each class, promoters were arranged by location of their maximum occupancy value in the range of the −1 to +1 nucleosome (−400:+250 with respect to the TSS; the diagonal pattern is thus implied by this ordering and not the data). WP and BP promoters clearly reflected the periodic H2A.Z nucleosomes flanking the NFR, especially downstream of the TSS. Between promoters, the strongest enrichments were often observed at different nucleosomes, likely due to the sparse nucleosome occupancy data.
doi:10.1371/journal.pgen.1001274.g002

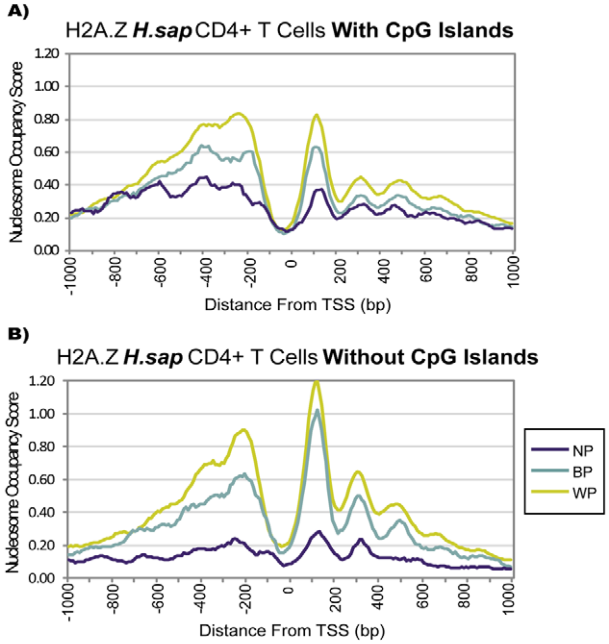

Figure 3. The Presence of a CpG Island Alone Does Not Imply Distinct Chromatin Architecture. When stratifying promoters according to the presence of CpG islands as defined by Takai and Jones [30], no deviation in the nucleosome organization of the promoter classes is observed; WP and BP promoters maintain a higher association to H2A.Z than NP promoters. This pattern is consistent in alternative definitions of CpG islands (cf. Figure S2).
doi:10.1371/journal.pgen.1001274.g003

specific differences of H2A.Z occupancy (stronger for dispersed promoters) and pol II (stronger for focused promoters) remained within each group of similarly expressed genes (Figure S4). Likewise, the reported coupling of H2A.Z with H3K4 trimethyl marks at TSSs [10,31] was maintained across expression levels (Figure S4), and promoter-class-specific differences were also observed for H3K4 mono-and dimethylation (Figure S5).

Computational Models of Promoter Classes Confirm the Different Contributions of Chromatin Features

As core promoters have traditionally been characterized and identified by the presence of regulatory sequence elements, we sought to quantify how informative the ensemble of chromatin features discussed so far would be to define human TSSs. Specifically, we were interested in how strongly the different promoter classes were defined by sequence versus chromatin features. To this end, we trained and applied computational models to classify between TSS versus non-promoter genomic locations. Our goal was to identify potential differences between promoter classes when comparing models under the same assumptions side-by-side, similar in spirit to recent splicing simulators integrating sequence and chromatin features [32].

We computed average profiles of the 2 kb upstream and downstream regions of each TSS for bulk and H2A.Z nucleosomes as well as H3K4 mono-, di-, and trimethylation, for a total of 10 representative profiles for each promoter class. The inner products of the representative profiles with those of a genomic test location were used as input features for sparse linear classifiers, trained separately for WP and NP promoters. Each model was then tested on independent data of WP, NP, and BP promoters (Figure 4), as well as negative samples from other genomic locations, including CpG islands without evidence of transcription. WP and BP classification was much more accurate than NP; this was consistent with our findings that chromatin features were more pronounced and less variable for classes with dispersed initiation (cf Figure 2).

Inspection of the model features showed that each class relied on similar features, selecting an informative subset of nucleosome profiles (Figure 4). The highest weight was assigned to the H3K4 trimethylation downstream profile, followed by the H2A.Z profiles, likely due to the strong periodic signal especially within the transcript. In fact, applying the WP model for the recognition of NP promoters was more successful than using the model trained on NP promoters themselves. Overall however, results stayed well below those obtained on both WP and BP promoters. When adding Fourier-transform based features to reflect the periodicity of nucleosomes, results were slightly improved but highly consistent (Figure S6).

We had previously demonstrated that NP promoters could be characterized with great success by ensembles of transcription factor binding sites based on their enrichment at specific locations relative to the TSS, using features beyond the strict core promoter sequence motifs (including factors such as E2F, CREB, YY1, etc) [33]. Following this example, and using the performance of the chromatin models as baseline, WP classifiers built on sequence features performed considerably worse than the WP chromatin model (Figure 5). The opposite was true for NP promoters, for which sequence models achieved higher success rates on NP and BP promoters than chromatin models. Combining sequence and chromatin features increased accuracy on all test sets, and demonstrated that WP TSSs relied much more on chromatin features than NP TSSs. This was seen in both the relative changes of classification accuracy as well as in the relative strength of features within the combined models, in which chromatin features accounted for stronger contributions for the WP compared to the NP model (Figure 5).

Profiles of Nucleosome Organization Are Conserved across Metazoans

In light of the above observations that distinct chromatin patterns were associated with different initiation patterns, we investigated whether these different modes would be conserved across species. The *D. melanogaster* genome was particularly instructive as its genome does not contain CpG islands, but has recently been found to exhibit the same distinct dispersed and focused initiation patterns.

D. melanogaster promoter classes were defined based on mixed stage embryonic libraries, and all available promoters were further filtered to transcripts present during hours 0–12 of embryogenesis (Figure S7). This matched them more precisely with the available chromatin data, and resulted in 26% NP, 21% BP, and 53% WP promoters (cf. Table 1). As in human, BP and WP promoters showed a significantly greater association with H2A.Z nucleosomes than NP promoters (Figure 6A, $p<10E-23$). BP and WP promoters also had a greater percentage of H2A.Z nucleosomes within 1 kb of the TSS (Figure S8, $p<10E-02$). The +1 H2A.Z nucleosome occurred at 125 bp, which is 10 bp upstream of the previous estimate in fruit fly [6]. An apparent difference between humans and flies was the absence of the H2A.Z association at the -1 nucleosome in *Drosophila*, which has been previously reported [6]. However, this absence does not coincide with a lower level of bulk nucleosomes at this location (Figure 6B). As this phenomenon was not observed in human, additional experiments would be beneficial to confirm any such putative species-specific difference.

Examining the locations of bulk nucleosomes led to an overall lower signal above background; this may at least partially be due

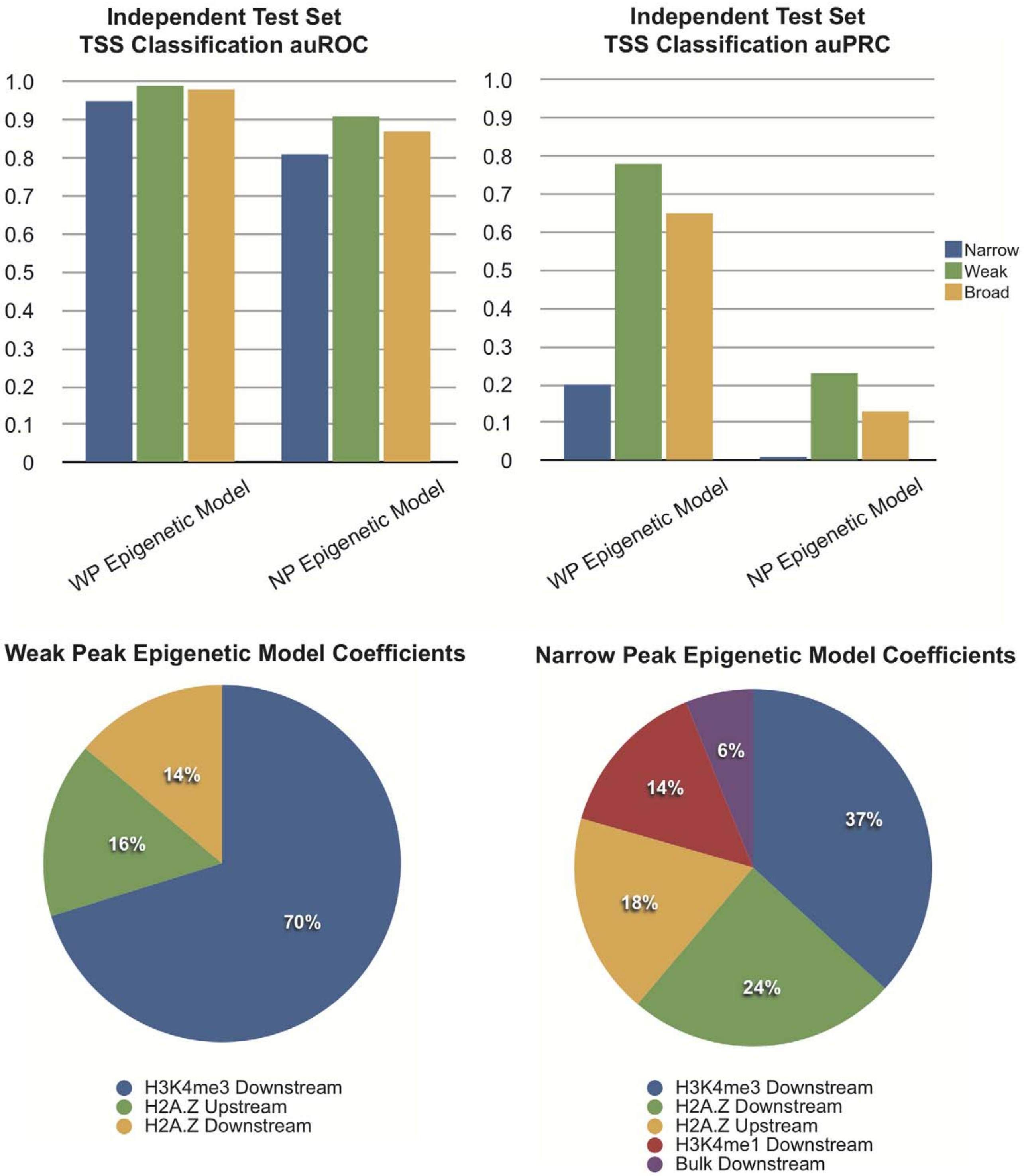

Figure 4. Computational Models Using Chromatin Features Show Different Accuracy for Promoter Classes. Classification accuracy of two epigenetic models (i.e., using chromatin features) was evaluated on test sets for each promoter class (evaluated with auROC and auPRC). Values of 1 indicate perfect classification; auROC values close to 0.5 and auPRC values close to 0 reflect random results. At the bottom, relative weights of chromatin profile features included in each model are depicted.
doi:10.1371/journal.pgen.1001274.g004

to the lower resolution of the tiling arrays used to measure the fly bulk profiles when compared to the deep sequencing data available for H2A.Z. Yet, the consistent difference between promoter patterns was confirmed (Figure 6B, p<10E-02 for NP vs. WP). Currently, data comparable to DNase hypersensitivity is not available for the fly genome; in its place, we used a recent model predicting bulk nucleosome occupancy from sequence features [34]. The computational model displayed some notable differences to *in vivo* bulk nucleosomes, in particular, a more 5′ location of the NFR and a predicted affinity for nucleosomes at the

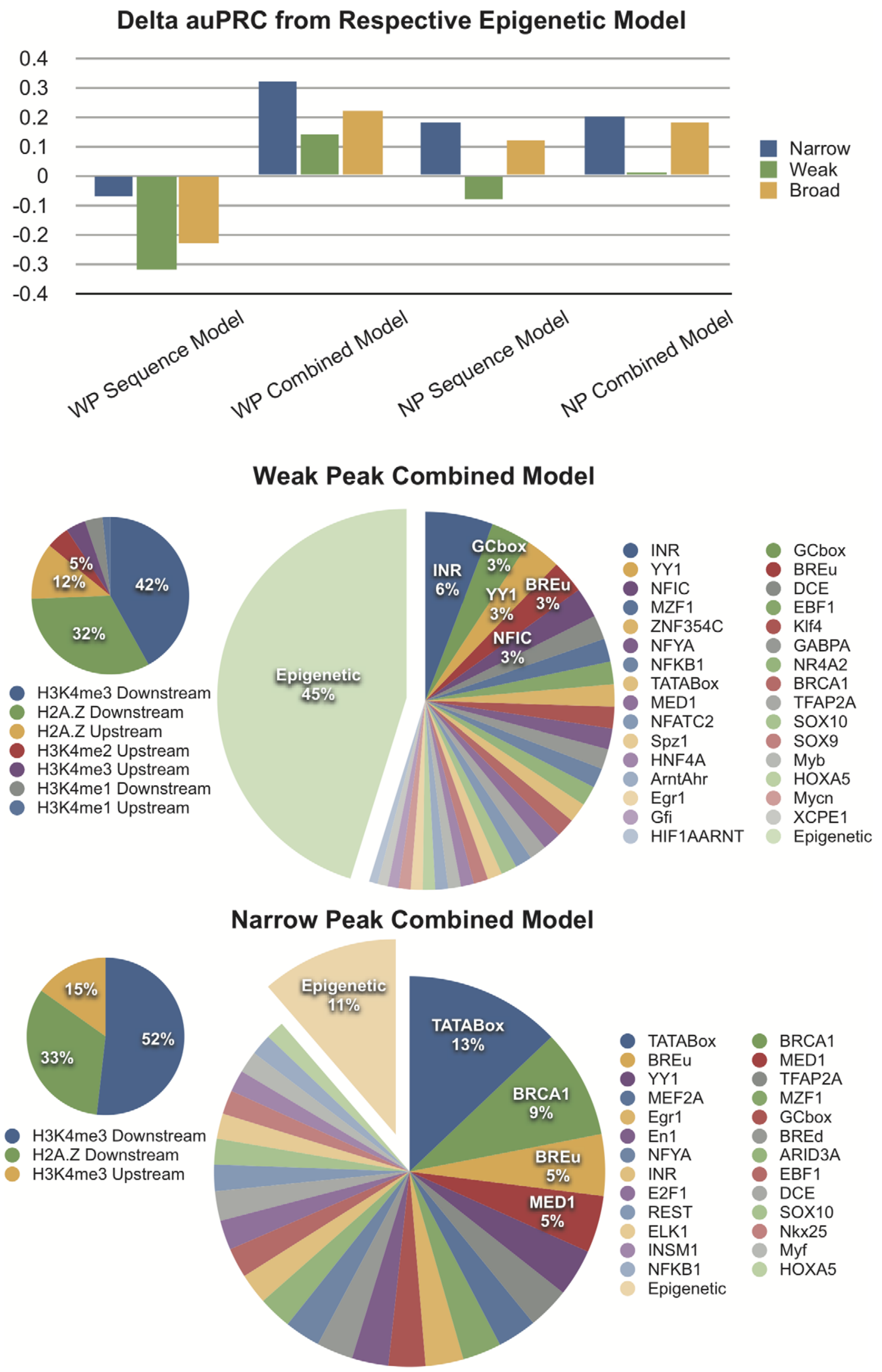
Delta auPRC from Respective Epigenetic Model
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
WP Sequence Model
WP Combined Model
NP Sequence Model
NP Combined Model
Narrow
Weak
Broad
Weak Peak Combined Model
42%
32%
12%
5%
H3K4me3 Downstream
H2A.Z Downstream
H2A.Z Upstream
H3K4me2 Upstream
H3K4me3 Upstream
H3K4me1 Downstream
H3K4me1 Upstream
Epigenetic 45%
INR 6%
GCbox 3%
YY1 3%
BREu 3%
NFIC 3%
INR
YY1
NFIC
MZF1
ZNF354C
NFYA
NFKB1
TATABox
MED1
NFATC2
Spz1
HNF4A
ArntAhr
Egr1
Gfi
HIF1AARNT
GCbox
BREu
DCE
EBF1
Klf4
GABPA
NR4A2
BRCA1
TFAP2A
SOX10
SOX9
Myb
HOXA5
Mycn
XCPE1
Epigenetic
Narrow Peak Combined Model
52%
33%
15%
H3K4me3 Downstream
H2A.Z Downstream
H3K4me3 Upstream
Epigenetic 11%
TATABox 13%
BRCA1 9%
BREu 5%
MED1 5%
TATABox
BREu
YY1
MEF2A
Egr1
En1
NFYA
INR
E2F1
REST
ELK1
INSM1
NFKB1
Epigenetic
BRCA1
MED1
TFAP2A
MZF1
GCbox
BREd
ARID3A
EBF1
DCE
SOX10
Nkx25
Myf
HOXA5

Figure 5. Computational Models Support the Stronger Contribution of Chromatin Features to the Definition of Dispersed TSSs. Changes in accuracy (auPRC) when using sequence models and combined (sequence and epigenetic feature) models are given, relative to the baseline performance in Figure 4. Below, the relative contribution of sequence and epigenetic features in the combined models is shown.
doi:10.1371/journal.pgen.1001274.g005

TSS. Overall, the model agreed well with the *in vivo* profiles; there was a higher association for BP and WP promoters compared to NP promoters at the +1 nucleosome (Figure 6C; p<10E-09). Moreover, the predicted occupancy at the NFR was significantly different from random only for BP and WP, but not for NP promoters. As in human, the increase in NFR accessibility was not accompanied by higher levels of pol II, given that NP and BP promoters had elevated amounts of pol II compared to WP (Figure 6D).

The fly genome contains a repertoire of validated core promoter elements [3,4], and TATA-containing promoters in particular were reported to display a 'very fuzzy' H2A.Z nucleosome organization [6,35]. High-resolution TSS maps have shown that the canonical core promoter elements including the TATA box largely occur in the NP class [2]. After stringent assignments of motifs, we found that NP promoters containing TATA boxes, Initiators, Downstream Promoter Elements (DPE), or Motif Ten Elements (MTE) were in fact completely devoid of any periodic

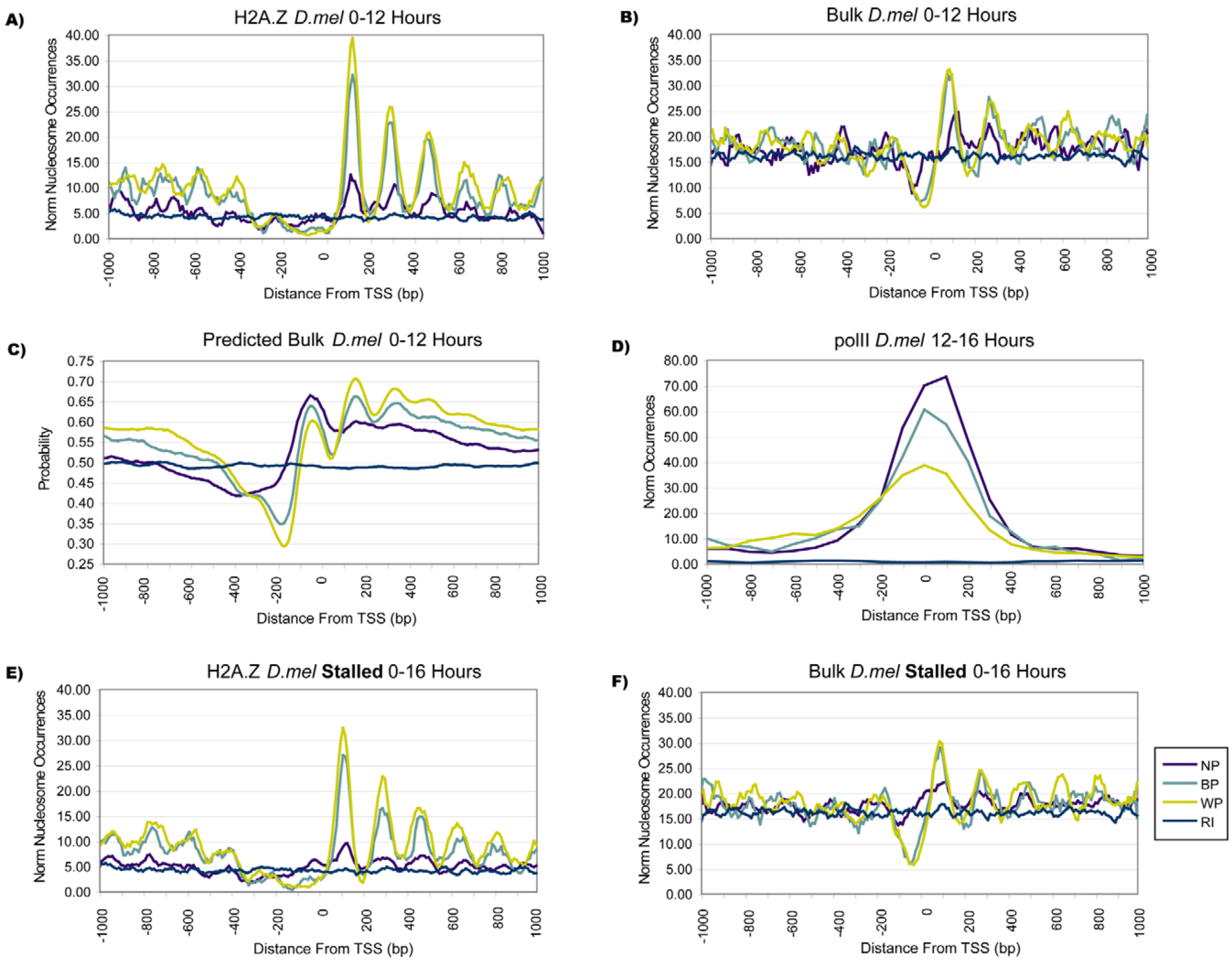

Figure 6. Distinct Nucleosome Organization Is Conserved in Insects. (A) Fruit fly H2A.Z profiles show that BP and WP promoters had increased H2A.Z levels (p<10E-07). Nucleosomes in BP and WP promoters had a more precise spacing, with an average separation of 170 bp and deviations of up to 10 bp, compared to a mean distance of 183 bp between H2A.Z peaks at NP promoters, with deviations of up to 33 bp. (B) Differences between promoter classes were less pronounced in the available lower-resolution *Drosophila* bulk nucleosome data, with a slight shift compared to H2A.Z as originally reported [6]. (C) Average bulk nucleosome occupancy profiles were computed by an *in silico* model, which assigned the predicted probability that a nucleosome was present at any given location [34]. An average occupancy score of .5 indicated no preference for nucleosome presence or absence, as reflected in the scores at random intergenic locations. A clear separation of NP, BP, and WP profiles was observed, and the NFR for NP promoters was clearly much less pronounced; all predicted profiles were significantly different from each other (p<10E-09). (D) NP promoters had noticeably higher levels of pol II binding than BP and WP promoters (12–16 hr embryos). (E,F) Stalled NP, BP, and WP promoters in *Drosophila* mixed stages embryos (0–16 hr) maintained the same associations to H2A.Z and bulk nucleosomes as observed for the set of all actively transcribed 0–12 hr promoters.
doi:10.1371/journal.pgen.1001274.g006

nucleosome positioning (Figure S9). In a notable exception, promoters with the TCT motif, which was recently validated to take the place of the Initiator in translation process genes such as ribosomal proteins [36], contained clearly positioned nucleosomes both up- and downstream of the TSS. This functional group obviously represents highly transcribed constitutive genes and is therefore different from typical NP promoters, which are enriched in precisely regulated genes such as developmental regulators [14,37].

Taken together, both *in vivo* and computational data showed that fly promoters exhibited the same dichotomy as human ones, despite large differences in sequence features such as the absence of CpG islands. Well-spaced nucleosomes and a well defined NFR were reflected in dispersed promoters, in contrast to the indistinct nucleosome positioning pattern of NP promoters.

Pausing of RNA Polymerase Is Not Limited to a Specific Chromatin Architecture

Initially demonstrated in *Drosophila*, RNA pol II can stall or pause 25 to 50 bp downstream of the TSS following transcription initiation [38]. The cause of the pausing is currently unknown, although it has recently been shown to occur at widespread locations across the genome, and to be present in other eukaryotes [39]. As the location of pol II pausing lies at the boundary of the +1 nucleosome, we examined whether stalled promoters exhibited different associations to nucleosome organization. To this aim, we clustered reads derived from short RNAs that corresponded to stalled polymerase in 0-16 h mixed staged embryos [5]. Stalled promoters have been implied with well positioned TSSs [5], and stalled-transcript clusters, defined in the same manner as those from total RNA, indeed contained a >2-fold larger fraction of NP promoters (55%). However, a considerable number of stalled promoters fell into the BP (16%) and WP (28%) classes as well. When we assessed H2A.Z and bulk nucleosomes for the different promoter classes within the stalled subset, we obtained profiles highly similar to those actively transcribed during hours 0–12 (Figure 6E, 6F): Stalled BP and WP promoters had H2A.Z profiles which were significantly different from NP promoters ($p<10E-12$), and exhibited a stronger periodic signal of nucleosomes within the transcript. Similar results were also obtained for stalled promoters from S2 cells (Figure S10), further demonstrating that the promoter classes reflect divergent nucleosome architectures, regardless of pol II stalling. Thus, nucleosome organization is not necessarily a cause or consequence of stalling *per se*; like CpG islands, stalling appears to be a feature enriched in a particular class of promoters. In this case, the nucleosome organization of stalled promoters reflects the overall highly regulated transcriptional program characteristic of focused promoters.

Insulator Classes Demarcate Initiator Patterns

Insulators separate differentially expressed genes, disrupt the communication between enhancers and promoters, and prevent the spreading of chromatin domains. Individual instances of insulator elements have been shown or suggested to play a role in chromatin remodeling near promoter regions [11,12]. Given the strong chromatin differences demonstrated between the promoter classes, we assessed whether associations to different insulators would support these differences.

The CCCTC-binding factor (CTCF) is one of the most prominent insulator proteins that is widely conserved across species [40]. It is known to interact with pol II, and has been implicated in the assistance of nucleosome positioning around its binding sites in human [12,41], as well as being particularly enriched at locations of H2A.Z and H3K4 methylation [10]. Supporting this, CTCF showed a higher association with human BP and WP promoters than NP promoters (Figure 7A, $p<10E-11$). The CTCF profile reached a maximum level at -125 bp upstream of the TSS. This organization placed CTCF in the proximity of the core promoter and just downstream of the -1 nucleosome, and agrees with observations that nucleosomes enriched for H2A.Z were well-positioned and flanked by CTCF [12]. Concordant results were observed between NP and BP promoters when *Drosophila* CTCF (dCTCF) binding was evaluated (Figure 7B, $p<10E-03$), albeit at broader enrichment due to the lower resolution of the tiling array.

The availability of genome-wide data on insulator binding elements as part of the modENCODE project [42] provided an opportunity to expand the observations made for dCTCF. The data was obtained from 0–12 hr mixed stage embryos, i.e. from the same material as the nucleosome data analyzed above [22]. Genomic analyses had defined two classes of insulator elements in fruit fly based on co-occurrence of binding events, and showed significant associations with genomic properties such as proximity and organization of genes and cis-regulatory elements. In addition to dCTCF, CP190 and BEAF32 comprise the Class I insulator elements in fruit fly [22]. In accordance with the frequent co-occurrence of their binding sites, these other Class I insulators also showed specific enrichments in WP and BP promoters (Figure 7C, 7D, $p<10E-03$). Class II insulators in fruit fly are comprised of Su(Hw) associated proteins [22]. Mod(mdg4) and CP190 have been shown to recruit Su(Hw) to the *gypsy* insulator, however, Su(Hw) is reportedly not enriched in promoters [22]. Mod(mdg4) had no significant differences across all promoter classes, which suggests similar functional roles across promoters (Figure 7E). As expected, Su(Hw) was absent from all promoters (Figure 7F).

Lastly, we investigated the GAGA binding factor (GAF) which did not cluster with factors in either Class I or Class II insulators [22]. GAF can regulate gene expression at multiple levels, mediating promoter-enhancer interactions and insulating chromosomal position effects [43]. For instance, at the *D. melanogaster hsp70* promoter, GAF works in combination with the Nucleosome Remodeling Factor (NURF) to disrupt histone octamers over the GAGA site [11] and promote pol II pausing [44]. Given the preference of stalling for NP promoters, we observed a corresponding prominent enrichment of GAF binding in NP promoters from −1400 bp to +1100 bp of the TSS (Figure 7G, $p<10E-03$). When scanning promoters for matches to the GAGA sequence motif, we found that NP promoters showed high levels of matches in a narrower area within the region bound by GAF, while BP and WP promoters had a pronouncedly lower level (Figure 7H, $p<10E-02$) – i.e., the opposite of Class I insulators. Therefore, at least in the case of GAF, the preference for a particular promoter class does not necessarily reflect a dynamic state (such as expression level), but rather is statically encoded in the DNA sequence. In summary, proteins from the recently defined insulator classes and the GAGA binding factor clearly separated among the promoter classes, and points to potential underlying mechanisms which help to define the different promoter classes.

Discussion

The high-throughput sequencing of 5′ capped sequence tags has clearly shown that eukaryotic promoters separate into at least two classes defined by focused and dispersed distributions of initiation events. Many recent studies have reported on the chromatin structure in eukaryotic genomes; our approach differed from most

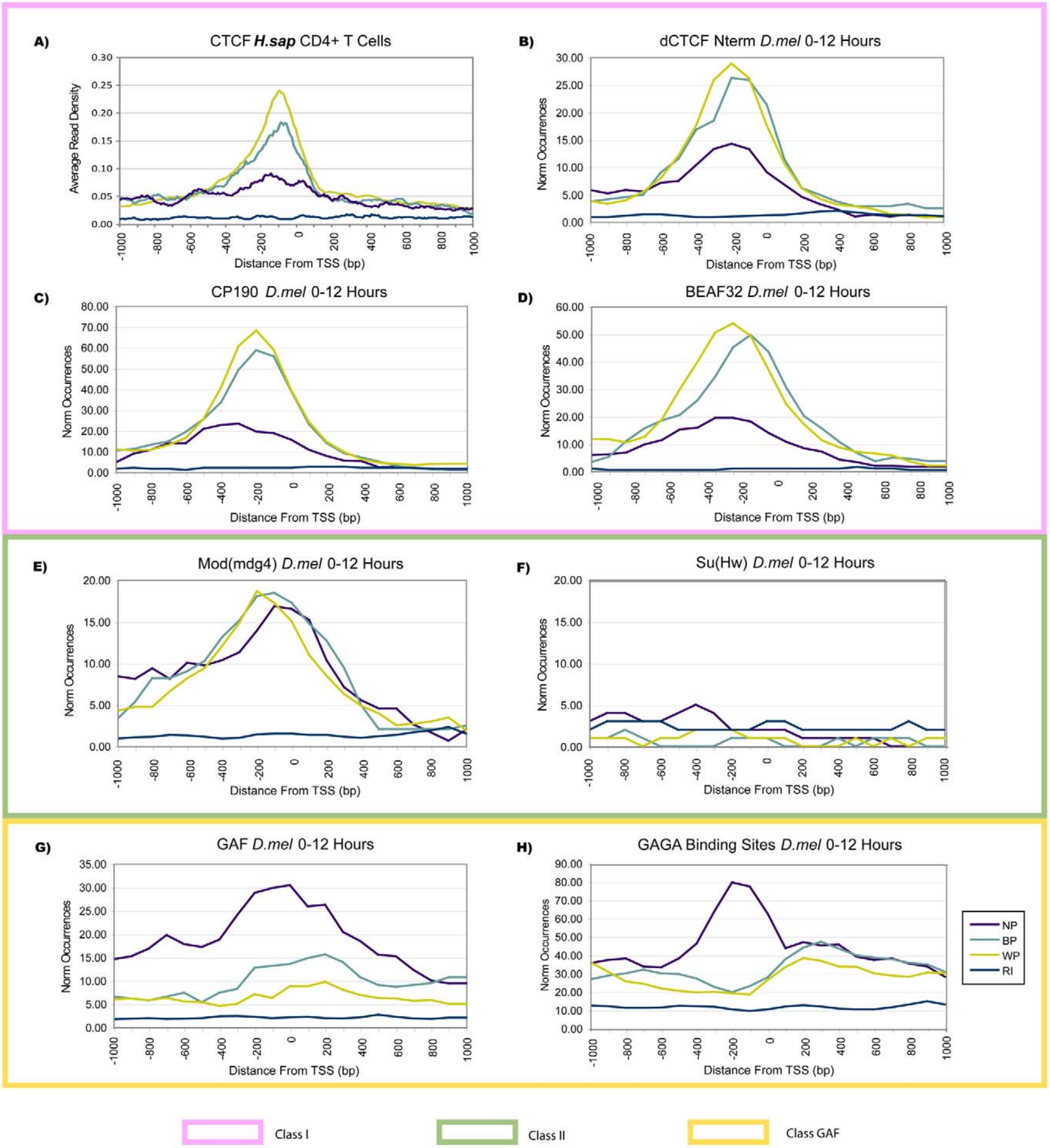

Figure 7. Insulator Classes Are Characteristic of Promoter Classes. (A) Human CTCF had higher occurrences in BP and WP promoters ($p<10E-11$). (B,C,D) Two classes of fruit fly insulators [22] were compared to promoters classes on embryonic data from 0–12 hr. Class I insulators (including dCTCF, CP190, and BEAF32) supported the same pattern of increased BP and WP levels as observed for human CTCF. (E,F) Class II insulators had equal occurrence across promoter classes, with Su(Hw) not being bound to proximal promoter regions. (G,H) ChIP-chip profiles of the chromatin-remodeling transcription factor GAF, as well as presence of GAGA binding sites in the genome, showed a clear enrichment at NP promoters ($p<10E-02$).

doi:10.1371/journal.pgen.1001274.g007

of these efforts by assessing chromatin features from the basis of transcription initiation as derived from 5′ tag data. In one exception, work concurrent to ours found differences on H3K9 acetylation based on different promoter classes [45]. Here, we have established that promoters from different classes not only contain different core promoter sequence features, but also reflect distinct patterns of nucleosome organization, chromatin structure, and insulator preferences (Figure 8).

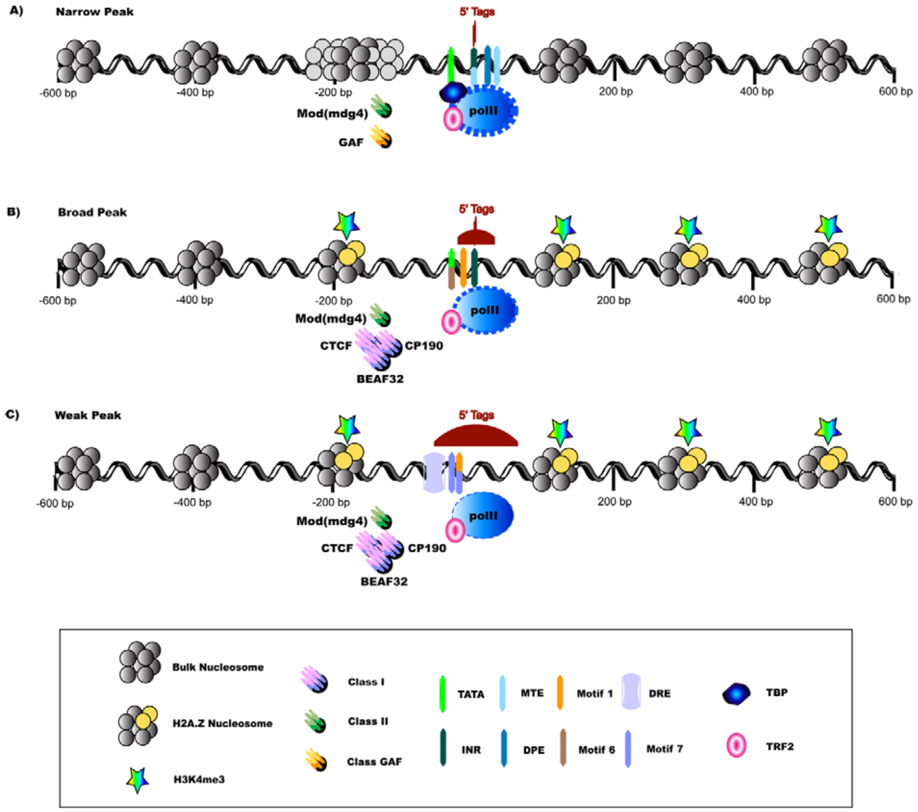

Figure 8. Promoter Classes Are Indicative of Divergent Strategies for Transcription Initiation. The aggregation of differences in transcription factor binding sites, nucleosome organization, histone variants and chromatin marks as well as insulator elements paint a picture of divergent strategies for transcription initiation in metazoans. (A) NP promoters are marked by a 'fuzzy' nucleosome organization [6] (noted by alternative bulk -1 nucleosome locations in the figure) yet precise positioning of transcription initiation, which is reflected in the presence of location specific core promoter motifs that interact with a canonical TBP-containing basal complex [2,23]. NP promoters show higher levels of pol II bound around the TSS, possibly due to an enriched presence of stalled polymerase. They are also associated with specific chromatin remodelers in fly, namely GAF. (C) Initiation events in WP promoters spread over a larger genomic span, reflected in the presence of motifs with lower positional enrichment that have been linked to remodeled basal complexes containing TRF2 in fly [60]. They exhibit a well-defined NFR and well-positioned H2A.Z nucleosomes as well as associated histone marks such as H3K4 tri-methylation. WP promoters in fly contain an enrichment of Class I insulators (CTCF, CP190, BEAF32). (B) BP promoters have a combination of features from both transcriptional programs. While chromatin organization is conserved, some of the known core promoter sequence elements depicted appear to be fly specific (Motif 1, DRE, Motif 6, Motif 7, MTE) [2,16,23]. Pol II and insulator proteins are depicted at the maximum binding locations; sizes of the transcriptional components are not drawn to scale.
doi:10.1371/journal.pgen.1001274.g008

Our findings revealed that the periodic distribution of nucleosomes in the vicinity of TSSs was strongest for dispersed promoters (classes BP and WP), which have defined NFRs and highly periodic H2A.Z-containing nucleosomes. In contrast, focused promoters (class NP) exhibited significantly lower occupancy and/or less organized nucleosomes. Furthermore, recently defined insulator classes showed distinct associations: class I insulators (which include CTCF) were associated with H2A.Z organization and H3K4 methylation at WP promoters, whereas class II insulators were evenly distributed. Conversely, GAF and pol II showed higher levels at NP promoters. The enrichment of the *Drosophila* GAF protein at NP promoters was intriguing, as it is a protein with many reported roles in transcription and chromatin remodeling [46], and may assist transcription initiation at NP promoters in the presence of unorganized nucleosomes. For instance, GAF forms a multimer in replacement of the NFR to establish proper nucleosome organization [47] and is enriched at genes with polymerase stalling [48].

NP and WP promoters in fruit fly and human likely correspond to two classes of promoters that have been recently characterized in yeast [13,21]. The first class has well-defined NFRs flanked by nucleosomes (Depleted Proximal Nucleosome, DPN), while the second class has variable nucleosome positioning without a clear NFR (Occupied Proximal Nucleosome, OPN). CAGE-like data is not available at a scale needed for the identification and assignment of promoter classes in yeast, but OPN promoters have a low association with H2A.Z, a high transcriptional plasticity, and are enriched for TATA boxes, while the opposite is true for DPN promoters. Our work supports and extends the yeast model, in which access to most eukaryotic focused/OPN promoters is highly regulated as the corresponding genes carry out specific functions in response to specific conditions, while expression from many dispersed/DPN promoters is constitutive because they perform housekeeping functions in the cell.

A separation of mammalian promoters has frequently been proposed based on the presence of CpG islands. Differential regulation of some promoters with CpG islands has been shown to result from unstable nucleosomes, contrary to the involvement of chromatin remodelers at non-CpG island promoters [20]. Somewhat differently, we found that CpG islands are present across all initiation patterns, which indicates that CpG islands are not a homogeneous class and do not all encode constitutively unstable arrangements of nucleosomes. The work by Ramirez-Carrozzi et al. [20] focused on a specific set of promoters, those adjacent to stimulus-response genes, in which nucleosomes are pre-organized to facilitate a regulated primary response. Such genes may form an intermediate class between constitutively expressed genes typically associated with CpG islands, and NP promoter genes, which contain genes like developmental TFs that are expressed in a precisely determined and highly regulated order. The conservation of our findings in Drosophila, as well as the previous studies in yeast, support that some CpG islands may provide an additional mechanism of sequence-encoded nucleosome propensities specifically found in mammals.

Multiple aspects may contribute to the relationship between the promoter classes and chromatin features. First, differences in chromatin architecture may be directly reflected in distinct initiation patterns, as illustrated by the nucleosome organization in constitutive versus regulated genes in yeast [49]. Thus, in fly and human, dispersed promoters result from a well-defined NFR increasing the accessibility of the DNA to the polymerase, causing initiation to occur at multiple locations over a large region. In turn, the lower accessibility of focused promoters provides for a more regulated transcription initiation due to the lack of a common NFR. Instead, TSSs of focused promoters are well-defined by position-specific sequence elements including the canonical core promoter motifs [2,33], which serve to actively recruit the core complex to precise TSS locations. Our computational models clearly support this idea: chromatin features contribute to NP promoter definition, but much less so than for other classes, and with little improvement on sequence information. Overall, the higher pol II level at the TSSs of actively expressed genes with NP promoters also suggests that polymerase stalling is involved as an additional regulatory step enriched but not restricted to these genes [5].

Second, the relationship between the promoter classes and chromatin profiles may also be influenced by the duration of active transcription. It has been suggested that nucleosomes are properly positioned through repeated rounds of active transcription [50,51]. As dispersed promoters, and focused promoters containing the TCT motif [36], are enriched in constitutively expressed genes [14], this would support the greater degree of nucleosome organization and the combinations of histone variants and chromatin marks (such as H2A.Z and H3K4me3) traditionally associated with active transcription. In turn, many focused promoters are associated with specific time points during embryogenesis [14], and the lack of constant transcription potentially leads to a reduced positioning of nucleosomes. Finally, promoters may have distinct chromatin patterns involving features we did not investigate. For instance, a higher rate of H3 turnover was observed at OPN promoters in yeast [21], and the presence of GAF has been associated with H3.3 replacement [52], suggesting the possibility that focused promoters may have a higher association with H3.3 replacement.

As more data becomes available through large-scale efforts such as the modENCODE and ENCODE projects, the presence of high-level divergent strategies of gene regulation established at the basal promoter will become better characterized throughout development and differentiation in model organisms as well as in human. Promoter classes may have associations to epigenetic inheritance, cellular memory, evolvability, and the development of disease [53,54]. Understanding initiation patterns does not only help deepening our knowledge of the core promoter sequence, but also provide insight into the epigenetic architecture of regulatory regions. Together, they illustrate the interplay between chromatin and sequence information to encode divergent strategies for gene expression.

Methods

Selection of Fruit Fly and Human Transcription Start Sites and Fly TSSs Associated with Polymerase Stalling

We used a recently published dataset, determined by clustering of >10 million aligned 5′ capped paired-end sequence tags from 0–24 hour mixed stage *D. melanogaster* embryos [2]. We selected strong clusters (>100 tags) located within initiation regions, which included annotated 5′UTRs and 250 bp upstream of the annotated TSS in Flybase [55]. This dataset comprised ~4,000 promoters which are classified by means of two features, genomic span of initiation events (as defined by the size of distinct 5′ tag clusters), and localization of initiation. For NP promoters, tag clusters have to be smaller than 25 nt, and at least 50% of tags align at the peak location (defined as the mode of the cluster ±2 nt). BP promoters exceed the 50% tag cutoff at the mode, but are spread out over a genomic range >25 nt. WP promoters are those which meet neither genomic span nor peak location cutoffs; they do however still show a distinct albeit lower peak, frequently associated with the presence of a minimal initiator sequence motif. The modes of the tag distributions were used as representative TSS locations for all promoter classes.

To match these TSS data to available chromatin resources, only those promoters with active transcription in at least one time point from 0–12 hours of fruit fly embryogenesis were used (517 NP, 406 BP, and 1,054 WP promoters). The temporal activity of each promoter was determined through Affymetrix tiling array data that measured RNA levels every 2 hours during the first 24 hours of *D. melanogaster* embryogenesis [56] (Table S1). The utilization of promoters at each time point was evaluated as described previously (see Text S1) [23]. Briefly, the median hybridization value of several tiles 3′ of the TSS, i.e. in a putatively transcribed region, was contrasted with the median of tiles 5′ of the TSS location. The significance of active promoter calls was evaluated by repeating the analysis on three sets of 1,000 randomly selected intergenic sites (Table S2).

To use consistent promoter classes, human CAGE tags and fly short RNAs associated with polymerase stalling were clustered using the same strategy and parameters as above [5,26]. Promoters of clusters in the initiation region as defined in ENSEMBL or Flybase, respectively, were again classified as NP, BP, and WP based on the shape of their tag distributions. In human, we started from the published alignments of 29 million tags generated by the FANTOM consortium and classified 1,409 NP, 1,759 BP, and 7,656 WP promoters falling in the initiation region that contained more than 100 reads. In fruit fly, we clustered ~6 million reads from short RNAs from 0–16 hour embryos, which separated into 2,176 NP, 645 BP, and 1,101 WP stalled promoters, and additionally ~16.5 million reads from S2 cells, resulting in 1,977 NP, 1,158 BP, and 2,530 WP stalled promoters, each with clusters that contained more than 100 reads.

Scoring Human Nucleosomes and Regulatory Factor Profiles

The nucleosome occupancy score for H2A.Z, H3K4 methylation, and bulk profiles was calculated according to Schones *et al*, using raw short aligned reads mapping to 5′ or 3′ nucleosome boundaries [7]. We divided each somatic chromosome into 10 bp non-overlapping windows, and read counts for a window were calculated by summing the number of reads that aligned in the 80 bp upstream (on the sense strand) or 80 bp downstream (on the anti-sense strand) windows, assuming that 5′ and 3′ reads mapping to the ends of the same nucleosome would be ~140–160 bp apart. Promoters were analyzed in windows from −1 kb to +1 kb of the TSSs identified by tag clustering; to reduce the noise in the bulk data, promoters with outlier read counts less than 8 or greater than 2,400 were removed from the analysis. A raw nucleosome occupancy score was determined for each promoter window by averaging the read counts across all of the individual promoters within one pattern (NP, BP, and WP). A moving average over five windows of raw nucleosome occupancy scores was taken for each promoter pattern to produce the smoothed nucleosome profiles shown. Window scores thus reflected nucleosome midpoints. A set of 5,000 random intergenic sites was chosen across Chromosome 1 for which nucleosome profiles were determined akin to that of the promoters.

For pol II, DHS, and CTCF profiles, raw read data was assigned to 10 bp non-overlapping windows regardless of strand. Within each promoter pattern, the read counts were averaged for windows covering ±1 kb with respect to their locations from the TSS, and a moving average over five windows was used for smoothing, resulting in the average read density shown in the figures. The same steps were applied to the set of random intergenic sites from Chromosome 1.

For a complete summary of human data sources, see Table S3.

Scoring Fruit Fly Nucleosome and Regulatory Factor Profiles

Mavrich *et al* determined nucleosome positions by deep sequencing of MNase digested DNA associated with nucleosomes containing the H2A.Z histone variant, as well as by tiling array hybridization of bulk- and pol II-associated nucleosomes. The published data had been processed to retain only peaks above background, reflecting the midpoints of nucleosomes. From this data, we calculated normalized nucleosome occurrences for the H2A.Z, bulk, and pol II bound data by first determining distances of the TSSs from the nucleosome midpoints with respect to the orientation of transcription, and adding them into 10 bp non-overlapping bins. The moving average of five neighboring bins within the window from −1 kb to +1 kb was then normalized to the number of nucleosome occurrences per 500 TSSs. Enrichments are contrasted with averaged results of profiles on three sets of 1,000 random intergenic (RI) sites. As in human, scores thus reflected nucleosome midpoints; unlike in human, profiles are based only on midpoints as determined by local maxima above background and not the complete data. For computationally predicted bulk nucleosome locations, the nucleosome occupancy scores were calculated from average occupancy probabilities and processed analogous to human data (see Scoring Human Nucleosome Profiles).

H3K4 methyl marks and insulator binding profiles were measured by hybridization to tiling arrays that were acquired from the modENCODE repository. For the pol II data generated in the S2 cells, read counts were summed within 25 bp windows [5] and those windows with at least 25 reads were used in the analysis. The distances of the mark and profile binding midpoints were calculated relative to the TSS locations and cumulated into 100 bp bins. The moving average over three neighboring bins within −1 kb to +1 kb was normalized to the number of occurrences per 500 TSSs. The same strategy was again repeated on sets of random intergenic sites.

For a complete summary of *Drosophila* data sources, see Table S4.

Assignment of Sequence Features

CpG islands were initially taken from the UCSC Genome Browser annotation, which follows the definition by Gardiner-Garden & Frommer [57]: a>200 bp stretch with a G+C content of at least 50% and an observed vs expected ratio of CG dinucleotides of >0.6. We then filtered this initial set by the more recent criteria of Takai & Jones [30], which led to a strict subset of regions with length >500 bp, G+C content >55%, and CG ratio >0.65.

Drosophila core promoter motifs were taken from Ni *et al* [2], which assigned them by position weight matrix matches to narrow sequence windows relative to the TSS in which they were significantly enriched. To be as comprehensive as possible, we used the largest p value cutoff for which matches were reported (p<10-2). Motif matches were therefore allowed to be comparatively weak but were based on precise distances to defined TSS locations.

Stratification by Human Expression Levels

The log values of gene expression from NimbleGen tiling arrays for CD4$^+$ T-cells generated in an earlier study [28] were mapped to corresponding TSSs via associated genes (Figure S1). The log2(expression) values of all genes, regardless of promoter pattern, were plotted and divided into four groups. As in the previous study, we declared genes below a cutoff of 4.5 as "silent", and divided the remaining genes into three groups by their expression level. Consequently, there were 948 genes with values below 4.5 that had 'no' expression, 2,504 genes above 4.5 and below 6.25 that had 'low' expression, 3,526 genes above 6.25 and below 8 that had 'medium' expression, and 3,846 genes with values higher than 8 that had 'high' expression. Within each expression group, the TSSs were then subdivided a second time according to their promoter pattern (NP, BP, WP). Expression levels across promoter patterns were thus based on the same cutoffs. Occupancy scores were then calculated as described above. As there were nearly twice as many promoters associated with genes in each group with expression than those with no expression, occupancy profiles for 'no' expression are less smooth.

Statistical Significance

We assessed differences in nucleosome occupancy at specific locations relative to the TSS. The significance between distributions of occupancy scores at the +1 nucleosome midpoint (defined as global maximum downstream-proximal of the TSS; maximum value within the 10 nt bin) and nucleosome free region (defined as global minimum upstream-proximal of the TSS; mean value within the 10 nt bin), as well as the number of nucleosomes within 1 kb of the TSS in fly, were determined using a Mann-Whitney U-test. A χ^2 test was used to compare the H2A.Z peaks in fly, as peaks from Mavrich *et al* corresponded to the filtered number of promoters above background rather than original read density or intensity values.

Additionally, we assessed the statistical significance between pairs of profiles, as measured by the set of differences between values observed at all locations along the profile, using a Wilcoxon Signed Rank test. We compared each pair of NP/WP/WP profiles, as well as each profile to random intergenic regions, for a total of 6 tests (a Bonferroni correction thus led to cutoff of significance at p<(.05/6) =.0083). Due to the pooling of observations at many genomic locations, we observed that comparisons generally led to small p values which particularly on the human data frequently exceeded the precision of the software (1.44E-34); in those cases, we primarily relied on tests at specific locations as described above. All of the tests were performed in Matlab; the exact p values for all tests can be found in Tables S5 and S6.

Computational TSS Models Using Chromatin and Sequence Features

To evaluate the contribution of chromatin features to the definition of different promoter classes, separate linear classifiers for NP and WP promoters were trained on chromatin features, or combinations of sequence and chromatin features. These classifiers were then tested to determine how well they were able to distinguish between TSSs from the three promoter classes and other genomic locations.

Training and test data. NP and WP TSSs were divided into training and test data, using two-thirds of each set for training and the remaining samples for testing. For each TSS in the training set, 20 intergenic locations were drawn at random from -4000 to -100 relative to the TSS. Additionally, one location was drawn from annotated CDS of human UCSC Known Genes, and two locations from annotated CpG islands without evidence of transcript activity (i.e. those without human CAGE aligned reads). Intergenic, CDS, and CpG island locations together comprised the negative examples. For each of the remaining independent TSSs in the test set, we further randomly selected 100,000 CpG island locations (again sampled from those without human CAGE tags) as well as locations from anywhere in the genome. To ensure that each sample contained sufficient data for chromatin feature extraction, all positive and negative training and test samples passed a filter of at least eight aligned reads of the bulk nucleosome data (cf. *Scoring Human Nucleosomes…* above). All analyses were also performed using unfiltered data, with consistent results (data not shown).

Feature generation. Chromatin or "epigenetic" features were designed to reflect similarity to the typical nucleosome profile surrounding a TSS. Epigenetic features were calculated as the inner product of an example's profile and a reference profile obtained from the respective training set. Reference profiles were generated by averaging the profiles of the respective TSS training set, split at the TSS in 2 kb upstream and 2 kb downstream regions. A total of 10 profiles were thus generated for each model, corresponding to Bulk, H2A.Z, and H3K4 monomethyl, dimethyl, and trimethyl profiles. The processed chromatin data was binned into 10 bp intervals, and the closest datapoint to the TSS location was used as the "0" location for relative profile coordinates. Each epigenetic profile was smoothed using a Discrete Fourier Transform Low Pass Filter with a low pass limit of 150 bp, eliminating noise at frequencies higher than an average nucleosome size.

To select informative sequence features, position weight matrices (PWMs) of transcription factors were obtained from the JASPAR Core Vertebrate and RNA pol II datasets [58]. We then followed the protocol described in [33], in which we previously described a classifier for murine NP promoters. Briefly, for each promoter class, TFs were filtered to those exhibiting match score enrichments in specific regions relative to the TSSs; these factor-specific enriched regions were each subdivided into seven windows. For every selected factor, background-normalized cumulative PWM scores were computed for each of the windows and used as features.

Model training, testing, and evaluation. Further following the example of [33], we used L1-regularized logistic regression to learn a sparse linear classifier for each promoter class, as implemented in the l1_logreg package [59]. Sparse logistic regression selects features by assigning coefficients to each, while penalizing the use of large numbers of features. Thus, coefficients of features that are not important for the classification problem are driven to zero and effectively excluded from the model.

L1-regularized logistic regression uses the L1 penalty parameter to set the balance of including features. We performed 10-fold cross-validation to select the optimal L1 parameter for each model. The training data was divided into 10 parts, each part having an equal number of positive, negative intergenic, negative CDS, and negative CpG island examples. For each round of cross-validation, 8 parts were used for training, one for testing and selection of the optimal L1 parameter, and one for independent testing with the optimal L1 parameter. The range of L1 parameters for each cross-validation ranged from 0.0001 to 0.01. All training was performed using the l1_logreg data standardization option, normalizing for potentially different scales between features. After cross-validation, a final model was created by training on the entire training set with the mean optimal L1 parameter.

The models were tested on the independent test data of each of the three classes, using the final NP and WP models generated on the full respective training sets. Classification performance was evaluated with two standard metrics: the receiver operating characteristics (ROC) and the precision recall curves (PRC), and the area under ROC (auROC) and PRC curves (auPRC), which summarize classifier performance when varying the true positive rate. While ROC effectively normalizes for differences in size of positives and negatives, PRC is sensitive to imbalanced datasets – as is the case for promoters in which a small number of TSS locations are outnumbered by the non-TSS locations in the genome. This implies that ROC curves are comparable for different classifiers (e.g. NP and WP), while PRC curves will reflect differences in the relative size of the positive class. This partially explains the larger differences we observed for auPRC values, which reflects the harder problem of identifying fewer NP than BP promoters within a large genomic background.

To visualize the importance of features for each class, a modified version of l1_logreg was used to obtain standardized coefficients, representing input features normalized to the same scale. From these standardized coefficients, we determined which features were consistently present during the ten-fold cross-validation training step. For each model, we determined the features whose absolute value was greater than 0.05 in at least 8 of the 10-fold cross-validations.

Supporting Information

Figure S1 Unsupervised Clustering of Chromatin Profiles. Each promoter profile with sufficient data for bulk and H2A.Z nucleosome occupancy was normalized to promoter-specific Z-scores (subtracting the mean and dividing by the standard deviation for that profile). Promoters were then clustered with MATLAB's Kmeans function (K = 3, Euclidean distance metric). Enriched promoter classes present within clusters were calculated using a hypergeometric test. The first cluster corresponded to unstructured nucleosome profiles enriched particularly for NP promoters ($p<10E-5$; $p<10-2$ for BP promoters; no significance for WP). The second and third cluster largely corresponded to two WP clusters (those with predominant read data for the +1 and -1 nucleosome, respectively, $p<0.05$ and $p<0.01$; no significance for NP and BP).

Found at: doi:10.1371/journal.pgen.1001274.s001 (10.10 MB EPS)

Figure S2 Chromatin Profiles for Promoters Classes Divided by CpG Island Presence, Using the Definition of Gardiner-Garden & Frommer.

Found at: doi:10.1371/journal.pgen.1001274.s002 (3.70 MB EPS)

Figure S3 Expression Levels of Human Genes by Promoter Class. Gene expression intensities from NimbleGen tiles [28] were assigned to human TSS clusters (see Methods), and their log(expression) values were binned separately for each promoter class and normalized to relative frequencies. BP and WP promoters had nearly identical expression, while NP promoters showed a skew towards lower expression.
Found at: doi:10.1371/journal.pgen.1001274.s003 (2.03 MB EPS)

Figure S4 Promoter Classes Separate Chromatin Profiles Even When Stratified by Expression Level. Human promoters were separated into 4 classes based on expression levels of associated genes in CD4$^+$ T-cells. (A) Across all expression levels, BP and WP promoters showed greater enrichments in +1 H2A.Z nucleosome occupancy ($p<10E-5$) than NPs. H2A.Z enrichments have been reported to be present in promoters of both active and inactive genes in yeast [9]. We also observed H2A.Z enrichments for BP and WP promoters at all expression levels; however, the H2A.Z association disappeared at NP promoters with low or no expression. (B) In addition to nucleosome positioning, several histone modifications preferentially occur at promoter regions. To validate this association across promoter classes and expression levels, we matched promoters to human H3K4 methylation data [10]. The positioning of H3K4me3 signals across all promoters and expression levels corresponded with the positioning and levels of H2A.Z nucleosomes. (C) Levels of pol II binding showed the opposite trend, with NP promoters being much more occupied by pol II at the TSS despite the much lower H2A.Z and H3K4me3 association. As such, the lower level of H2A.Z or H3K4me3 at focused promoters did not correspond to a reduced presence of the polymerase at the TSS.
Found at: doi:10.1371/journal.pgen.1001274.s004 (9.06 MB EPS)

Figure S5 WP and BP Promoters Have Stronger Associations to H3K4 Methylation. (A, B) Average profiles of H3K4me3 occupancy in *Drosophila* and human promoters showed an overall similar pattern, with significant differences between NP and the other classes ($p<10E-03$). (C, D) The lower association of H3K4 methylation for NP promoters was retained in human H3K4me1 and H4K4me2 profiles, which consistently showed relative enrichments further within transcribed regions for WP and BP promoters ($p<10E-21$).
Found at: doi:10.1371/journal.pgen.1001274.s005 (7.57 MB EPS)

Figure S6 Including Fourier Transform–Based Chromatin Features in a Computational TSS Model. We explored the effect of adding Discrete Fourier Transform (DFT) coefficients as features, in addition to the epigenetic profile features. The Fourier transform decomposes a signal into its spectral components, and coefficients reflect the presence of periodicities within the data. The DFT was computed in Matlab, on the data pre-processed as described in the main text. As with the profile features, DFT coefficients were computed for the 2 kb upstream and 2 kb downstream regions relative to the TSS, for the whole 2 kb windows as well as smaller 500 bp sliding windows, moved within the 2 kb regions 250 bp at a time. DFT coefficients were computed for Bulk, H2A.Z, and H3K4 monomethyl, dimethyl, and trimethyl profiles, and coefficients reflecting periodicity in the range of a nucleosome turn were added to the features for model training as described in the main text.
Found at: doi:10.1371/journal.pgen.1001274.s006 (1.81 MB EPS)

Figure S7 Fruit Fly Promoter Classes Show Different Temporal Trends at the Same Magnitude of Expression. Specific time points of utilization for each promoter were determined using the differences in median fluorescence intensity values of the Affymetrix tiling arrays [23]. The number of promoters with utilization at each time point were added by pattern and normalized per 1,000 TSSs. (A) The overall progression of expression agreed with previous results [23]: higher numbers of BP and WP promoters were utilized during the earlier stages of embryogenesis, while the opposite was true for NP promoters. (B) Promoters with utilization in at least one time point from 0–12 hours were assigned to expression levels based on array fluorescence (differences in median fluorescence of tiles downstream of a TSS vs. upstream, discretized in bins of size 10). Promoter numbers in each bin were divided by the total number of differences, resulting in the frequency of expression as shown. A line graph was used to smoothly join the discrete bin densities. While quantities of promoter patterns changed throughout embryogenesis (A), the distribution of expression levels was highly consistent across all promoters. (C) The expression analysis was repeated for promoters with utilization in at least 1 time point from 7 to 8 (hours 12–16, to match pol II occupancy data from developmental stage 12). Again, a similar distribution of expression levels from the tiling arrays was observed across all promoter patterns.
Found at: doi:10.1371/journal.pgen.1001274.s007 (5.65 MB EPS)

Figure S8 Density of H2A.Z Nucleosomes Is Higher in BP and WP Promoters than in NP Promoters. The midpoints of all H2A.Z nucleosomes were taken from Mavrich et al and mapped to the locations of the 0–12 hour NP, BP, and WP promoters. There were 95% of WP and 89% of BP promoters that had at least one H2A.Z nucleosome within 1 kb of a TSS, compared to 79% of NP and 71% of random intergenic sites. Greater differences in percentages were observed for BP and WP promoters with more than one nucleosome within 1 kb of the TSS ($p<.05E-01$). This illustrates the stronger connection of BP and WP promoters to the positioning and quantity of H2A.Z nucleosomes within the immediate vicinity of the TSS.
Found at: doi:10.1371/journal.pgen.1001274.s008 (0.68 MB EPS)

Figure S9 NP Chromatin Profiles Separated by Presence of Core Promoter Motifs. The NP H2A.Z profile (cf. Figure 6) was split into (possibly overlapping) subsets by presence of regulatory sequence motifs. Promoters with canonical motifs (TATA, MTE, DPE, INR) exhibited virtually no periodic nucleosome organization. The remaining signal in average NP plots originated from two subgroups, those without canonical motifs (<10%) which possibly represent a small fraction of false assignments to initiation patterns, and those with the TCT motif, which has recently been identified as a hallmark of ribosomal and other basal translation proteins in non-TFIID-initiated promoters [36].
Found at: doi:10.1371/journal.pgen.1001274.s009 (0.11 MB EPS)

Figure S10 Chromatin Profiles for Stalled Polymerase in S2 Cells. To further evaluate the influence of pol II stalling on the divergent patterns of nucleosome organization, TSSs at stalled promoters in S2 cells were compared to H2A.Z and bulk nucleosome locations. Like total TSS data from 0–12 hour, and stalled 0–16 hr promoters, BP and WP promoters had higher associations to H2A.Z (A) and bulk (B) nucleosomes. (C) Similar to the later stage of development (12–16 hr; cf Figure 6D), S2 cell NP promoters had higher levels of pol II binding than BP and WP promoters. This corresponded with the observed higher number of NP promoters utilized during later stages of embryogenesis (Figure S7), and was consistent with with possible pol II stalling frequently observed at NP promoters in more specialized cell types.
Found at: doi:10.1371/journal.pgen.1001274.s010 (8.09 MB EPS)

Table S1 Tile Conversion Statistics for Mapping the Affymetrix Tiling Arrays from Release 4 to Release 5. Column 1 notes the chromosome, and columns 2 and 3 list the number of tiles in Release 4 and Release 5, respectively. Column 4 contains the number of tiles that were removed because they were mapped to multiple locations, or did not map to within 5 bp of the Release 4 tile size. Column 5 and column 6 cite the genomic locations of the first and last tiles in Release 5. Promoters identified using PEAT that were located outside of the scope of the Release 5 Affymetrix tiling array were excluded from the evaluation of temporal utilization using the 2-hr time course.
Found at: doi:10.1371/journal.pgen.1001274.s011 (0.04 MB DOC)

Table S2 False Positive Rates of Expressed Transcript Calls at TSSs. For each time point (column 1) corresponding to a 2 hour interval (column 2), a previously determined difference threshold (column 3) was used to determine false positive rates (column 4) for TSS utilization from background noise [23]. Relative false positive rates were obtained and averaged across three random intergenic sets. FP rates were consistently below 0.04.
Found at: doi:10.1371/journal.pgen.1001274.s012 (0.04 MB DOC)

Table S3 Summary of Data Sources Used For Promoter Comparisons in Human. The data type (column 1) and publication source (column 2) are listed with the total size of the dataset (column 3) and the cell type in which it was generated (column 4). Column 5 refers to the type of experiment performed, and column 6 denotes the figure in which the data is used. S = Supplementary Figure.
Found at: doi:10.1371/journal.pgen.1001274.s013 (0.04 MB DOC)

Table S4 Summary of Data Sources Used For Promoter Comparisons in Fruit Fly. The data type (column 1) and publication source (column 2) are listed with the total number of Release 5 locations (column 3). The sample source (time window during embryogenesis), the type of experiment, and the figure are summarized in columns 4, 5 and 6, respectively. S = Supplementary Figure.
Found at: doi:10.1371/journal.pgen.1001274.s014 (0.07 MB DOC)

Table S5 Summary of Statistical Significance of Differences between Human Nucleosome Profiles. The dataset and type of comparison (profile, peaks, NFR) are listed in column 1 and row 1. "Peaks" refers to the window corresponding to the +1 nucleosome midpoint location; NFR to the window at the center of the nucleosome free region. Profiles were evaluated using Wilcoxon Sum Ranks, and all peak and NFR windows were evaluated using the Mann-Whitney U-test. The p-value of each comparison is listed; those not significant are highlighted. Black boxes correspond to comparisons not evaluated or applicable. All calculations were performed in Matlab.
Found at: doi:10.1371/journal.pgen.1001274.s015 (0.07 MB DOC)

Table S6 Summary of Statistical Significance of Differences between Fly Nucleosome Profiles. The dataset and type of comparison (profile, peaks, NFR) are listed in column 1 and row 1. "Peaks" refers to the window corresponding to the +1 nucleosome midpoint location; NFR to the window at the center of the nucleosome free region. Profiles were evaluated using Wilcoxon Sum Ranks, number of H2A.Z nucleosomes within 1 kb, bulk peaks, and bulk NFR were evaluated using Mann-Whitney U-test, and H2A.Z peaks were evaluated using χ^2. The p-value of each comparison is listed; those not significant are highlighted. All calculations were performed in Matlab.
Found at: doi:10.1371/journal.pgen.1001274.s016 (0.06 MB DOC)

Text S1 Supplementary Methods on Mapping Drosophila TSS to Tiling Arrays.
Found at: doi:10.1371/journal.pgen.1001274.s017 (0.02 MB PDF)

Acknowledgments

The authors thank Kevin White and his lab at the University of Chicago for making the H3K4 and insulator data available ahead of publication, as part of the modENCODE Consortium; Gregory Crawford at Duke University for use and assistance with DHS data generated as part of the ENCODE Consortium; and Molly Megraw at Duke University for providing and assisting with software for computational promoter models. In addition, the authors are in gratitude to Franklin Pugh at Pennsylvania State University for his support in using the *D. melanogaster* nucleosome data.

Author Contributions

Conceived and designed the experiments: EAR DRW JZ UO. Performed the experiments: EAR DRW AMB. Analyzed the data: EAR DRW AMB DLC. Contributed reagents/materials/analysis tools: DLC TN JZ. Wrote the paper: EAR DRW UO.

References

1. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635.
2. Ni T, Corcoran DL, Rach EA, Song S, Spana EP, et al. (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods 7: 521–527.
3. Juven-Gershon T, Kadonaga JT (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol 339: 225–229.
4. Ohler U, Wassarman DA (2010) Promoting developmental transcription. Development 137: 15–26.
5. Nechaev S, Fargo DC, Dos Santos G, Liu L, Gao Y, et al. (2009) Global Analysis of Short RNAs Reveals Widespread Promoter-Proximal Stalling and Arrest of Pol II in Drosophila. Science 327: 335–338.
6. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the Drosophila genome. Nature 453: 358–362.
7. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132: 887–898.
8. Jin C, Zang C, Wei G, Cui K, Peng W, et al. (2009) H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. Nat Genet 41: 941–945.
9. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, et al. (2005) Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. Cell 123: 233–248.
10. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.
11. Tsukiyama T, Becker PB, Wu C (1994) ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. Nature 367: 525–532.
12. Fu Y, Sinha M, Peterson CL, Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet 4: e1000138. doi:10.1371/journal.pgen.1000138.
13. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF (2006) Nucleosome positions predicted through comparative genomics. Nat Genet 38: 1210–1215.
14. Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 17: 1898–1908.
15. Ganapathi M, Srivastava P, Das Sutar SK, Kumar K, Dasgupta D, et al. (2005) Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. BMC Bioinformatics 6: 126.

16. Ohler U (2006) Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. Nucleic Acids Res 34: 5943–5950.
17. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103: 1412–1417.
18. Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, et al. (2010) High nucleosome occupancy is encoded at human regulatory sequences. PLoS ONE 5: e9129. doi:10.1371/journal.pone.0009129.
19. Ponger L, Duret L, Mouchiroud D (2001) Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res 11: 1854–1860.
20. Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, et al. (2009) A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell 138: 114–128.
21. Tirosh I, Barkai N (2008) Two strategies for gene regulation by promoter nucleosomes. Genome Res 18: 1084–1091.
22. Negre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, et al. (2010) A comprehensive map of insulator elements for the Drosophila genome. PLoS Genet 6: e1000814. doi:10.1371/journal.pgen.1000814.
23. Rach EA, Yuan HY, Majoros WH, Tomancak P, Ohler U (2009) Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. Genome Biol 10: R73.
24. Davey C, Pennings S, Allan J (1997) CpG methylation remodels chromatin structure in vitro. J Mol Biol 267: 276–288.
25. Davey CS, Pennings S, Reilly C, Meehan RR, Allan J (2004) A determining influence for CpG dinucleotides on nucleosome positioning in vitro. Nucleic Acids Res 32: 4322–4331.
26. Kawaji H, Severin J, Lizio M, Waterhouse A, Katayama S, et al. (2009) The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. Genome Biol 10: R40.
27. Tolstorukov MY, Kharchenko PV, Goldman JA, Kingston RE, Park PJ (2009) Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. Genome Res 19: 967–977.
28. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132: 311–322.
29. Wang X, Xuan Z, Zhao X, Li Y, Zhang MQ (2009) High-resolution human core-promoter prediction with CoreBoost_HM. Genome Res 19: 266–275.
30. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 99: 3740–3745.
31. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40: 897–903.
32. Spies N, Nielsen CB, Padgett RA, Burge CB (2009) Biased chromatin signatures around polyadenylation sites and exons. Mol Cell 36: 245–254.
33. Megraw M, Pereira F, Jensen ST, Ohler U, Hatzigeorgiou AG (2009) A transcription factor affinity-based code for mammalian transcription initiation. Genome Res 19: 644–656.
34. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458: 362–366.
35. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, et al. (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature 446: 572–576.
36. Parry TJ, Theisen JW, Hsu JY, Wang YL, Corcoran DL, et al. (2010) The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. Genes Dev 24: 2013–2018.
37. FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C (2006) Comparative genomics of Drosophila and human core promoters. Genome Biol 7: R53.
38. Zeitlinger J, Stark A, Kellis M, Hong JW, Nechaev S, et al. (2007) RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet 39: 1512–1516.
39. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, et al. (2010) c-Myc regulates transcriptional pause release. Cell 141: 432–445.
40. Smith ST, Wickramasinghe P, Olson A, Loukinov D, Lin L, et al. (2009) Genome wide ChIP-chip analyses reveal important roles for CTCF in Drosophila genome organization. Dev Biol 328: 518–528.
41. Chernukhin I, Shamsuddin S, Kang SY, Bergstrom R, Kwon YW, et al. (2007) CTCF interacts with and recruits the largest subunit of RNA polymerase II to CTCF target sites genome-wide. Mol Cell Biol 27: 1631–1648.
42. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, et al. (2009) Unlocking the secrets of the genome. Nature 459: 927–930.
43. Mahmoudi T, Katsani KR, Verrijzer CP (2002) GAGA can mediate enhancer function in trans by linking two separate DNA molecules. Embo J 21: 1775–1781.
44. Lis J (1998) Promoter-associated pausing in promoter architecture and postinitiation transcriptional regulation. Cold Spring Harb Symp Quant Biol 63: 347–356.
45. Kratz A, Arner E, Saito R, Kubosaki A, Kawai J, et al. (2010) Core promoter structure and genomic context reflect histone 3 lysine 9 acetylation patterns. BMC Genomics 11: 257.
46. Adkins NL, Hagerman TA, Georgel P (2006) GAGA protein: a multi-faceted transcription factor. Biochem Cell Biol 84: 559–567.
47. Katsani KR, Hajibagheri MA, Verrijzer CP (1999) Co-operative DNA binding by GAGA transcription factor requires the conserved BTB/POZ domain and reorganizes promoter topology. Embo J 18: 698–708.
48. Hendrix DA, Hong JW, Zeitlinger J, Rokhsar DS, Levine MS (2008) Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. Proc Natl Acad Sci U S A 105: 7762–7767.
49. Cairns BR (2009) The logic of chromatin architecture and remodelling at promoters. Nature 461: 193–198.
50. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, et al. (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol 16: 847–852.
51. Henikoff S, Ahmad K (2005) Assembly of variant histones into chromatin. Annu Rev Cell Dev Biol 21: 133–153.
52. Mito Y, Henikoff JG, Henikoff S (2007) Histone replacement marks the boundaries of cis-regulatory domains. Science 315: 1408–1411.
53. Tirosh I, Barkai N, Verstrepen KJ (2009) Promoter architecture and the evolvability of gene expression. J Biol 8: 95.
54. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128: 669–681.
55. Wilson RJ, Goodman JL, Strelets VB, Flybase Consortium (2008) FlyBase: integration and improvements to query tools. Nucleic Acids Research 36: D588–D592.
56. Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, et al. (2006) Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nat Genet 38: 1151–1158.
57. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261–282.
58. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, et al. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 38: D105–110.
59. Koh K, Kim S-J, Boyd S (2007) An interior-point method for large-scale L1-regularized ligistric regression. J Mach Learn Res 8: 1519–1555.
60. Hochheimer A, Zhou S, Zheng S, Holmes MC, Tjian R (2002) TRF2 associates with DREF and directs promoter-selective gene expression in Drosophila. Nature 420: 439–445.