Detailed Methods and Materials

Growth and development of *D. discoideum*

Strain AX2-214 was grown in liquid nutrient medium containing 1.8% maltose (Watts and Ashworth 1970). Shaking was done at 160 rpm (22 °C) until the cells were in the exponential growth phase and had reached a density of ~4 × 10^6 cells/ml. Crosslinking with formaldehyde (1% final concentration) was done for 15 minutes at 22 °C while shaking. 2.5 M glycine solution was added to 0.125 M final concentration to quench fixation and shaking continued for 5 min. After quenching cells were kept on ice. Cells were pelleted (2 min, 2000 rpm) and gently washed two times with Soerensen phosphate buffer (17 mM Na^+/K^+-phosphate buffer, pH 6.0). They were then resuspended in lysis buffer (50 mM Tris/HCl, pH 7.4, 100 mM NaCl, 5 mM $MgCl_2$, 250 mM sucrose, 1 mM benzamidine, 1 mM PMSF, aprotinin, leupeptin, pepstatin at 10 mg/ml each) and lysed by the addition of 1 % Triton X-100. Lysis was microscopically controlled. Equivalents of 5 × 10^8 cells were distributed into tubes, pelleted at 16,000 × g, the lysis buffer discarded, the samples snap frozen in liquid nitrogen and kept at minus 80°C until further use.

To isolate nucleosomes from cells in a specific developmental stage we chose the aggregation stage in which cells homogeneously acquire new cell surface markers, change their cell shape and become sensitive to cAMP. This is associated with changes in gene expression. Cells were grown in shaking suspension as described above, harvested, washed twice in Soerensen phosphate buffer and resuspended in the same buffer at a density of 1 × 10^7 cells/ml and shaken at 160 rpm and 22 °C for six hours when they had formed aggregates as monitored by microscopy. Crosslinking and lysis was done as above.

Preparation of nucleosomes

Preparation of mononucleosomal DNA is similar to what has been previously described (Albert et al. 2007; Mavrich et al. 2008b). The chromatin pellet from 1 × 10^9 crosslinked and lysed cells (equivalent to 200 ml of cell culture) was washed once in 4 ml NP-S Buffer and spun down at 16,000 × g for 5 minutes at 4°C. The supernatant was discarded, and the chromatin pellet was resuspended in 2 ml NP-S Buffer + 1 mM β-mercaptoethanol. The sample was treated with 2,000 units of Micrococcal Nuclease for 20 minutes at 37 °C to digest the chromatin down to predominantly mononucleosomal size, and the digestion was quenched by adding EDTA to a final concentration of 10 mM and chilling the sample in ice for 10 minutes. MNase has known sequence bias, with A/T-rich dinucleotides being preferred. However, at a high degree of digestion this bias is greatly diminished, since even unfavorable sites are cleaved.

Correction for this bias does not alter nucleosome patterns (Albert et al. 2007). Nonetheless, as the *Dictyostelium* genome is highly A/T-rich we expect even less potential bias.

The sample was treated with Proteinase K at 65 °C overnight and nucleosomal DNA was extracted twice with phenol:chloroform:isoamyl alcohol, treated with RNase for 2 hr, extracted once more with phenol:chloroform:isoamyl alcohol, precipitated with 100 % ethanol, and resuspended in TE. Nucleosomal DNA from 1.5×10^7 cells (equivalent to 64 ml of cell culture) was separated by electrophoresis in a 2 % agarose gel for approximately 1.5 hr at 135 V, and using a long wavelength UV transilluminator DNA fragments ranging in size from 100 to 200 bp were excised. DNA was purified from the agarose using a QIAquick Gel Extraction Kit and was subject to pyrosequencing using the Roche GS20/FLX in accordance with the manufacturer's instructions.

Nucleosome mapping

Due to the A/T-richness of the *Dictyostelium* genome, and thus potentially lower sequence complexity of its DNA, and to inherent sequencing biases of some deep sequencing platforms (Kozarewa et al. 2009; Hoeijmakers et al. 2011), we utilized Roche/454 long-read sequencing technology to sequence entire nucleosomal DNA fragments. This allowed both ends of each nucleosomal DNA fragment to be identified and mapped to the *Dictyostelium* genome with the highest accuracy.

Nucleosome locations on the genome were identified by aligning sequencing reads to the reference genome. Sequencing reads were mapped to the *Dictyostelium discoideum* AX4 reference genome which was released from dictyBase (Kreppel et al. 2004) in February 2008. We used SHRiMP version 1.3.2 (SHort Read Mapping Package) (Rumble et al. 2009) for read mapping. The local alignments produced with the sequencing reads were retained, whose alignment length covers at least 90% of the length of the sequencing read and whose percent identity was 90% or more. Only uniquely aligned reads were used for nucleosome maps. Candidate nucleosome dyads were determined by calculating the coordinate of each sequenced read midpoint. We found that the same specific loci have an abnormally high number of reads in both the vegetative and aggregation read mapping data. Their read numbers were as extreme statistical outliers in terms of the distribution of the mapping read number per each genomic coordinate. Since we detected no special genomic feature surrounding them, we concluded they might be artifacts owing to biased PCR amplification by unknown reasons and filtered them out in the subsequent data analysis.

We collected 274,046 nucleosomes in 246,513 positions across the genome—i.e. 88,265 nucleosomes in 84,441 unique positions from vegetative *Dictyostelium* and 185,781 nucleosomes in 162,072 unique positions from multicellular aggregates. These numbers represent a genome-wide sampling, rather than complete coverage. We assumed that the digested DNA fragment is centered at the nucleosome midpoint (or dyad) and examined the composite distribution of single nucleotides and dinucleotides of sequenced fragments around each nucleosome midpoint.

Transcriptome sequencing and analysis

In order to identify transcription start sites (TSS) and transcription end sites (TES), transcriptome sequencing with the Roche GS20/FLX platform was employed in this study. As the transcribed regions are defined by independent experimental measurements, it seems unlikely that the patterns in Fig. 1 arose from systematic errors. Moreover, it seems unlikely the reverse transcriptase used in mapping the RNA TSS prematurely stopped at poly-T tracts that randomly occur throughout the transcribed region since other mapped features (e.g. TATA box, ORF start, and nucleosomes) all showed a canonical and non-random distribution relative to the measured TSS.

For *D. discoideum* several axenic laboratory strains are available. Most common are AX4 and AX2. AX4 was the strain which was sequenced. It is characterized by a 750 kb duplication on chromosome 2, which might influence expression at least in this duplicated region. We therefore decided to use the wild type strain NC4 from which the sequenced AX4 strain as well as AX2 were derived. NC4 grows only in association with bacteria. We used *Klebsiella planticola* as food source.

NC4 cells were mixed with a culture of these bacteria in logarithmic growth phase and plated onto NA medium of the following composition: a mixture of 0.5 g glucose, 0.5 g peptone, 20 ml 50× phosphate buffer pH6.5, and diluted with 1000 ml H_2O, then 20 ml phosphate buffer, 99.86 g KH_2PO_4, 17.8 g $NaHPO_4$× $2H_2O$, and 1000 ml H_2O added again. When the bacteria were consumed the amoebae were recovered from the plates and washed twice with the phosphate buffer to remove residual bacteria. The amoebae were then plated onto phosphate agar plates (buffered plates without nutrients). The plating time point was taken as the initiation of the developmental cycle. Cells were harvested and RNA was extracted using the RNeasy Mini Kit from Qiagen at 0, 4, 8, 12, 16, 20, and 24 hours during development. The RNA was converted to cDNA using the Evrogen Mint Kit. The sequencing libraries were prepared from this cDNA and subject to pyrosequencing by Roche GS20/FLX system according to the instructions of the manufacturer on experimental setup and design.

The resulting sequencing reads were sorted into "gene bins" according to the annotated genome of *D. discoideum* AX4 (http://dictybase.org, released in February 2008) using BLAT (Kent 2002), and we mapped the positions of 5' and 3' UTRs of protein-coding genes. In order to find intron/exon boundaries the best hits were aligned to the genome using EXALIN (Zhang and Gish 2006). To obtain an unbiased location of transcripts, all seven RNA-seq data of 0, 4, 8, 12, 16, 20, and 24 hours were compiled together, and the 5` and 3' most base of the matching region of a gene were defined as TSS and TES, respectively. As a result, the TSS of 5,468 protein-coding genes and the TES of 5,400 protein-coding genes were annotated in this study (**Supplemental Table 3**), and were compared to TSS locations that pre-exist in the literature (**Supplemental Table 4**). Furthermore, the read counts for each gene at the indicated time point of development were taken as a proxy for expression levels. Normalization of the data was done with the following formula: (number of mapped reads to an individual CDS ÷ (all RNAseq reads × gene length)) × 10^8. Differential expression was assumed if the read numbers between start point of development and 4 and/or 8 hours into the cycle were at least 5 fold different. Among the 5,468 genes with the annotated TSS, 325 genes were finally grouped as developmentally upregulated genes, which were highly expressed during the first 8 hours of starvation (**Supplemental Table 3**).

Data analysis

The consensus sequence of TATA element, TATAAA[AT][AT], *for D. discoideum* was determined based on the transcriptome mapping data in this study, which allows us to analyze most significant TATA-containing DNA sequences. In order to identify over-represented TATA sequence motifs in core promoters, we initially used the *D. discoideum* consensus TATA box TATAAA[T/A]A, which was reported in 15 genes (Kimmel and Firtel 1983). We located TATA boxes with this consensus as a preliminary screen and aligned their locations by the TSS of 5,468 genes annotated in this study. We extracted the DNA sequences in the core promoter region from -37 to -21 bp upstream of the TSS, where bona fide TATA boxes are highly likely to be located. Since the purpose of this analysis was to identify the most authentic TATA sequence, we considered the TATA sequences only in the sense (or sense) strand (Basehoar et al. 2004). We looked for over-represented motifs with those sequences using the MEME software (Bailey and Elkan 1994). A significantly over-represented TATA motif was identified by the MEME algorithm for *D. discoideum* and its consensus TATA sequence, TATAAA[AT][AT], was used in this study.

In accordance with the systematic nomenclature of nucleosome positions in the study of Jiang and Pugh (Jiang and Pugh 2009), we demarcated the border of the -1, NFR, and +1 zones around transcription start sites by using the level of nucleosome occupancy. As a result, we categorized nucleosome positions relative to the TSS for both life stages, where the nucleosome dyad is defined as +1 nucleosomes if positioned between +60 and +187 and -1 nucleosomes from -294 to -110, and thus the boundary of 5' NRF is defined as the region from -109 (as the right border of -1) to +59 (as the left border of +1). The nucleosomes between the +1 nucleosome and the transcription end site of each gene are grouped as other genic nucleosomes such as +2, +3, etc. The TES of some genes was not available, and the ORF end site was used instead. Since the nucleosome occupancy of the -1 nucleosome was moderate, a wider range was chosen by our visual inspection (compared with Yeast (Jiang and Pugh 2009), i.e. from -307 to -111). As a result, 73,396 nucleosomal core particle DNAs were categorized as -1, +1, and the other genic nucleosomes as of 5,468 protein coding genes with the annotated TSS.

In order for the NPS (nucleosome-positioning sequence) analysis with nucleosomal core particle DNA sequences, the frequency of 10 dinucleotide sets (Satchwell et al. 1986) was compiled as a function of the distance from the nucleosome dyad. Each dinucleotide is defined by reading their sequence in the 5' to 3'-end direction, and all plots in this study were prepared according to what has been described elsewhere (Albert et al. 2007; Mavrich et al. 2008a; Mavrich et al. 2008b). Furthermore, the expected frequency of each dinucleotide was estimated by random sampling of any DNA sequence in the entire *Dictyostelium* genome.

External data sources

The nucleosome organization of seven major eukaryotes, human, *D.melanogaster*, *S.pombe*, *S.cerevisiae*, *A.thaliana*, and *C.elegans,* were chosen by considering availability of their nucleosome mapping data in public repositories and the evolutionary position in the tree of life (Dacks and Doolittle 2001). For human, mononucleosome mapping data by MNase digestion was obtained in Schones et al. (Schones et al. 2008), which is for CD4+ T cells without activation by TCR signaling. We used 3,859 TSSs of the expressed genes (courtesy of Dr. Schones) that were screened by microarray experiments in the study and produced the composite distribution of nucleosome positions at the 5' end of genes.

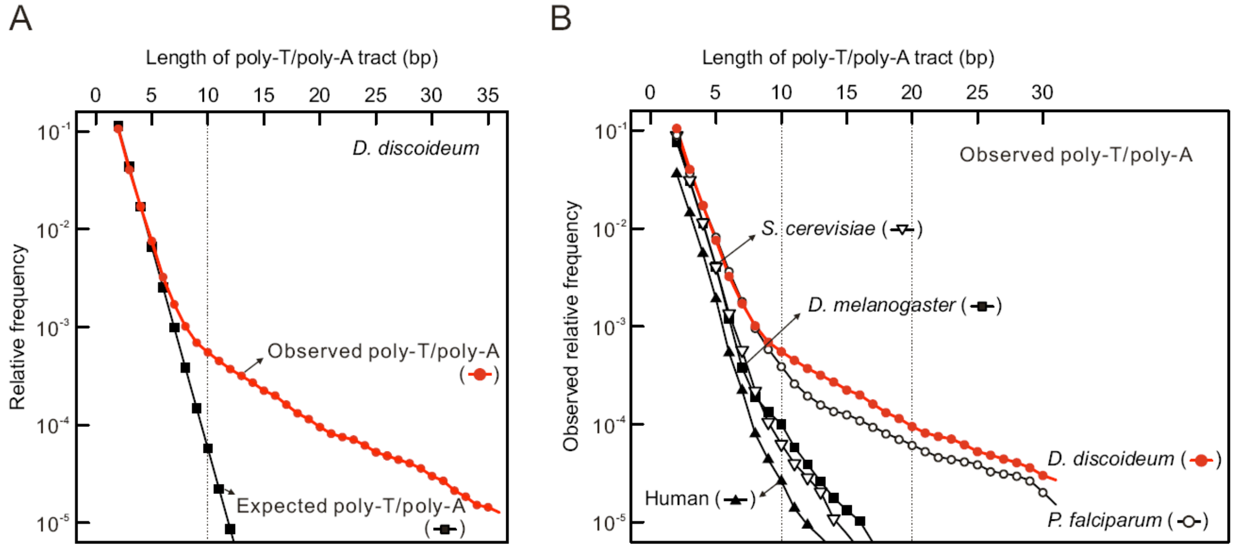
The nucleosome occupancy data of *S. pombe* was obtained from Lantermann et al. (Lantermann et al. 2010) (courtesy of Dr. Korber), which were determined at 20-bp resolution by Affymetrix S. pombe Tiling 1.0FR array. The nucleosome organization was obtained according to the method described in the study with the TSS of around 4,000 genes. The high resolution maps of nucleosome positions for *S. cerevisiae* and *D. melanogaster* were obtained in our previous publications (Albert et al. 2007; Mavrich et al. 2008a; Mavrich et al. 2008b). The composite distributions of nucleosome positions for these species were produced using the same TSS annotation sets.

For *A. thaliana*, the nucleosome mapping data was obtained as of shoots of three-week old *Arabidopsis* plants in Chodavarapu et al. (Chodavarapu et al. 2010). The nucleosome distribution around the 5' end of genes was derived with the TSS annotation curated by The *Arabidopsis* Information Resource (TAIR). Also, we obtained the mapping data for *C. elegans* in mixed stages from Valouev et al. (Valouev et al. 2008). 5-pile data set was used to show the nucleosome organization among the full data set that was downloaded from the UCSC genome browser.

Data visualization and statistical significance test

Data plots were generated using the statistical package of R (version 2.11.1) and PERL scripts with GD graphics library. The density for the locations of specific features (e.g. poly-T/poly-A tracts in the sense strand) were calculated by Gaussian kernel function with bandwidths described, which is implemented in R. For the statistical significance test of each pattern in the plots (e.g. Figs. 1 and 3), the expected frequency of mononucleotide or dinucleotide was estimated by randomly sampling the same number of DNA sequences at any location in the *Dictyostelium* genome. This simulation assumed hypothetical genomic features (e.g. nucleosomal DNA, TSS, or TES) which are randomly positioned in the sequence-independent manner. Two randomization simulations were independently performed and shown in all plots.

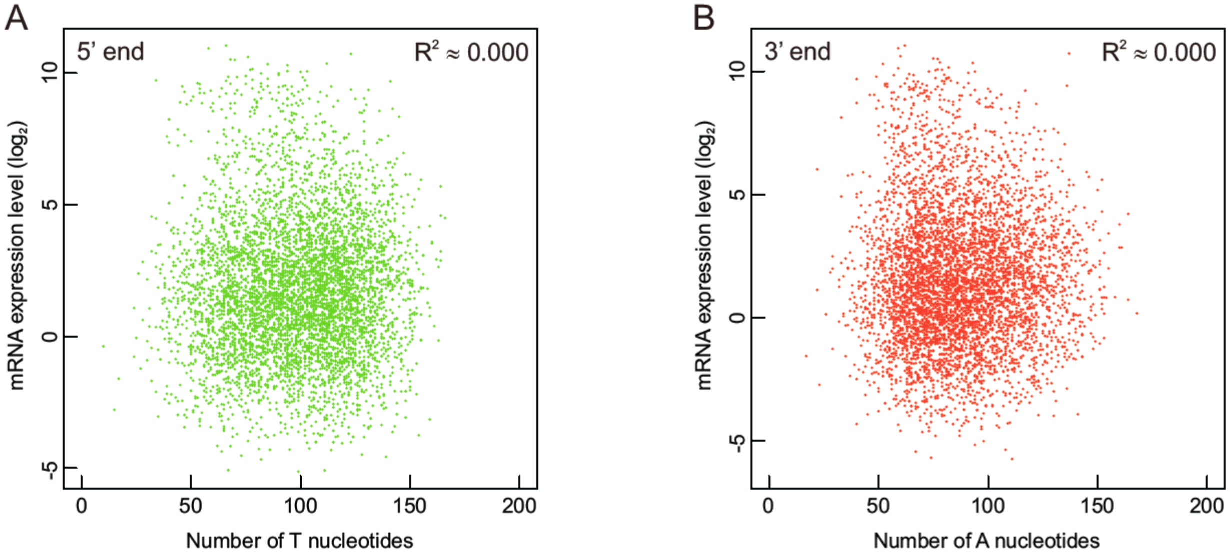
Supplemental Figures



Supplemental Figure 1. Overrepresentation of poly-T/poly-A tracts in the *Dictyostelium* genome.

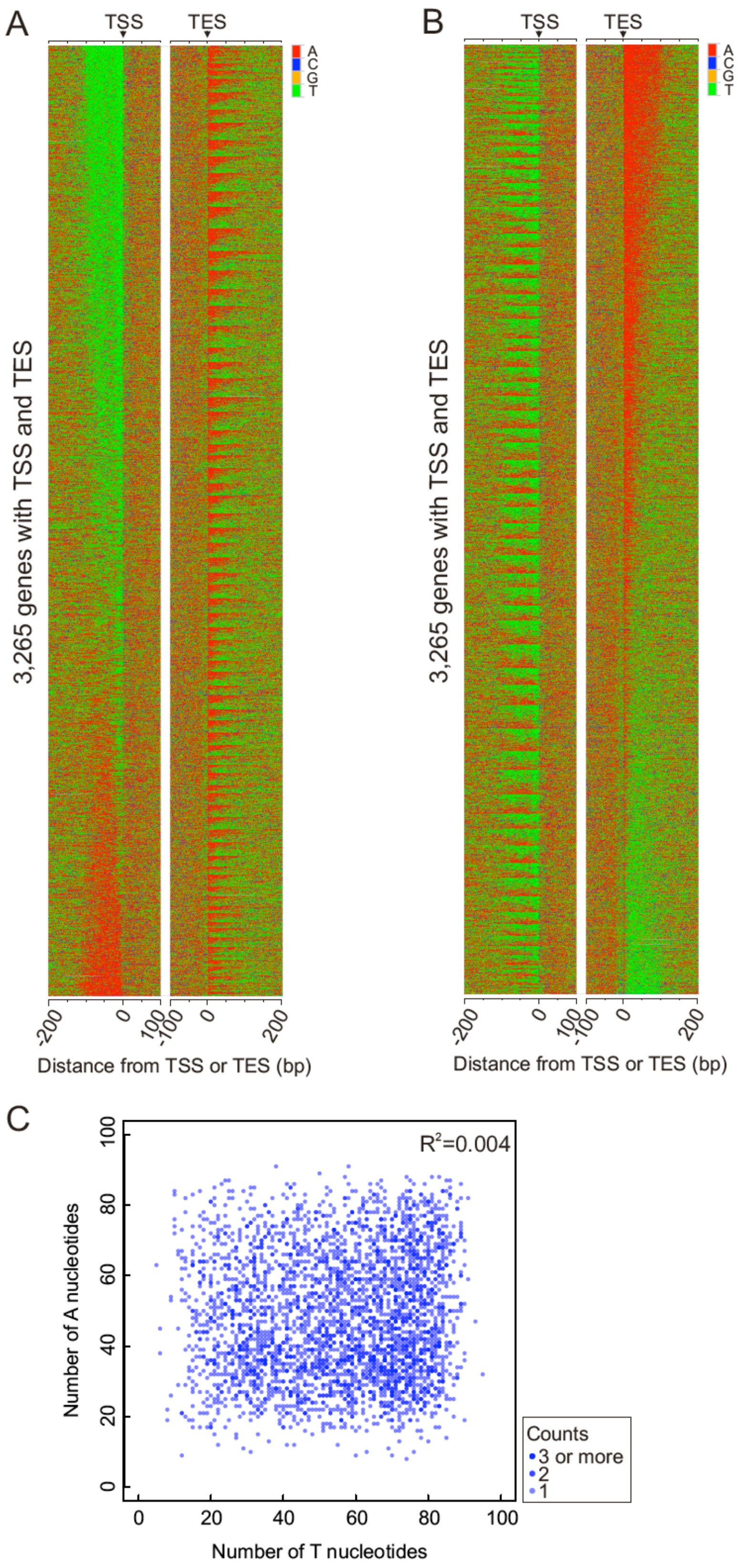
(A) Frequency distribution of poly-T/poly-A tracts in the *Dictyostelium* genome. The occurrence of non-overlapping poly-T/poly-A tracts of length 2 or more (n ≥ 2) were counted in all 6 chromosomes of *D. discoideum* (total 33,929,503-bp genome size), and their relative frequencies were displayed as a function of the tract length (n) in the figure. The observed relative frequency of each tract of a given length was calculated as a ratio of their count number to the genome size, which is the number of observed tracts normalized to the genome size. For instance, n=10 was observed to occur once every 1,810 bp in the genome, whereas the expected frequency of n=10 was only once every 17,400 bp. The expected frequency of the occurrence of a non-overlapping poly-T/poly-A tract was calculated as a zero-order Markov chain (Dechering et al. 1998).

(B) Frequency distribution of poly-T/poly-A tracts in five eukaryotic genomes. The observed relative frequency of poly-T/poly-A tracts was shown as a function of the tract length after normalization by genome size, as described. As a result, the high incidence of long poly-T/poly-A tracts is unique to A/T-rich genomes, i.e. *D. discoideum* and *P. falciparum*. For example, 90,201 (n ≥ 12) tracts were found in the *Dictyostelium* genome, which means n ≥ 12 was observed to occur once every 376 bp in average. In *P. falciparum*, 34,798 found for n ≥ 12, thus once every 669 bp. Moreover, 1,117 were found, occurring on average once every 10,807 bp in *S. cerevisiae*, 15,617 were found once every 7,708 bp in *D. melanogaster*, and 97,541 found once every 31,581 bp in the human genome.

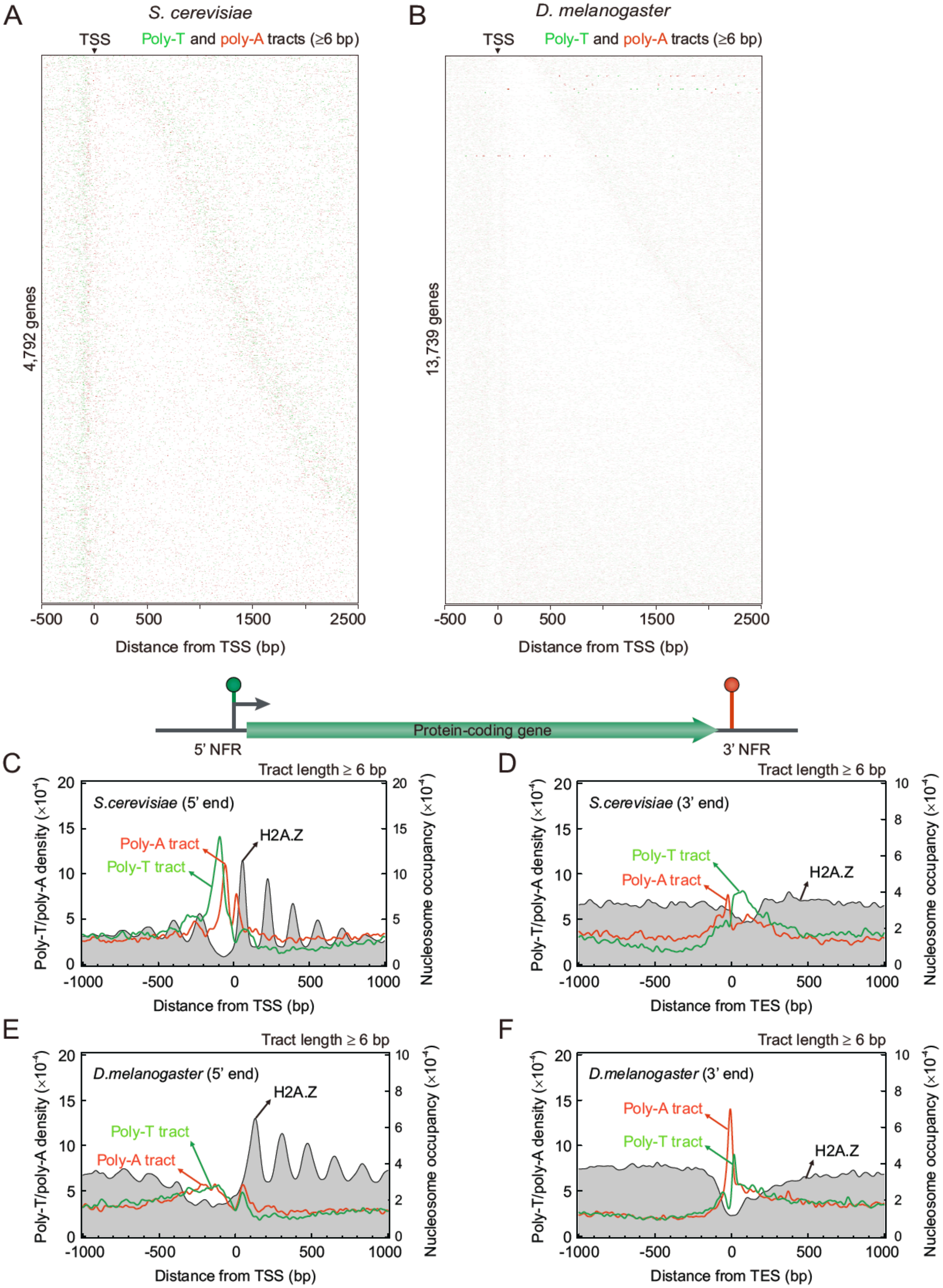


Supplemental Figure 2. Lack of significant correlation of transcript abundance and the enrichment of poly-T/poly-A tracts at the ends of *Dictyostelium* genes.

Correlation analysis between poly-T tracts at the 5' end (from -200 to TSS) of 5,468 protein-coding genes and their mRNA expression level **(A)** and between poly-A tracts at the 3' end (from TES to 200) of 5,400 protein-coding genes and their expression level **(B)**. Each scatter plot represents the number of T's at the 5' end and A's at the 3' end of those *Dictyostelium* genes versus the mRNA expression level. The mRNA transcript abundance of a gene was calculated in the transcriptome analysis after data normalization. As shown with $R^2 \approx 0.000$ and $R^2 \approx 0.000$, no significant correlation was found between poly-T/poly-A tracts and transcript abundance of *Dictyostelium* genes.



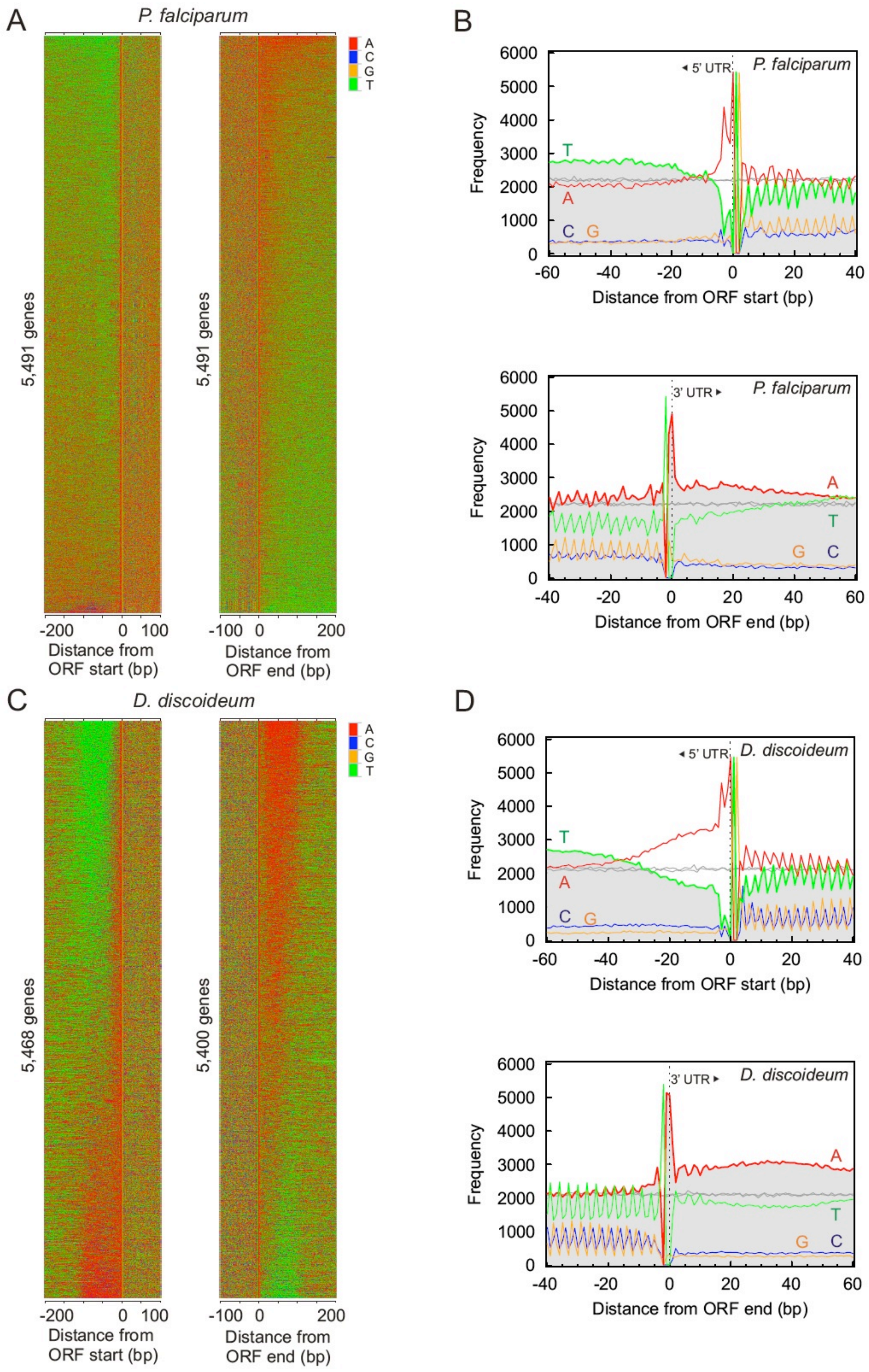
Supplemental Figure 3. Lack of significant correlation between poly-T enrichment near the TSS and poly-A enrichment near the TES in the *Dictyostelium* genome. **(A,B)** DNA sequence track views around the 5' (TSS) and 3' (TES) end of *Dictyostelium* genes. Both TSS and TES were identified for 3,265 protein-coding genes of *D. discoideum* in this study. Two track views were produced akin to **Figure 1C** of the main text, and each paired tracks were linked at the 5' and 3' end. Plots were primarily sorted by "T" nucleotide content in the 5' region (from -100 to TSS) in panel **A**, and by "A" nucleotide content in the 3' region (from TES to +100) in panel **B**.
(C) Correlation analysis between poly-T at the 5' end (from -100 to TSS) and positionally-linked poly-A at the 3' end (from TES to +100). As a result, 3,265 data points were used to extrapolate the linear regression line in the figure, and the coefficient of determination (R^2) was 0.004. When a 20-bp region was chosen near the TSS and TES instead of 100-bp, no significant correlation was detected (R^2 = 0.003, data not shown).


A
S. cerevisiae
TSS
Poly-T and poly-A tracts (≥6 bp)
4,792 genes
-500
0
500
1500
2500
Distance from TSS (bp)
B
D. melanogaster
TSS
Poly-T and poly-A tracts (≥6 bp)
13,739 genes
-500
0
500
1500
2500
Distance from TSS (bp)
Protein-coding gene
5' NFR
3' NFR
C
Tract length ≥ 6 bp
S.cerevisiae (5' end)
Poly-A tract
Poly-T tract
H2A.Z
Poly-T/poly-A density (×10⁻⁴)
Nucleosome occupancy (×10⁻⁴)
0
5
10
15
20
-1000
-500
0
500
1000
Distance from TSS (bp)
D
Tract length ≥ 6 bp
S.cerevisiae (3' end)
Poly-T tract
Poly-A tract
H2A.Z
Poly-T/poly-A density (×10⁻⁴)
Nucleosome occupancy (×10⁻⁴)
0
5
10
15
20
0
2
4
6
8
10
-1000
-500
0
500
1000
Distance from TES (bp)
E
Tract length ≥ 6 bp
D.melanogaster (5' end)
H2A.Z
Poly-T tract
Poly-A tract
Poly-T/poly-A density (×10⁻⁴)
Nucleosome occupancy (×10⁻⁴)
0
5
10
15
20
0
2
4
6
8
10
-1000
-500
0
500
1000
Distance from TSS (bp)
F
Tract length ≥ 6 bp
D.melanogaster (3' end)
Poly-A tract
Poly-T tract
H2A.Z
Poly-T/poly-A density (×10⁻⁴)
Nucleosome occupancy (×10⁻⁴)
0
5
10
15
20
0
2
4
6
8
10
-1000
-500
0
500
1000
Distance from TES (bp)

Supplemental Figure 4. Distribution of poly-T/poly-A tracts around the 5' and 3' end of Pol II-transcribed genes of *S. cerevisiae* and *D. melanogaster*.

(A,B) Distribution of the locations of poly-T/poly-A tracts in the DNA sequence of all genes with the known TSS. The figures were generated by the same protocol described in **Figure 1A**.

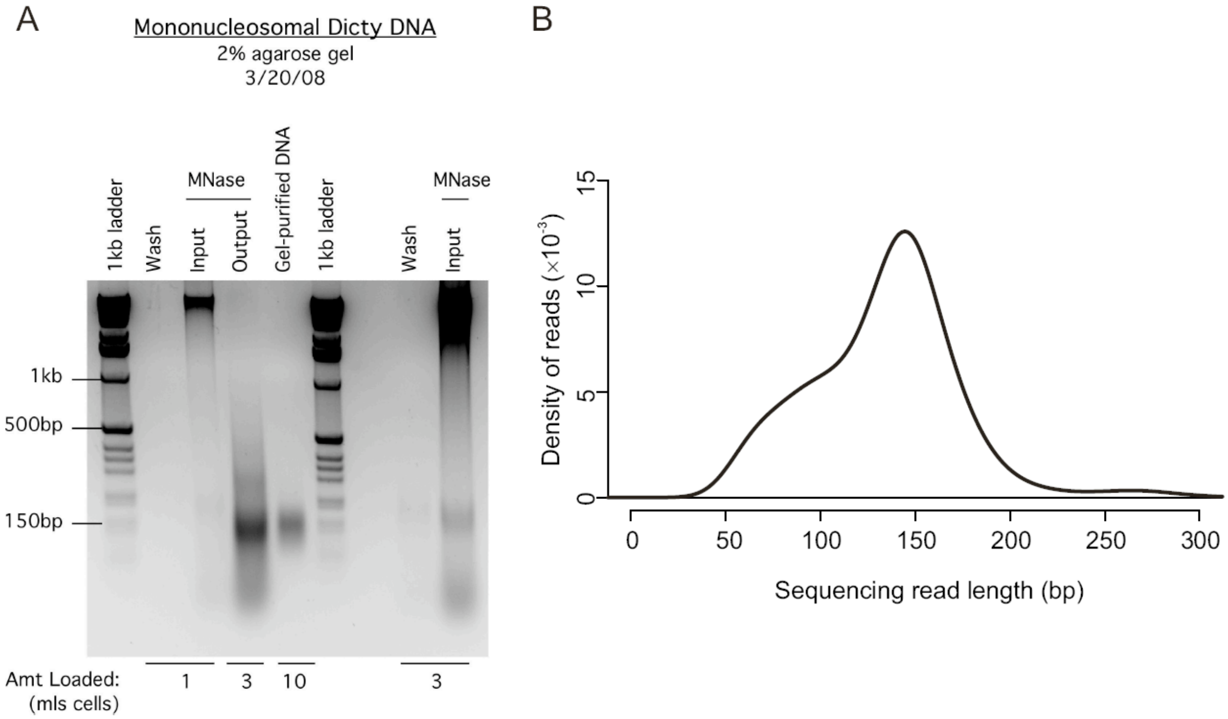
(C-F) Composite distribution of the location of poly-T/poly-A tracts around the TSS **(C,E)** and TES **(D,F)** of all annotated genes of *Saccharomyces* **(C,D)** and *Drosophila* **(E,F).** The density of poly-T/poly-A tracts (n ≥ 6) was estimated as a function of the distance (bp) between the middle location of each tract and the given TSS as described in **Figure 1B**, except the smoothing bandwidth of 10 bp used. Tracts were counted from the sense (nontemplate) strand. The composite distributions of nucleosome positions were added as gray-filled traces (Albert et al. 2007; Mavrich et al. 2008b).


A
P. falciparum
A
C
G
T
5,491 genes
5,491 genes
-200
0
100
Distance from ORF start (bp)
-100
0
200
Distance from ORF end (bp)
B
6000
5000
4000
3000
2000
1000
0
Frequency
◂ 5' UTR
P. falciparum
T
A
C
G
-60
-40
-20
0
20
40
Distance from ORF start (bp)
3' UTR ▸
P. falciparum
A
T
G
C
-40
-20
0
20
40
60
Distance from ORF end (bp)
C
D. discoideum
A
C
G
T
5,468 genes
5,400 genes
-200
0
100
Distance from ORF start (bp)
-100
0
200
Distance from ORF end (bp)
D
6000
5000
4000
3000
2000
1000
0
Frequency
◂ 5' UTR
D. discoideum
T
A
C
G
-60
-40
-20
0
20
40
Distance from ORF start (bp)
3' UTR ▸
D. discoideum
A
T
G
C
-40
-20
0
20
40
60
Distance from ORF end (bp)

Supplemental Figure 5. Comparison of the DNA content around the 5’ and 3’ end of protein-coding genes between *P. falciparum* and *D. discoideum*. We did not have high resolution maps of TSS and TES for *Plasmodium*, and so could only examine distributions around ORF start and end sites. Therefore we also generated similar ORF-centered maps in *Dictyostelium*. In comparison to *Dictyostelium*, similar A and T nucleotides biases did not exist in the expected regions upstream of *Plasmodium* ORF start sites.

(A,C) DNA sequence track view around the 5’ and 3’ end of *Plasmodium* **(A)** or *Dictyostelium* **(C)** ORFs. We curated 5,491 protein-coding genes that were annotated in *Plasmodium* Genomics Resource (7.2 release) (Aurrecoechea et al. 2009). Genes were aligned by the ORF start/end and sorted by the T or A content in the whole fetched DNA sequences (301 bp).

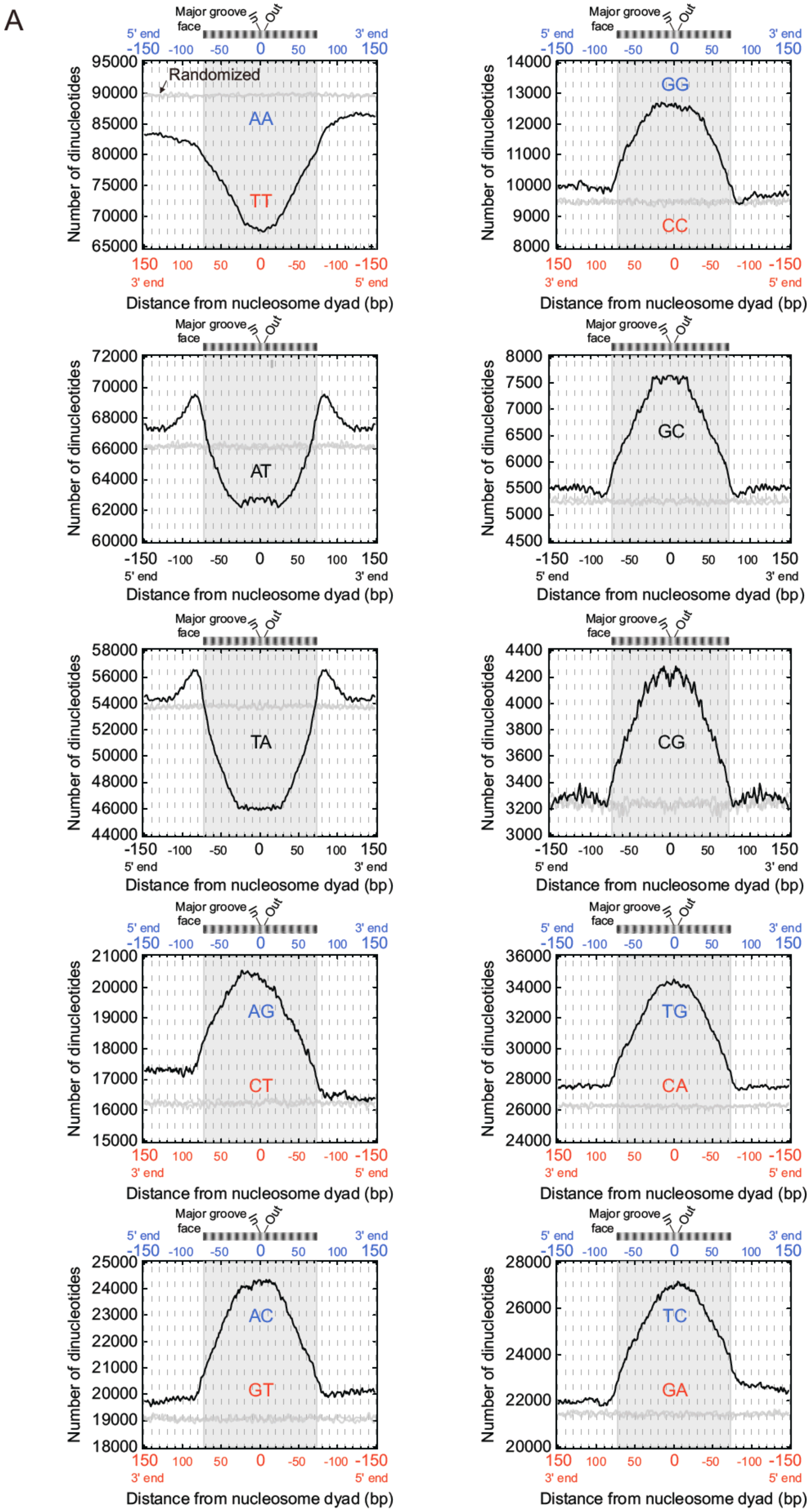
(B,D) Frequency distribution of A, C, G and T nucleotide at every position across the DNA sequence including the start/end of *Plasmodium* **(B)** or *Dictyostelium* **(D)** ORFs. The same frequency distribution of each nucleotide was obtained for the *Plasmodium* genome, as described previously except that the DNA sequences were fetched in the region from -60 to 40 relative to the ORF start (upper panel) and from -40 to 60 relative to the ORF end of 5,491 genes (lower panel). Also, two random simulations were performed to estimate the expected frequency of T and A nucleotides, shown as two gray traces in the figures.

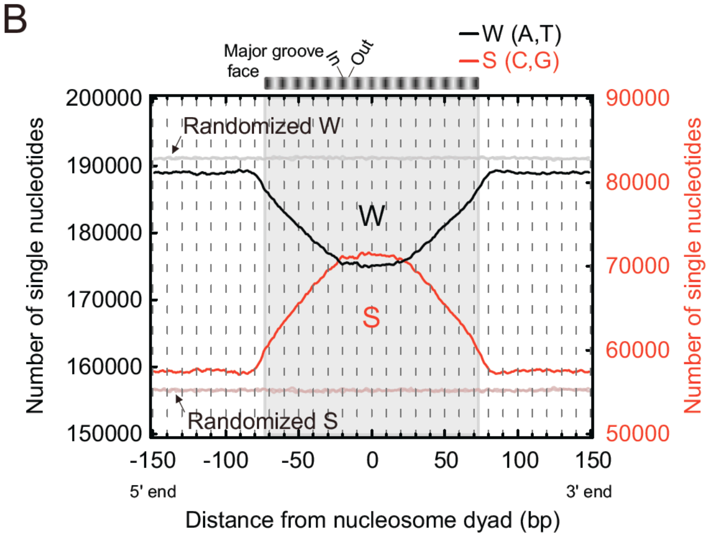


Supplemental Figure 6. Isolation of nucleosome core particle DNA from *D. discoideum*.

(A) Ethidium bromide stained agarose gel of samples in mononucleosome preparation for sequencing. After crosslinked and lysed, omatin pellets were treated with MNase and digested to predominantly mononucleosomal size. Finally, resulting nucleosomal DNAs were size selected by gel purification. The nucleosomal DNA sample prepared in the vegetative stage was shown in the figure.

(B) Frequency distribution of the length of sequencing reads for nucleosomal DNA. The frequency distribution was calculated with nucleosome core particle DNA prepared and sequenced from multicellular *D. discoideum* aggregates. This figure included all sequencing reads and the mode of the read length was 144 bp. Also, a 133-bp mode was calculated in the vegetative stage.

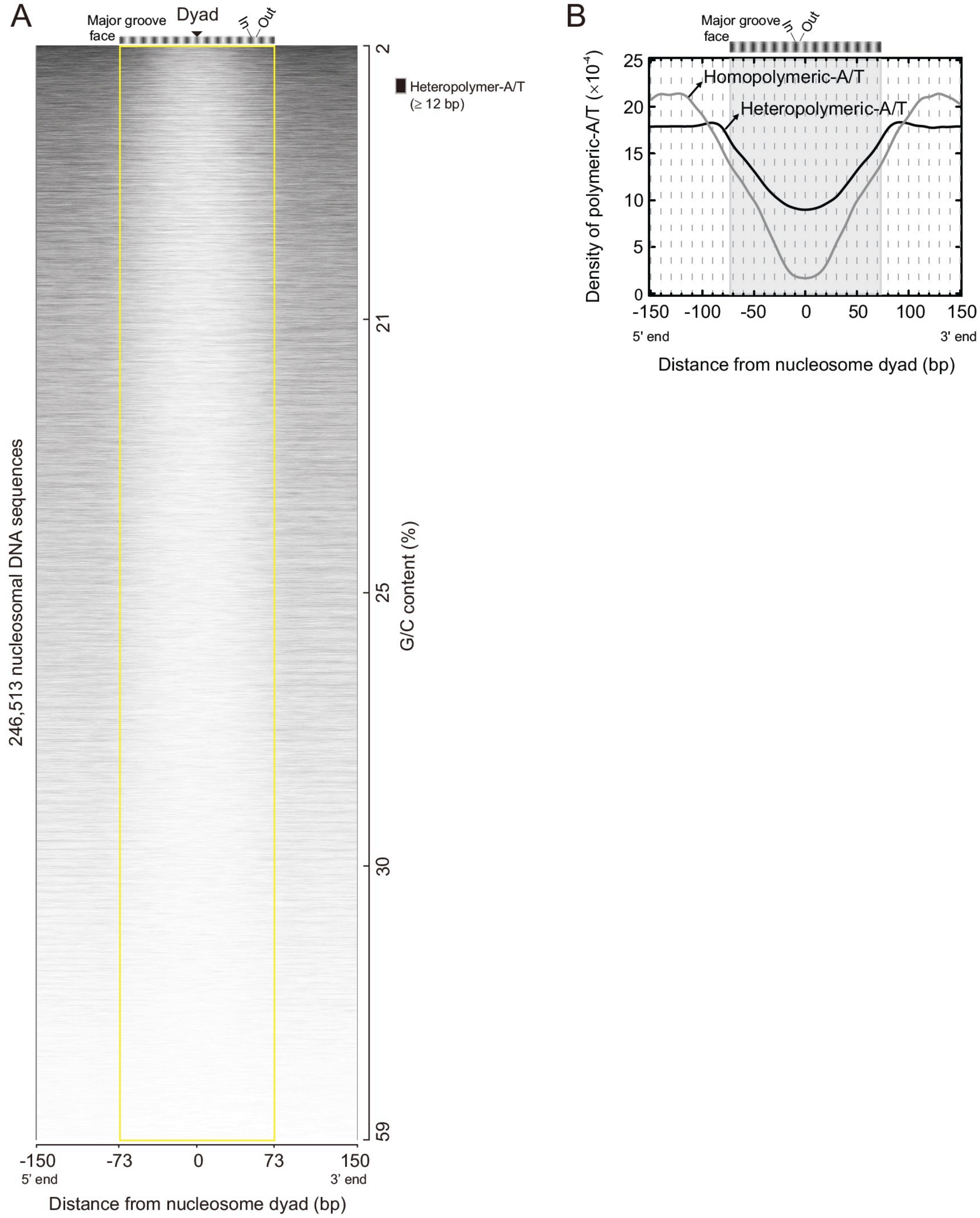

A
Major groove face
In
Out
5' end
3' end
-150 -100 -50 0 50 100 150
150 100 50 0 -50 -100 -150
Distance from nucleosome dyad (bp)
Number of dinucleotides
Randomized
AA
TT
95000 90000 85000 80000 75000 70000 65000
GG
CC
14000 13000 12000 11000 10000 9000 8000
AT
72000 70000 68000 66000 64000 62000 60000
GC
8000 7500 7000 6500 6000 5500 5000 4500
TA
58000 56000 54000 52000 50000 48000 46000 44000
CG
4400 4200 4000 3800 3600 3400 3200 3000
AG
CT
21000 20000 19000 18000 17000 16000 15000
TG
CA
36000 34000 32000 30000 28000 26000 24000
AC
GT
25000 24000 23000 22000 21000 20000 19000 18000
TC
GA
28000 26000 24000 22000 20000



Supplemental Figure 7. Frequency distribution of the indicated dinucleotide pairs and W, S nucleotides at every position across the nucleosomal DNA.

(A) Frequency distribution of the indicated dinucleotide pairs at every position across the 147-bp nucleosomal DNA and flanking regions. Each dinucleotide is defined in the 5' to 3'-end direction, for example 5'-AT-3'. The number of each dinucleotide was counted by reading both strands in the 5' to 3' direction. Only 10 unique frequency distributions exist from 16 dinucleotides—AA (= TT), GG (= CC), AT, GC, TA, CG, AG (= CT), TG (= CA), AC (= GT), and TC (= GA)(Satchwell et al. 1986; Albert et al. 2007). 246,513 positions of nucleosome dyads were identified in the two genome-wide nucleosome maps of vegetative and aggregating cells. Their nucleosomal DNA sequences were aligned by their nucleosome dyads and extended to include the linker DNA (i.e. from -150 to 150 relative to the dyad). The frequency count of each dinucleotide was shown in the y axis after smoothing (see the Methods). The schematic bar in the upper part of each plot represents the rotational orientation of the major groove against the histone octamer surface. For the significance test of each pattern, the expected frequency of each dinucleotide was calculated by randomly picking up the same numbers of DNA sequences at any location from the *Dictyostelium* genome. This simulation assumed the dinucleotide frequency of hypothetical nucleosomes, which are randomly positioned in a sequence-independent manner. Two computational simulations were independently carried out and shown as two gray traces in each plot.

(B) Frequency distribution of the W and S single nucleotides at every position across the 147-bp nucleosomal DNA. The frequency distribution of W (A or T) and S (G or C) was calculated by reading the 246,513 nucleosomal sequences. The count number of a given nucleotide was obtained in the forward strand, which was finally summed up from all the sequences in a composite manner and smoothed using a three-base moving average algorithm. The frequency of W was shown in black and S in red in the figure, and their expected frequencies were also calculated described as in **A**, displayed as gray for W and light red for S. As shown, *in vivo* nucleosome placement was associated by being G/C-rich near the nucleosomal dyads.

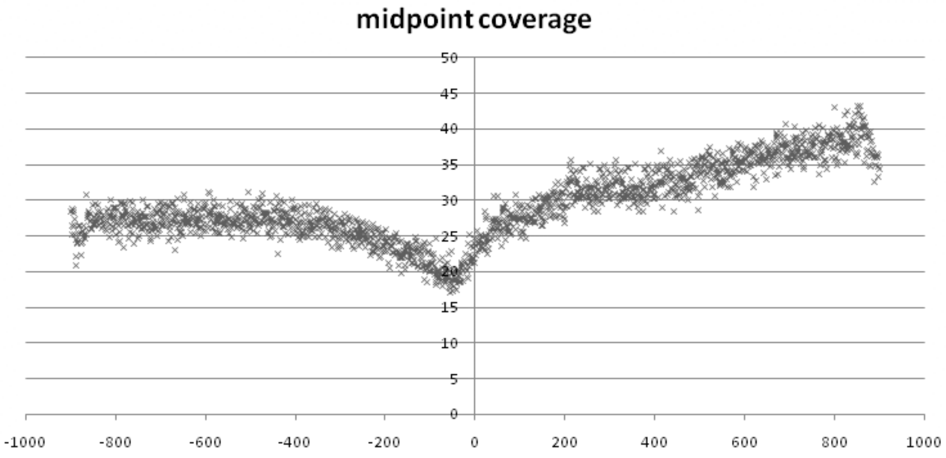

A
Major groove face
Dyad
In
Out
2
21
25
30
59
G/C content (%)
■ Heteropolymer-A/T (≥ 12 bp)
246,513 nucleosomal DNA sequences
-150
5' end
-73
0
73
150
3' end
Distance from nucleosome dyad (bp)
B
Major groove face
In
Out
Density of polymeric-A/T (×10^-4)
25
20
15
10
5
0
Homopolymeric-A/T
Heteropolymeric-A/T
-150
5' end
-100
-50
0
50
100
150
3' end
Distance from nucleosome dyad (bp)

Supplemental Figure 8. *In vivo* nucleosome placement with polymeric-A/T sequences near nucleosome peripheries.

(A) Distribution of the locations of heteropolymeric-A/T tracts across nucleosomal core particle DNAs. 246,513 nucleosomal DNA sequences were sorted by G/C content that varied from 2% to 59%. Each track represents an individual nucleosomal DNA sequence, in which the dyad of the 147-bp nucleosomal core DNA was centered as shown in the yellow box. All DNA sequences in the forward strand were aligned by their dyad and shown in the 5' to 3' direction. They included linker DNAs by being extended from -150 to 150 relative to the nucleosome dyad. The G/C content was measured by taking the whole 301-bp DNA sequence, and all tracks were displayed in ascending order of G/C content in the figure. If any heteropolymeric-A/T tracts (n ≥ 12) were found along the nucleosomal DNA sequence, all the bases were colored as black in the tract. The other bases were colored in white. The homopolymeric stretches of A or T (called as poly-T/poly-A tracts) were not included as heteropolymeric-A/T.

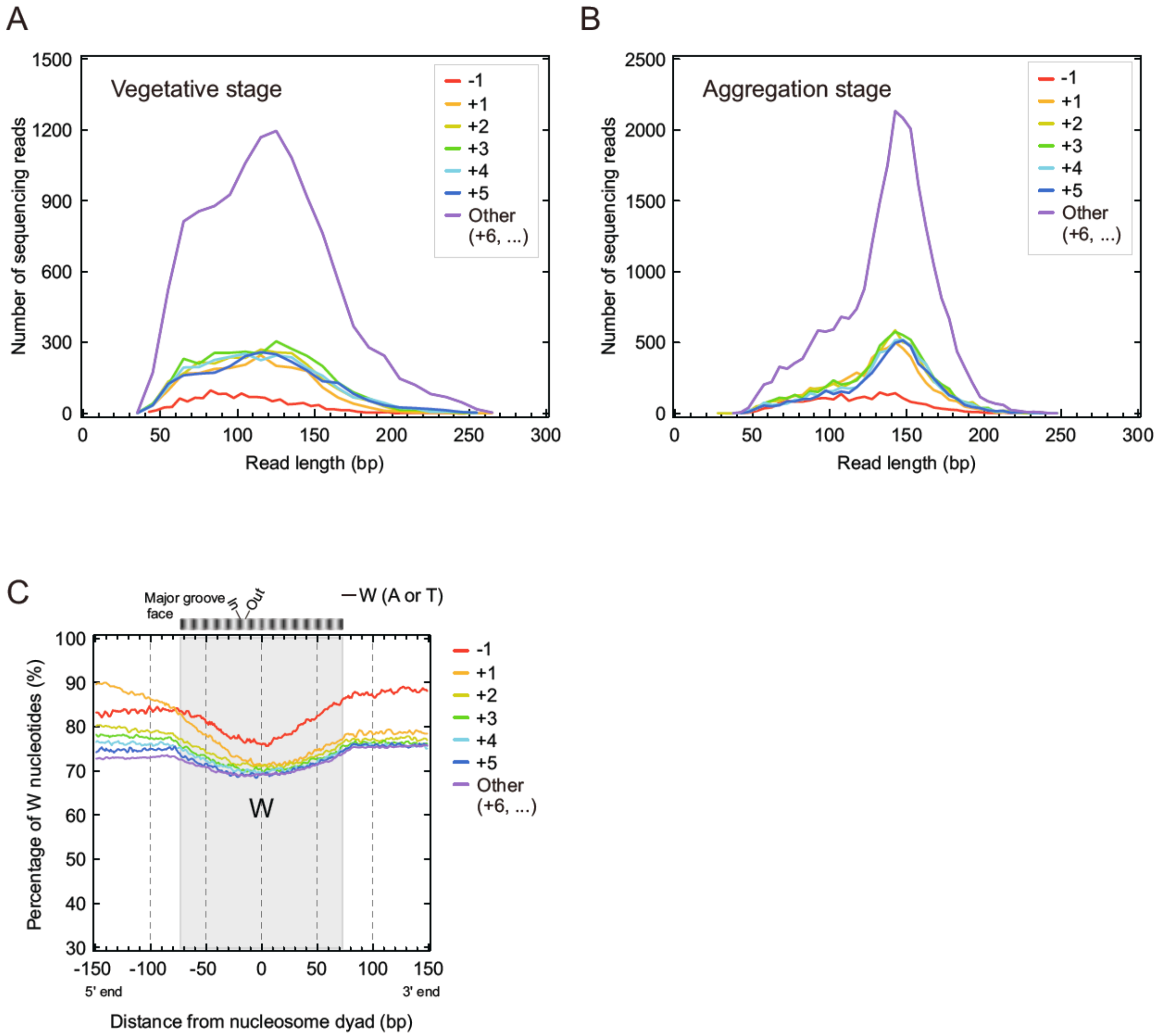
(B) Composite distribution of the locations of polymeric-A/T tracts across nucleosomal DNAs. Heteropolymeric-A/T sequences were counted per each position across the 246,513 nucleosomal DNA sequences. The density of the occurrence of heteropolymeric-A/T tracts (n ≥ 12) was estimated using Gaussian kernel with the smoothing bandwidth of 5, and plotted as a function of the distance between the tract and dyad. This relative distance was calculated by the location of the center of each heteropolymeric-A/T sequence and their frequency was calculated in the forward strand, shown as a black trace in the figure. Also, the frequency distribution of homopolymeric-A/T sequences (n ≥ 12) was obtained in the same manner and displayed as grey in the figure. The density of polymeric-A/T tracts towards the nucleosome border is relatively higher than in the dyad region, which indicates the higher incidence of polymeric-A/T tracts in the linker DNA.

D. lacteum coding sequence start site coverage with 454 reads



Mean coverage of genome: 28X

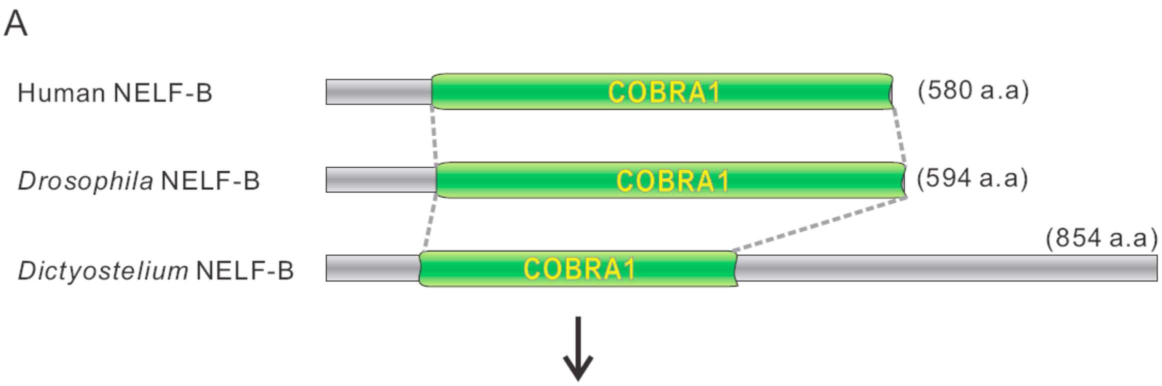
Supplemental Figure 9. Distribution of 454 sequencing tag midpoints (in bp) around the ORF start of *Dictyostelium lacteum*. This genome is approximately as AT rich as the *D. discoideum*. Only *D. lacteum* had available whole genome sequencing done on the 454 instrument. Note that the moderate depletion just upstream of the ORF start and the overall pattern of coverage was not as strong in magnitude nor reflected the pattern as seen for *D. discoideum* nucleosome maps, indicating that the nucleosome patterns were not caused by differences in extraction efficiency.



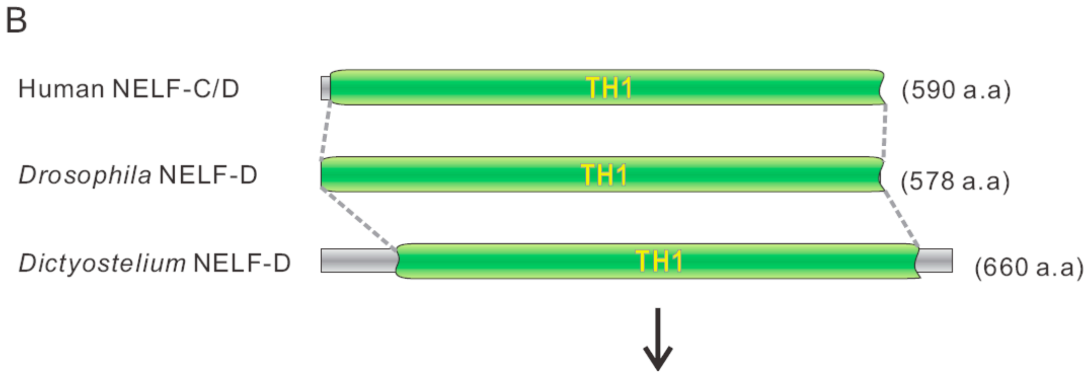
Supplemental Figure 10. Nucleosomal DNA properties of *Dictyostelium* intergenic nucleosomes and genic nucleosomes.

(A,B) Distribution of the size of nucleosome protected DNA fragments which are grouped based on nucleosome position. The sequencing reads mapped to the reference genome were grouped by their dyad (or middle) locations, as -1, +1, +2, +3, +4, +5, and all other genic nucleosomes (such as +6, +7, …). The distribution of the length of reads is displayed for each group, and plotted in the vegetative *Dictyostelium* (**A**) and in the multicellular aggregate (**B**) separately. The frequency of read length was binned at 10-bp interval for the vegetative stage and 5-bp interval for the aggregation stage, and shown for -1 in red, +1 in orange, +2 in yellow, and so on.

(C) Distribution of the sequence composition for nucleosome protected DNA for each position. 73,396 nucleosome dyad locations were grouped as -1 (N=3,230), +1 (N=7,285), +2 (N=8,288), +3 (N=8,792), +4 (N=7,605), +5 (N=7,082), and all other genic nucleosomes (N=31,922), which were combined from the data of the vegetative and aggregation stage. The nucleosomal DNA sequences were fetched as in 301-bp length including linker DNA, and the W (A or T) nucleotide percentage was calculated at every position on the sense (nontemplate) strand (from 5’ left to 3’ right with the same directionality of transcription), and the frequency of W was plotted by the relative distance term from the dyad and smoothed using a three-bp moving average algorithm.



```
Conservation:                         9    9              9       999     9999999  9  99     99  999
NELFB_CORBRA1_hsa_106-580       1  MPSLQPVVMCVMKHL-PKVPEKKLKLVMADKELYRACAVEVKRQIWQDNQALFGDEVSPLLKQYILE---   66
NELFB_CORBRA1_dme_115-590       1  VKSLRPVVMAILRNT-QHIDDKYLKILVRDRELYADTDTEVKRQIWRDNQSLFGDEVSPLLSQYIRE---   66
NELFB_CORBRA1_ddi_101-414       1  --ELQEIPMNVMKRLTPEVPVAFLLKLAEAEELYEQCPIEVKRQIWIVNEELFKEKIKPLINQYIEDPIF   68
Consensus_ss:                         hhhhhhhhhhh      hhhhhhhhhhhhhhhh  hhhhhhhhhh hhhhhhhhhhhhhhhhhh

Conservation:                           9  9          9       9 9    9 9      9     99 9        9    9
NELFB_CORBRA1_hsa_106-580      67  ----KESALFSTELSVLHNFF--SPSPKTRRQGEVVQRLTRMVG-KNVKLYDMVLQFLRTLFLRTRNVHY  129
NELFB_CORBRA1_dme_115-590      67  ----KEHILFDH-TNLNNLFF--HPTPKVRRQGEVVQKLANMIG-TSVKLYDMVLQFLRTLFLRTRNVHY  128
NELFB_CORBRA1_ddi_101-414      69  IQDMNEQLLLSSSTTVDPLFAVSHLPAKRRENNSVLQEMVELFGTKSPDLYQMFINQIKRSFADSGNYLL  138
Consensus_ss:                          hhhhhh       hhh      hhhhh  hhhhhhhhhh    hhhhhhhhhhhhhhhh    eee

Conservation:                      9 9999 99  99    99   9  9   99999999    9  9  9 9  9           999  9
NELFB_CORBRA1_hsa_106-580     130  CTLRAELLMSLHDLDVGEICTVDPCHKFTWCLDACIRERFVDSKRARELQGFL-DGVKKGQEQVLGDLSM  198
NELFB_CORBRA1_dme_115-590     129  CTLRAELLMALHDLEVQEIISIDPCHKFTWCLDACIREKNVDIKRSRELQGFL-DNIKRGQEQVLGDLSM  197
NELFB_CORBRA1_ddi_101-414     139  CNLRAEILMAIHDKSIPEIYDTDASHNIAWCLDACIRDNTLDARRIKEIQTNLSSSVHHQNSSTLGDTAM  208
Consensus_ss:                      hhhhhhhhhhhh   hhhh     hhhhhhhhhhhhhhhhhhhhhhhhhhhhh hhhhhhhhhhhhhhhh

Conservation:                          9 9 9         9         9 9      9   9 9      99    99 9         9
NELFB_CORBRA1_hsa_106-580     199  ILCDPFAINTLALSTVRHLQELVGQETLPRDSPDLLLLLRLLALGQGAWDMIDSQVFKEPKMEVELITRF  268
NELFB_CORBRA1_dme_115-590     198  TLCDPYAINFLATSAIKILHHLINNEGMPRDNQILILLLRMLALGLSAWVMIDSQDFKEPKLDCQVVTKF  267
NELFB_CORBRA1_ddi_101-414     209  VFANPIAVNCIVRNILIQLKEVVKRKQIPKDDESIKFLTYLLVLALKSHEMIKESKFKIPHVKKHILQTF  278
Consensus_ss:                      hhh hhhhhhhhhhhhhhhhhhh         hhhhhhhhhhhhhhhhhhhhhhhhh   hhhhhhhhhh

Conservation:                       9 9     99                           9
NELFB_CORBRA1_hsa_106-580     269  LPMLMSFLVDDYTFNVDQKLPAEEKA----PVSYPNTLPESFTKFLQEQRMACEVGLYYVLHITKQRNKN  334
NELFB_CORBRA1_dme_115-590     268  LPALMSLMVDDQCRSLHAKLPPDERESALTTIEHSGPAPDAVEAYIQESSVASILAMYYTLHTARLKDRV  337
NELFB_CORBRA1_ddi_101-414     279  YPLLATQILDDITREQTSTISLSSSS----TSNTSSTNPA------------------------------  314
Consensus_ss:                      hhhhhhhhhhhhhhhh                      hhhhhhhhhh  hhhhhhhhhhhhh   hhhh

Conservation:
NELFB_CORBRA1_hsa_106-580     335  ALLRLLPGLVETFGDLAFGDIFLHLLTGNLALLADEFALEDFCSSLFDGFFLTASPRKENVHRHALRLLI  404
NELFB_CORBRA1_dme_115-590     338  GVLRVLAILSACKDDRAYEDPFLHSLIALLIPMSEEFATEDFCTTLFDEFIFAGLTRENVTSRHMLKLLW  407
NELFB_CORBRA1_ddi_101-414          ----------------------------------------------------------------------
Consensus_ss:                      hhhhhhhhhhhh        hhhhhhhhhhh         hhhhhhhhhhhhhh   hhhhhhhhhhhhh

Conservation:
NELFB_CORBRA1_hsa_106-580     405  HLHPRVAPSKLEALQKALEPTGQSGEAVKELYSQLGEKLEQLDHRKPSPAQAAETPALELPLPSVPAPAP  474
NELFB_CORBRA1_dme_115-590     408  YVHNKLPAGRLATLMKAMQPTTAHNEHIHKLYEILQERIGTGA--AETPVIEAPPMEFDSPLKSVPTPGP  475
NELFB_CORBRA1_ddi_101-414          ----------------------------------------------------------------------
Consensus_ss:                      hh     hhhhhhhhhh        hhhhhhhhhhhhhh

Conservation:
NELFB_CORBRA1_hsa_106-580     475  L  475
NELFB_CORBRA1_dme_115-590     476  H  476
NELFB_CORBRA1_ddi_101-414          -
Consensus_ss:
```



```
Conservation:                                                         9       9 99        9 9   99 9
NELFD_TH1_hsa_10-589          1  MDEDYYGSAAEWGDEADGGQQEDDSGEGEDDAEVQ-QECLHKFSTRDYIMEPSIFNTLKRYFQAGGSPEN   69
NELFD_TH1_dme_1-578           1  MEVEYDDSG--WQGRAKGQTNPEETLE-DNPQKTI-QECLEKFLTPDYIMEPGIFTQLKRYFQSGGSPEE   66
NELFD_TH1_ddi_80-621          1  ----------------------------KTPTDLIKEKCESSLQQPDSIMSIEASSLMKEYITVGGRPQD   42
Consensus_ss:                       hh   hhh                   hhhhhhhhhhhhh     ee hhhhhhhhhhhh    hhh

Conservation:                    99  999 9   99  999  99   9      9    9999      999  99 999     9 99
NELFD_TH1_hsa_10-589         70  VIQLLSENYTAVAQTVNLLAEWLIQTGVEPVQVQETVENHLKSLLIKHFDPRKADSIFTEEGETPAWLEQ  139
NELFD_TH1_dme_1-578          67  VISMLSENYKAVAQMANLLAEWLILAGVKVTEVQAMVENHLKEMILKSFDPKKADTIFTEEGETPDWLTE  136
NELFD_TH1_ddi_80-621         43  VITFLSESYRGYAQMCNLLCYWLKIAGIPEEQIQTQFSNHLKEIIINKFDPKVADTIFTSSSTPPHWLDE  112
Consensus_ss:                    hhhhhhhhhhhhhhhhhhhhhhhhh    hhhhhhhhhhhhhhhhhh   hhhhhhhh      hhhhhh

Conservation:                    99    99 9 9 9 9    999999    999 9 9 99 9 999     99  99  9
NELFD_TH1_hsa_10-589        140  MIAHTTWRDLFYKLAEAHPDCLMLNFTVKLISDAGYQGEITSVSTACQQLEVFSRVLRTSLATILDGGEE  209
NELFD_TH1_dme_1-578         137  MIDHYTWRSLIYRLAEEYPDCLMLNFTIKLISDAGFQSEITSISTAAQQIEVFSRVLKTSIVKFLN-NPD  205
NELFD_TH1_ddi_80-621        113  MIVDSQWRSLIYQLSEEHKNCLMLNFIIQRISDSGYQNEIASLSTASTYYRVFNKVLFDSLADLVTLDEI  182
Consensus_ss:                    hh    hhhhhhhhhhh    eeehhhhhhhhhhhh  hhhhhhhhhhhhhhhhhhhhhhhhhhh   hh

Conservation:                           9     9   999   9                  9  99    9         9
NELFD_TH1_hsa_10-589        210  NLEKNLPEFAKMVCHGEHTYLFAQAMM-SVLAQEEQGGSAVRRIAQEVQRFAQEKGHDASQITLALGTAA  278
NELFD_TH1_dme_1-578         206  DVHGAIQECARMVCHGQHTYVYSQVLI-QVLSQEQKGGFNMKRLSQEIIKYALQNNQNVTPITMALNGSA  274
NELFD_TH1_ddi_80-621        183  SFRQQLDEFKKSCCQHQHTYLYAQATIQNLIKNNPQKAYNLKRISQELEIEATKKNKIAQRISLMMSNIS  252
Consensus_ss:                    hhhhhhhhhhhhhh    hhhhhhhhhhhhhh      hhhhhhhhhhhhhhhhh     eeeee   hh

Conservation:                     99         9        99  9        999    9 9  9 9    99  9  9   99 9
NELFD_TH1_hsa_10-589        279  SYPRACQALGAMLSKGALNPADITVLFKMFTSMDPPPVELIRVPAFLDLFMQSLFKPGARINQDHKHKYI  348
NELFD_TH1_dme_1-578         275  VYPQACQALTSMLTRNTLNPADITVLFRNYSGSDPPPIDLIRNPQFLELLVDALFRSGVKINPEHKPKYM  344
NELFD_TH1_ddi_80-621        253  QYPLISSSITSLLMSNTTTTGDIIKLYNEYTKAQPPPVDFLRYPNLLELLLSDLFCPGKVIQSVHKSKWI  322
Consensus_ss:                    hhhhhhhhhhhhhh     hhhhhhhhhhh         ee  hhhhhhhhhhh            hhhh

Conservation:                        9                        9    9 9  9           99     99      9
NELFD_TH1_hsa_10-589        349  HILAYAASVVETWKKNKRVS---INKDELKSTSKAVETVHNLCCNENKGASELVAELSTLYQCIRFPVVA  415
NELFD_TH1_dme_1-578         345  FLLAYASAVIDQPAKKRPMTERMLNKEELKSTIQAIEKAHTI-CNVDQGSTELIAELQTLYNCIKYPVVG  413
NELFD_TH1_ddi_80-621        323  YLVTYSVCINSSSSSNEK------DNNVFKDITRAIEIVHNICQ-SNPLGSELQSVVGTLKDHLDFPIIS  385
Consensus_ss:                    hheeehhhhee      eee     hhhhhhhhhhhhhhhhhh       hhhhhhhhhhhhh     ee

Conservation:                     9   9               9   9  9  9      9       9  9   99     9     99
NELFD_TH1_hsa_10-589        416  MGVLKWVDWTVSEPRYFQL-QTDHTPVHLALLDEISTCHQLLHPQVLQLLVKLFETEHSQLDVMEQLELK  484
NELFD_TH1_dme_1-578         414  VGVIRWIENVVMEPSYFKL-STDSCPTHLAVLDEVAAVHPTLQQQILFLLIRLFESKQDELEILVQLEMK  482
NELFD_TH1_ddi_80-621        386  MGIMQWIKHNLTDTTNLSSYNTLCVPIYLDFLRDICLRHPLHRSVVLNILVSHFEVEAQDLESLAILEMR  455
Consensus_ss:                    eeeeeeeeeeee  hhhhh      hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh    hhhhhhhhhhh

Conservation:                    9   9  9 9   9 9 99   9         9 99 999   99  9  99   9            9
NELFD_TH1_hsa_10-589        485  KTLLDRMVHLLSRGYVLPVVSYIRKCLEKLDTDISLIRYFVTEVLDVIAPPYTSDFVQLFLPILENDSIA  554
NELFD_TH1_dme_1-578         483  KMILDRMVNLLTRGCVVPVLRYIKQCCAIEDTDISLIRYFVTEVLETITHPYSPEFVQLFLPMVENEEIT  552
NELFD_TH1_ddi_80-621        456  KNVIDNIVYLFSCGYVIPVLDTIESWAPK--IDPSLTRYFINQVLDIIEPPYSQSFMDKMKSIIK---II  520
Consensus_ss:                    hhhhhhhhhhh   hhhhhhhhhhhhhhh  h hhhhhhhhhhhhhhh     hhhhhhhhhhhh hhhh

Conservation:                              9 9  9   99
NELFD_TH1_hsa_10-589        555  GTIKTEGEHDP-VTEFIAHCKSNFIMV  580
NELFD_TH1_dme_1-578         553  GTMRGEGDNDP-VSEFIVHCKAHYTTV  578
NELFD_TH1_ddi_80-621        521  NPFGESTKLEPSVQSFLDQCKN-----  542
Consensus_ss:                    hhh         hhhhhhhhhh
```

Supplemental Figure 11. Multiple sequence alignment with NELF-B and NELF-D subunit to generate the NELF homology model between human, *Drosophila*, and *Dictyostelium*.

NELF-B and NELF-D subunits in the NELF complex were reported as evolutionarily conserved between human, fruit fly, and mouse (Narita et al. 2003). Structure-based sequence alignments of homologous COBRA1 domains (Cofactor of BRCA1, Pfam accession number: PF06209) in NELF-B **(A)** and TH1 domain (Trihydrophobin 1, Pfam accession: PF04858) of NELF-D **(B)** are shown. The amino acid sequence of NELF-B was obtained as NCBI Accession number NP_056271 (human), NP_572402 (*D. melanogaster*), and XP_638637 (*D. discoideum*), and their multiple alignment was produced by using PROMALS (Pei and Grishin 2007; Pei et al. 2007) with default parameters. The first line in each block is conservation indexes associated with each position (an integer between 0 and 9, with 9 corresponding to highest conservation (Pei and Grishin 2001)). A predicted secondary structure was reported as a colored sequence (red: alpha-helix, blue: beta-strand), and together the consensus secondary structures is shown in the last line (h: alpha-helix, e: beta-strand). Also, the amino acid sequence of the TH1 domain was fetched as NP_945327 (human), NP_573123 (*D. melanogaster*), and XP_646857 (*D. discoideum*) and their multiple alignment is shown in **B**. Furthermore, we have tried to detect homology of the B and D subunit with the proteomes of *A. thaliana*, *C. elegans*, *S. cerevisiae*, and *S. pombe* by using PSI-BLAST, but no sequence-level homology was detected from those eukaryotes.

Supplemental Tables

Supplemental Table 1. Dinucleotide compositions in five eukaryotic genomes including A/T-rich *D. discoideum* and *P. falciparum* genome.

The dinucleotide compositions were calculated based on all omosome DNA sequences of 5 eukaryotes. If the reverse complementary sequence is considered for each of all possible 16 dinucleotides, several dinucleotides are equivalently assigned, for example AA=TT, GG=CC, AG = CT, etc., which results in unique 10 dinucleotide sets (or steps). The percent composition (%) of the 10 dinucleotide sets was calculated per each species in the table. Compared with budding yeast, human, and fruit fly, the most common dinucleotide in the A/T-rich *Dictyostelium* and *Plasmodium* genome was WW. (WW: AA, TT, AT, TA, SS: GG, CC, GC, CG, SW: GA, GT, CA, CT, WS: AG, AC, TG, TC).

	Dinucleotide content (%)*				
	D.discoideum	***P.falciparum***	***S.cerevisiae***	**Human**	***D.melanogaster***
AA + TT	36.4	32.3	21.6	19.6	20.2
GG + CC	3.8	2.9	7.8	10.4	9.4
AT	13.4	17.5	8.9	7.7	8.1
GC	1.1	0.9	3.7	4.3	5.7
TA	10.9	15.9	7.3	6.6	6.3
CG	0.7	0.7	2.9	1.0	4.2
AG + CT	6.6	6.4	11.7	14.0	10.8
TG + CA	10.7	8.7	13.0	14.5	13.8
AC + GT	7.7	7.0	10.5	10.1	10.4
TC + GA	8.7	7.7	12.5	11.9	11.1
WW	60.8	65.8	37.9	33.9	34.5
SS	5.6	4.5	14.5	15.7	19.3
SW + WS	33.7	29.8	47.7	50.5	46.1

* The ambiguity code "N" in DNA sequences was excluded in the content.

Supplemental Table 2. (A) Consensus location of nucleosome positions downstream of the TSS of seven eukaryotes. We curated *in vivo* mononucleosome mapping data of human, *D. melanogaster*, *S. pombe*, *S. cerevisiae*, *A. thaliana*, and *C. elegans* from the literatures and produced the composite distributions of nucleosome locations at the 5' end of genes. According to the systematic naming of each predominant nucleosome position(Jiang and Pugh 2009), the consensus location of each nucleosome position (such +1, +2, and so on) was obtained as the location of the highest nucleosome occupancy based on the composite distribution of nucleosome locations relative to the TSS. Some nucleosome positions were not clearly identified from the composite distributions due to lack of high resolution maps of TSS.

(B) Averages of inter-nucleosomal spacing intervals of 7 species. The spacing distances between genic nucleosomes (i.e. +1, +2, +3, +4, and +5) were obtained as the peak-to-peak distance (bp) from the composite distribution of nucleosome positions each species, for example the distance between +1 and +2, and between +2 and +3. These spacing intervals were averaged in the table. These averages are consistent with the literatures. For example, (Lantermann et al. 2010) reported nucleosome repeat lengths as 154 bp. However, we were not able to calculate robust estimates for *A. thaliana* and *C. elegans* due to lack of high resolution map of TTS annotation. In case of *A. thaliana*, (Chodavarapu et al. 2010) computed approximately 175-bp spacing between inter-nucleosomes, and (Valouev et al. 2008) reported that *C. elegans* nucleosomes are uniformly distributed at the 175-bp interval.

A. Consensus location of the nucleosome dyads relative to the TSS of genes

Species	+1	+2	+3	+4	+5
D.discoideum (vegetative)	116	265	435	588	762
D.discoideum (aggregation)	109	264	441	601	786
Human	120	308	488	663	841
D.melanogaster	130	306	473	649	832
S.pombe	60	210	360	510	660
S.cerevisiae	56	225	391	558	724
A.thaliana	126	256	413	575	736
C.elegans	61	--*	--	--	--

* Not determined in our composite distribution plots.

B. Averaged spacing interval (bp) of inter-nucleosomes

Species	Interval (bp)
D.discoideum (vegetative)	162
D.discoideum (aggregation)	169
Human	180
D.melanogaster	176
S.pombe	150
S.cerevisiae	167
A.thaliana	175*
C.elegans	175*

* The spacing distances estimated in the literatures.

Supplemental Table 3. Annotation of *D. discoideum* transcription start sites (TSS) and transcription end sites (TES). The reference genome sequence of *Dictyostelium discoideum* AX4 was obtained from dictyBase (http://dictybase.org), which was distributed in February 2008. This annotation data was generated by the transcriptome sequencing and analysis (described the Method) based on this genome build.

* Separate EXCEL file included, “Supplemental Table 3.xls”

Supplemental Table 4. Comparison of literature TSS with RNA-seq TSS

							5'	3'	5'		
			CDS	CDS			UTR	UTR	UTR		
ID	Chr	Str	start	end	TSS	TES	(bp)	(bp)	lit.[1]	D[2]	Reference and comments
DDB0191129	1	-	1094419	1093913	1094526	1093768	107	145	92	-15	Mol Cell Biol (1988) 8,8; 3458-3466
DDB0191242	2	-	2308523	2307782	2308627	2307766	104	16	46	-58	Differentiation (2001) 68:92–105
DDB0215400	2	-	2461805	2461044	2461876	2460985	71	59	62	-9	Nucleic Acids Res. (1982) 10, 4; 1231-1241; Sequence of described UTR not entirely the same sequence as from genome project
DDB0214998	3	-	2570624	2569081	2570652	2568934	28	147	93	65	EMBO Journal (1987) 6 ,1; 195 -200 ; ends in polyA stretch
DDB0215010	3	-	4543611	4540912	4543905	4540889	294	23	134	-160	J. Gen Micr (1993), 139, 3043-3052; No experiment described in this paper
DDB0191431	4	+	411746	413523	411593	413537	153	14	130	-23	Mechanisms of Development (1994) 45; 59-72
DDB0191479	4	-	1959735	1956975	1960901	1956917	1166	58	650	-516	Mol. Cell. Biol. 1990, 10(5):1921; This gene has an extraordinarily long and complex promoter region with at least 3 initiation sites. Adjacent gene is far away.
DDB0191156	5	-	2344339	2342795	2344481	2342703	142	92	142	0	J. Biol Chem (1992) 267, 27; 19655-19664
DDB0216195	6	-	2531851	2530991	2532212	2530958	361	33	361	0	Nucleic Acids Res. (1992) 20, 6; 1325-1332
DDB0219974	4	-	3931286	3929775	NA	3929722	ND	53	437	ND	Devel Biol (2003) 255; 373–382

[1] 5’ UTR length (bp) from literature.

[2] Bases upstream (-) or downstream (+) of TSS in literature

Supplemental References

Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. 2007. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**: 572-576.

Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS et al. 2009. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* **37**: D539-543.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* **2**: 28-36.

Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**: 699-709.

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ et al. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**: 388-392.

Dacks JB, Doolittle WF. 2001. Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell* **107**: 419-425.

Dechering KJ, Cuelenaere K, Konings RN, Leunissen JA. 1998. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res* **26**: 4056-4062.

Hoeijmakers WA, Bartfai R, Francoijs KJ, Stunnenberg HG. 2011. Linear amplification for deep sequencing. *Nature protocols* **6**: 1026-1036.

Jiang C, Pugh BF. 2009. A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. *Genome Biol* **10**: R109.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Kimmel AR, Firtel RA. 1983. Sequence organization in Dictyostelium: unique structure at the 5'-ends of protein coding genes. *Nucleic Acids Res* **11**: 541-552.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**: 291-295.

Kreppel L, Fey P, Gaudet P, Just E, Kibbe WA, Chisholm RL, Kimmel AR. 2004. dictyBase: a new Dictyostelium discoideum genome database. *Nucleic Acids Res* **32**: D332-333.

Lantermann AB, Straub T, Stralfors A, Yuan GC, Ekwall K, Korber P. 2010. Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. *Nat Struct Mol Biol* **17**: 251-257.

Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. 2008a. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**: 1073-1083.

Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC et al. 2008b. Nucleosome organization in the Drosophila genome. *Nature* **453**: 358-362.

Narita T, Yamaguchi Y, Yano K, Sugimoto S, Chanarat S, Wada T, Kim DK, Hasegawa J, Omori M, Inukai N et al. 2003. Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex. *Mol Cell Biol* **23**: 1863-1873.

Pei J, Grishin NV. 2001. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**: 700-712.

-. 2007. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* **23**: 802-808.

Pei J, Kim BH, Tang M, Grishin NV. 2007. PROMALS web server for accurate multiple protein sequence alignments. *Nucleic Acids Res* **35**: W649-652.

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. 2009. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* **5**: e1000386.

Satchwell SC, Drew HR, Travers AA. 1986. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191**: 659-675.

Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**: 887-898.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**: 1051-1063.

Watts DJ, Ashworth JM. 1970. Growth of myxameobae of the cellular slime mould Dictyostelium discoideum in axenic culture. *The Biochemical journal* **119**: 171-174.

Zhang M, Gish W. 2006. Improved spliced alignment from an information theoretic approach. *Bioinformatics* **22**: 13-20.