Supplementary Methods

Antibody specificity

Peptide competition assays to verify antibody specificity (Supplementary Fig. S1B) were performed by blocking 2 µg of each antibody with a 10 fold excess of its corresponding peptide over night at 4°C with agitation. 1 pmol to 100 pmol of each peptide were blotted on a nitrocellulose membrane and decorated with blocked and unblocked antibodies.

Sequencing, pre-processing and read alignment

Prepared libraries were quantified using the dsDNA HS assay for Invitrogen Qubit 2.0 Fluorometer (Thermo Fisher Scientific) and size distribution was measured with the Bioanalyzer High Sensitivity DNA Kit (Agilent). DNA libraries resulting from MNase digestion and ChIP were sequenced on an Illumina HiSeq 2500 in high output run mode. All histone ChIP-seq reads were first trimmed for adapter sequence and low quality tails ($Q < 20$) with Trim Galore (v.0.4.2) [1].

Alignments were performed using the local mode of Bowtie 2 software [2] with default parameters except the seed alignment mismatch parameter which was set to 1 (-N 1).We used these alignments for the subsequent steps described in sections *Peak Calling* and *Comparative Pol II analysis and pausing index*.

For the steps described in *Segmentation analysis of chromatin marks*, we used histone ChIP-seq alignments performed using the default parameters of the GEM mapper [3] and then duplicated reads were annotated with Picard tools (v1.115) (http://broadinstitute.github.io/picard).

ClustalW alignments

Amino acid sequences from RPB1 subunits were aligned by ClustalW and visualized in Geneious Prime 2020.2.2 and BioEdit [4] (Accession Nos.: H.s. P24928; S.p. NM001021568; S.c. YDL140C; T.t. 00538940; P.t. PTET.51.1.P1370127).

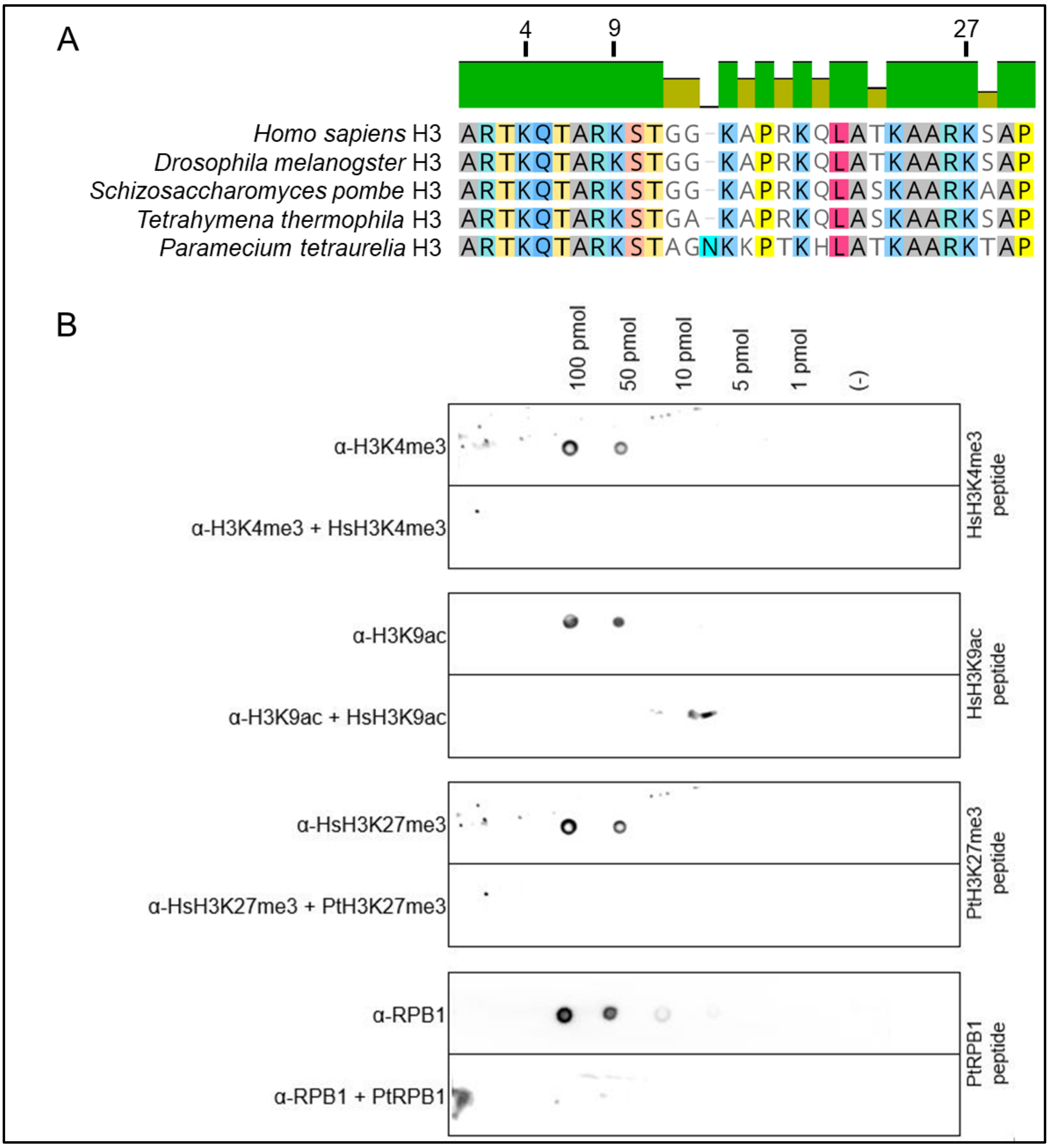
[1] *Krueger, F. (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 516:517.*

[2] *Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4):357.*

[3] *Marco-Sola, S., Sammeth, M., Guig´o, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nature methods, 9(12):1185.*

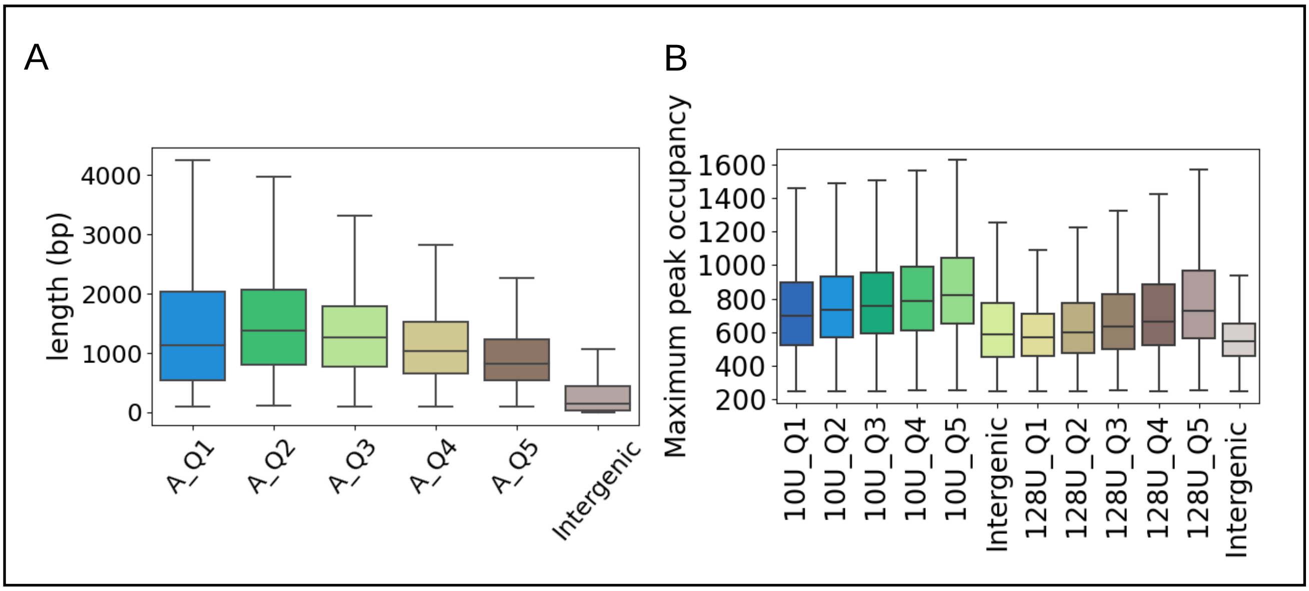
[4] *Hall, T., Biosciences, I., and Carlsbad, C. (2011). BioEdit: an important software for molecular biology. GERF Bull Biosci, 2(1):60–61.*

Supplementary Fig. S1



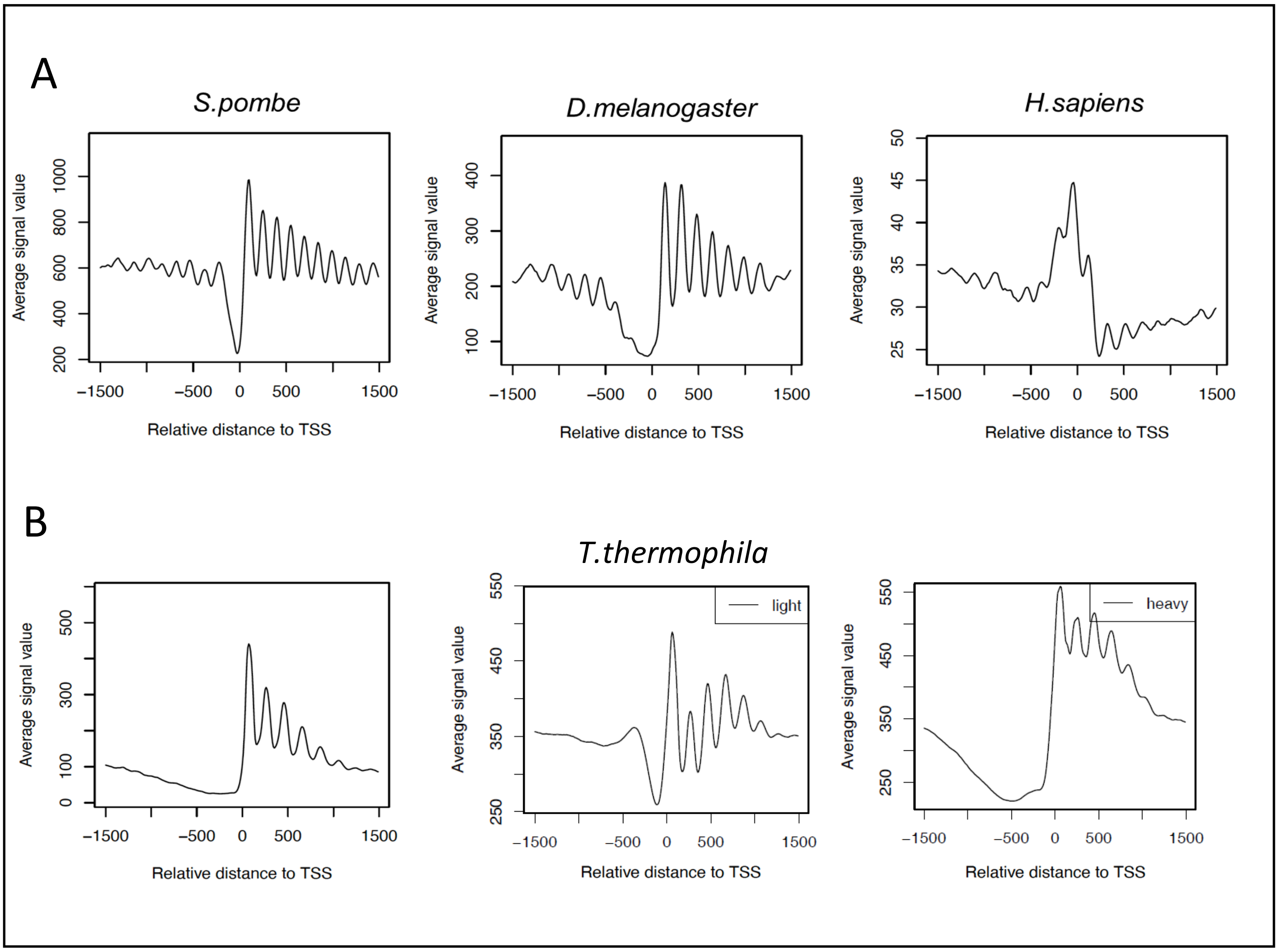
Supplementary Fig. S1 A) Sequence alignment of the N-terminal amino acid residues (1-30) of histone H3 proteins of indicated species. Positions analyzed for their modifications are highlighted (H3K4, H3K9 and H3K27). Accessions: CAB02546 (*Homo sapiens*), NP_001027285 (*Drosophila melanogaster*), P09988 (*Schizosaccharomyces pombe*), XP_001016594 (*Tetrahymena thermophila*) and PTET.51.1.P1080178, H3P1 (*Paramecium tetraurelia*). B) Competition assay. 100 pmol to 1 pmol of each peptide were spotted and either probed with the corresponding antibody or with antibody that was blocked with the indicated peptide in a 10 fold excess in advance. For all antibodies, blocking results in a loss of specific binding to the membrane bound peptide. Blocking the α-HsH3K27me3 with the Paramecium PtH3K27me3 peptide also results in complete competition.

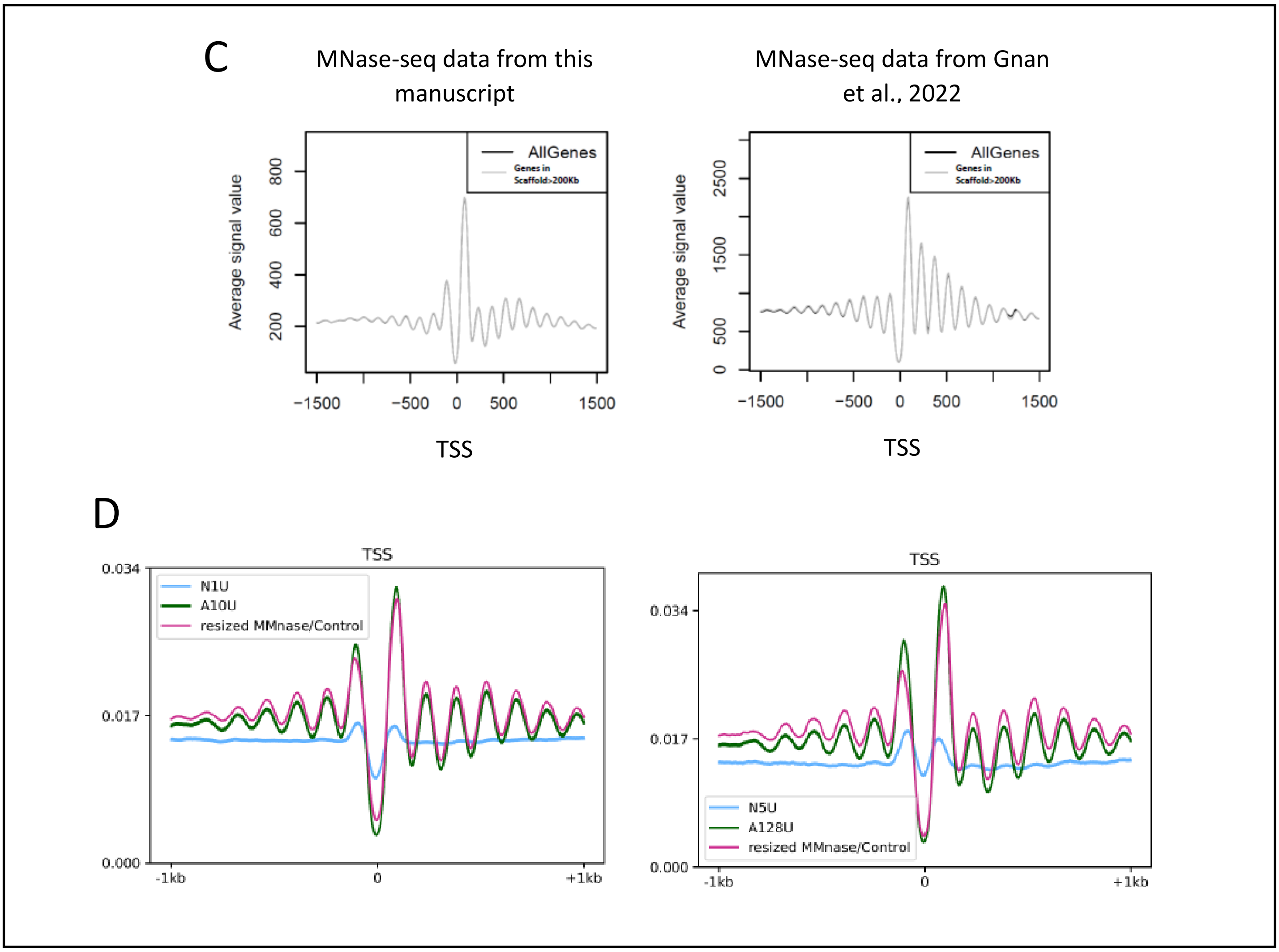
Supplementary Fig. S2



Supplementary Fig. S2 A) Gene length distribution for all genes in each expression quantile and intergenic regions plotted in Figure 3C. B) Maximum peak occupancy along the gene body for all genes in each expression quantile and intergenic regions for 10U and 128U MNase digest.

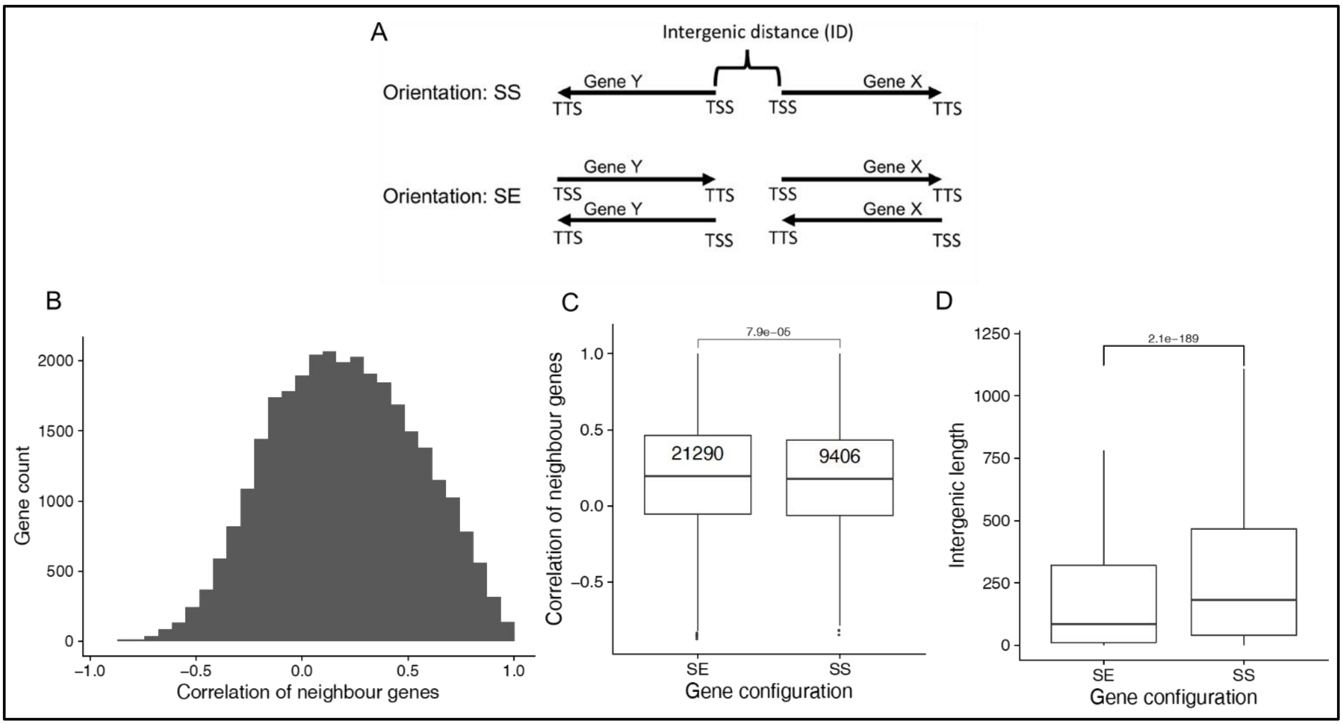
Supplementary Fig. S3





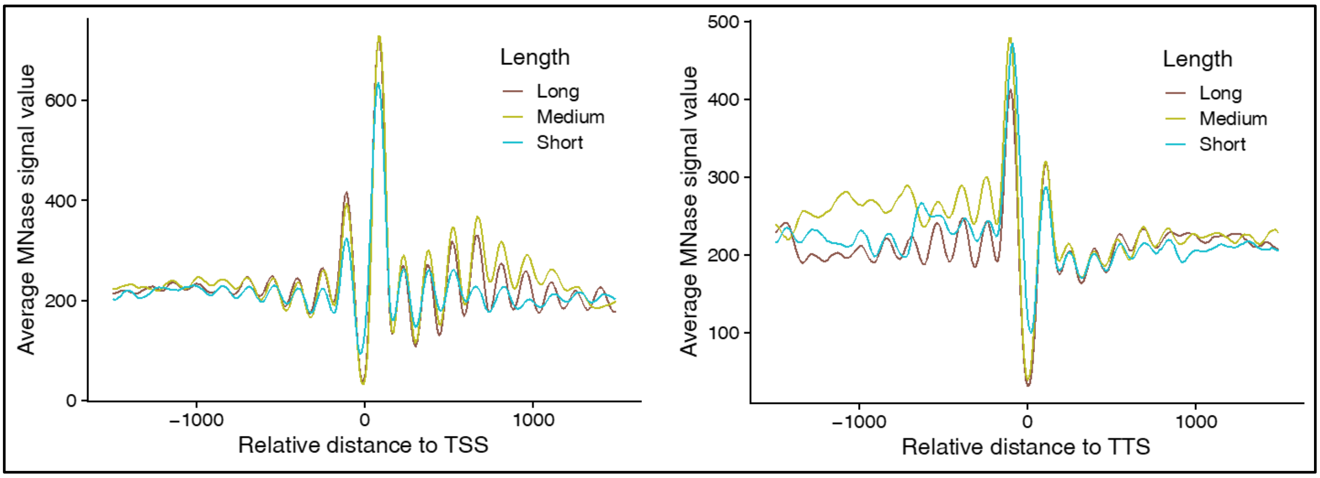
Supplementary Fig. S3 A) Profile plot of nucleosome distribution at the transcription start site (TSS) for all analyzed genes of *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*. Signals for 1500 bp up- and downstream of the TSS are shown. TSS were annotated by EST analysis (*S.pombe*, Wood et. al. (2002); *D.melanogaster* (Adams et al. 2000); *H.sapiens*, Lander, E. S. et al. (2001)). *P.tetraurelia* TSS annotation (Fig. 3) is predicted from CAP-seq data (Arnaiz et. al 2017). B) Same plot as in A, but for *Tetrahymena thermophila* MNase-seq analyses from varying fixation/digestion protocols. Left: Yang et al. (2019) (SRR2041661), mild fixation+light digest; middle: Xiong et al. (2016), no fixation+light digest (Rep1, GSM2055775); right: no fixation+heavy digest (Rep1, GSM2055773). Data set information is found in Supplementary Table 1 (Comparative MNase analysis). C) Nucleosome distribution at the transcription start site (TSS) obtained by plotting data from this manuscript (left) and the data from Gnan et al., 2022 (right), both analyzing nucleosome profiles of *Paramecium tetraurelia* chromatin. Plots were created for all genes and genes on scaffolds >= 200 kb using the DANPOS2 pipeline. D) Profile plot created by Gnan et al., 2022 using their nucleosome analysis pipeline and MNase-seq data from this manuscript (Drews et al.) to allow for comparisons of different pipeline approaches.

Supplementary Fig. S4



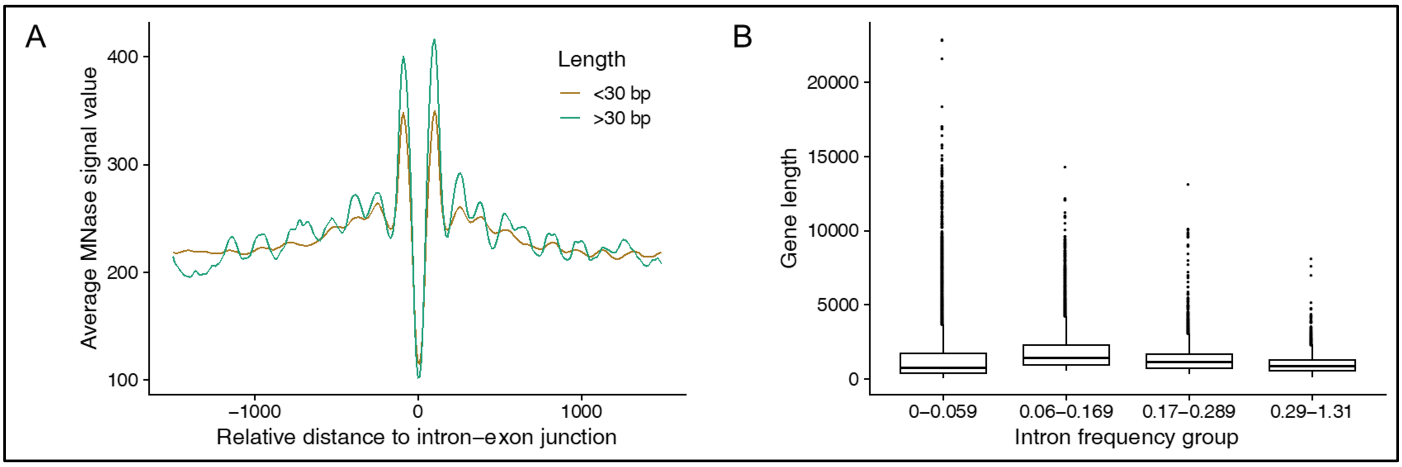
Supplementary Fig. S4 A) An illustration of the gene orientation-based grouping. SS = bidirectional, SE = unidirectional B) Distribution of the Pearson's correlation coefficient of neighboring genes expression from different serotypes/temperatures. To calculate the correlation of neighboring genes expression,15 RNA seq samples from our previous publication (Cheaib et al, DNA Res 2015) including different serotypes and temperatures, were analyzed. C) Pearson's correlation coefficient of neighboring genes expression for genes with different configurations. D) Length of intergenic regions is shown for genes with the same configurations as in B. Y-axis is zoomed in and outliers are not shown. The P-values shown are based on a two-tailed Wilcoxon test.

Supplementary Fig. S5



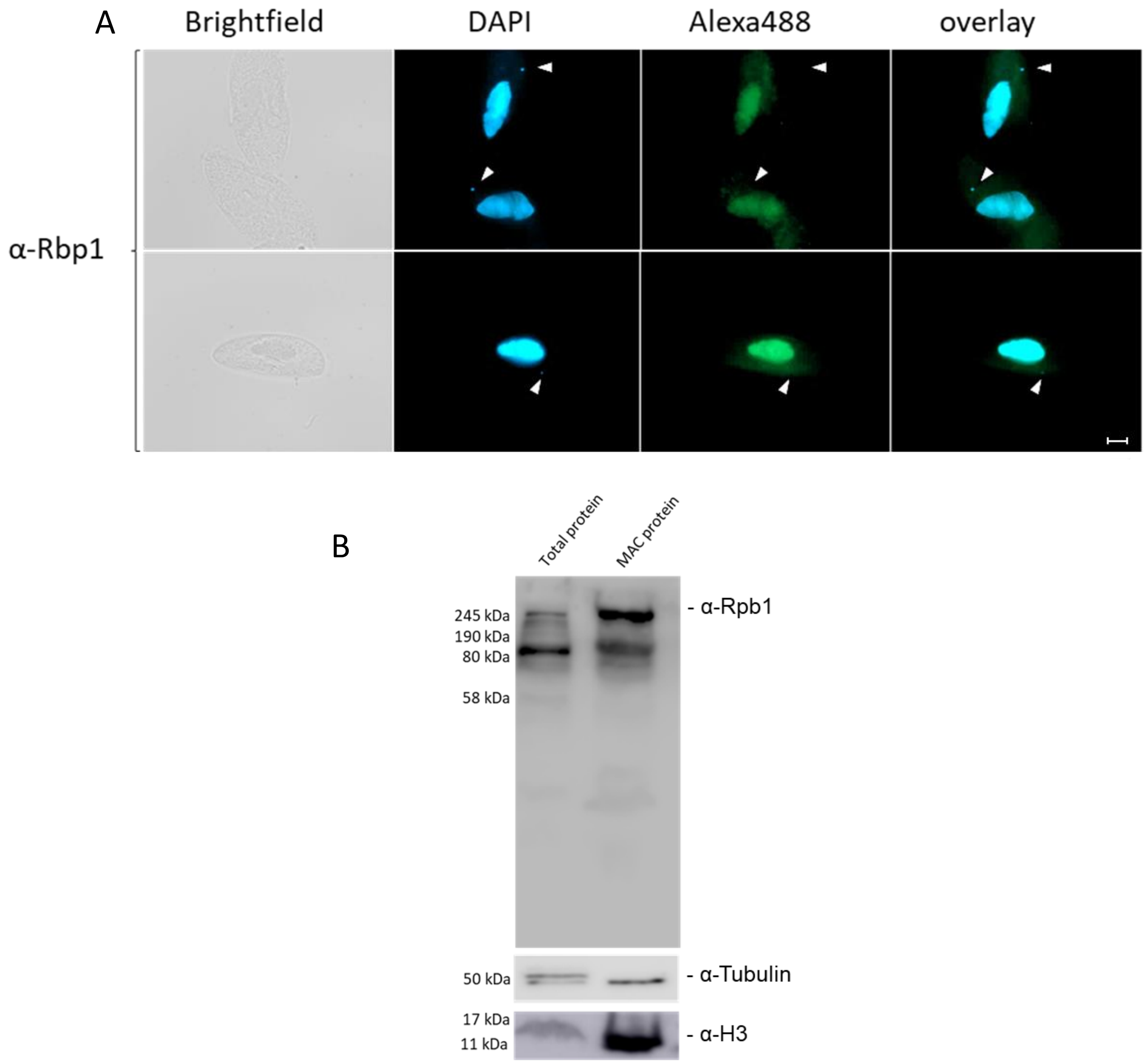
Supplementary Fig. S5 Profile plot of the nucleosome distribution at the TSS and TTS, stratified according to gene length groups (Short: 50-765 bp; Medium: 765-1470, Long: >1470).

Supplementary Fig. S6



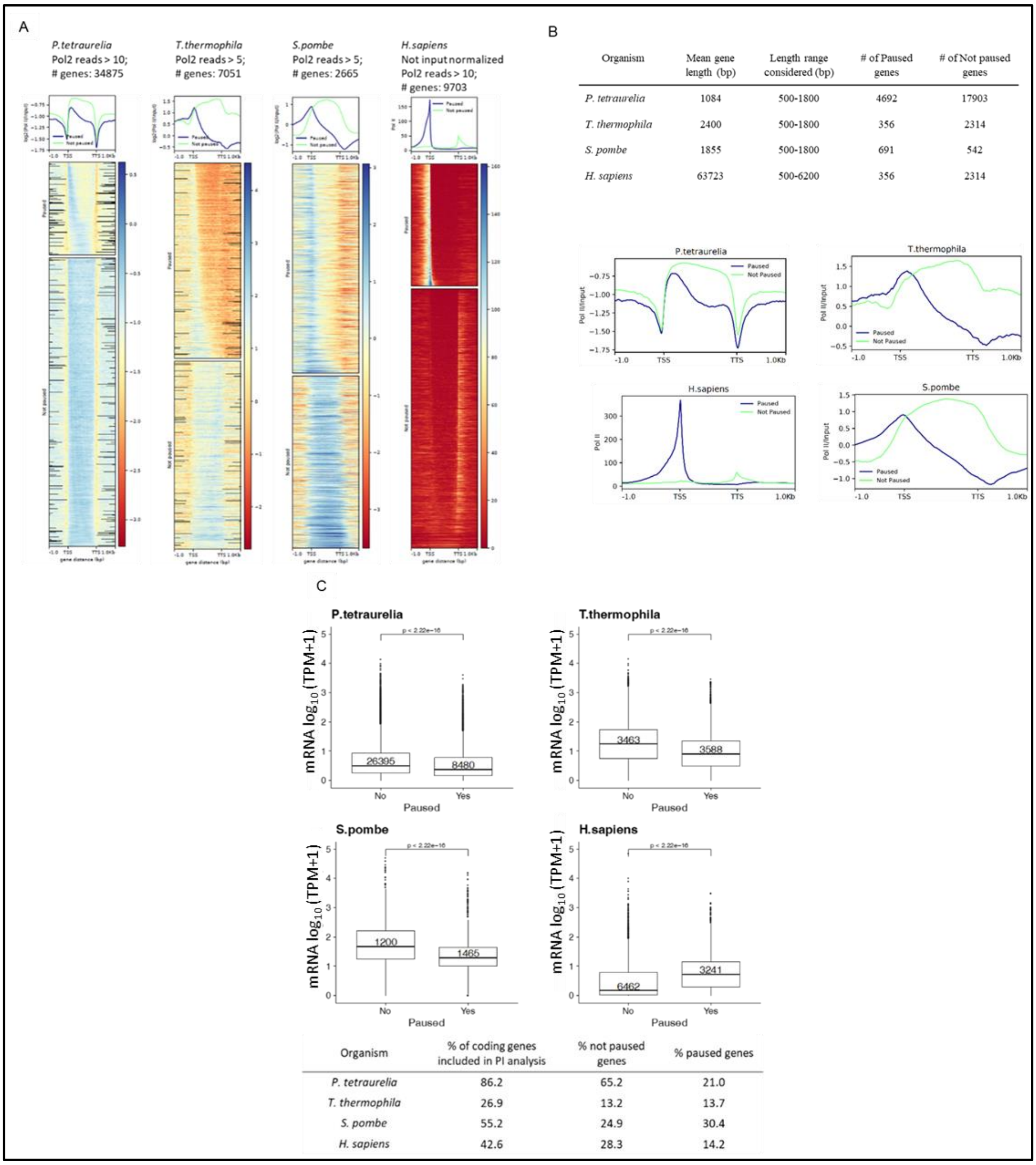
Supplementary Fig. S6 A) Profile plots of nucleosome distribution around the intron-exon junctions, categorized based on intron length groups. B) Box plot showing the distribution of gene length across different intron frequency groups (as shown in Fig.4). A Krustal-Wallis test showed that the expression distribution of all intron frequency groups is statistically significant (P < 2.2 x 10^{-16}).

Supplementary Fig. S7



Supplementary Fig. S7 Localization of Polymerase II by immunofluorescent staining and western blots. A) Vegetative Paramecium cells were stained by indirect immunofluorescent staining using custom antibody directed against Rpb1 as the largest subunit of Polymerase II. Primary antibody was labeled with Alexa488-conjugated secondary antibody (green) and nuclei were stained with DAPI (blue). Arrowheads point at micronuclei. Representative overlays of Z-Stacks are shown. Scale bar is 10 µm. B) Western blots using custom antibody against Rpb1 were performed as previously described (Klöppel et al., 2009). Protein lysate from whole cells (total protein) and protein from fractions of enriched macronuclei (MAC protein) were blotted and the membrane was decorated with antibodies against Rpb1 (200 kDa), α-Tubulin (50 kDa) and Histone H3 (15 kDa) as loading control.

Supplementary Fig. S8

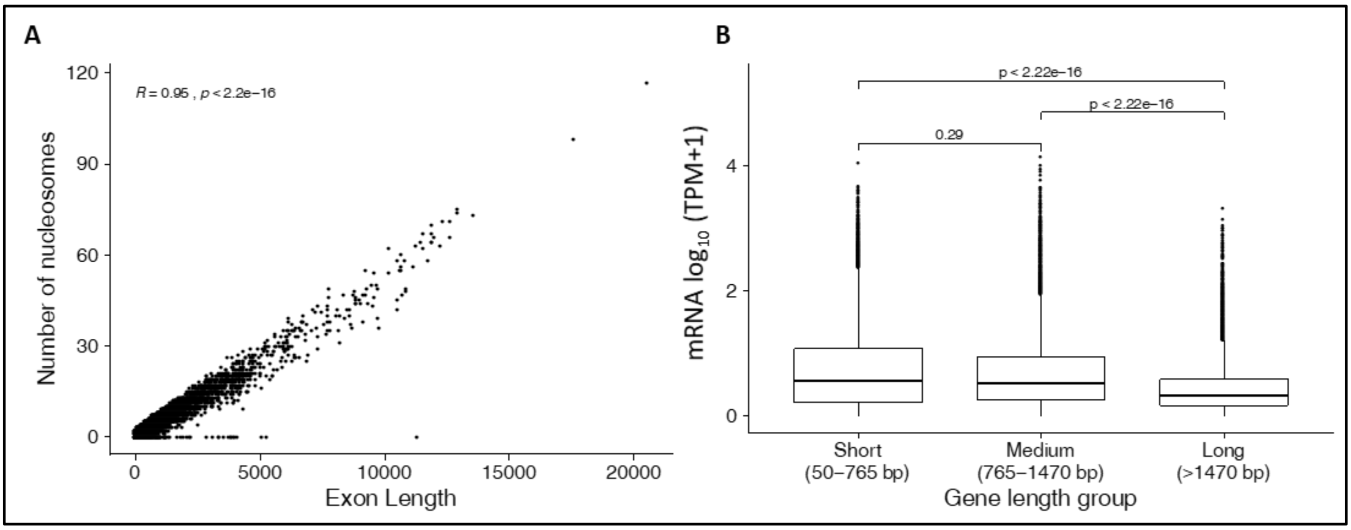


Organism	Mean gene length (bp)	Length range considered (bp)	# of Paused genes	# of Not paused genes
P. tetraurelia	1084	500-1800	4692	17903
T. thermophila	2400	500-1800	356	2314
S. pombe	1855	500-1800	691	542
H. sapiens	63723	500-6200	356	2314

Organism	% of coding genes included in PI analysis	% not paused genes	% paused genes
P. tetraurelia	86.2	65.2	21.0
T. thermophila	26.9	13.2	13.7
S. pombe	55.2	24.9	30.4
H. sapiens	42.6	28.3	14.2

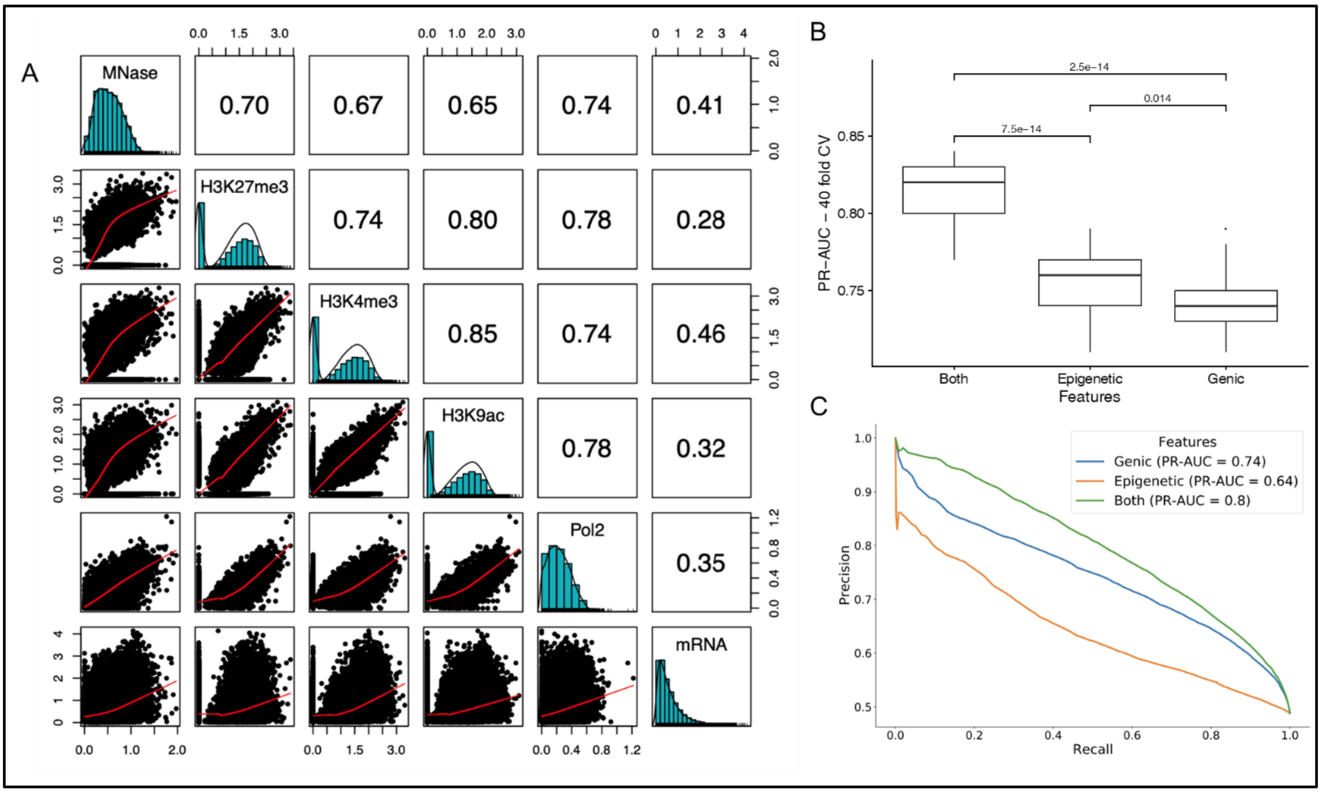
Supplementary Fig. S8 A) A heat map and profile of the Pol II enzyme in different organisms (see Supplementary Table 1: Comparative Pol II analysis sheet). The number of genes used in the plots are shown above each plot, after applying filters on the number of reads to avoid bias from silent genes. The genes are categorized as paused, if they had a pausing index more then 1.5 (see Methods). B) Pol II profiles for subsets of genes with approximately the same defined gene length in all organisms. Due to the mean gene length in *H.sapiens*, the gene length cut off had to be adjusted. Profiles are shown for paused and not paused genes as in Figure 6D. C) A box plot of the gene expression stratified according to their paused status in different organisms is shown. Table below shows, how many genes are included in the pausing index analysis.

Supplementary Fig. S9



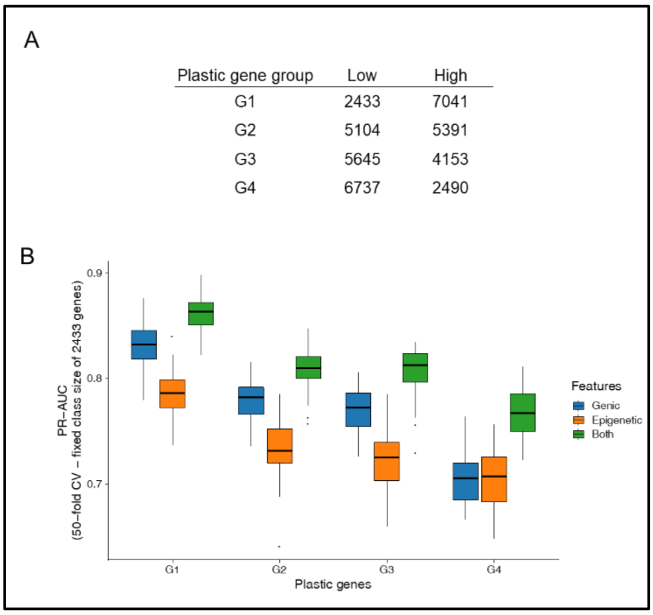
Supplementary Fig. S9 A) Scatter plot shows the increasing number of nucleosomes (y-axis) proportional to the increasing exon length (x-axis). Pearson's correlation coefficient (and its statistical significance value) between the exon length and number of nucleosomes, is embedded in the plot. B) Gene expression stratified according to different gene length groups.

Supplementary Fig. S10



Supplementary Fig. S10 A) The distribution plot of each epigenetic mark and mRNA are shown along the diagonal. The Pearson's correlation coefficients (above the diagonal) are shown for the respective variables mentioned along the x- and y-axis of each box. The y-axis of scatter plots (below the diagonal) belongs to the variable mentioned along the horizontal line of that plot. mRNA is measured in TPM units. All values shown are $\log_{10}$ transformed with a pseudocount of 1. B) A box plot showing the distribution of PR-AUC values of all 40-fold CV (y-axis) for random forests model using different feature subsets. C) A precision recall curve showing the average precision and recall values from a 40-fold cross validated (CV) random forests algorithm employing different feature sets (colors) to classify gene expression as high or low using the length normalized epigenetic signals from TSS+300 bp window, as opposed to whole gene body signals in the main Figure 7.

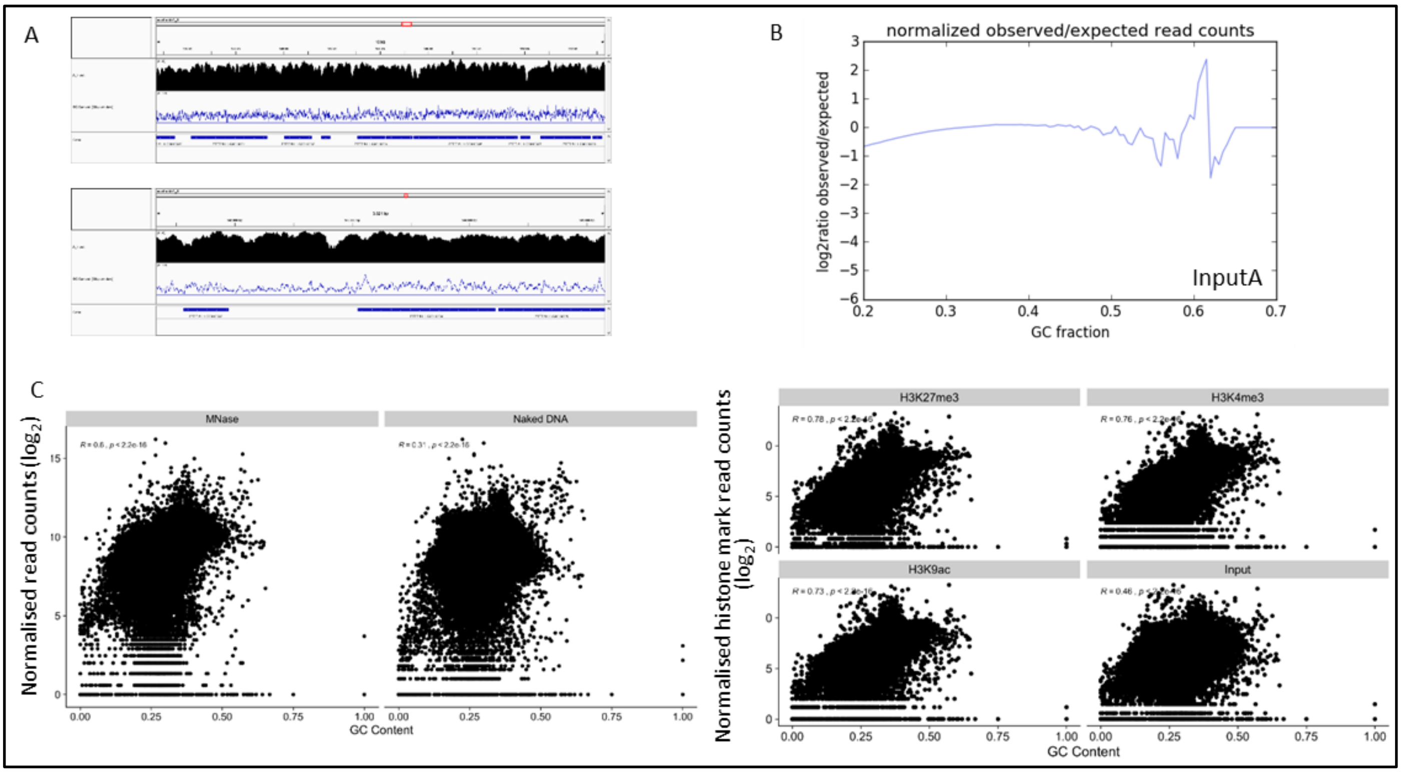
Supplementary Fig. S11



Plastic gene group	Low	High
G1	2433	7041
G2	5104	5391
G3	5645	4153
G4	6737	2490

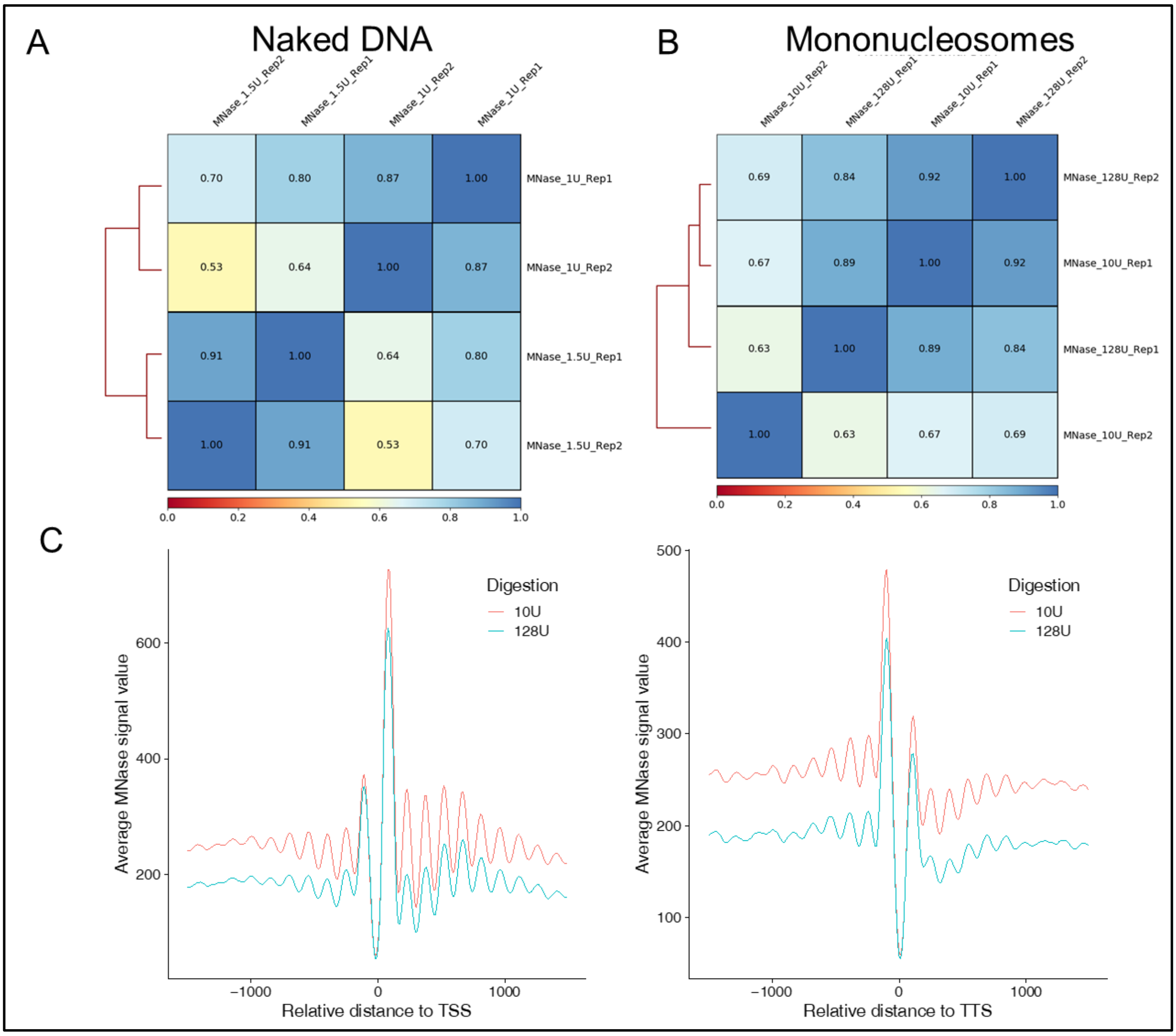
Supplementary Fig. S11 A) The number of high and low expressed genes in each plastic gene group is shown. B) Box plot showing the distribution of a 50-fold cross-validation based PR-AUC of random forests models with different feature subsets. The number of genes in each plastic gene group were randomly subsampled to have 2433 genes in high and low expressed category. We chose 2433, as it is the lowest number of genes among the high/low expressed plastic genes.

Supplementary Fig. S12



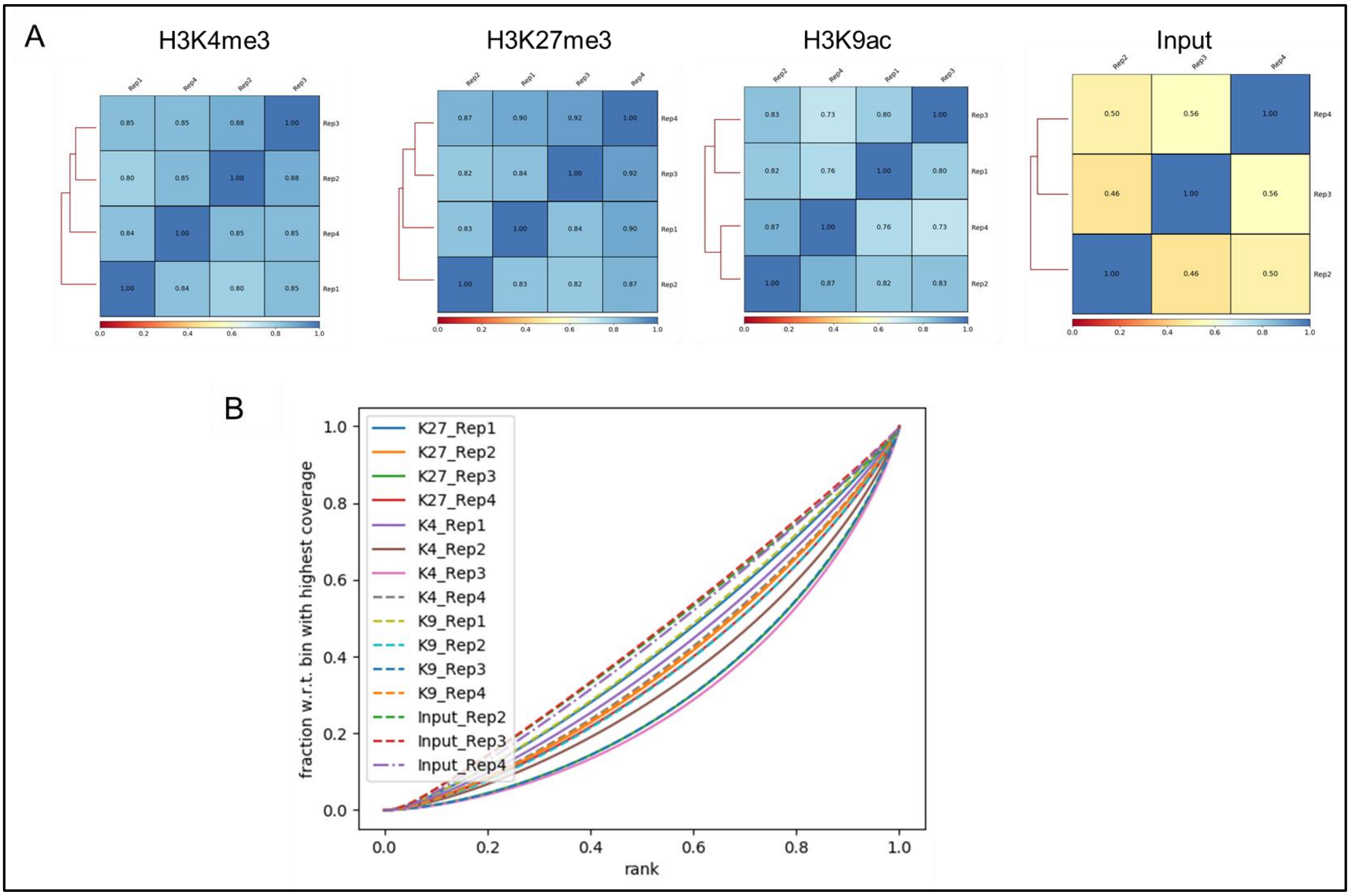
Supplementary Fig. S12 A) Snapshots from Genome browser showing coverage track of ChIP input along scaffold51_1(black) and GC content of the underlying DNA sequence in 20 bp windows (blue line). Genes are shown in dark blue in the third row. The top panel shows the view in a 18 kb window while the lower panel shows the zoomed view in a 3 kb window including an intergenic region. B) Correlation of reads from ChIP experiments to GC content. After binning the genome in 147 bp bins, the reads in each bin were counted, normalized and correlated against the GC content. B) The expected GC profile by counting the number of DNA fragments of a fixed size (by default is 300bp) per GC fraction is compared to the observed profile which is generated by counting the number of reads per GC fraction (the lower panel). C) Scatter plots of the GC content (x-axis) vs bin-length normalized read counts (y-axis; $\log_2$) of raw MNase, naked DNA (left), H3K27me3, H3K4me3, H3K9ac, and Input (right) measured in 147 bp bins of the genome. The Pearson's correlation coefficient values are mentioned on the top left corner of the plots.

Supplementary Fig. S13



Supplementary Fig. S13 Pearson's correlation coefficient of read counts between naked DNA (A) and MNase-seq (B) replicates of mononucleosomes obtained by different digestion units. C) Profile plot of nucleosome distribution at the transcription start site (TSS), and the transcription termination site (TTS) respectively.

Supplementary Fig. S14



Supplementary Fig. S14 A) Subsampled Pearson's correlation of read counts between histone mark and input replicates. B) Fingerprint plot, generated with deeptools, showing the quality of replicates.

Histone ChIP-seq				
Sample/# of reads	**Rep1**	**Rep2**	**Rep3**	**Rep4**
H3K27me3	18.552.130	35.134.413	16.927.429	11.967.159
H3K4me3	15.977.332	7.576.189	5.751.651	25,682,022 *
H3K9ac	10.750.224	8.635.166	9.492.335	20,214,521 *
Input	NA	11.837.934	11.024.554	28,621,594 *
* Downsampled to 10,000,000				

Pol II ChIP-seq	
Sample/# of reads	**Rep1**
Pol II	14.559.173

Mononucleosomal DNA		
Sample/# of reads	**Rep1**	**Rep2**
MNase_10U	28.992.372	71.437.557
MNase_128U	32.262.035	22.393.596
Naked DNA		
Sample/# of reads	**Rep1**	**Rep2**
MNase_1U (=Input 10U normalization)	33.211.870	29.357.641
MNase_1.5U (=Input 128U normalization)	40.011.353	16.110.036

Organism	Sequence type	Nucleosomes	MNase Units used	# of replicates	Strain/Type	GEO accession number	SRR accession number
Schizosaccharomyces pombe	naked DNA			1	WT	GSE140920	SRR10528270
Schizosaccharomyces pombe	MNase	80 mono: 20 di	45	1	WT	GSE52170	SRR1821723
Schizosaccharomyces pombe	MNase			1	WT	GSE141676	SRR10611800
Drosophila melanogaster	naked DNA			2	S2cells	GSE69177	SRR2038260, SRR2038261
Drosophila melanogaster	Mnase - High 1N	Mono		2	S2cells	GSE69177	SRR2038276, SRR2038277
Homo sapiens	naked DNA			1	HeLa cells	GSE100401	SRR5749438
Homo sapiens	MNase - 1000U	Mono, di, and tri	100-2000	2	HeLa cells	GSE100401	SRR5749432, SRR5749433
Tetrahymena thermophila	MNase heavy digest			1	CU428	GSM2055773	SRX1590945
Tetrahymena thermophila	MNase light digest			1	CU428	GSM2055775	SRX1590947
Tetrahymena thermophila	naked DNA			1	SB210	GSE64061	SRR2041674
Tetrahymena thermophila	MNase light digest	80 mono: 20 di		1	SB210	GSE64061	SRR2041661

Organism	Sequence type	Description	# of replicates	Strain/Type	GEO accession number
Tetrahymena thermophila	Pol II	pull down of c-terminal HA-tagged RPB7	2	WT	GSE77583
Tetrahymena thermophila	Input DNA		2	WT	GSE77583
Tetrahymena thermophila	mRNA		3	WT (CU428)	GSE130336
Schizosaccharomyces pombe	Pol II	pull down with anti-Pol II Ab (Abcam ab5408) directed against CTD repeat YSPTSPS (phospho ser-5)	2	WT (4H8)	GSE115636
Schizosaccharomyces pombe	Input DNA		2	WT	GSE115636
Schizosaccharomyces pombe	mRNA		2	WT (Pem2)	GSE115636
Homo sapiens	Pol II	pull down with anti-Pol II Ab (Covance # MMS-128P) directed against CTD repeat YSPTSPS (phospho ser-5)	1	Resting	GSE98368
Homo sapiens	mRNA		2	Untreated	GSE98368

Supplementary Tab. S1:

Above: Read numbers of ChIP and MNAse libraries including information about downsampling

Below: References for published datasets used for comparison to other species