Supplementary Information

Mavrich et al.

Supplementary Tables

Supplementary **Table S1**. Nucleosome list. Each row corresponds to a coarse-grain nucleosome call and its associated properties. EXCEL spreadsheet.

Supplementary **Table S2**. List of chromosomal features used in this study. EXCEL spreadsheet.

Supplementary **Table S3**. NPS patterns. A list of AA and TT frequencies at each nucleosomal position. EXCEL spreadsheet.

Supplementary **Table S4** List of dinucleotides and associated calculations. EXCEL spreadsheet.

Supplementary **Table S5** List of low and high NPS scoring genes, and associated cis-regulatory element count. EXCEL spreadsheet.

Supplementary **Table S6**. List of potential TFIIB-looped genes. EXCEL spreadsheet.

Supplemental Data

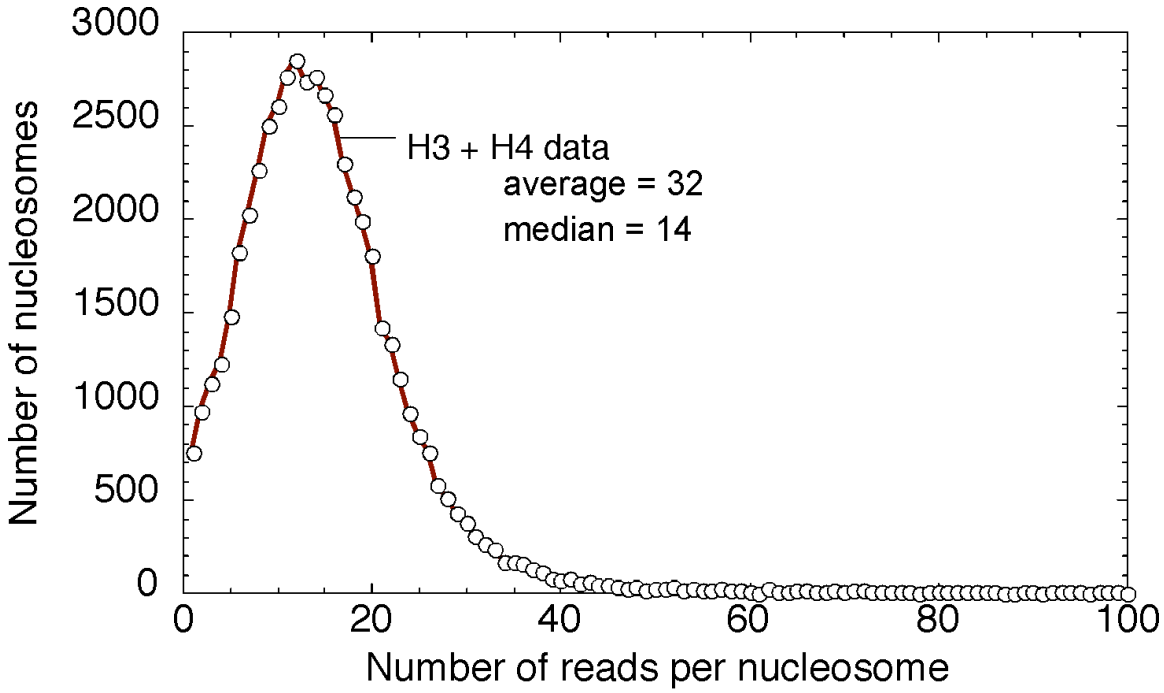


Figure S1 Distribution of sequencing reads per nucleosome. 53,026 consensus locations having 3 or more reads were identified using the dataset derived from the immunopurified H3 and H4 nucleosomes (54,753 with 1 or more reads). Extrapolation of the left side of the curve, based upon the right suggests that coverage is >93%.

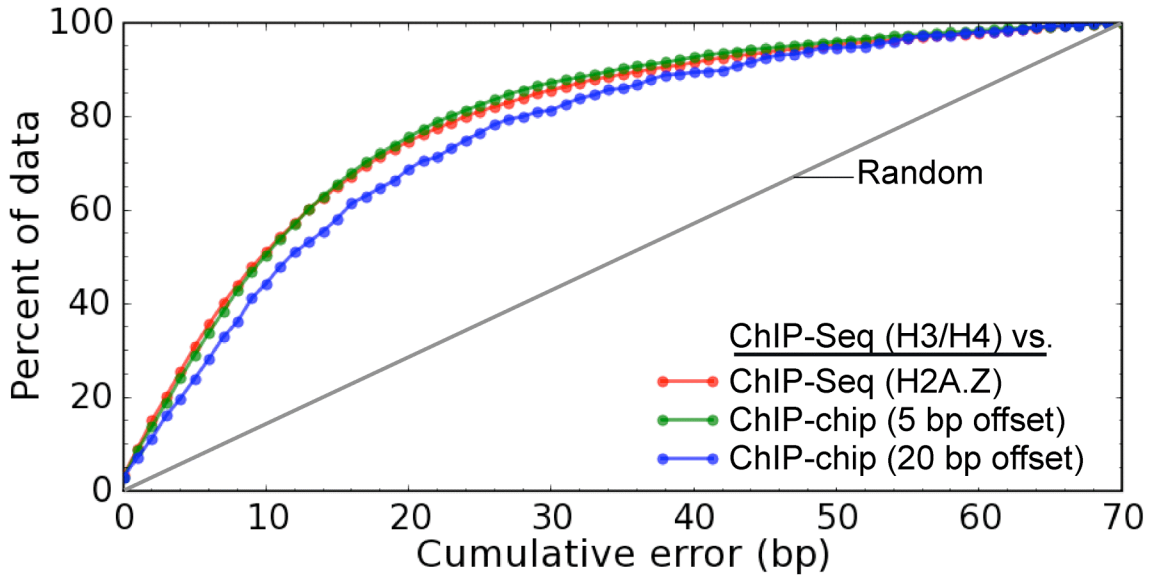


Figure S2. Accuracy of nucleosome calls compared to other genome-wide experimental data sets (Albert et al., 2007; Lee et al., 2007; Raisner et al., 2005). Shown is a comparison of nucleosome distances from the H3/H4 data set for those nucleosomes located between +1 and +500 relative to the TSS. Nucleosomes outside of this region were sufficiently delocalized to preclude informative comparison. The basis for Chip-seq and high resolution Chip-chip having equivalent accuracy might have more to do with the fact that even the most highly positioned nucleosomes are delocalized over 10-20 bp. Thus, biological noise rather than technical limitations limits the accuracy. 5 bp and 20 bp offset refer to the distance between the midpoints of adjacent microarray probes. The gray diagonal line represents the distribution expected for randomly placed calls. All datasets involve MNase digestion of chromatin crosslinked in vivo. H2A.Z dataset utilized Roche/454 GS20 DNA sequencing; 5 bp offset probes represent the Affymetrix high density tiling microarrays; and the 20 bp offset represent a custom designed microarray platform.

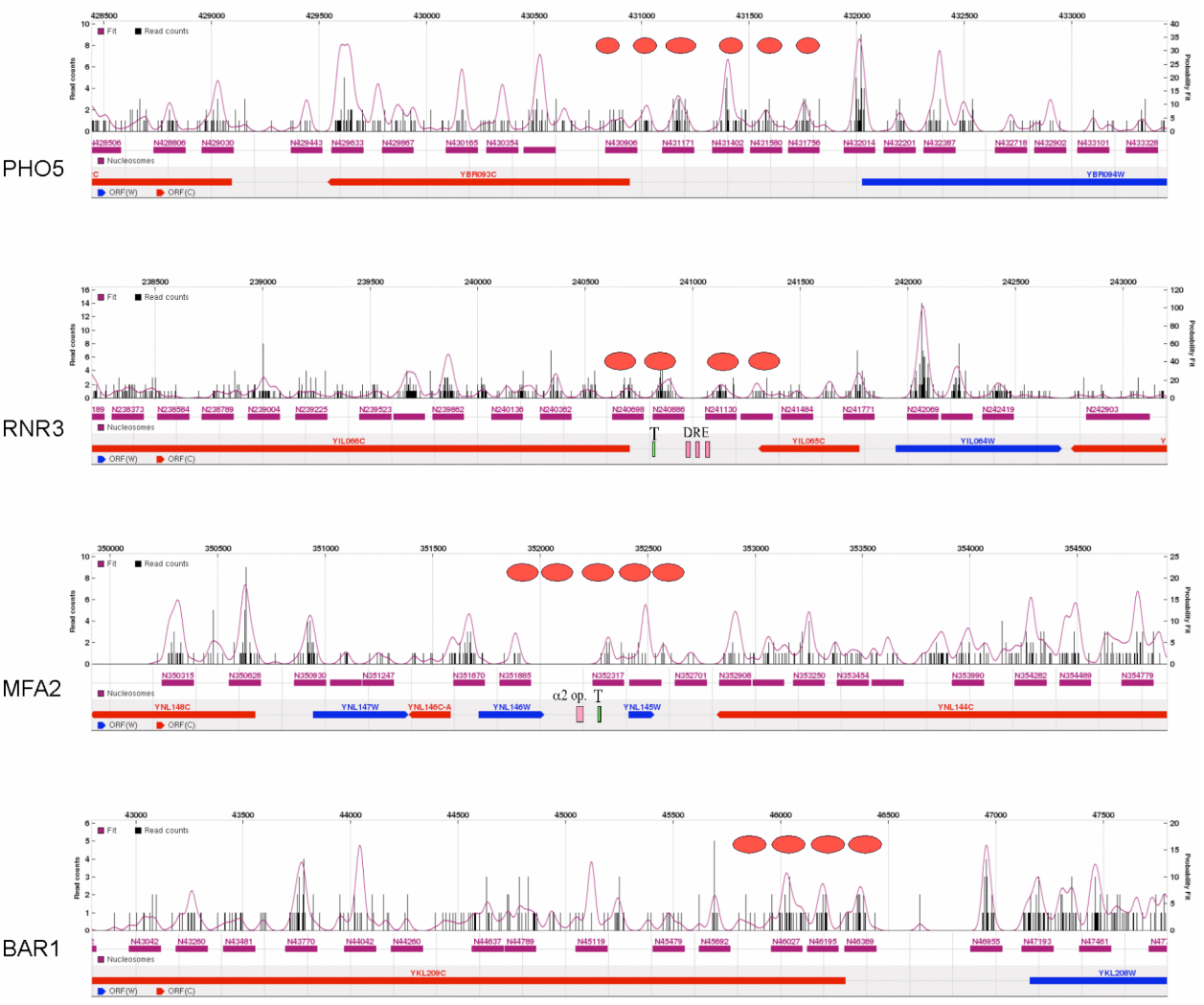


Figure S3. Browser shots of *PHO5, RNR3, MFA2*, and *BAR1*. Red ovals indicate locations of previously determined nucleosomes (Almer and Horz, 1986; Li and Reese, 2001; Shimizu et al., 1991; Teng et al., 2001). Note that *MFA2* and *BAR1* have distinct chromatin structures in MATa (BY4741) vs MATα strains (Shimizu et al., 1991). The former was used in the current study, and the latter to map genes-specific structure.

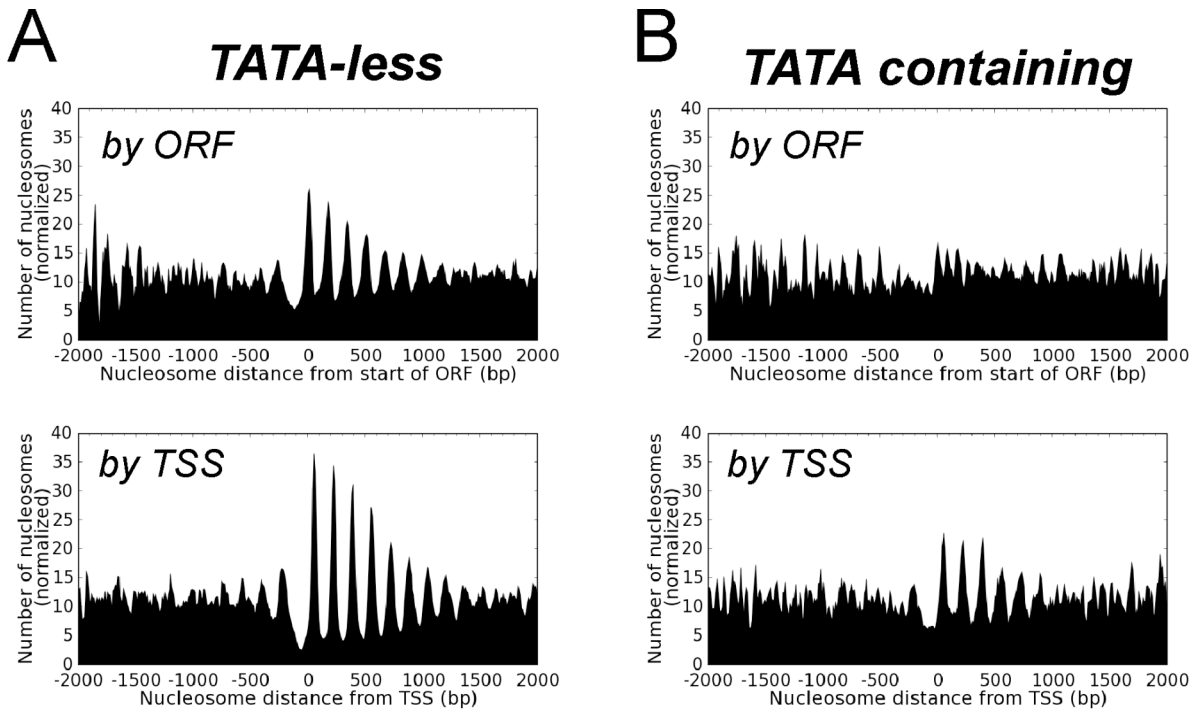


Figure S4A,B. Nucleosome distribution at the 5' end of yeast genes, aligned by either the TSS or the ORF start site. Panel **A** shows TATA-less genes and panel **B** shows TATA-containing. Note a stronger alignment to the TSS (bottom panels).

The TATA box is a core promoter element at ~20% of all genes (Basehoar et al., 2004). TATA-containing genes tended towards individualized nucleosome organization compared to TATA-less genes, reflecting a greater gene-specific role for chromatin regulation at TATA-containing promoters (Basehoar et al., 2004; Ioshikhes et al., 2006).

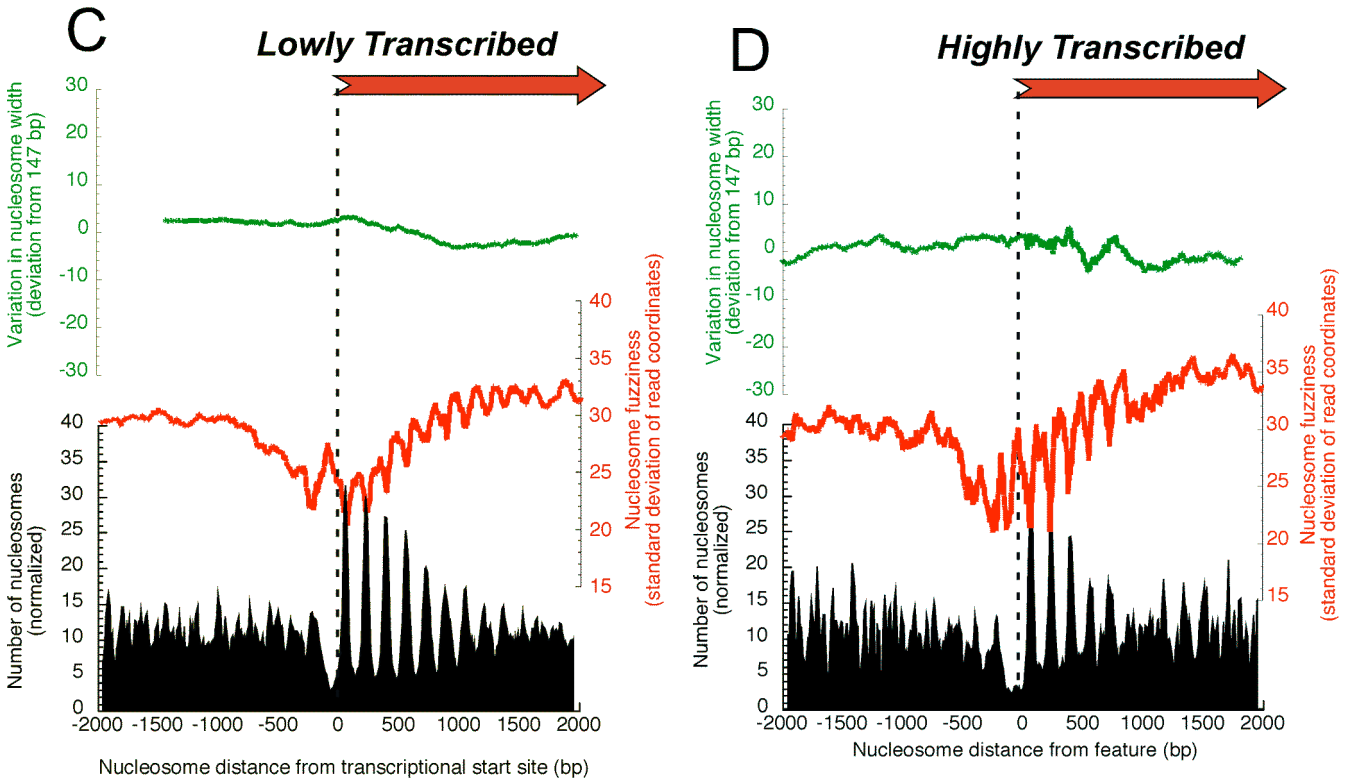


Figure S4C, D. Lowly and highly transcribed genes have similar nucleosome organization. Lowly and highly transcribed genes are those in the lowest 50th percentile (~2000 genes) and highest 15th percentile (~600 genes) of transcription frequency, respectively (Holstege et al., 1998). Filled black plots denote the distribution of nucleosome distances from the TSS (David et al., 2006), normalized to the number of regions analyzed (see Figure 2C legend of main text for more details). Red and green traces are moving averages of nucleosome fuzziness, and deviations in width from the canonical 147 bp, respectively, for the indicated class of genes. Highly transcribed genes are expected to be depleted of nucleosomes. The modest reduction in nucleosome content in panel D might be attributed to using a modestly stringent cutoff (15th percentile). A more stringent cutoff is shown in Figure S4E, where highly expressed ribosomal protein genes are displayed.

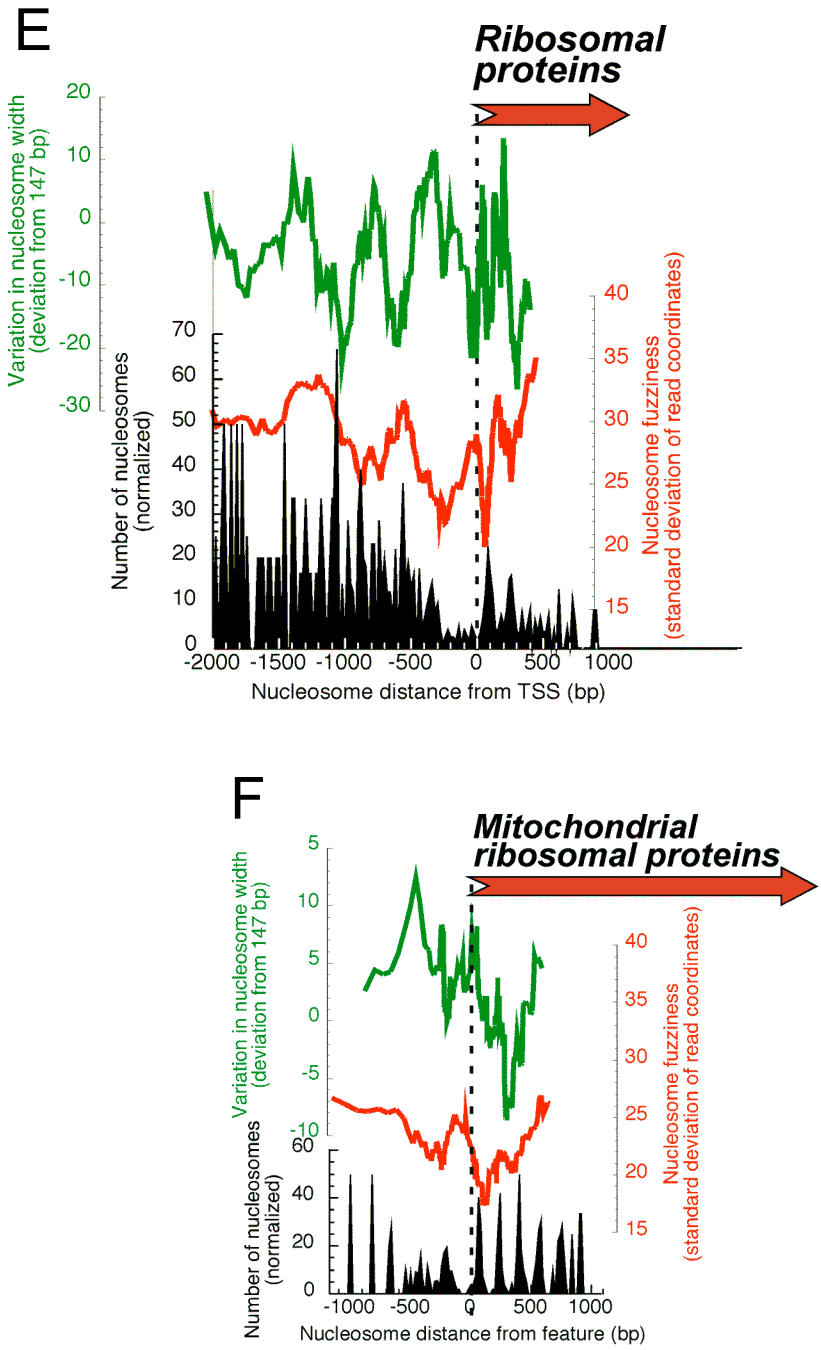


Figure S4E,F (*continued*). (E) Cytosolic ribosomal protein genes lack a -1 nucleosome and have an overall lower nucleosome density. (F) Mitochondrial ribosomal protein genes have a strong consensus nucleosome organization. See Figure S4C,D for figure description.

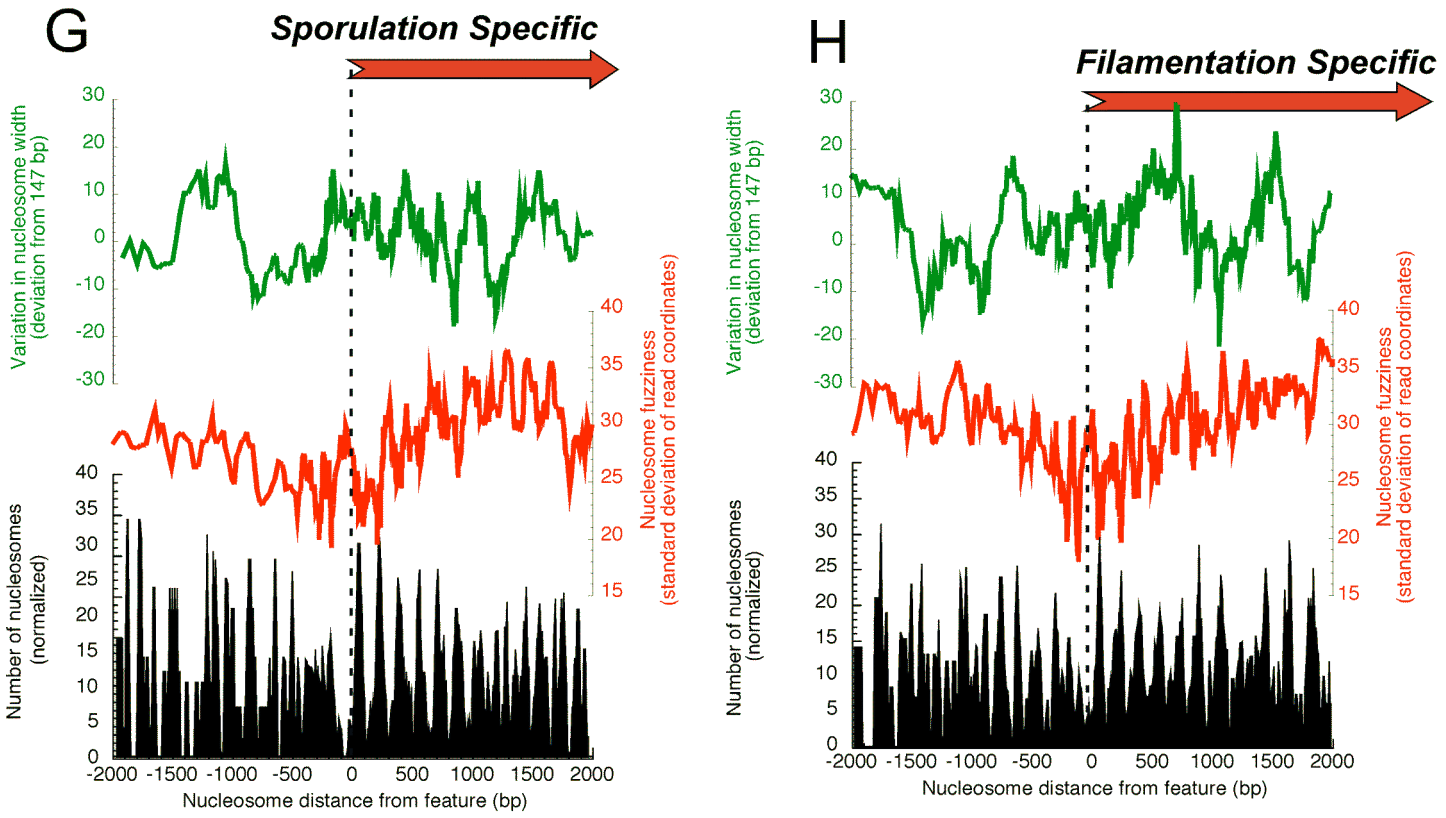


Figure S4G,H (*continued*) Developmentally programmed genes tend to have a canonical nucleosome organization. See Figure S4C,D for figure description. Mid-sporulation and filamentation-induced active genes are defined in (Gasch et al., 2000).

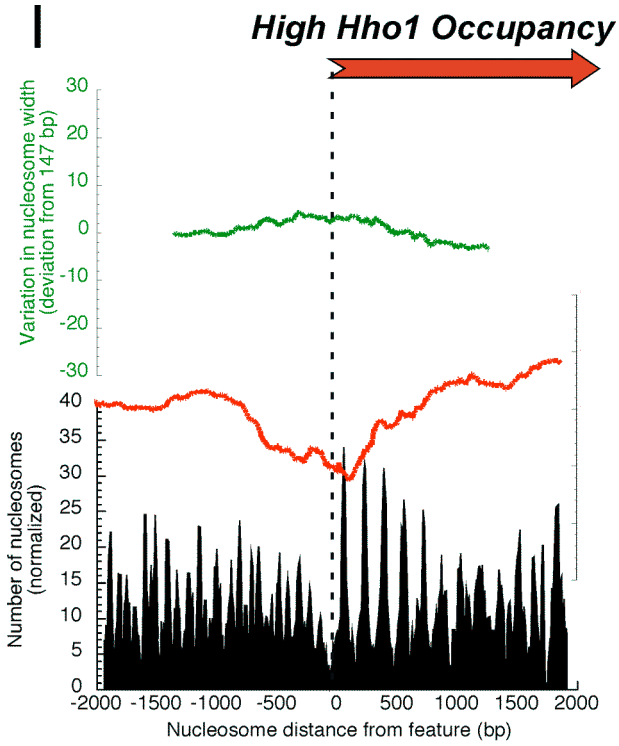


Figure S4I (*continued*) Histone H1-targeted genes have a canonical nucleosome organization, but lack a uniformly positioned -1 nucleosome. Here, feature denotes TSS of H1-targeted genes. Histone H1-targeted genes represent those in the highest 10 percentile of occupancy in ChIP-chip analysis (Zanton and Pugh, 2006). See Figure S4C,D for figure description. Note the lack of a well defined peak around -200, which suggests that histone H1(Hho1) might affect the organization of nucleosomes in promoter regions rather than in genes.

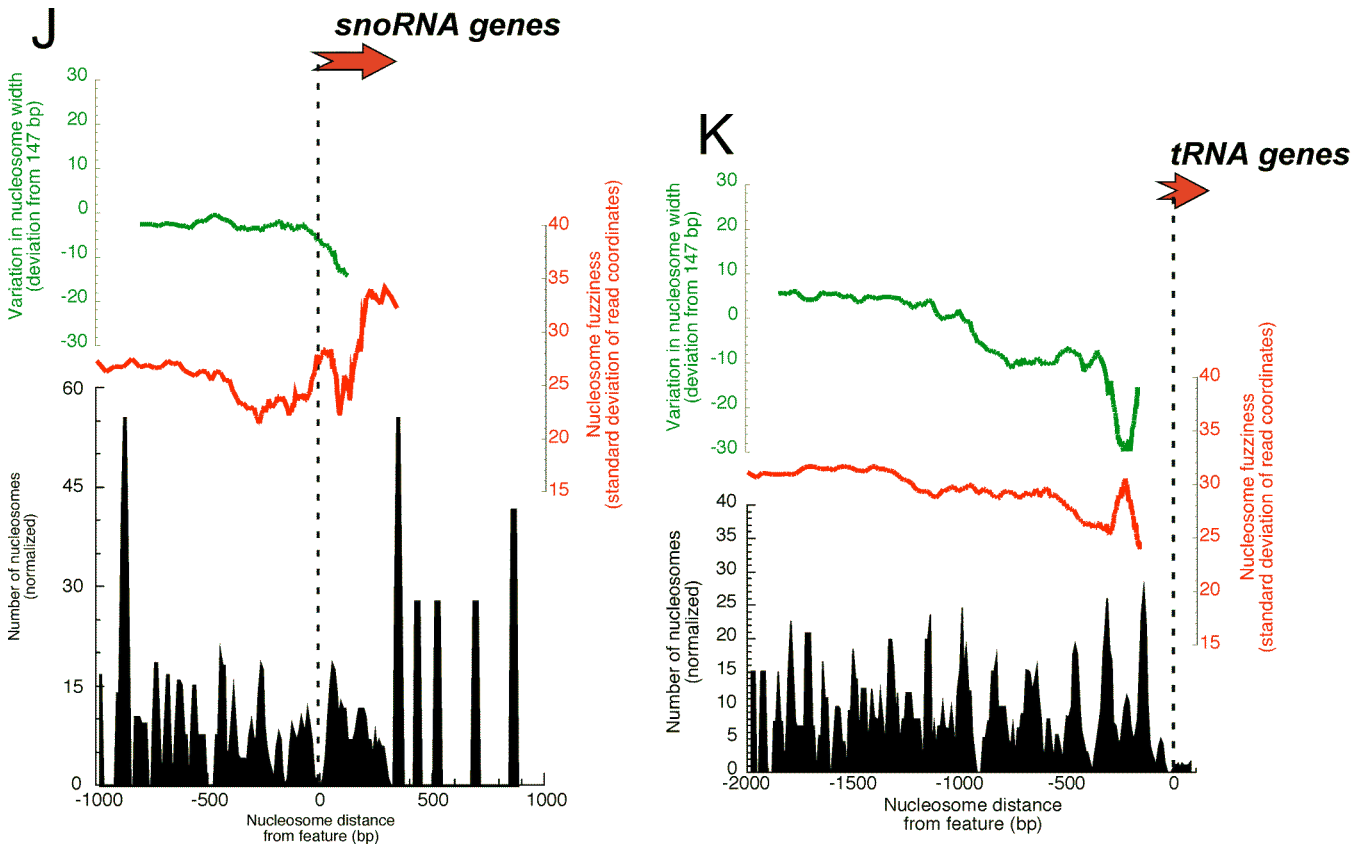


Figure S4J,K (*continued*) (J) snRNAs and snoRNAs lack a canonical nucleosome organization. (K) tRNA genes are depleted of nucleosomes. See Figure S4C,D for figure description. See Table S2 for a listing of features used. Here, feature denotes TSS of the indicated gene class.

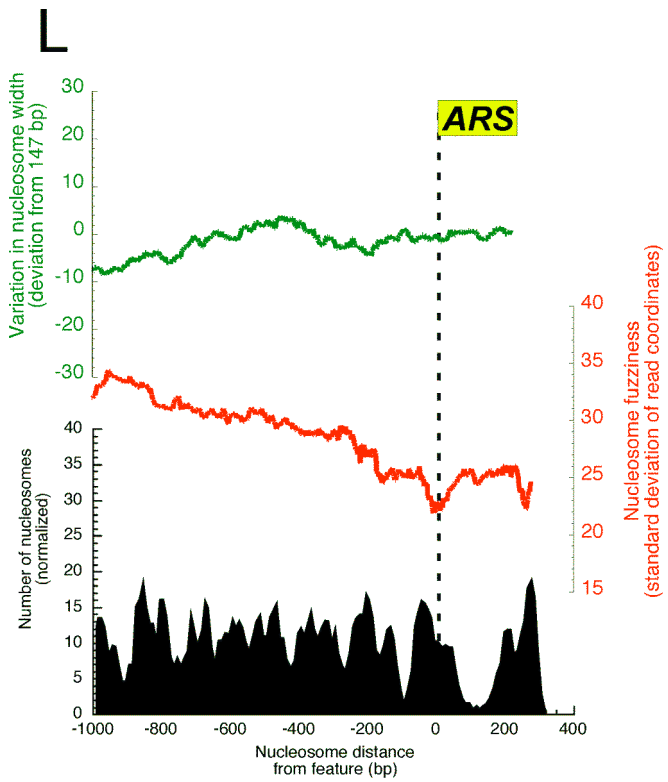


Figure S4L (*continued*) Replication origins are depleted of nucleosomes, although are bordered by nucleosomes. See Figure S4C,D for figure description. Here, feature denotes ARS start site, as defined in Table S2.

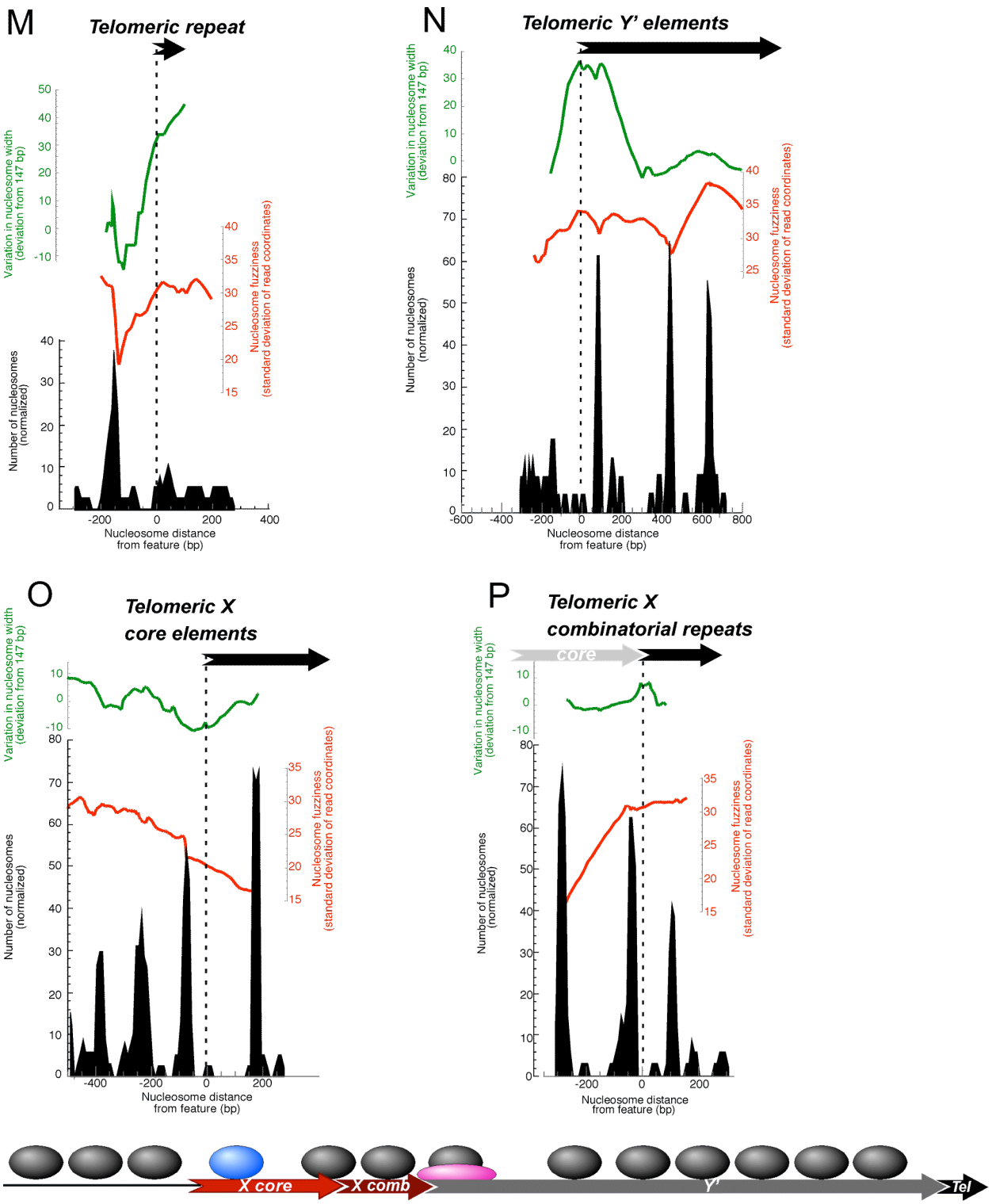


Figure S4M-Q (*continued*) Telomeric regions display a wide range of non-canonical nucleosome spacing, reflecting a telomere-specific organization. See Figure S4C,D for figure description. Here, feature denotes its start site as defined in Table S2. The cartoon provides a schematic of nucleosome organization at the left arm of chromosome IX (as shown in Figure 1D).

Estimation of maximal background in linker regions. Raw nucleosomal calls (read counts) found in consensus linker regions (as defined by consensus distance from the TSS) are considered to be outside of their expected location. This could be attributed to some combination of misplaced nucleosomes (or misplaced TSS) and contaminating nonspecific DNA. If most of the reads in consensus linker regions were from contaminating DNA then the increased fuzziness attributed to "misplaced" nucleosomes (Figure 2A) would be artifactual. However, several factors suggest that is not likely. Contaminating nonspecific DNA is expected to be minimal in nucleosome assignments because the nucleosomes were immunopurified and gel purified (~150 bp), and were required to be assigned by 3 or more reads.

To assess the maximum level of potential contamination in consensus linker regions, we first determined the background level of sequencing reads in an area that is not supposed to contain nucleosomes (described below). While there may not be such a region that absolutely contains no nucleosomes, we can estimate the maximum level of contamination by picking a region that we *a priori* would believe should be deficient in nucleosomes. (Note we are not searching the data and looking for regions that lack nucleosomes.) Such a potential "negative control" region would be the NFR of highly expressed genes since this region is generally nucleosome-free, and is particularly deficient at highly expressed genes (Figure S4E). The question then becomes: How does the per bp read density in this "negative control" region compare to the linker region. We looked at the linker regions between nucleosome +1 and +2, because this region provides the most stringent test. We found that nucleosome read density in the "negative control" NFR region is 15% of the read density in the +1/+2 linker region (Figure S5), which indicates that at most 15% of the fuzziness fluctuation in Figure 2A can be attributed to contamination.

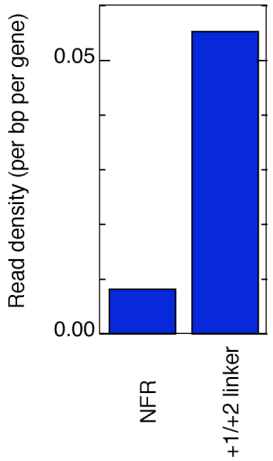


Figure S5. Estimation of maximal background associated with read counts in linker regions. The number nucleosomes (defined here as 73 bp downstream of the 5' end of each sequencing read) located in the NFR (from -110 to -10 relative to the TSS) of the top 5% of gene transcription frequencies (Holstege et al., 1998) were counted (count = 198). The count was normalized to the number of genes (243) examined and the bp range of the NFR (100 bp), and plotted as shown. This read density in the NFR of highly expressed genes was taken to represent the maximum level of background contamination. The number of nucleosomes (defined above) located in linkers between nucleosomes +1 and +2 (from +140 to +160 relative to the TSS) were counted for all genes (count = 5,297), and normalized to the number of genes (4,792) and the bp range of the linker (20).

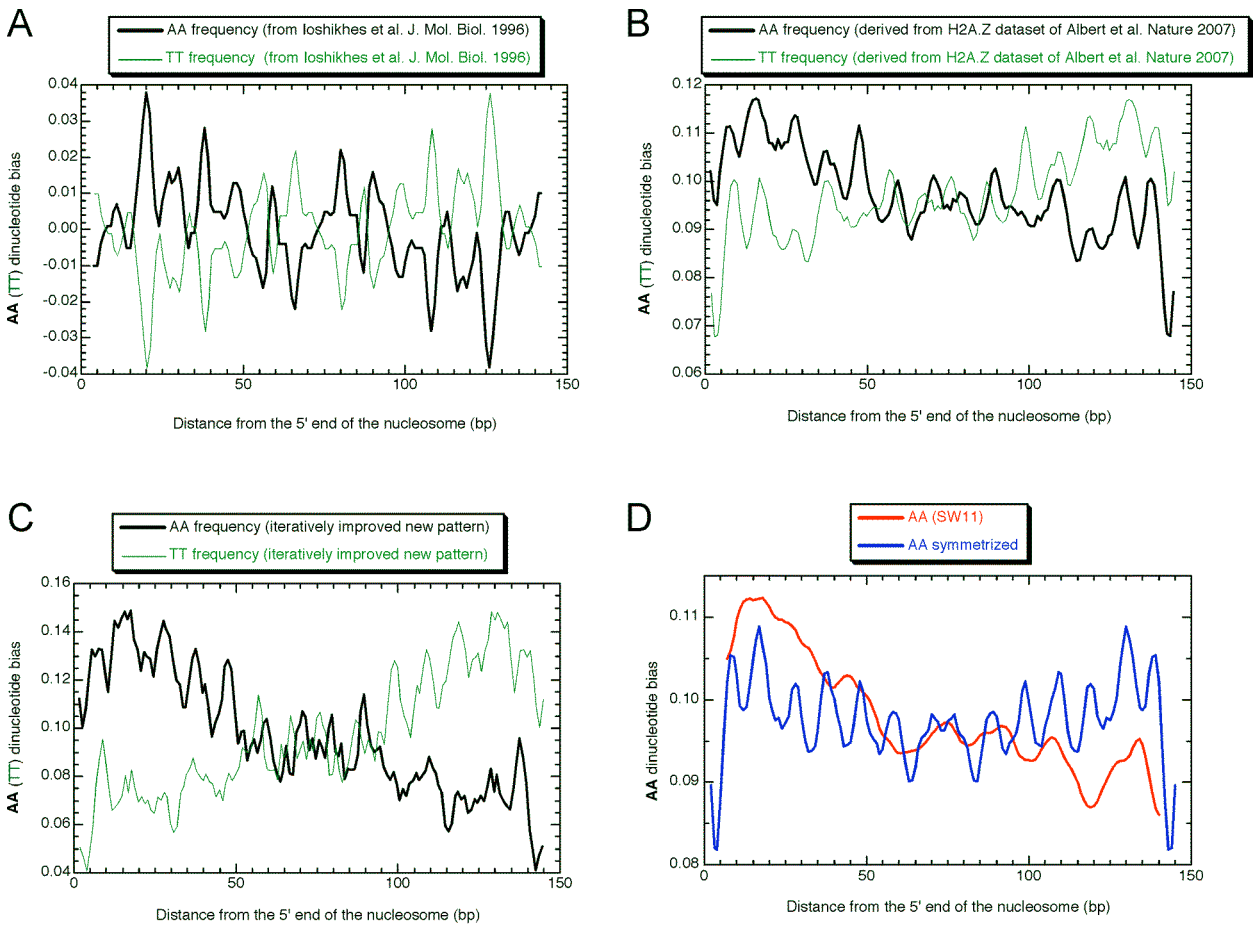


Figure S6. NPS patterns used to measure NPS correlations. (A) AA/TT NPS pattern derived from [Ioshikhes, 1996 #500]. (B) AA/TT NPS pattern derived from [Albert, 2007 #669]. (C) AA/TT NPS pattern used in this study. Its derivation from panel B is described in the methods section. Frequency distributions of AA (black) and TT (green) dinucleotides across highly positioned nucleosomes are shown. Panels A, B, and C are smoothed using 7, 3, and 3 bp moving averages. All the patterns are dyad 180˚ symmetrical as a result of considering of both DNA strands. Panel (A) y-axis is centered to 0. In panel (D), the blue trace shows the pattern symmetrized by averaging the 5'-3' ordered AA pattern in panel B with the 3'-5' reverse ordered AA pattern (blue) so as to remove the slope. The red trace shows the pattern in panel B smoothed using an 11 bp sliding window so as to remove the 10 bp periodicity.

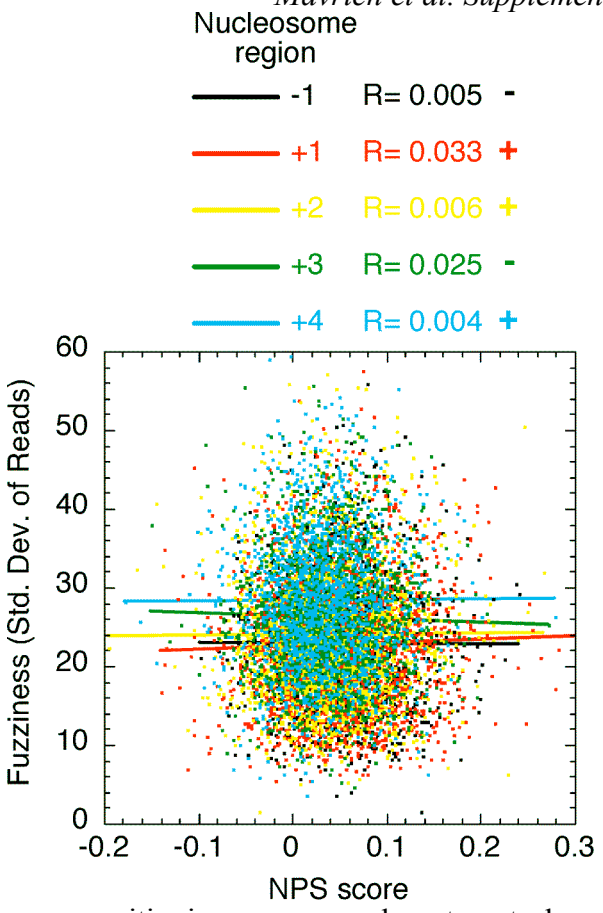
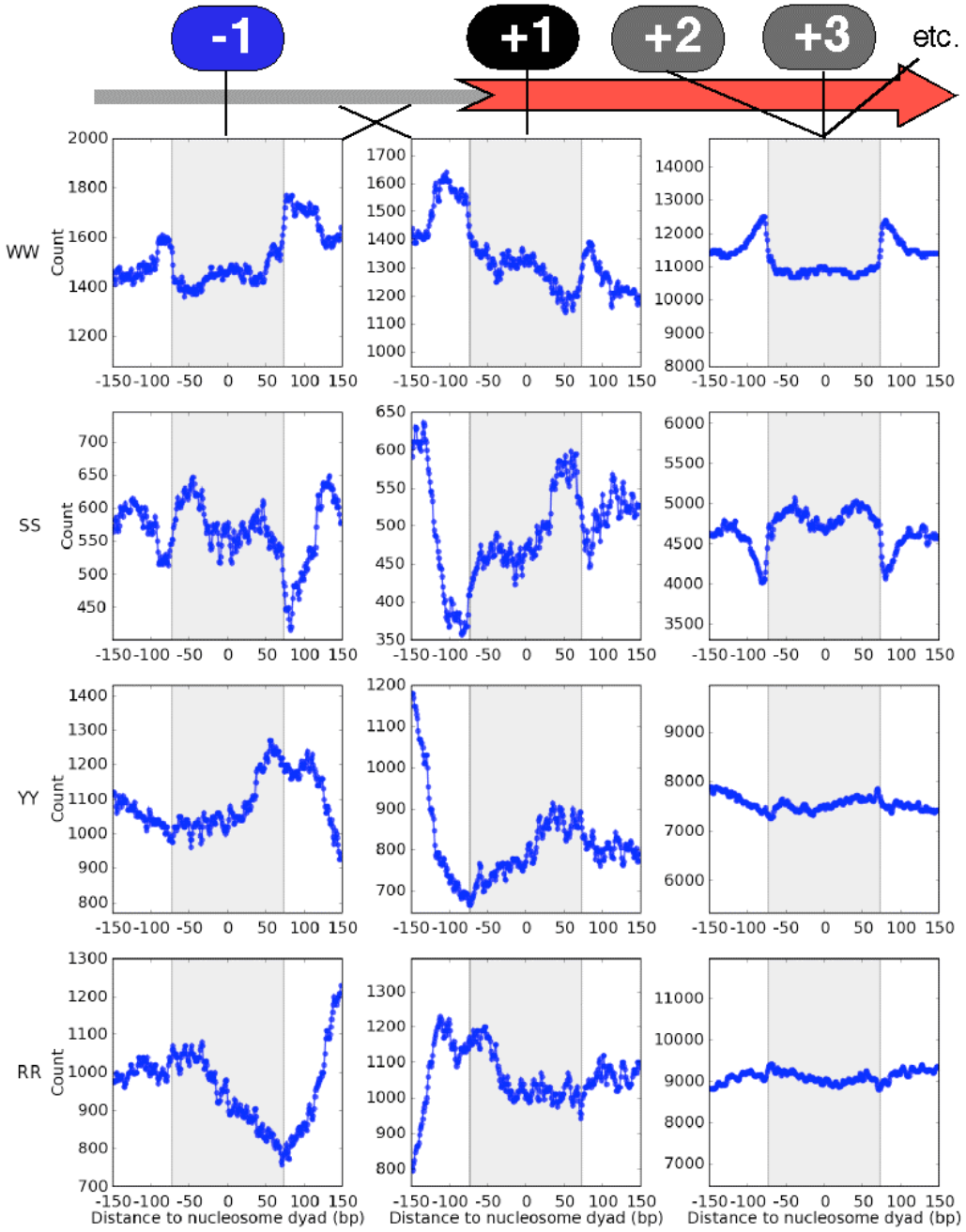


Figure S7. AA/TT nucleosome positioning sequences do not control nucleosome fuzziness. Peak NPS correlation scores were obtained for each gene from (Ioshikhes et al., 2006) and parsed into nucleosome zones (-1, +1, +2, +3, and +4 based upon the pattern shown in Figure 1C). For each zone, the peak NPS score was plotted against the fuzziness value (standard deviation of read locations) calculated for the measured nucleosome position within that zone.



WW = AA, TT, AT, TA	YY = TT, CC, TC, CT
SS = GG, CC, GC, CG	RR = AA, GG, AG, GA

Figure S8A (*continued*)

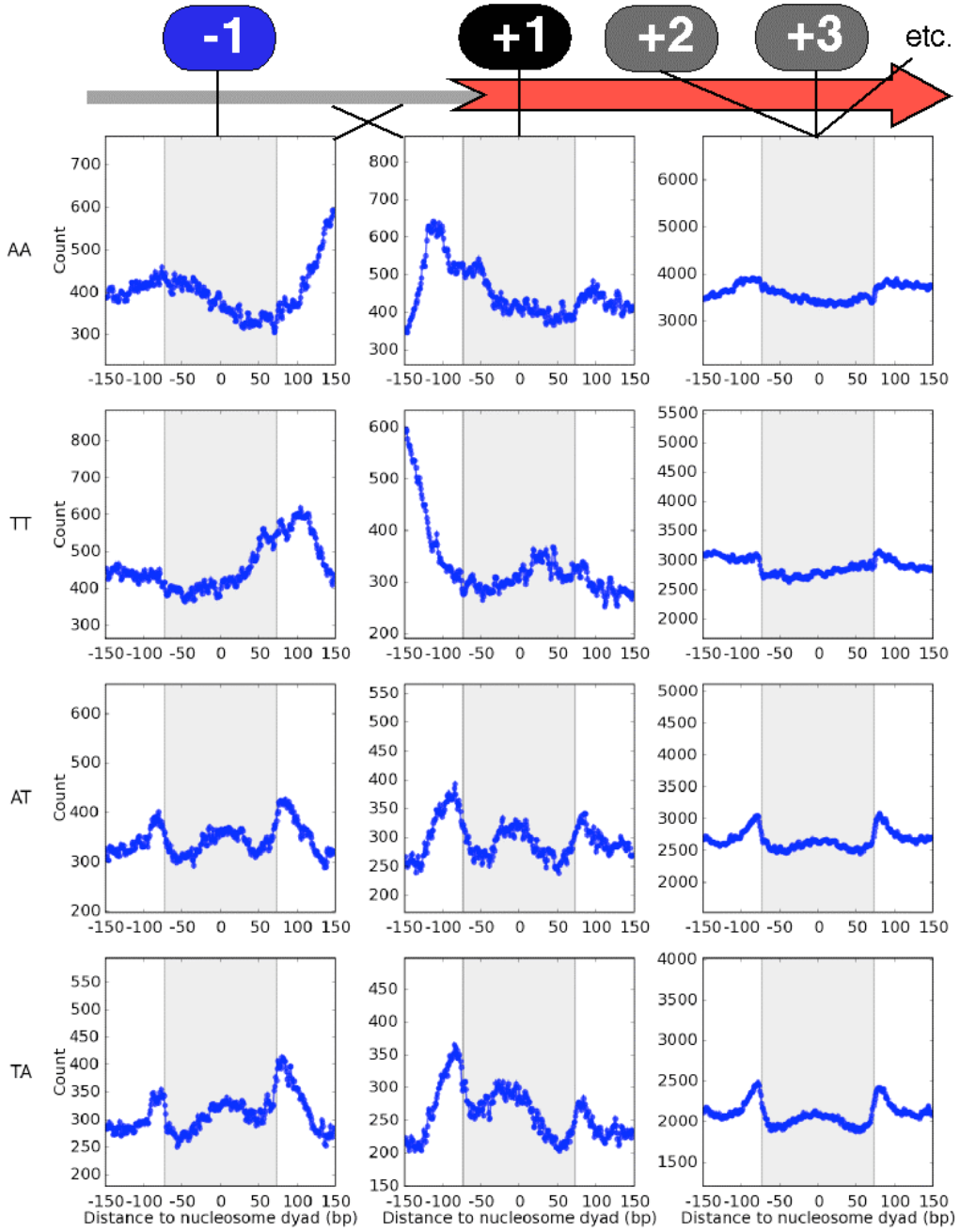


Figure S8B (*continued*)

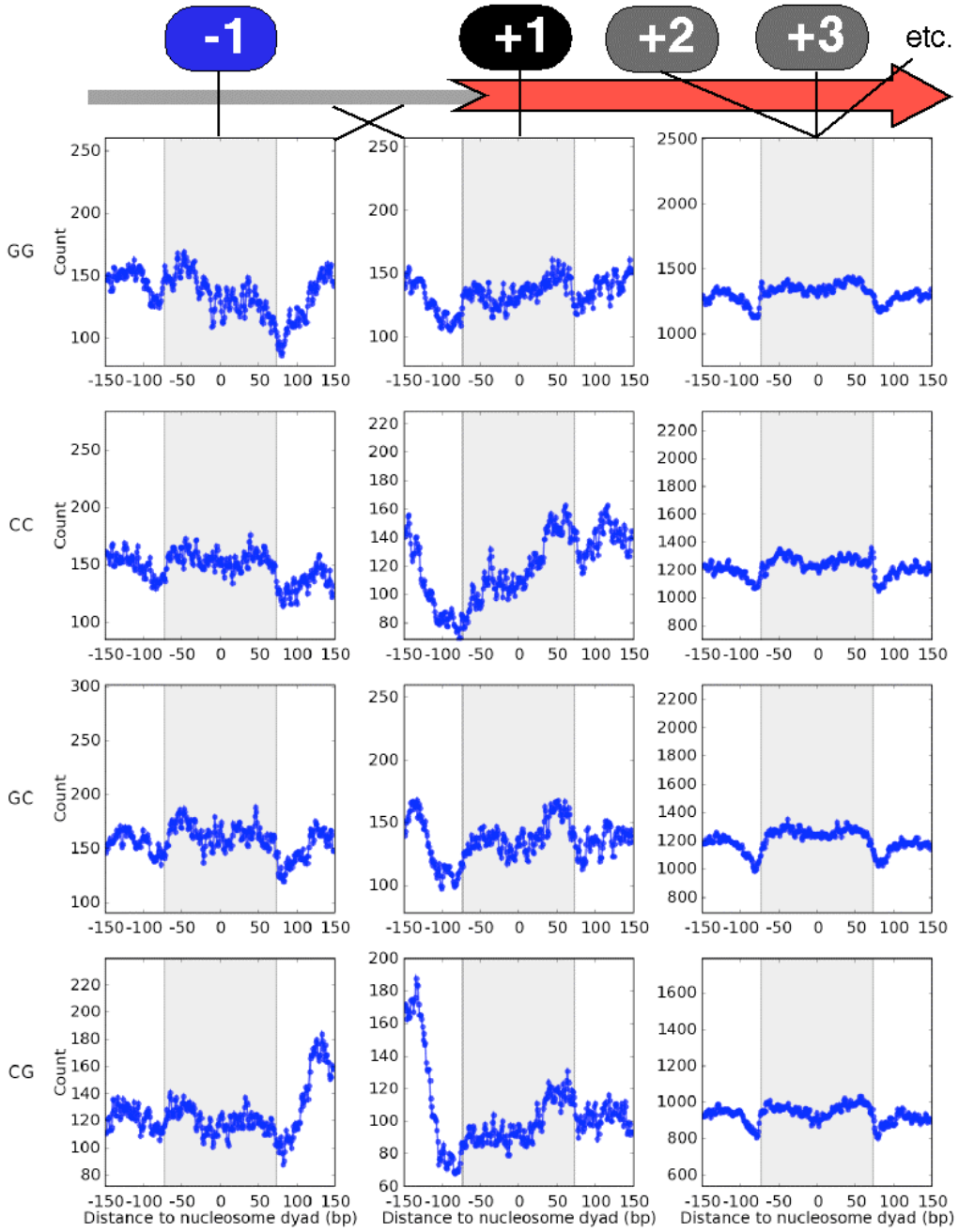


Figure S8C (*continued*)

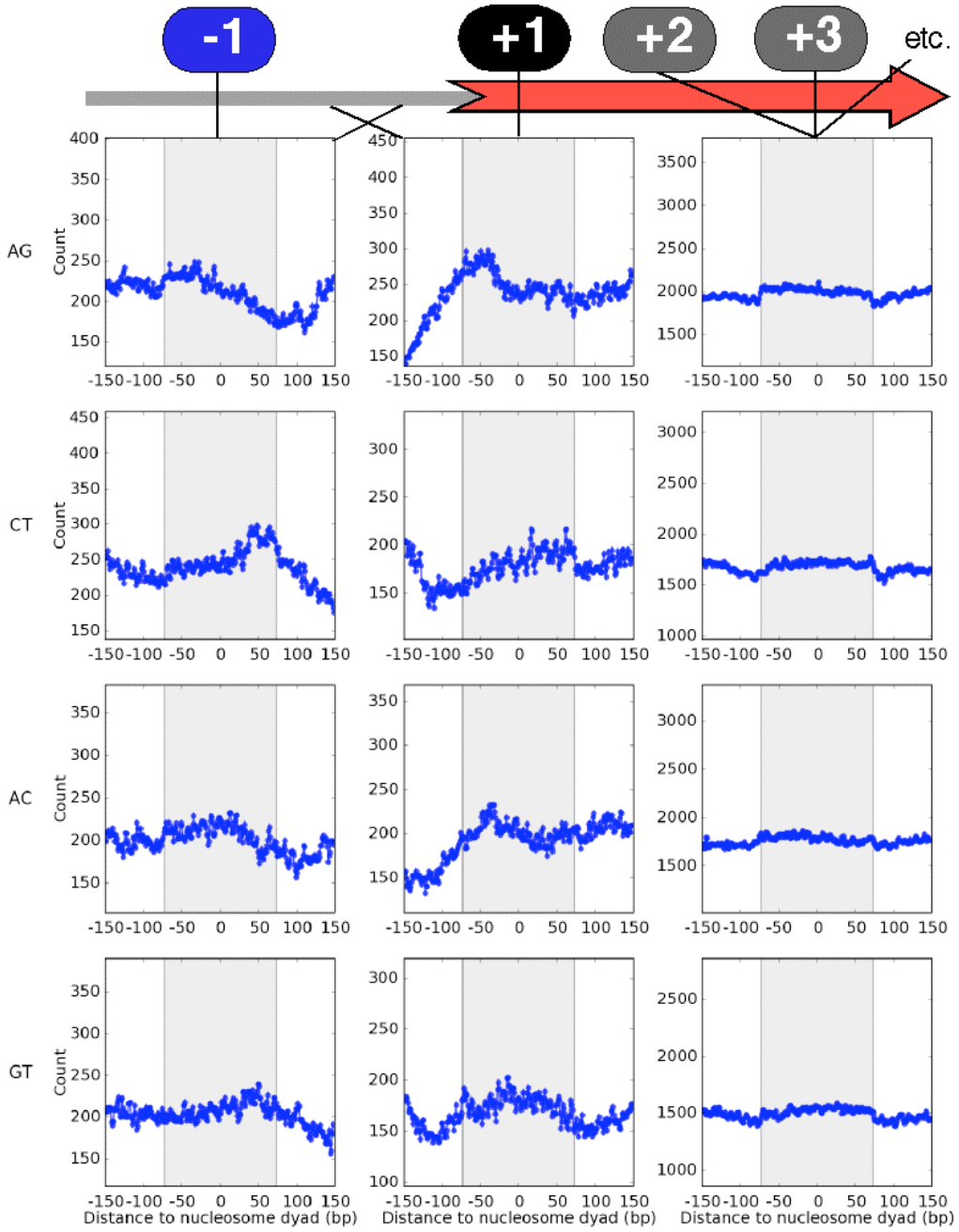


Figure S8D (*continued*)

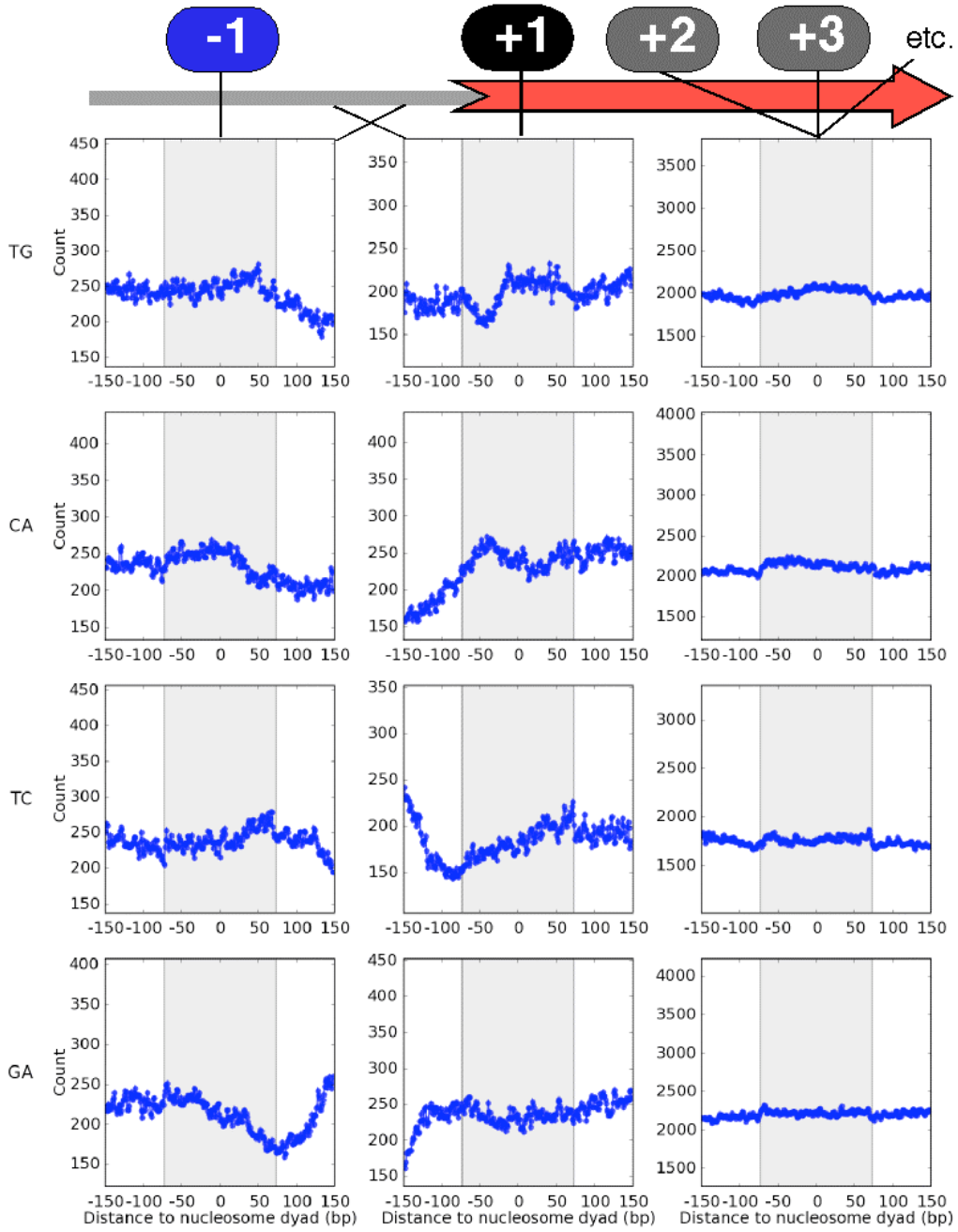


Figure S8E (*continued*)

Figure S8. Dinucleotide frequencies across nucleosomal DNA. The sum total count of each dinucleotide present at the indicated distance from the nucleosome midpoint was determined for three class of nucleosomes (-1, +1, and all other genic nucleosomes). Dinucleotides were counted only on the transcribed strand, so as to maintain directionality of the patterns relative to the TSS. Binned data were smoothed using a 3-bin moving average. The -1 nucleosomes is defined as being located between -300 and -100 relative to the TSS; +1 is defined as being located between -50 and +150 relative to the TSS. The y-axis is scaled such that the ratio of the upper and lower range are the same in all individual dinucleotide plots and in all grouped dinucleotide plots, with two groups having a different scale. Group plots include WW (AA, TT, AT, TA), SS (GG, CC, GC, CG), YY (CC, TT, CT, TC), and RR (AA, GG, GA, AG).

As expected, AA dinucleotide frequencies decreased across nucleosomal DNA in the 5' to 3' direction, and TT increased in frequency, leaving the total AA/TT content relatively constant.

MNase bias. MNase prefers to cleave at A/T-rich dinucleotides. Therefore, the 5' end of sequencing reads will be enriched with A/T dinucleotides that span the 5' end. In mapping the nucleosome midpoint we weight individual read locations to correct for that bias, as described previously (Albert et al., 2007). In addition, nucleosome midpoints were derived by the weighted average location of both nucleosome borders. From these midpoints the nucleosome borders were determined as the midpoint location minus (or plus) 73 bp. As an example, if MNase cuts at coordinate 1000, yielding a W strand 5' end at coordinate at 1000, and also cuts at position 1160, yielding a C strand 5' end at 1160, then (without correcting for MNase bias) the nucleosome midpoint will be called at coordinate 1080 = (1000 + 1160)/2, and the borders at (1080 -/+ 73, or 1007 and 1053), which are different coordinates than the original and thus does not return the original MNase bias.

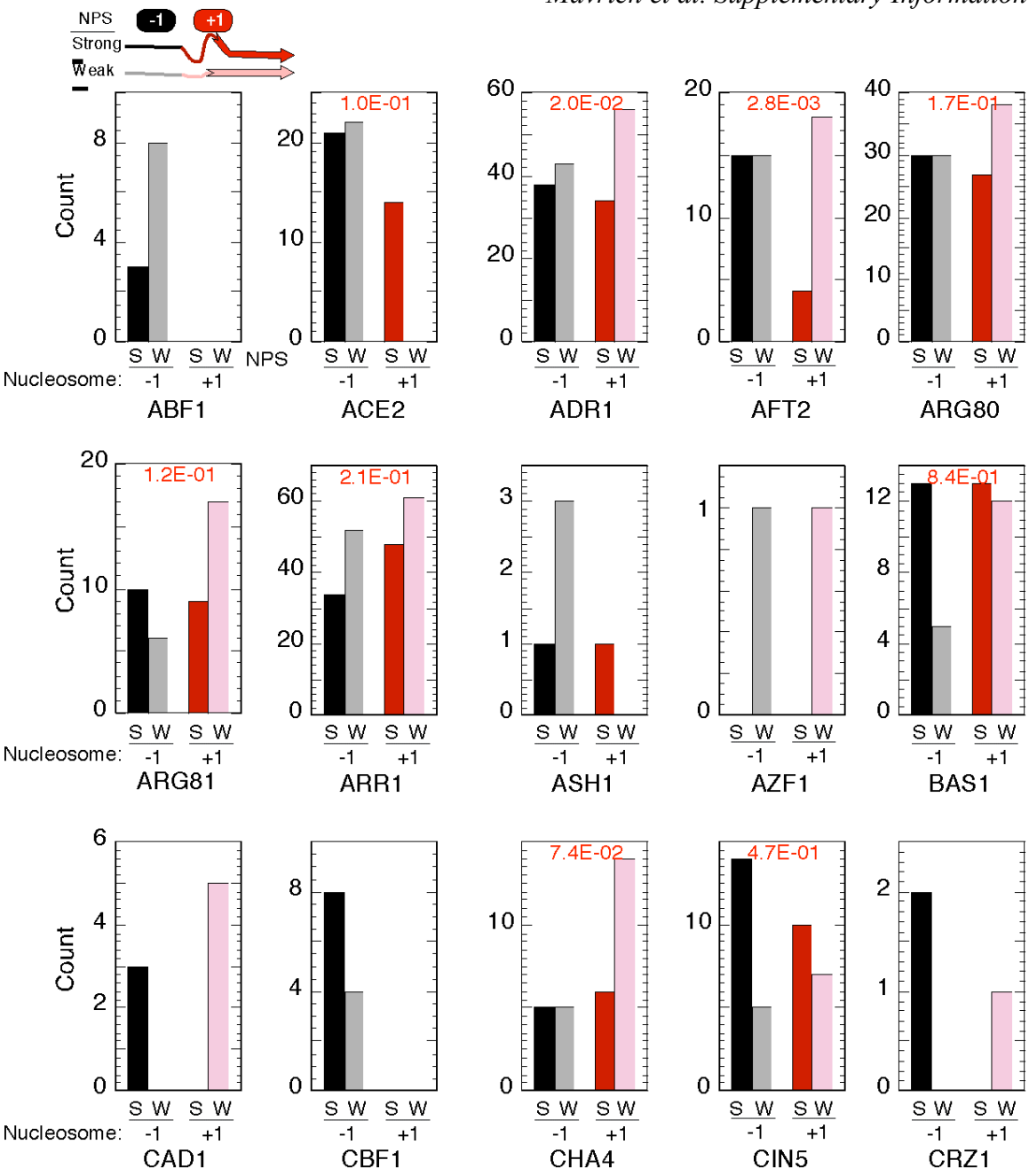


Figure S9 (*continued*)

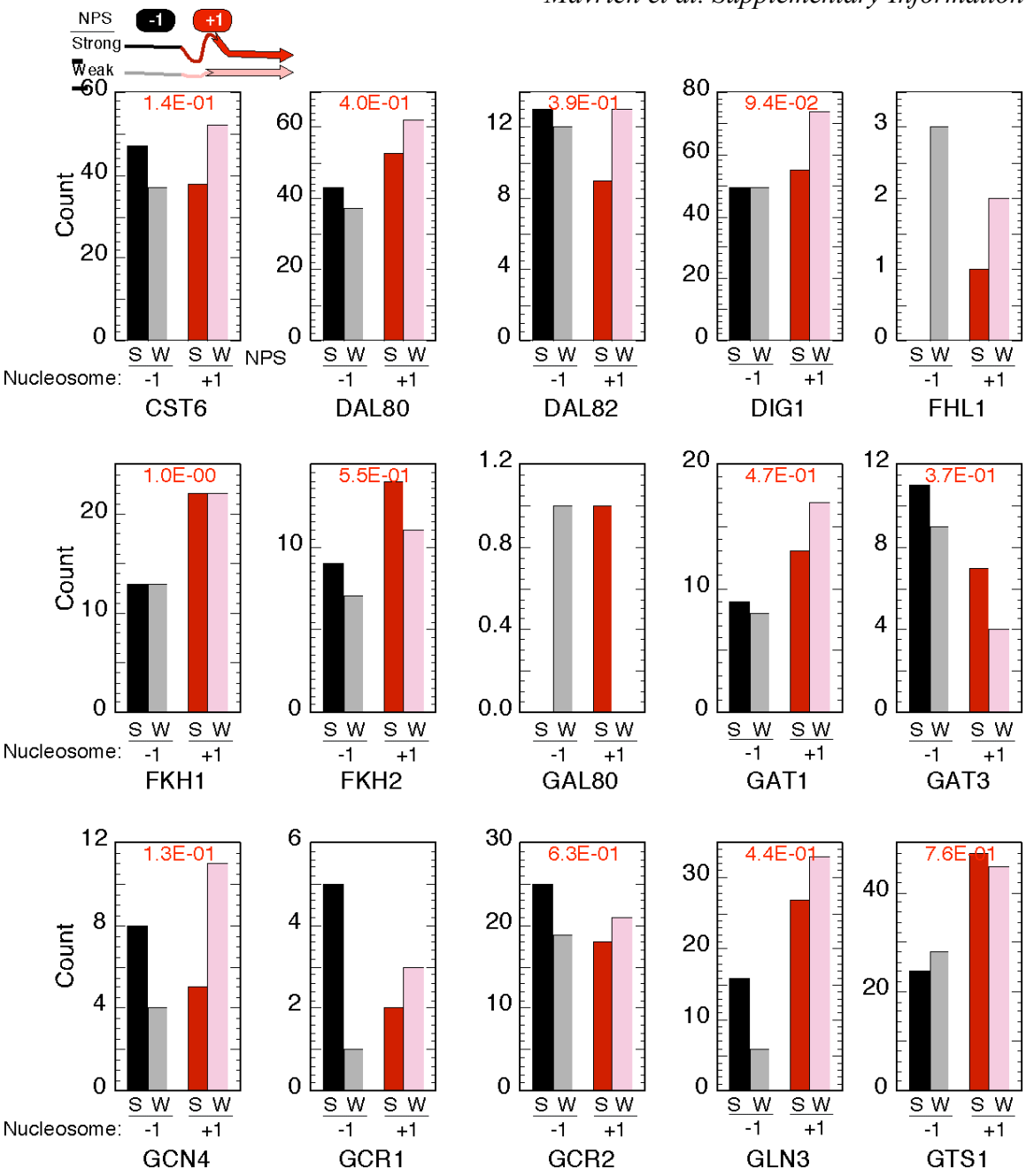


Figure S9 (*continued*)

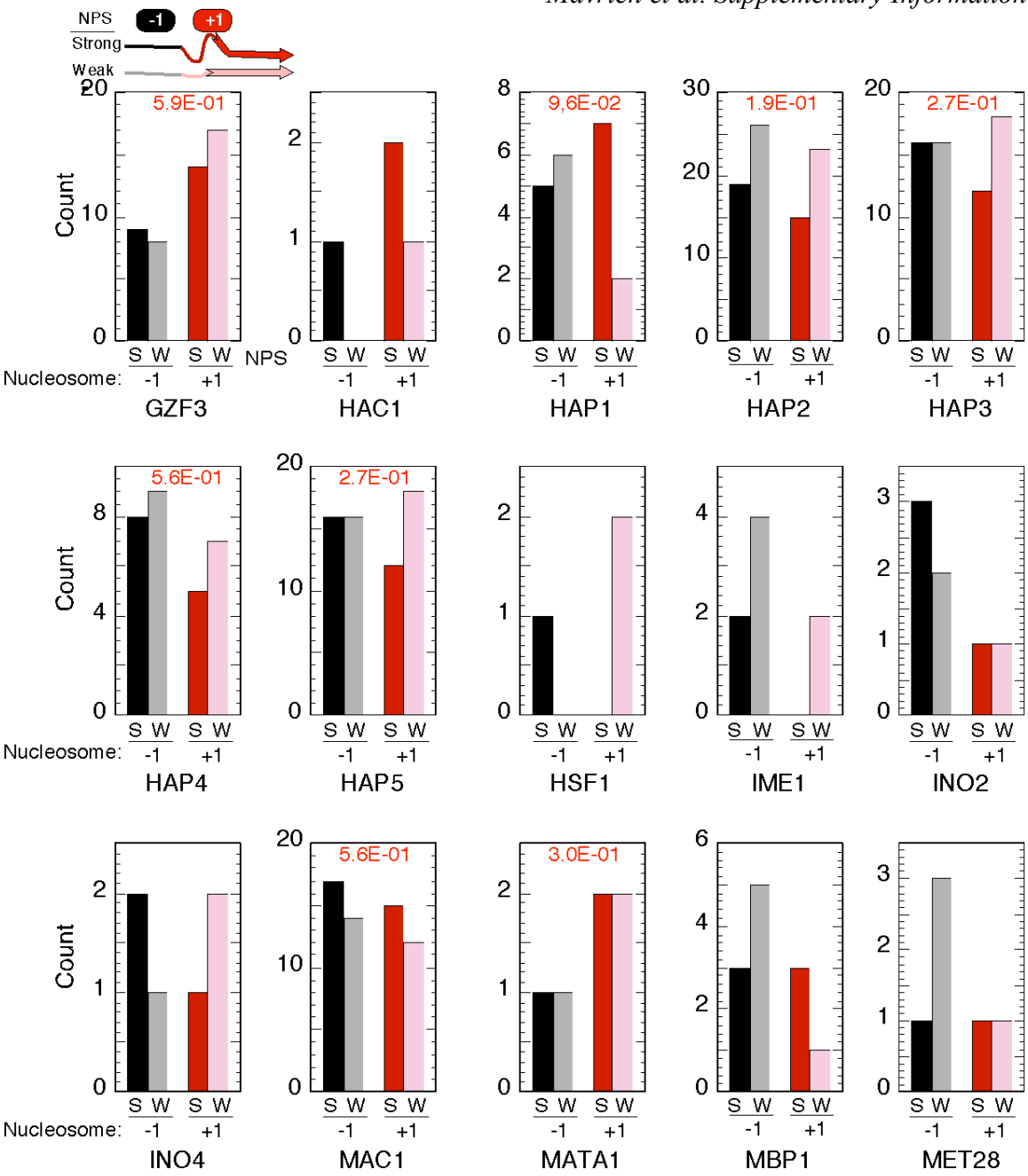


Figure S9 (*continued*)

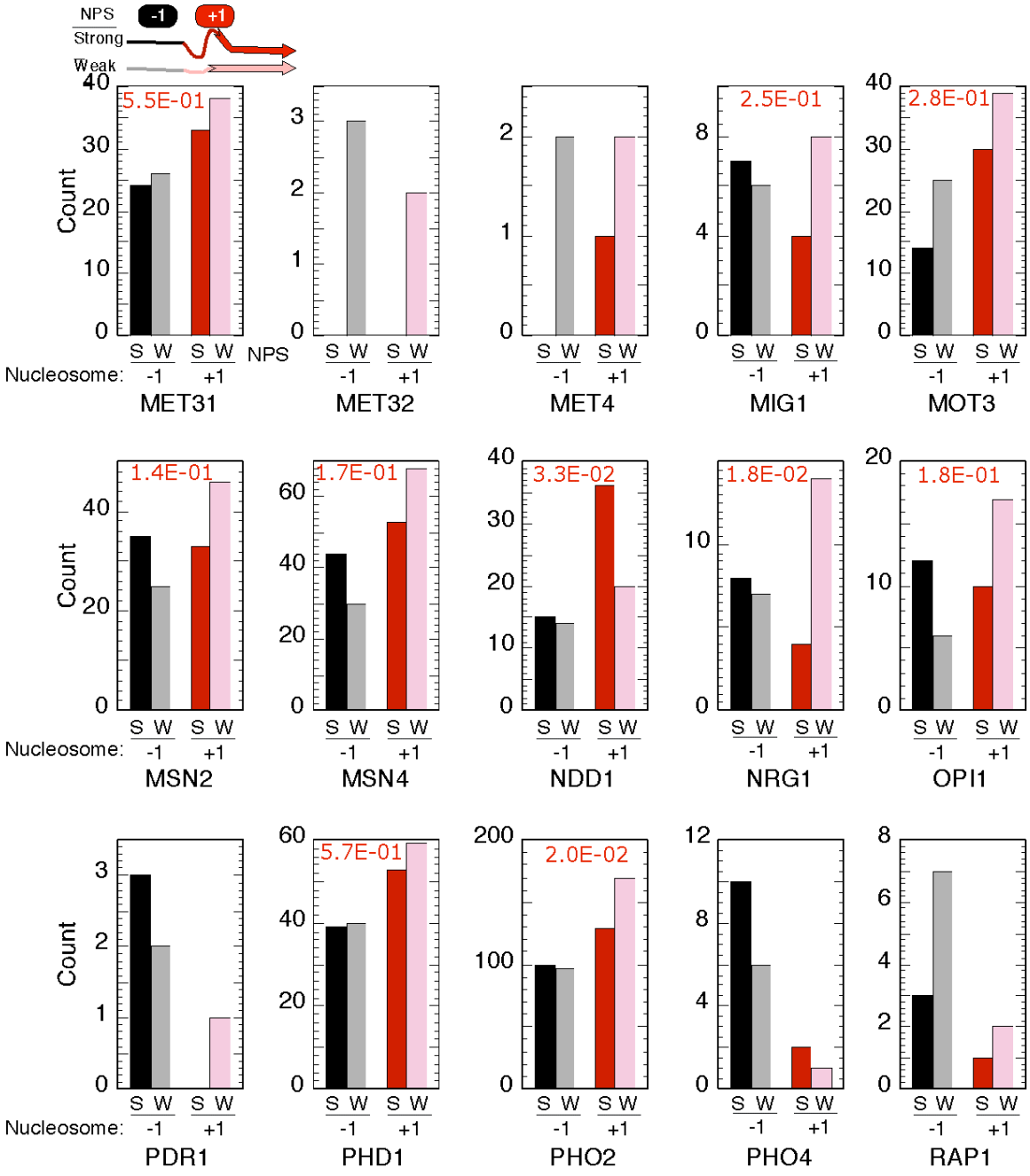


Figure S9 (*continued*)

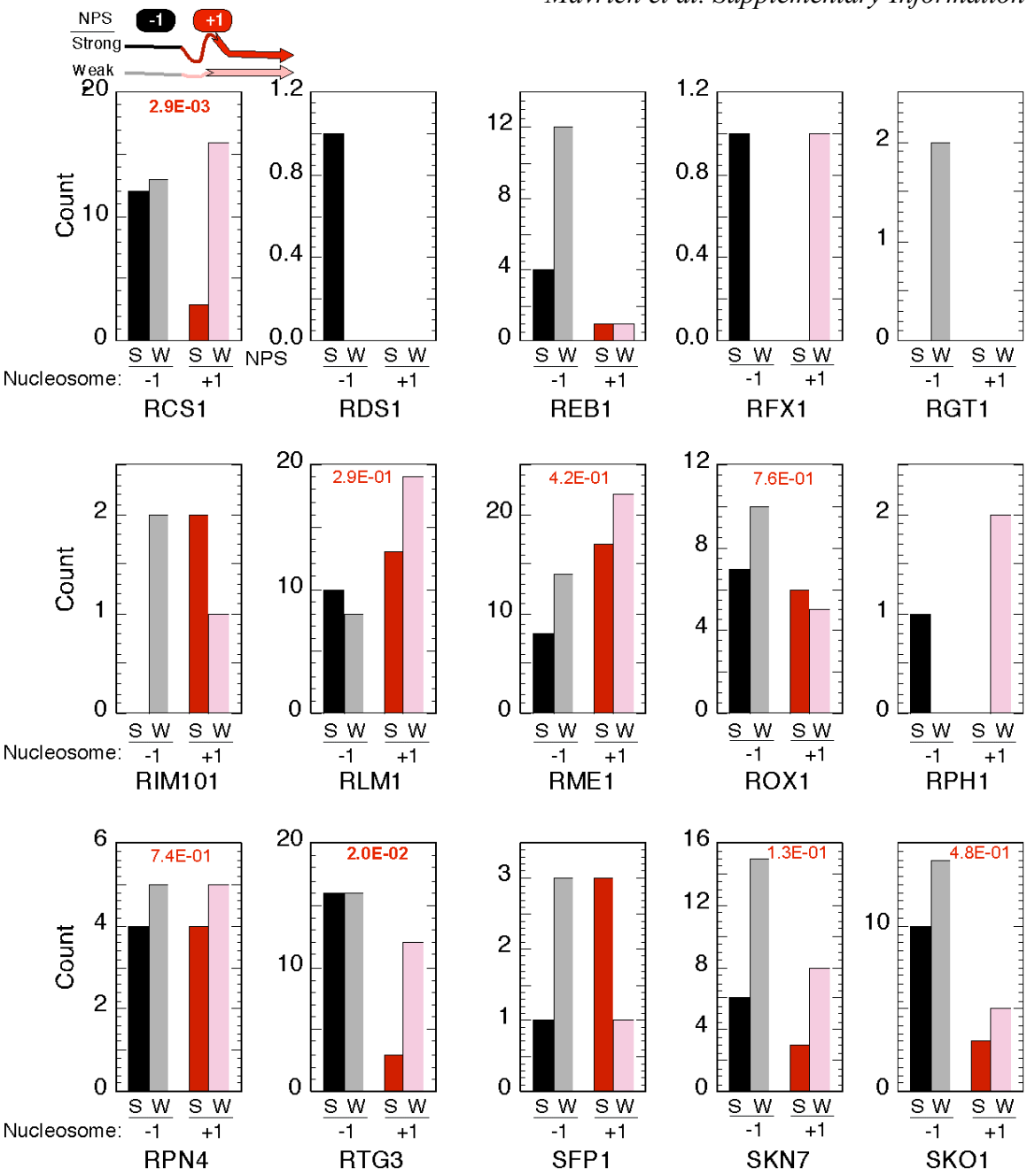


Figure S9 (*continued*)

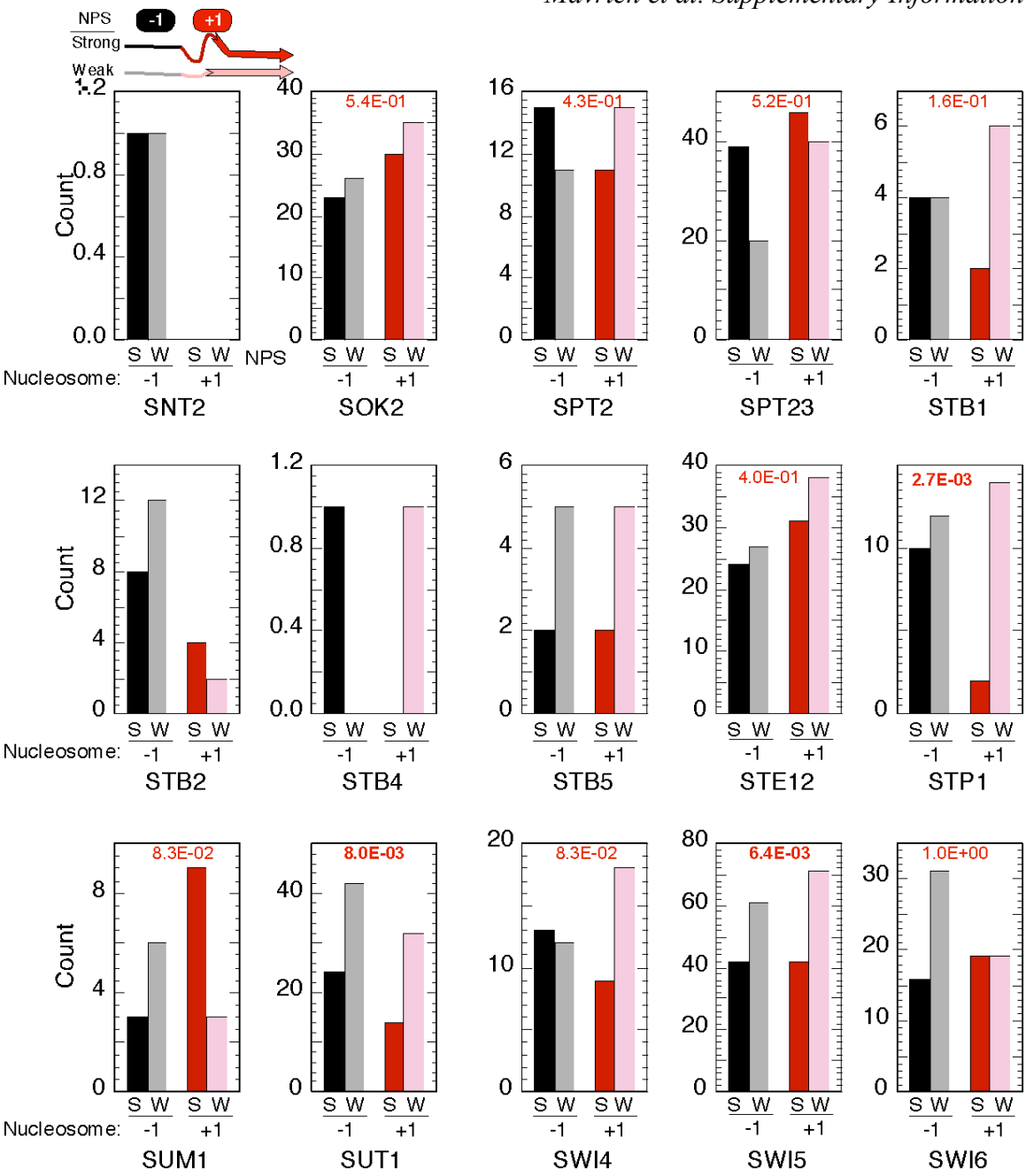


Figure S9 (*continued*)

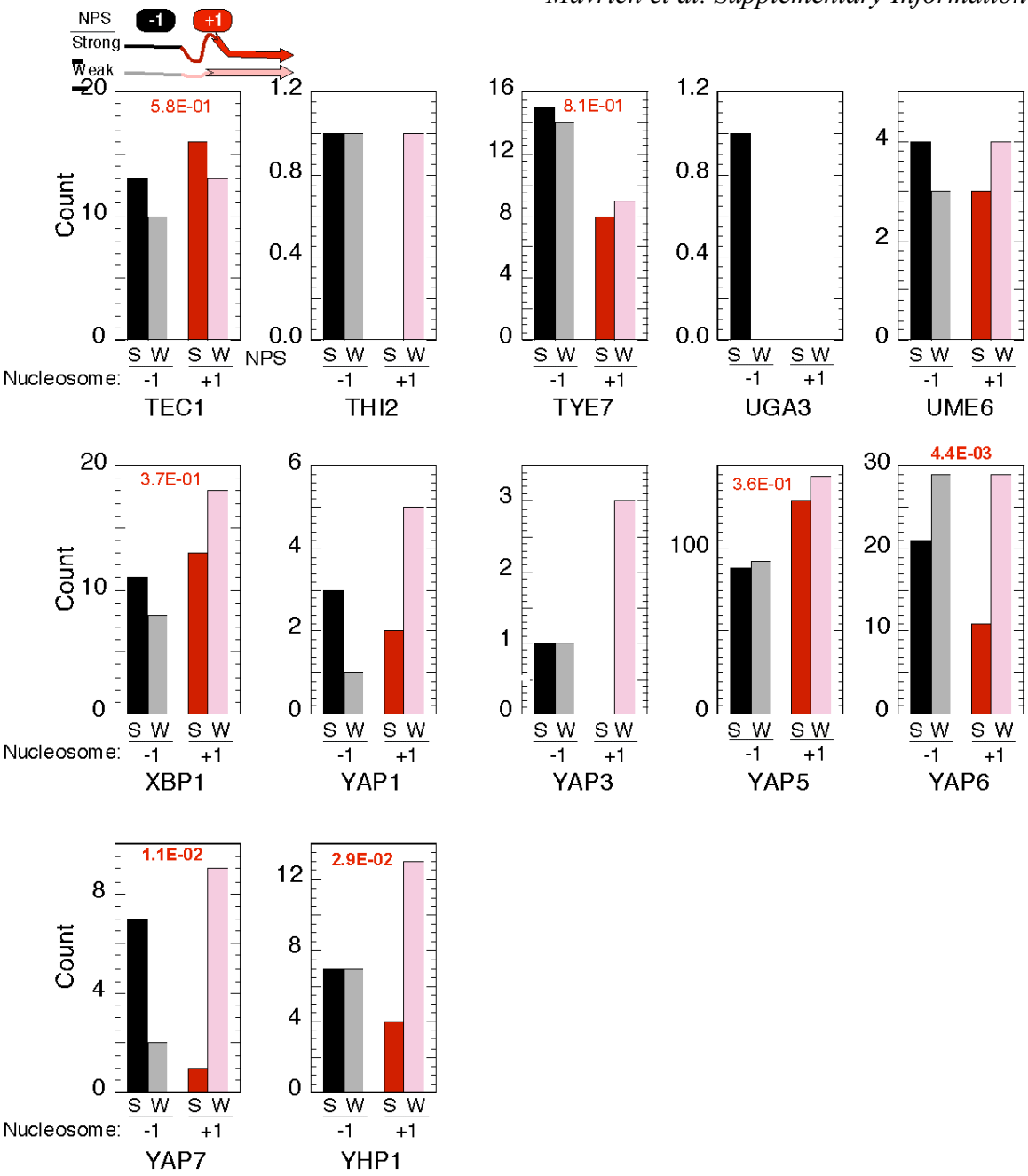


Figure S9. Distribution of transcription factor binding sites at the -1 (-350 to -150 relative to the TSS) and +1 (-150 to +150) nucleosomes, in which the +1 nucleosomal region (from -100 to +100) contains or lacks a strong NPS correlation (see methods). P-values were calculated using CHITEST in EXCEL. Actual counts and other P-value calculations including Bonferroni, FDR, and Benjamin-Hochberg corrections are included in Table S5

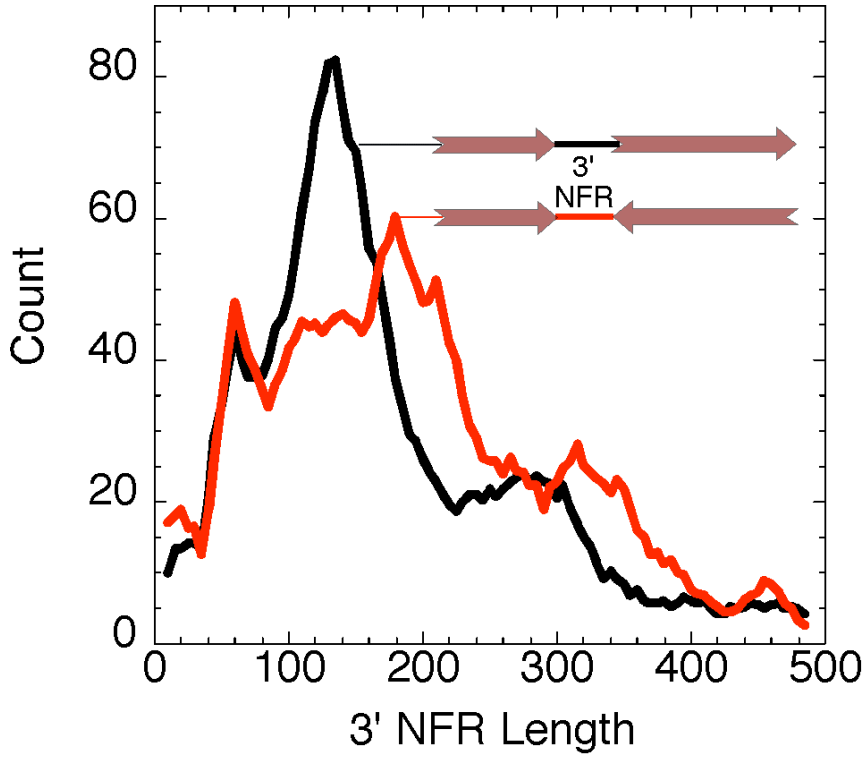


Figure S10. Distribution of 3' NFR lengths at genes with the indicated terminal orientation relative to the downstream gene. NFR length is defined here as the border to border distance between two nucleosomes.

A 3' NFR is defined here as the "closest" linker DNA border to the 3' end of the ORF. For purposes of "closeness" we masked those linkers that were <50 bp in length. If no linker >50 bp long was within 500 bp of the ORF end point, then the closest linker that was <50 bp was used as the 3' NFR. This strategy therefore preferentially looks for NFRs that are >50 bp, but if it does not find one, it will allow one that is <50 bp in length.

NFR lengths were binned in 5 bp bins and plotted as a 5 bin moving average

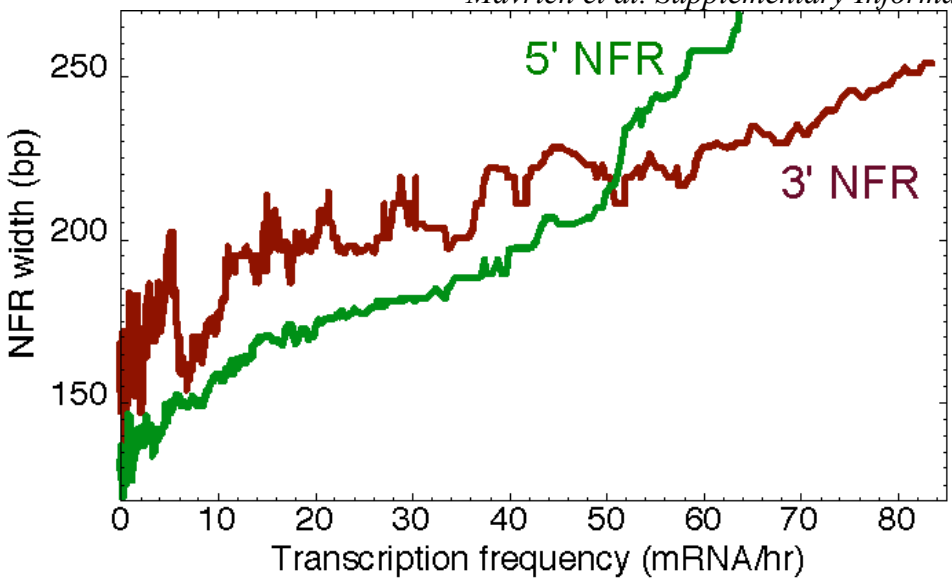


Figure S11. NFR width as a function of transcription frequency (Holstege et al., 1998). NFR width represents the distance between the flanking nucleosome borders. Data are shown as a moving average of 200-gene windows.

SUPPLEMENTARY REFERENCES

Albert, I., Mavrich, T.N., Tomsho, L.P., Qi, J., Zanton, S.J., Schuster, S.C., and Pugh, B.F. (2007). Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature *446*, 572-576.

Almer, A., and Horz, W. (1986). Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the PHO5/PHO3 locus in yeast. Embo J *5*, 2681-2687.

Basehoar, A.D., Zanton, S.J., and Pugh, B.F. (2004). Identification and distinct regulation of yeast TATA box-containing genes. Cell *116*, 699-709.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci USA *103*, 5320-5325.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell *11*, 4241-4257.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717-728.

Ioshikhes, I.P., Albert, I., Zanton, S.J., and Pugh, B.F. (2006). Nucleosome positions predicted through comparative genomics. Nat Genet *38*, 1210-1215.

Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R., and Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet *39*, 1235-1244.

Li, B., and Reese, J.C. (2001). Ssn6-Tup1 regulates RNR3 by positioning nucleosomes and affecting the chromatin structure at the upstream repression sequence. J Biol Chem *276*, 33788-33797.

Raisner, R.M., Hartley, P.D., Meneghini, M.D., Bao, M.Z., Liu, C.L., Schreiber, S.L., Rando, O.J., and Madhani, H.D. (2005). Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell *123*, 233-248.

Shimizu, M., Roth, S.Y., Szent-Gyorgyi, C., and Simpson, R.T. (1991). Nucleosomes are positioned with base pair precision adjacent to the alpha 2 operator in Saccharomyces cerevisiae. Embo J *10*, 3033-3041.

Teng, Y., Yu, S., and Waters, R. (2001). The mapping of nucleosomes and regulatory protein binding sites at the Saccharomyces cerevisiae MFA2 gene: a high resolution approach. Nucl Acids Res *29*, E64-64.

Zanton, S.J., and Pugh, B.F. (2006). Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock. Genes Dev *20*, 2250-2265.