Supporting Material

Genome-wide organization of eukaryotic pre-initiation complex is influenced by nonconsensus protein-DNA binding

Ariel Afek and David B. Lukatsky*
Department of Chemistry, Ben-Gurion University of the Negev, Beer-Sheva 84015, Israel

**Corresponding author:*
Email: lukatsky@bgu.ac.il

Supporting Figure Legends

Figure S1. This figure demonstrates the robustness of the computed free energy of nonconsensus TF-DNA binding with respect to the global variability of the nucleotide content along the yeast genome (**A**); the robustness with respect to the width of the sliding window, L (**B**); and the robustness with respect to the number of contacts, M, that the TF makes with DNA (**C**). (**A**) The average free energy of nonconsensus TF-DNA binding per bp, $\langle f \rangle = \langle \langle F \rangle_{\mathrm{TF}} \rangle_{seq} / M$, (red), as compared with the corresponding normalized free energy per bp, $\langle \delta f \rangle = \langle \langle \delta F \rangle_{\mathrm{TF}} \rangle_{seq} / M$, (blue), where $\delta F = F - F_{rand}$. For a given TF, F is computed as described in the main text, and F_{rand} is the free energy computed for a randomized sequence (in the same sliding window as F), and averaged over 25 random realizations. The described procedure removes the bias in the free energy, stemming from the global variability of the nucleotide content. (**B**) The normalized, average free energy, $\langle \delta f \rangle$, computed using different values of the width of the sliding window, $L = 30$ (red), $L = 50$ (black), and $L = 80$ (blue). We used $L = 50$ for all the calculations described in the main text. (**C**) The normalized, average free energy, $\langle \delta f \rangle$, computed using different values of the TF size, $M = 6$ (red), $M = 8$ (black), and $M = 10$ (blue). We used $M = 8$ for all the calculations described in the main text. In all plots, (**A**), (**B**), and (**C**), the average free energies are computed using 6,045 yeast transcripts.

Figure S2. This figure further demonstrates the statistical significance of the correlation between the free energy of nonconsensus TF-DNA binding and the PIC occupancy in the vicinity of the TSS. (**A**) Correlation between the minimal value of the free energy of nonconsensus TF-DNA binding, $f_{\min} = \min(f)$, where $f = \langle F \rangle_{\mathrm{TF}} / M$, and the combined occupancy of all GTFs, computed for individual genes in non-overlapping windows of 80 bp within the entire interval $(-400, 400)$ around the TSS for each of these 3,945 genes. The data are binned into 50 bins. The notation $\langle \mathrm{PIC} \rangle$ describes the average, combined occupancy profile of all nine GTFs. (**B**) Correlation between the minimal value of the free energy of nonconsensus TF-DNA binding, $f_{\min} = \min(f)$, and the maximal combined occupancy of eight GTFs (all GTFs less the Pol II occupancy), computed for individual genes within the entire interval $(-150, 0)$ around the TSS for each of these 3,945 genes. The data are binned into 25 bins.

Figure S3. (**A**) Correlation between the minimal value of the free energy of nonconsensus TF-DNA binding, $f_{\min} = \min(f)$, where $f = \langle F \rangle_{\mathrm{TF}} / M$, and the average, combined GTF occupancy computed for individual genes in non-overlapping windows of 80 bp within the entire interval $(-2990, 2070)$ around the open reading frame (ORF) ends for 2,903 mRNA genes. The data are binned into 50 bins. (**B**) Similarly computed correlation of $f_{\min}$ with the nucleosome occupancy. (**C**) The heat maps represent the free energy of nonconsensus TF-DNA binding, f, the combined occupancy of the GTFs, and the nucleosome occupancy, respectively, for 1,860 tandem mRNA genes aligned with respect to the ORF ends. The genes are sorted by intergenic length.

Figure S4. Correlation between the minimal value of the free energy of nonconsensus TF-DNA binding, $f_{\min} = \min(f)$, where $f = \langle F \rangle_{\mathrm{TF}} / M$, and the average GTF occupancy

computed for individual genes, in non-overlapping windows of 80 bp within the entire interval $(-2990,2070)$ around the ORF ends for 2,903 mRNA genes. The data are binned into 50 bins.

Figure S5. (**A**) Correlation between the minimal value of the free energy of nonconsensus TF-DNA binding, $f_{\min} = \min(f)$, where $f = \langle F \rangle_{\mathrm{TF}} / M$, and the average GTF occupancy of 676 TATA-containing genes. The correlation is computed for individual genes in non-overlapping windows of 80 bp within the entire interval $(-990,990)$ around the TSS. The data are then binned into 50 bins. (**B**) Similar to (**A**), but now $f_{\min}$ is correlated with the nucleosome occupancy of these TATA-containing genes. (**C**) Correlation between $f_{\min}$, and the average GTF occupancy of 3,269 TATA-less genes. The correlation is computed for individual genes in non-overlapping windows of 80 bp within the entire interval $(-990,990)$ around the TSS. The data are binned into 50 bins. (**D**) Similar to (**C**), but now $f_{\min}$ is correlated with the nucleosome occupancy of these TATA-less genes. (**E**) The average free energy of nonconsensus TF-DNA binding per bp, $\langle f \rangle$, for a larger set of 5,034 TATA-less genes (blue), and 1,011 TATA-containing genes (red), around the TSSs.

Figure S6. This figure is supplementary to **Figure 7** of the main text. It demonstrates the robustness of our conclusions with respect to an alternative definition of the TATA-like box occupancy score. (**A**) The TATA-like box occupancy scores were computed based on the PWM taken from V.X. Jin et al., BMC Bioinformatics **7**:114 (2006). We used 1,432 genes with double-peak in the nucleosome occupancy (red) and 2,513 genes with single-peak in the nucleosome occupancy (blue), as described in **Figure 7** of the main text. The obtained TATA-like box occupancy profiles are similar to those presented in **Figure 7** of the main text, and these profiles are statistically indistinguishable between those two groups of genes. (**B**) We also computed the initiator (INR) element PWM occupancy score for the selected two groups of genes, as described in (**A**). The obtained occupancy profiles are statistically indistinguishable between these two groups of genes.

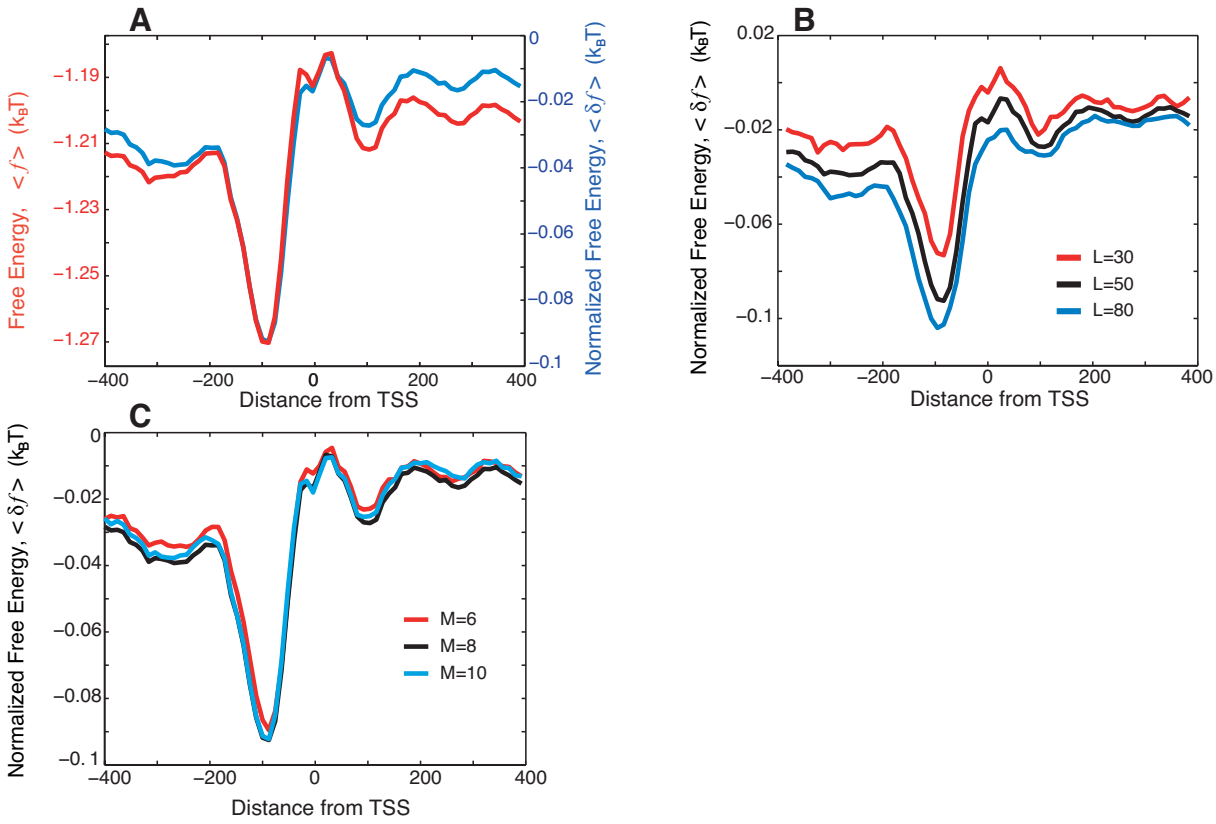

Figure S1

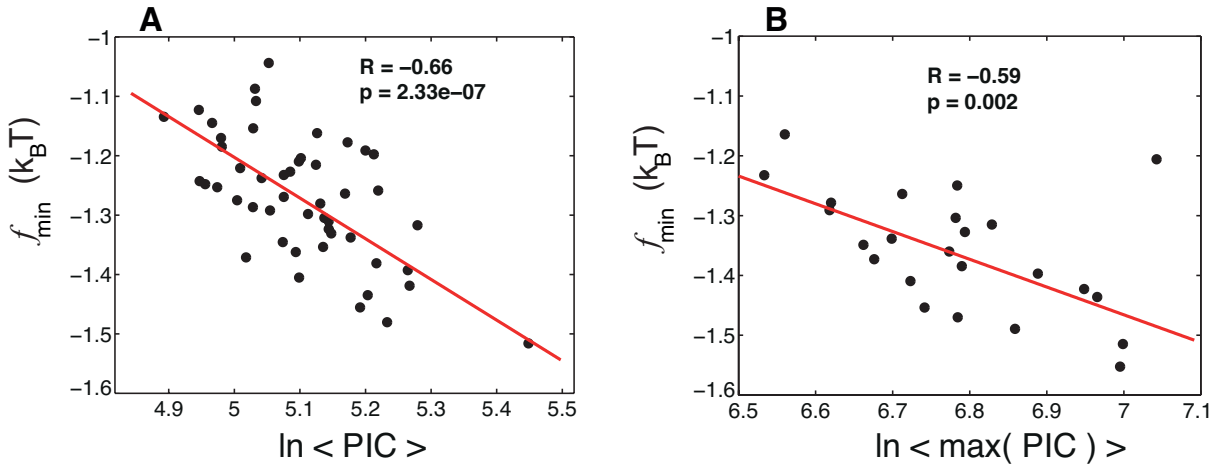

Figure S2

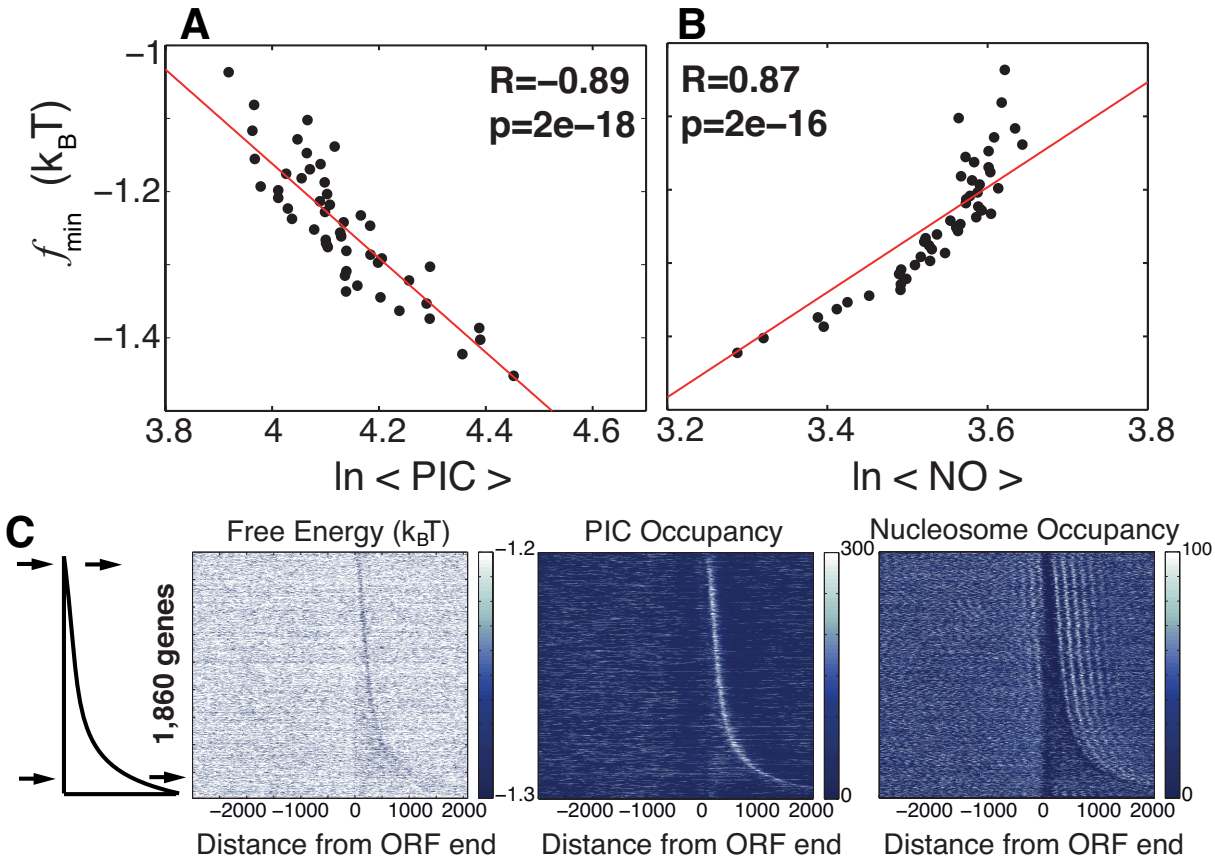
A
B
C
R=−0.89
p=2e−18
R=0.87
p=2e−16
f_{min} (k_BT)
−1
−1.2
−1.4
3.8
4
4.2
4.4
4.6
3.2
3.4
3.6
3.8
ln < PIC >
ln < NO >
Free Energy (k_BT)
PIC Occupancy
Nucleosome Occupancy
1,860 genes
−1.2
−1.3
300
0
100
0
−2000 −1000 0 1000 2000
Distance from ORF end

Figure S3

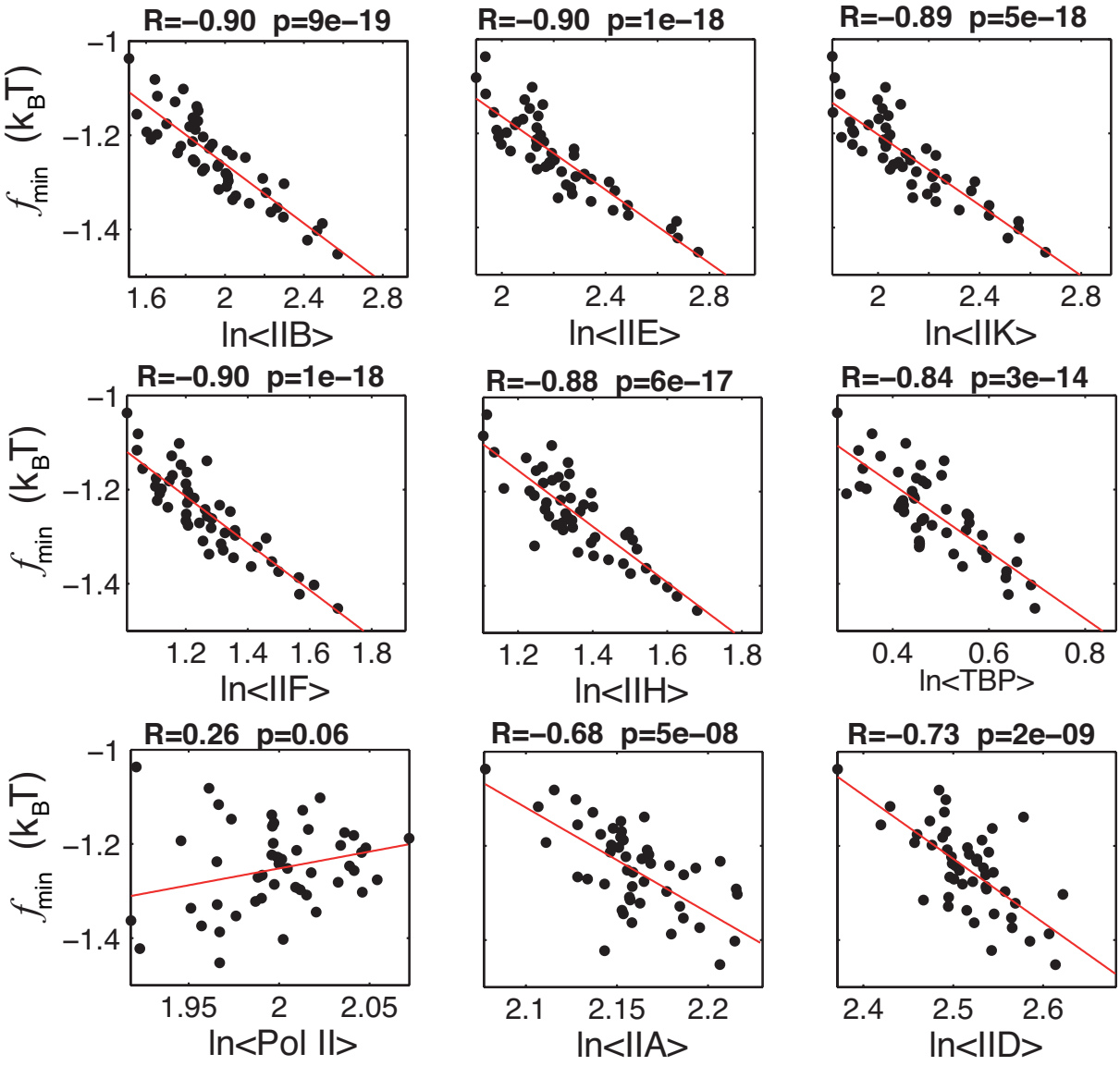
R=−0.90 p=9e−19
R=−0.90 p=1e−18
R=−0.89 p=5e−18
R=−0.90 p=1e−18
R=−0.88 p=6e−17
R=−0.84 p=3e−14
R=0.26 p=0.06
R=−0.68 p=5e−08
R=−0.73 p=2e−09
f_{min} (k_BT)
ln<IIB>
ln<IIE>
ln<IIK>
ln<IIF>
ln<IIH>
ln<TBP>
ln<Pol II>
ln<IIA>
ln<IID>

Figure S4

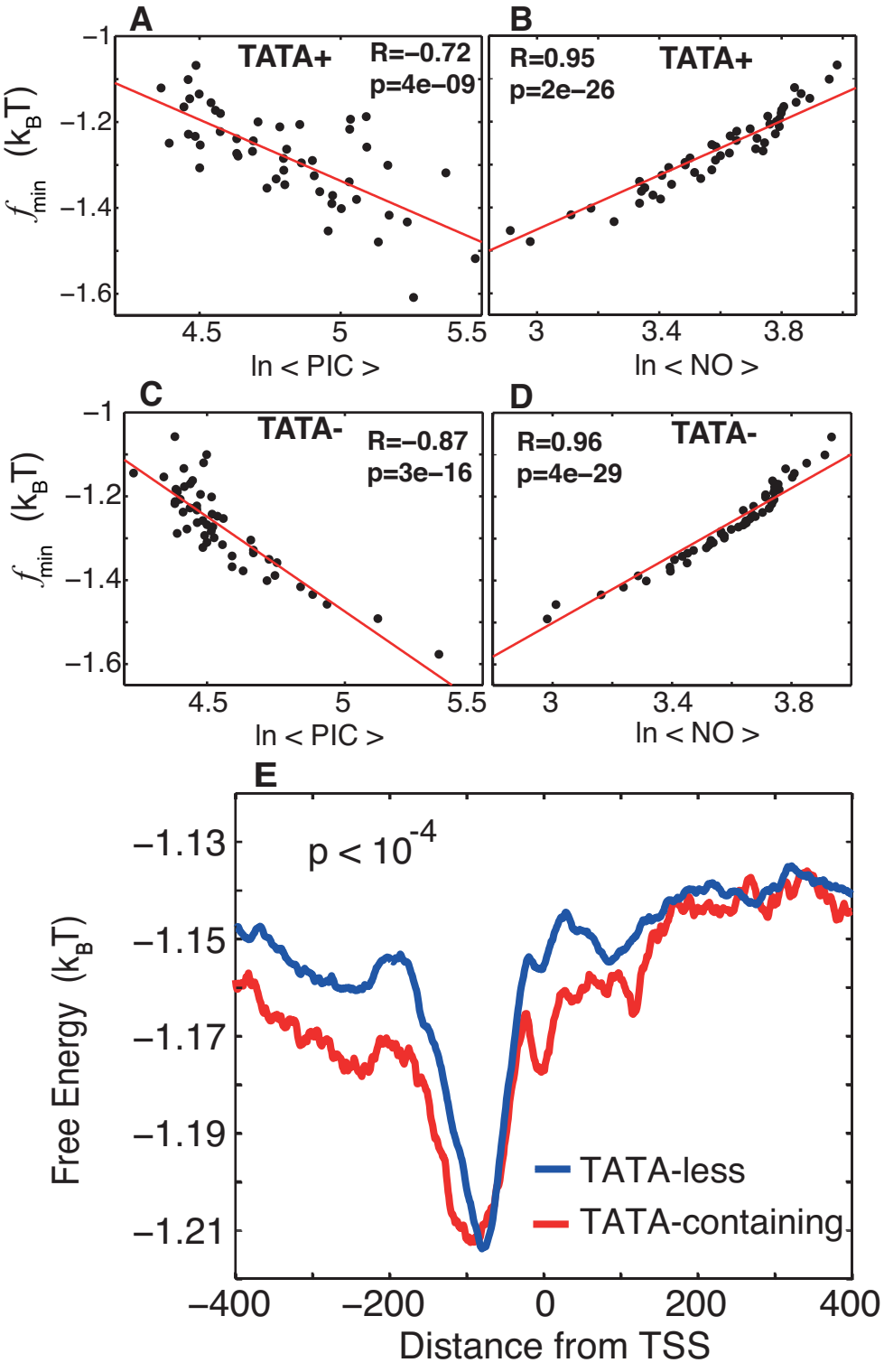
A
TATA+
R=−0.72
p=4e−09
B
R=0.95
p=2e−26
TATA+
C
TATA-
R=−0.87
p=3e−16
D
R=0.96
p=4e−29
TATA-
ln < PIC >
ln < NO >
E
p < 10^-4
Free Energy (k_BT)
TATA-less
TATA-containing
Distance from TSS

Figure S5

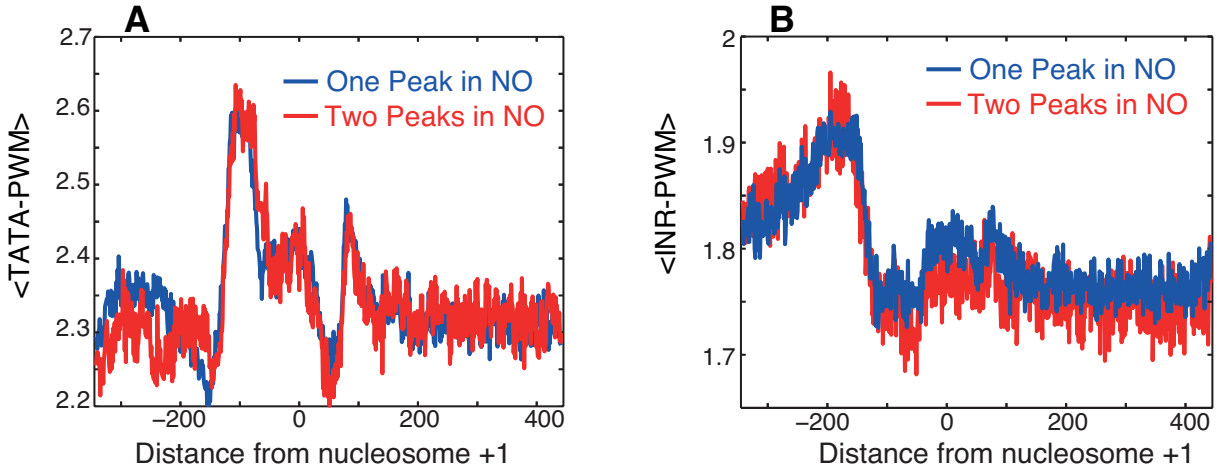
A
B
One Peak in NO
Two Peaks in NO
<TATA-PWM>
<INR-PWM>
Distance from nucleosome +1
Distance from nucleosome +1

Figure S6