Uniform transitions of the general RNA polymerase II transcription complex

Andreas Mayer,[*1] Michael Lidschreiber,[*1] Matthias Siebert,[*1] Kristin Leike,[1] Johannes Söding,[1] and Patrick Cramer[1]

[1]*Gene Center Munich and Department of Biochemistry, Center for Integrated Protein Science CIPSM, Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 Munich, Germany.*

[*]*These authors contributed equally to this work.*

Supplementary Figures

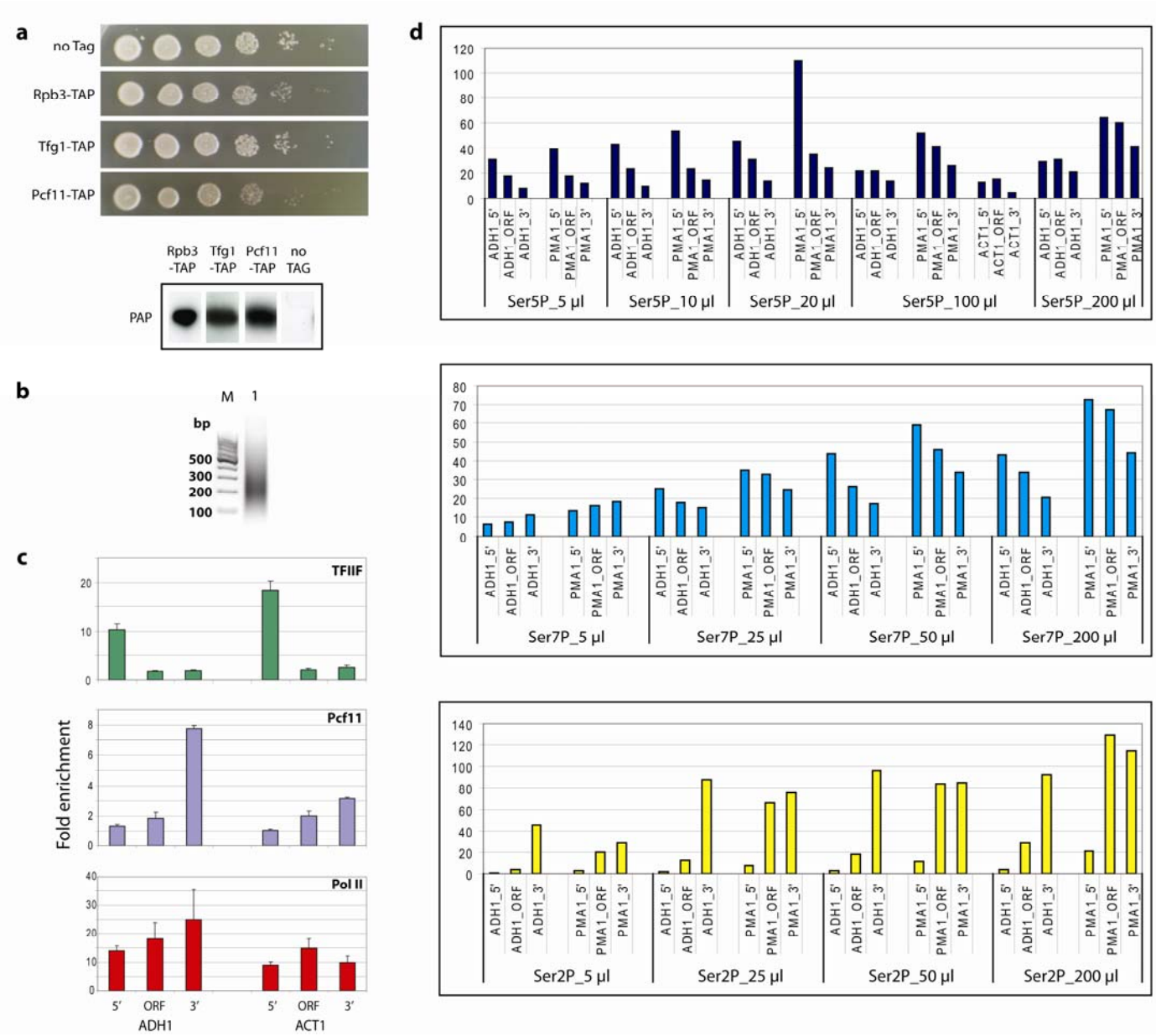
Supplementary Tables

Supplementary Methods

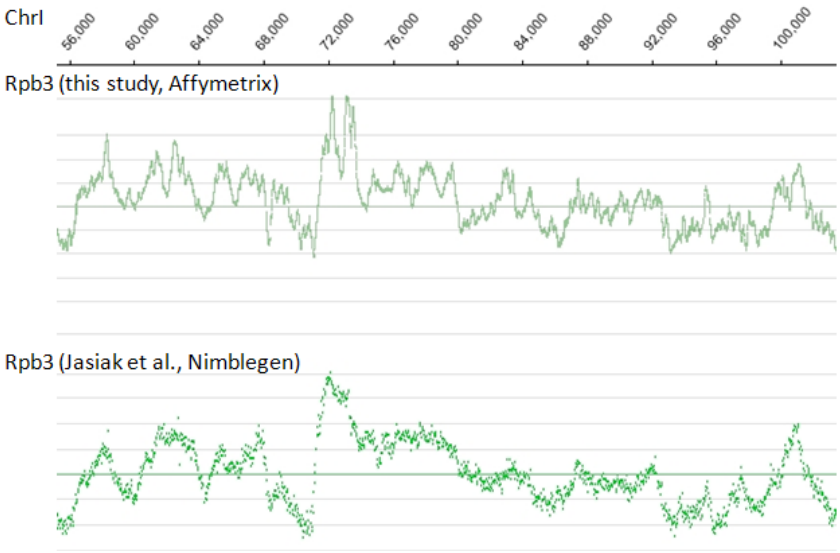
Supplementary References

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1903

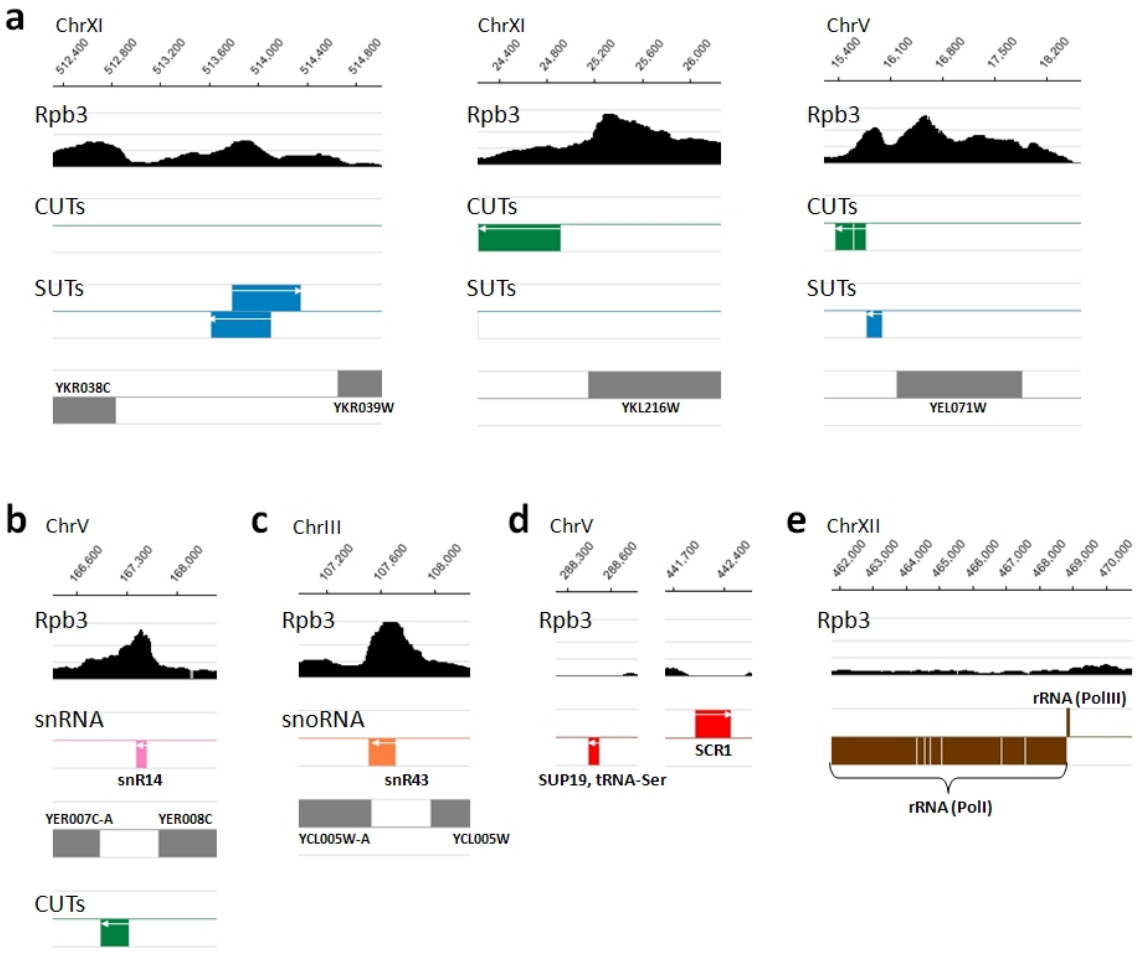
Supplementary Figures



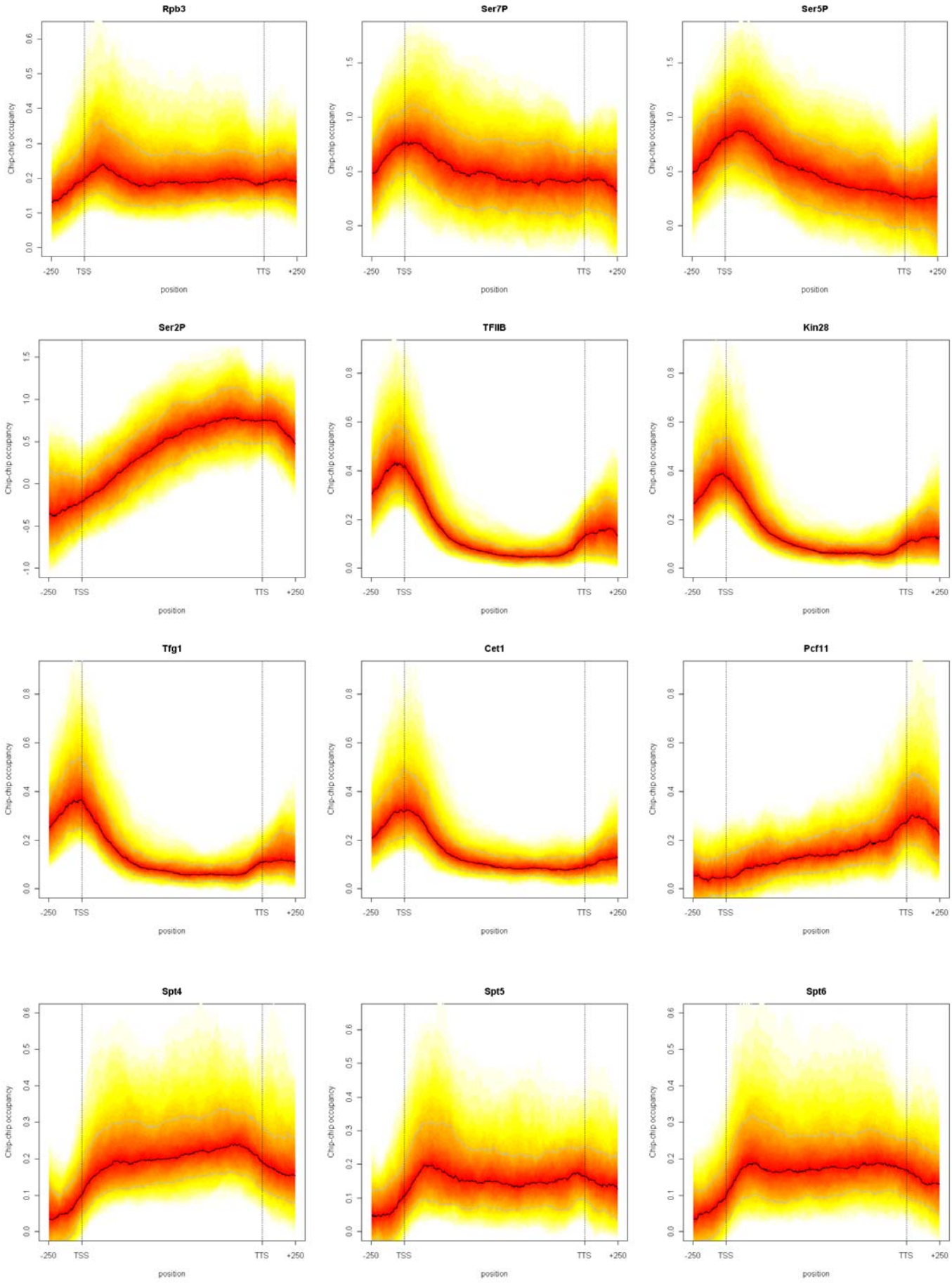
Supplementary Figure 1 ChIP-chip quality controls. (a) Validation of TAP-strains by (i) western blot analysis using the PAP-antibody (Sigma P1291) directed against the TAP-Tag, and by (ii) growing serial dilutions of yeast strains on YPD plates at 30°C. This is exemplarily shown for yeast strains having a C-terminally TAP-tagged version of Rpb3, Tfg1 or Pcf11. (b) Measurement of the average chromatin fragment size. Chromatin samples prepared by sonication in a Bioruptor (Diagenode) were decrosslinked, treated with Proteinase K, purified, treated with RNase A/T1 mix, purified again and resolved in 1.5% agarose gel. Lane M shows the molecular weight marker and lane 1 the chromatin sample after fragmentation. Sizes of the molecular weight marker are indicated. (c) Occupancy of TFIIF (Tfg1), Pcf11 and Pol II (Rpb3) at the promoter (5'), coding (ORF, open reading frame) and terminator region (3') of the two housekeeping genes *ADH1* and *ACT1*. Notice the different scales of the y-axes. Error bars show SD from at least two independent experiments of biological replicates. (d) ChIP experiments for Pol II phospho-isoforms. Ser5P (3E8 antibody), Ser7P (24E12) and Ser2P (3E10). Occupancies are shown for three different regions of the two housekeeping genes ADH1 and PMA1. The different amounts of antibodies tested in the ChIP experiments are given below. The fold enrichments over an open reading frame (ORF)-free heterochromatic region on chromosome V are indicated on the y-axes and are calculated as described in the methods part.

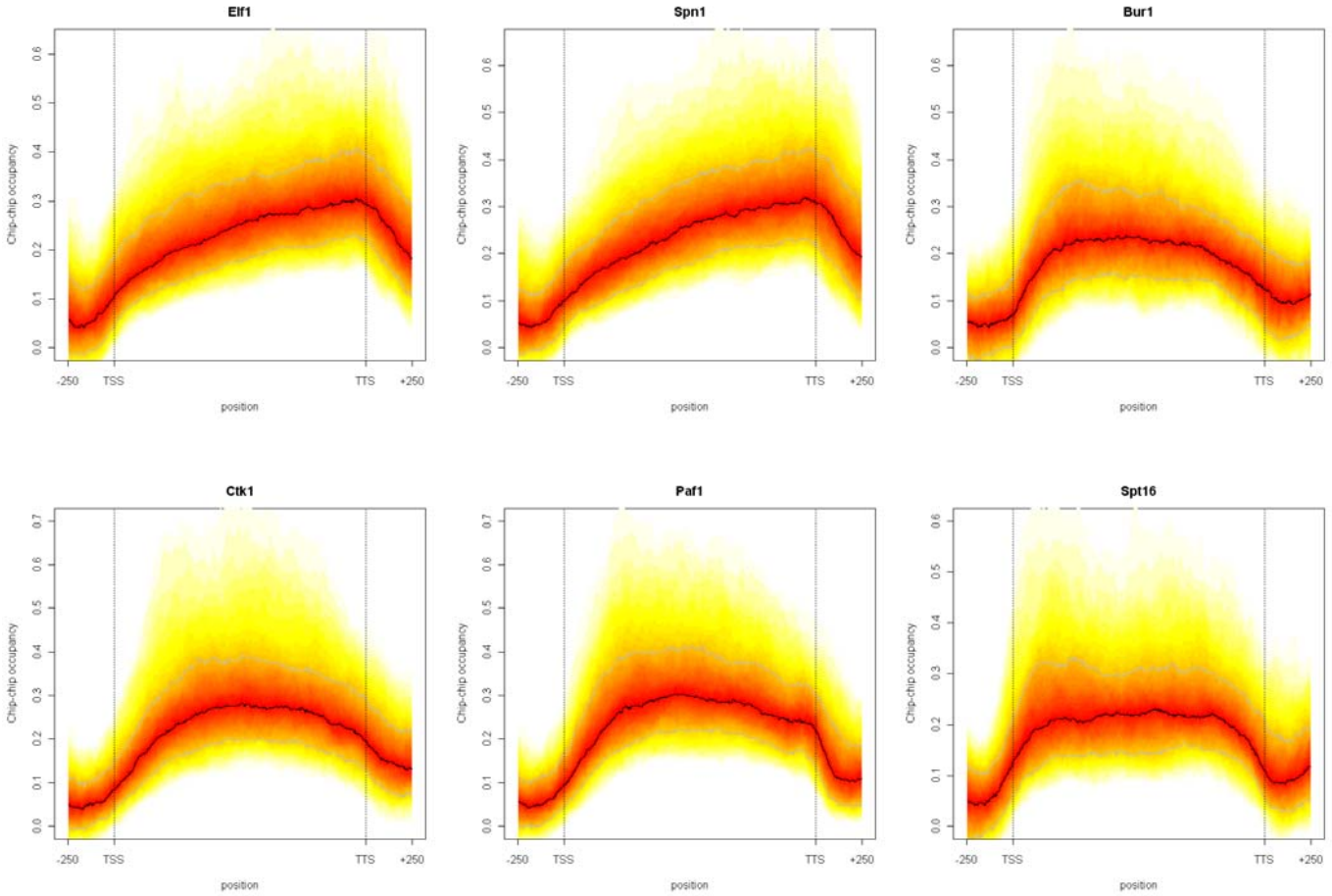


Supplementary Figure 2 Comparison of Rpb3 ChIP enrichment profile on chromosome 1 from this study with the profile from Jasiak *et al.*[23]. The profiles demonstrate the high reproducibility of the ChIP data across different array platforms (Nimblegen two-color tiling arrays versus Affymetrix tiling arrays) and ChIP protocols. The upper trace was smoothed by using a running median approach with a window size of 150bp. Note the better resolution of the present study.

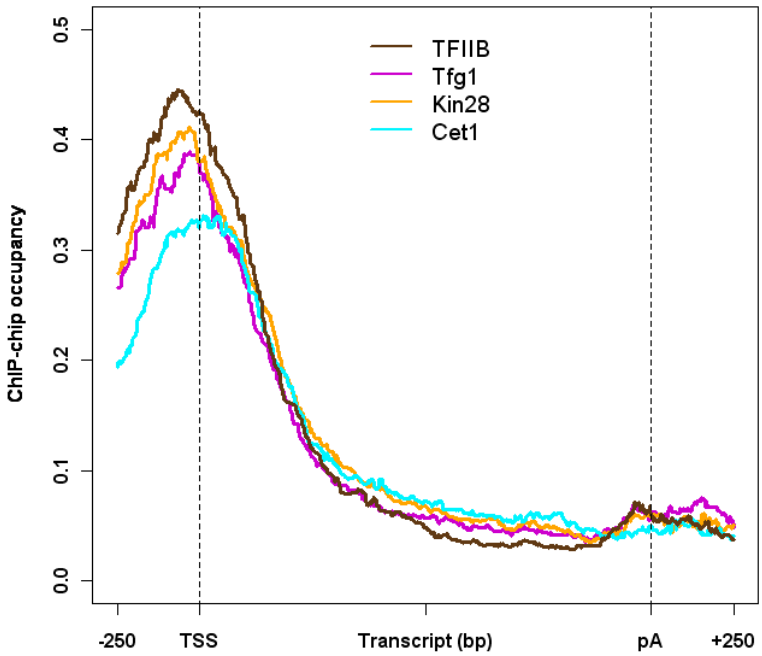


Supplementary Figure 3 Pol II (Rpb3) occupancy at CUTs, SUTs, sn(o)RNAs, and Pol I/III genes. Rpb3 occupancy on selected (a) CUTs (green) and SUTs (blue), (b) snRNAs (pink), (c) snoRNAs (orange), (d) Pol III (red) and (e) Pol I genes (brown). The genomic position with the corresponding chromosome numbers are shown above each panel. The location of transcripts[7] are indicated as grey boxes and are given below where present. The transcripts are shown for the Watson (upper line) and Crick strand (lower line). Each bar in the occupancy profile represents the normalized signal from a 150 bp sliding window.

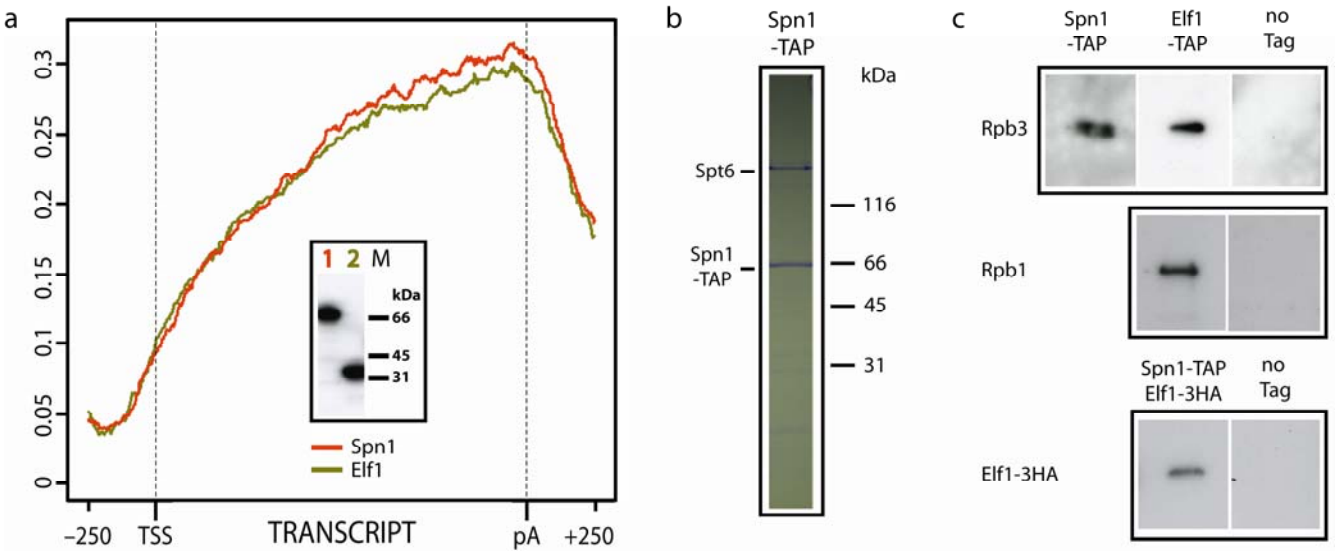

Rpb3
Ser7P
Ser5P
Ser2P
TFIIB
Kin28
Tfg1
Cet1
Pcf11
Spt4
Spt5
Spt6
Chip-chip occupancy
position
-250
TSS
TTS
+250



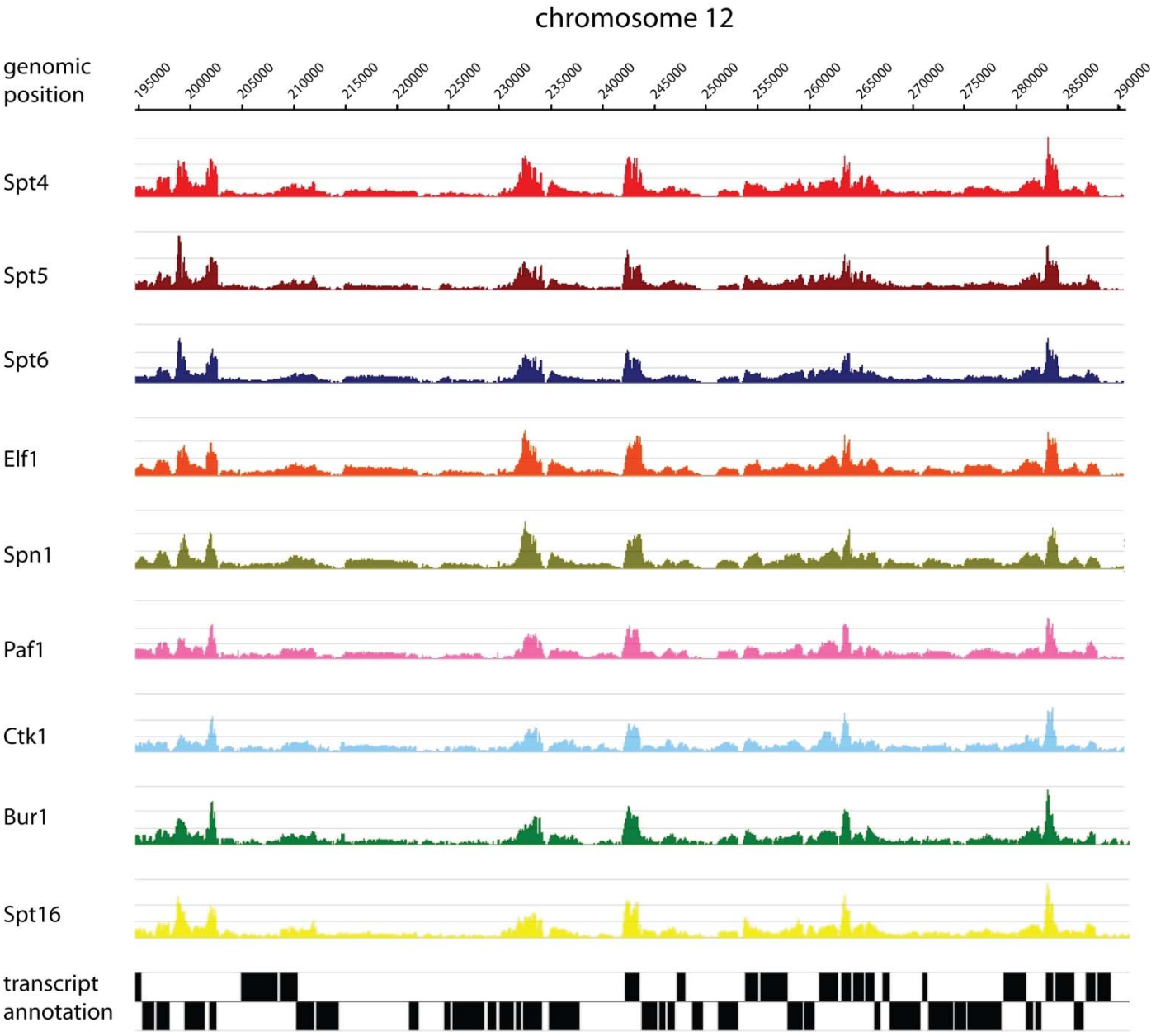
Supplementary Figure 4 Gene-averaged quantile profiles. Profile quantiles for length class M illustrate the variance of the occupancy profiles over genes. Black: median, grey: first and third quartile, color gradient: quantiles ranging from 5 to 95 %. Ser7/5/2P occupancies correspond to S7/5/2P $\log_2$ enrichments. See Fig. 1 and Supplementary Methods for details. TTS = transcript termination site (corresponds to the polyadenylation (pA) site).



Supplementary Figure 5 Gene-averaged profiles restricted to convergently transcribed genes. Gene-averaged median profiles for initiation (TFIIB, -F, -H) and 5′-capping (Cet1) factors (genes from length class M as shown in Fig. 1d) showing only convergently transcribed genes (99 out of 339 genes) in order to diminish spill-over effects.

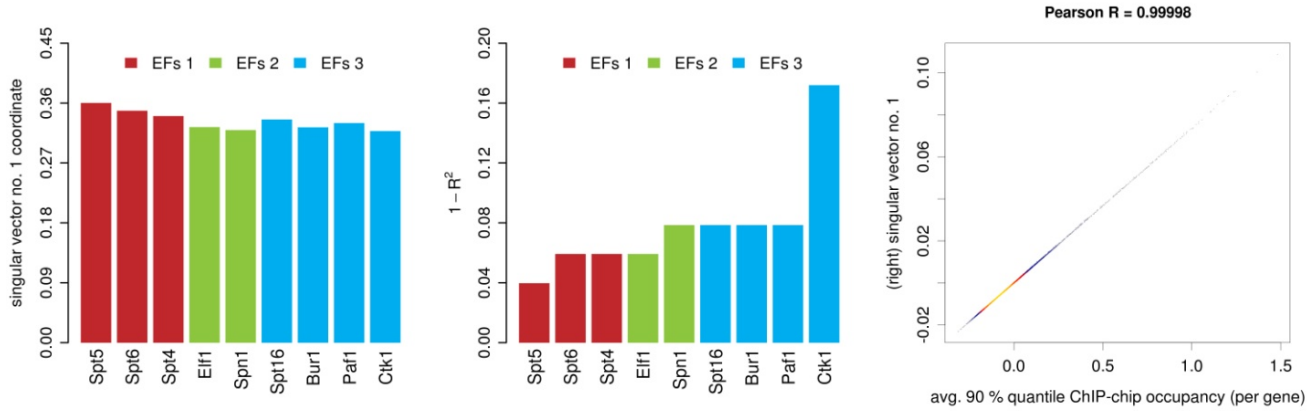


Supplementary Figure 6 Elf1 copurifies with Spn1 and Pol II. (a) Gene-averaged occupancy profiles of Elf1 (orange) and Spn1 (green). Western blotting analysis was performed to validate the TAP strains (inset). The results for the Spn1- and Elf1-TAP strains are indicated on lane 1 and 2, respectively. The molecular weight marker is given on lane M. (b) Spn1 TAP: the purified protein sample was analyzed by SDS-PAGE and Coomassie staining. Spt6 could be identified by mass spectrometry. The molecular weight marker (kDA) is given on the right. Spn1 copurifies with Spt6. (c) Elf1 and Spn1 TAPs: the protein samples were analyzed by SDS-PAGE and Western blotting. The TAP-strains as well as the control strain (with no TAP tag) are given above the panels. The proteins identified by Western blotting are indicated on the left. Elf1 copurifies with two Pol II subunits Rpb1 and Rpb3. Spn1 copurifies with Rpb3 and Elf1 bearing a C-terminally 3HA tag. For the latter experiment, a yeast strain was generated that expressed both, a C-terminally 3HA-tagged version of Elf1 as well as a C-terminally TAP-tagged version of Spn1. The ORF coding for the 3HA epitope tag together with the kanMX6 module originally amplified from the plasmid pFA6a-3HA-kanMX6[18] was inserted before the stop codon of *ELF1* in a Spn1-TAP background strain. Transformants were selected on Geneticin (G418) containing plates (0.4 g/L).

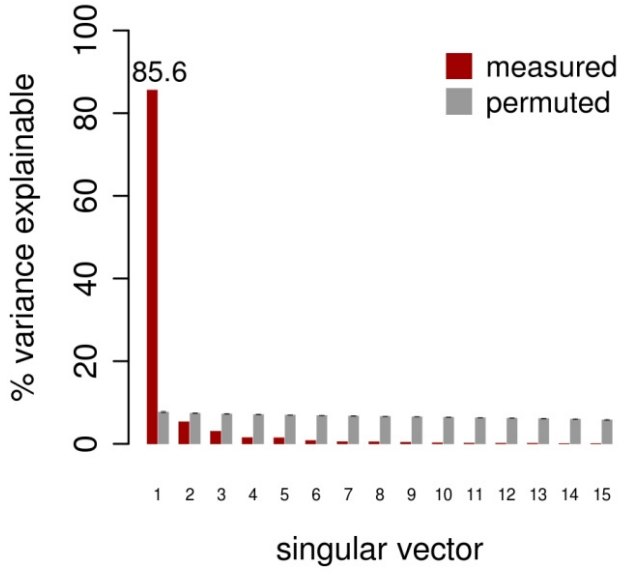


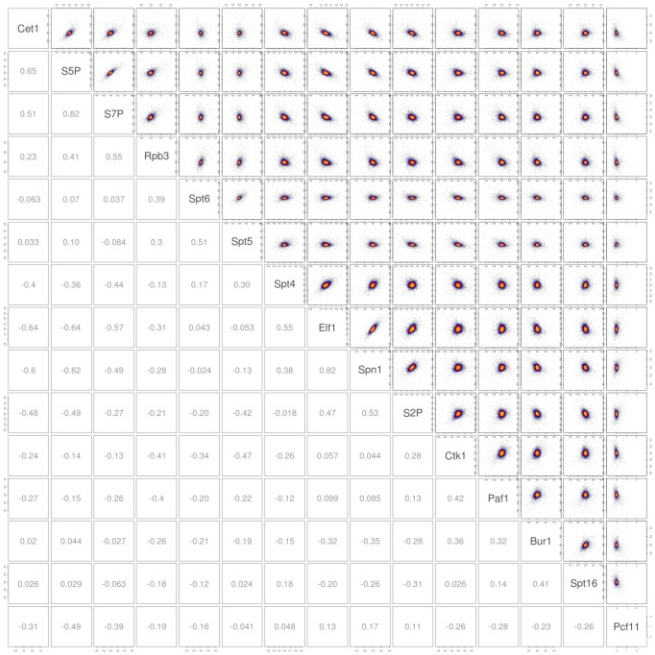
Supplementary Figure 7 Representative region on chromosome 12 showing occupancy profiles of elongation factors. Note the high degree of covariation among all factor occupancy profiles.

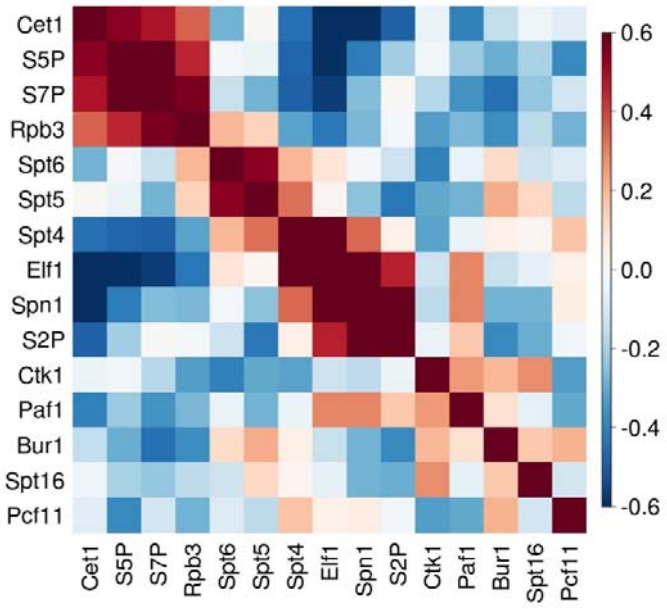
a


singular vector no. 1 coordinate
EFs 1
EFs 2
EFs 3
Spt5
Spt6
Spt4
Elf1
Spn1
Spt16
Bur1
Paf1
Ctk1
1 − R²
Pearson R = 0.99998
(right) singular vector no. 1
avg. 90 % quantile ChIP-chip occupancy (per gene)

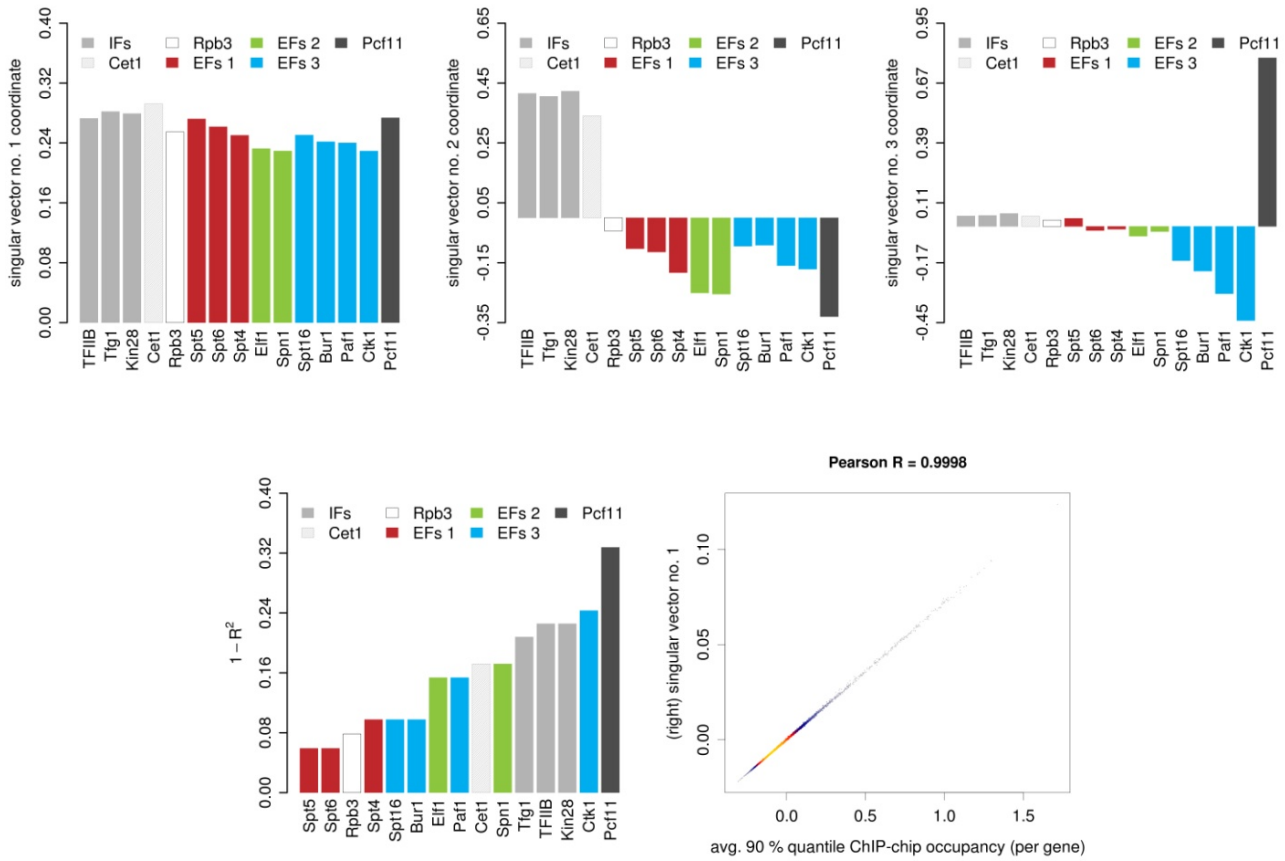
b


% variance explainable
85.6
measured
permuted
singular vector


Cet1
S5P
S7P
Rpb3
Spt6
Spt5
Spt4
Elf1
Spn1
S2P
Ctk1
Paf1
Bur1
Spt16
Pcf11

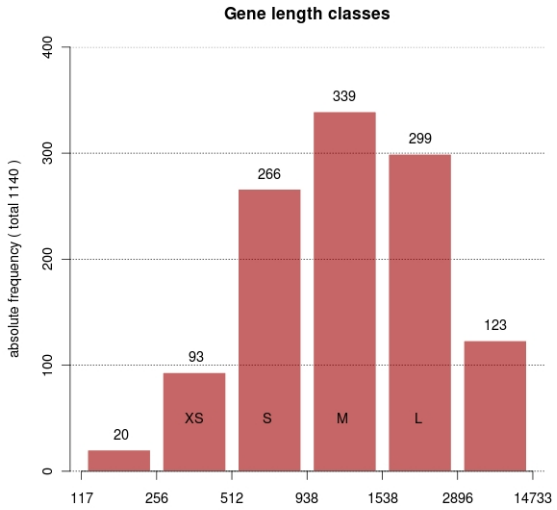

Cet1
S5P
S7P
Rpb3
Spt6
Spt5
Spt4
Elf1
Spn1
S2P
Ctk1
Paf1
Bur1
Spt16
Pcf11

c

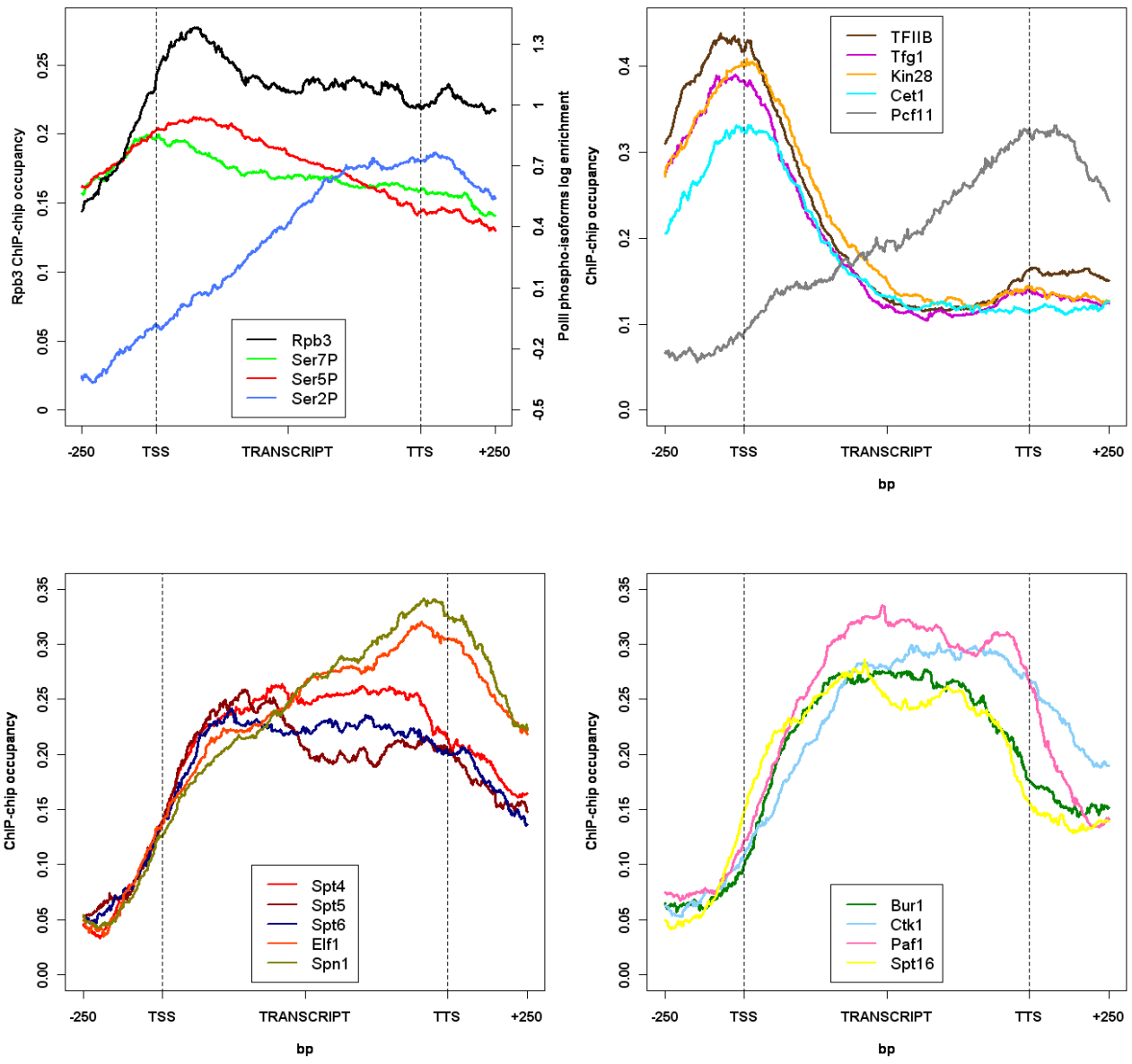


Supplementary Figure 8 Singular value decomposition (SVD). (a) SVD on nine elongation factors. Left: Coefficients of first left-singular vector for the SVD of the occupancy matrix for the nine elongation factors. The first eigenvector with its similar-sized coefficients describes the co-variation of all factors. Middle: One minus squared correlation coefficient for peak occupancy of factor with first right-singular vector. Right: scatter plot of average 90% quantile ChIP-chip occupancies versus the coefficients of right-singular vector for 4366 genes with defined TSS and pA shows a very high correlation. (b) SVD on twelve factors and three Pol II phospho-isoforms. Top left: percent variance explained by singular vectors in SVD (analogous to Fig. 2c) for Rpb3, Cet1, nine elongation factors, Pcf11, and the Pol II phospho-isoforms S2P, S5P, and S7P. Top right: residual correlations between the 15 factors and isoforms. (Upper right diagonal: scatter plots. Lower left diagonal: Pearson correlation coefficients.) Bottom: Matrix of residual correlations for SVD on 97 stringently filtered genes, analogous to Fig. 2d. (c) SVD on three initiation factors (TFIIB, Tfg1, Kin28), the capping factor Cet1, Rpb3, nine elongation factors, and the polyadenylation and cleavage factor Pcf11. Top row: first three left-singular vectors (u_{1f} ,u_{2f} ,u_{3f} ,f=1,..,15). The first eigenvector with its similar-sized coefficients describes the co-variation of all factors. Bottom left: One minus squared correlation coefficient for peak occupancy of factor with first right-singular vector. Spt5 peak occupancy is best described by the first term of the SVD, whereas Pcf11 is described worst. Bottom right: scatter plot of average 90% quantile ChIP-chip occupancies versus the coefficients of right-singular vector for 4366 genes with defined TSS and pA shows a very high correlation.

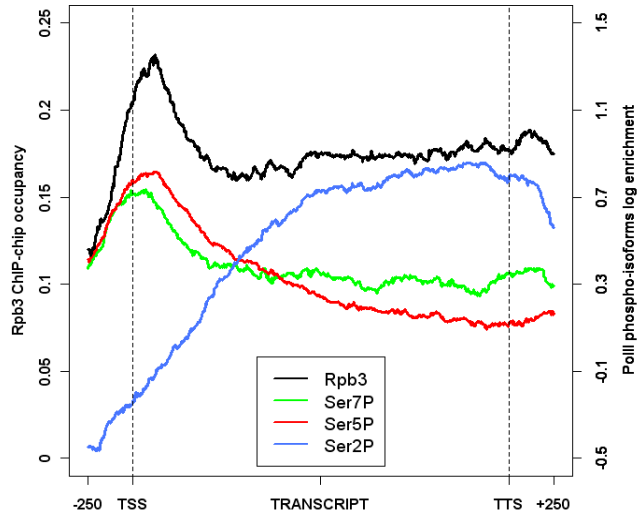
a

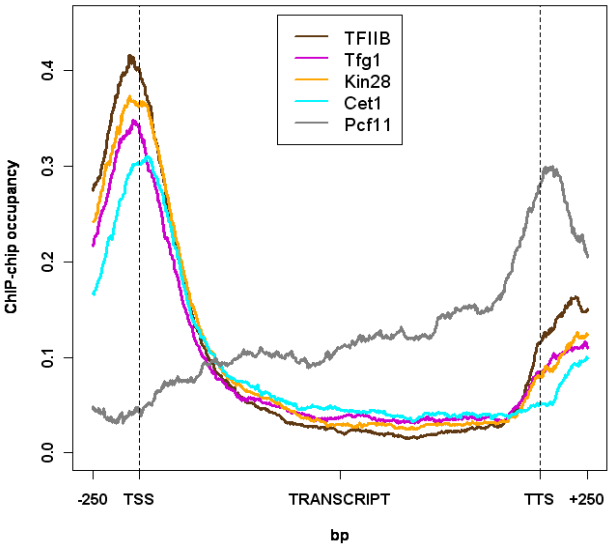

Gene length classes
absolute frequency (total 1140)
0
100
200
300
400
20
93
266
339
299
123
XS
S
M
L
117
256
512
938
1538
2896
14733

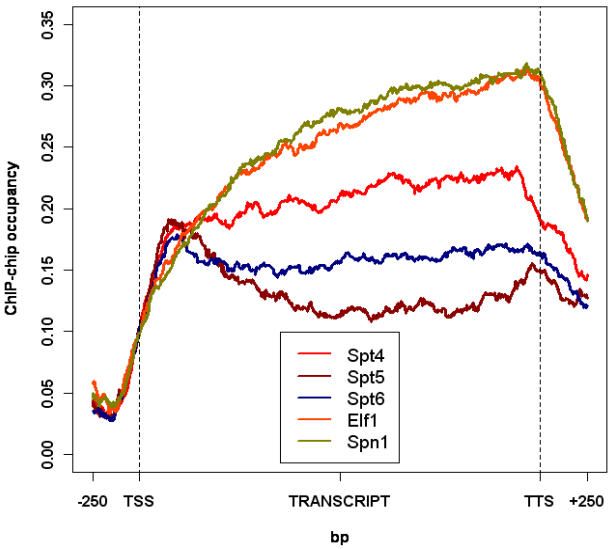
b

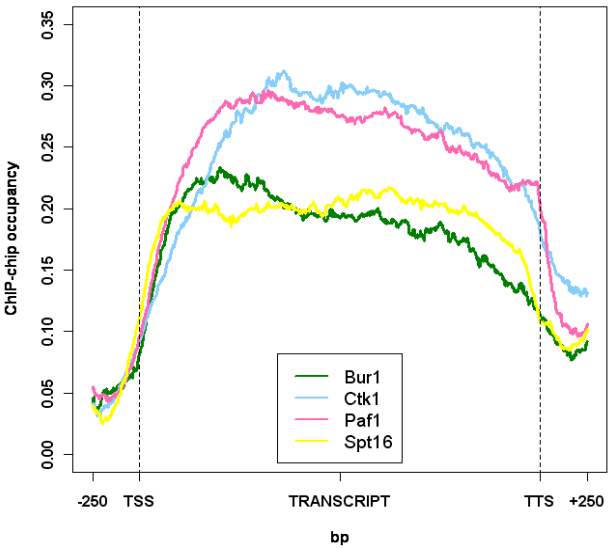

Rpb3 ChIP-chip occupancy
PolII phospho-isoforms log enrichment
Rpb3
Ser7P
Ser5P
Ser2P
-250
TSS
TRANSCRIPT
TTS
+250
ChIP-chip occupancy
TFIIB
Tfg1
Kin28
Cet1
Pcf11
bp
Spt4
Spt5
Spt6
Elf1
Spn1
Bur1
Ctk1
Paf1
Spt16

c

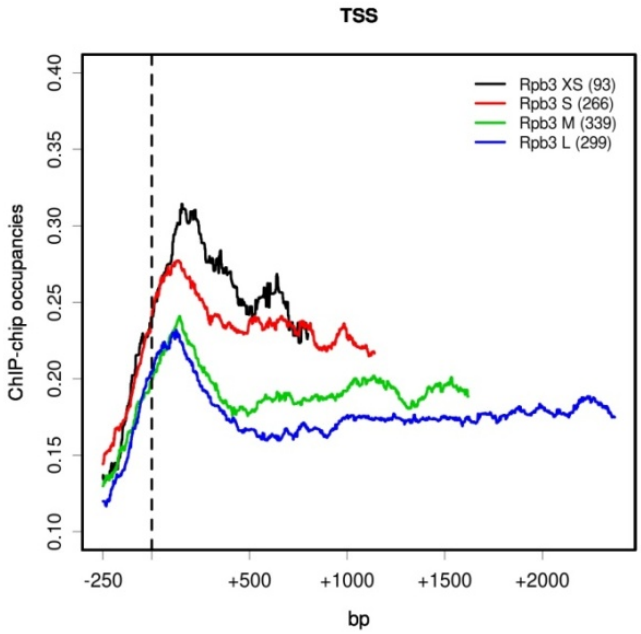

Rpb3 ChIP-chip occupancy
Polll phospho-isoforms log enrichment
Rpb3
Ser7P
Ser5P
Ser2P
-250
TSS
TRANSCRIPT
TTS
+250

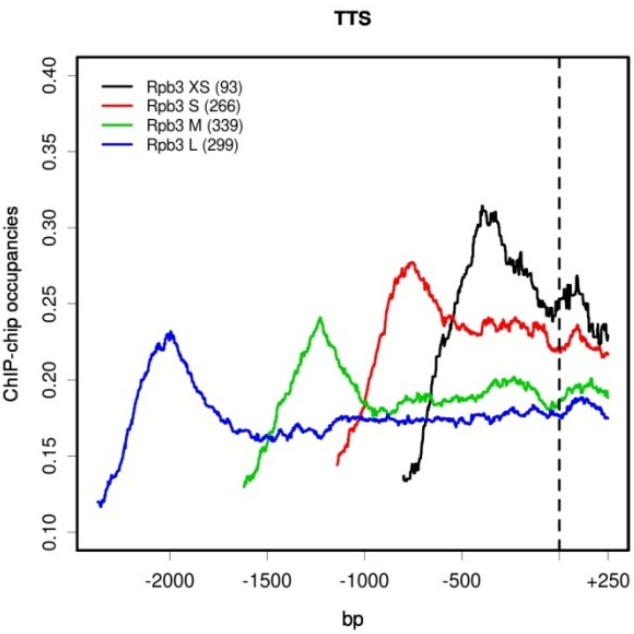

ChIP-chip occupancy
TFIIB
Tfg1
Kin28
Cet1
Pcf11
-250
TSS
TRANSCRIPT
TTS
+250
bp

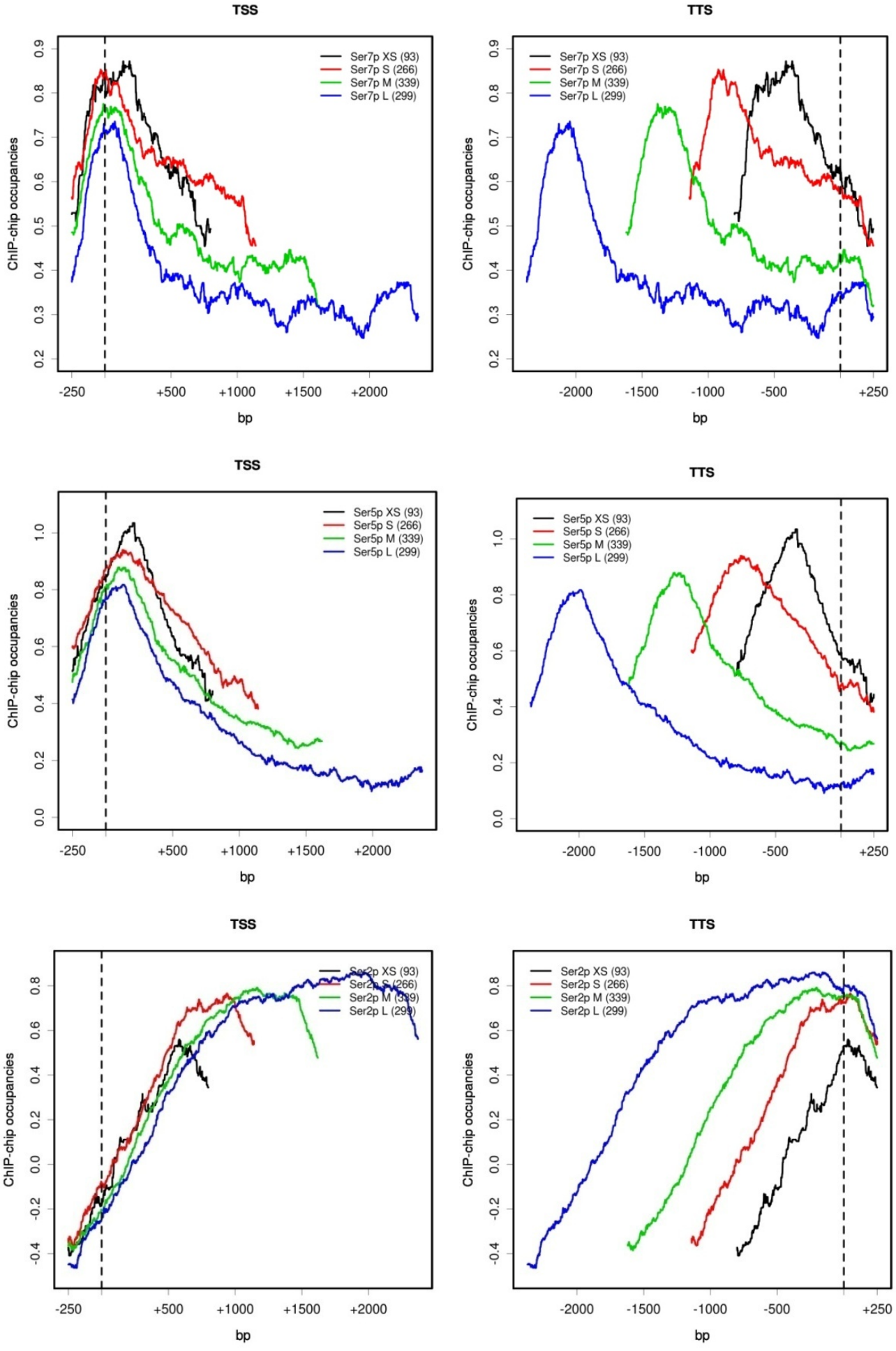

ChIP-chip occupancy
Spt4
Spt5
Spt6
Elf1
Spn1
-250
TSS
TRANSCRIPT
TTS
+250
bp

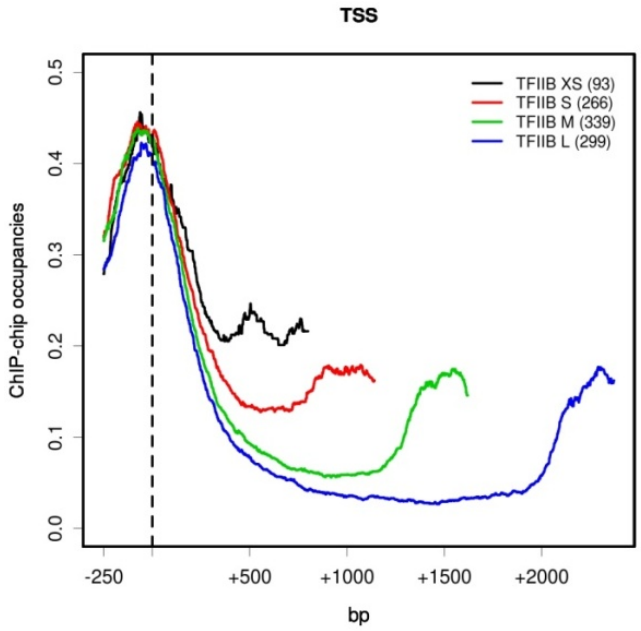

ChIP-chip occupancy
Bur1
Ctk1
Paf1
Spt16
-250
TSS
TRANSCRIPT
TTS
+250
bp

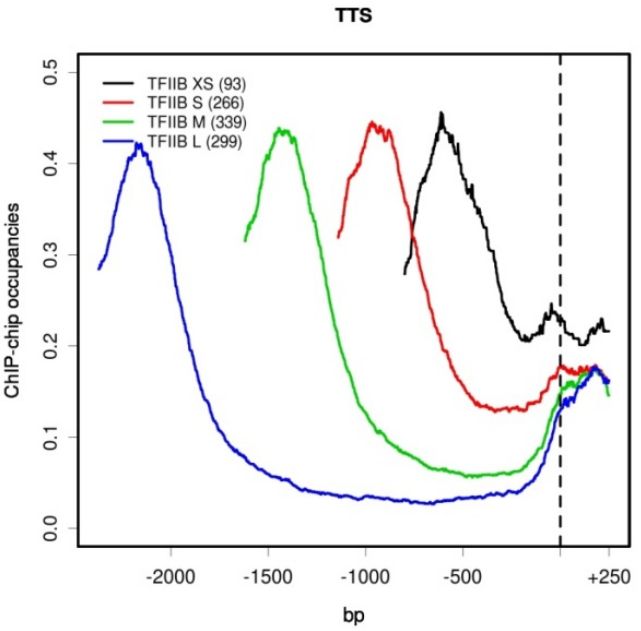
d

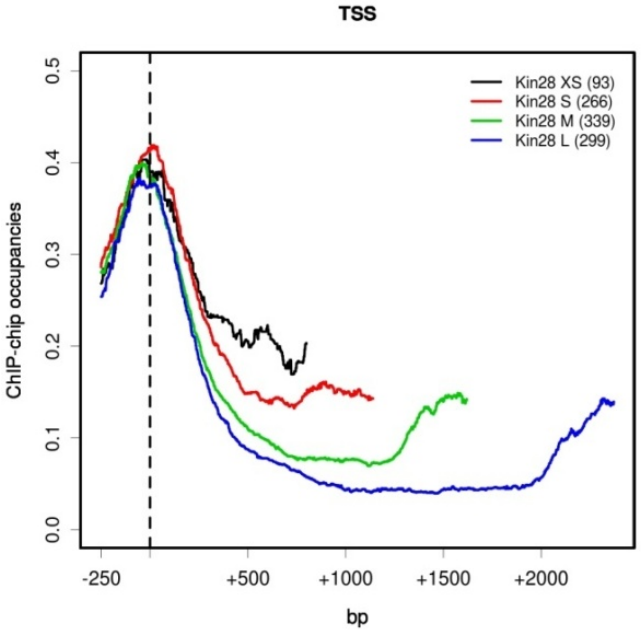

TSS
ChIP-chip occupancies
Rpb3 XS (93)
Rpb3 S (266)
Rpb3 M (339)
Rpb3 L (299)
-250
+500
+1000
+1500
+2000
bp

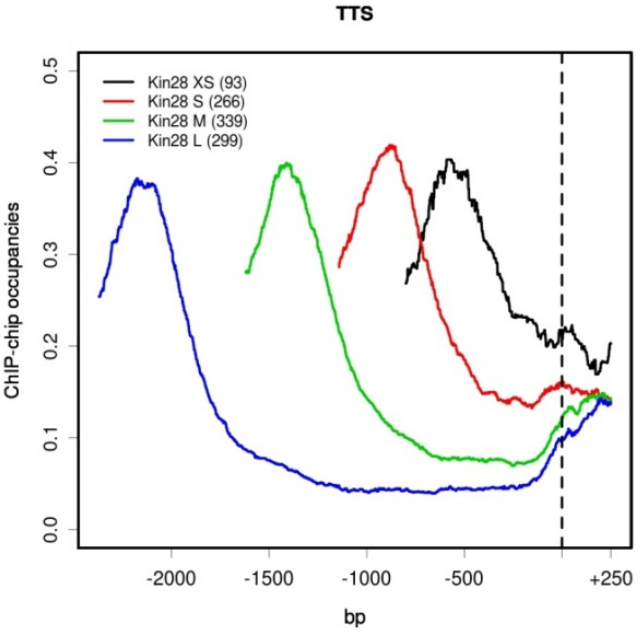

TTS
ChIP-chip occupancies
Rpb3 XS (93)
Rpb3 S (266)
Rpb3 M (339)
Rpb3 L (299)
-2000
-1500
-1000
-500
+250
bp

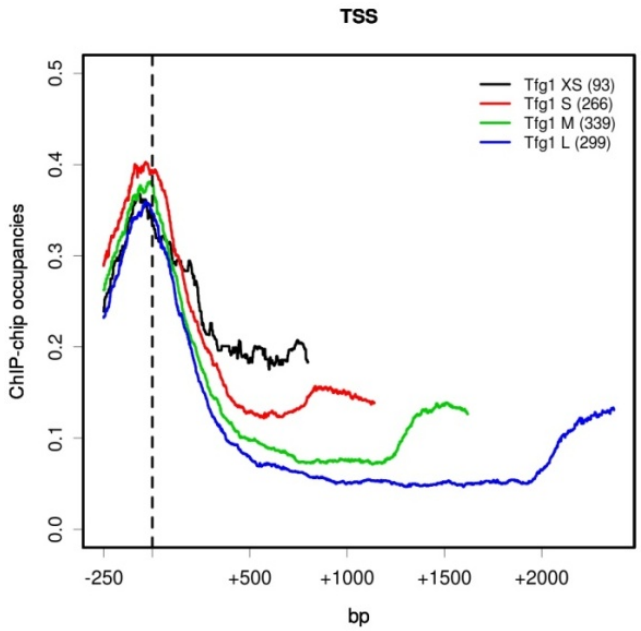

TSS
TTS
Ser7p XS (93)
Ser7p S (266)
Ser7p M (339)
Ser7p L (299)
Ser5p XS (93)
Ser5p S (266)
Ser5p M (339)
Ser5p L (299)
Ser2p XS (93)
Ser2p S (266)
Ser2p M (339)
Ser2p L (299)
ChIP-chip occupancies
bp
-250
+500
+1000
+1500
+2000
-2000
-1500
-1000
-500
+250

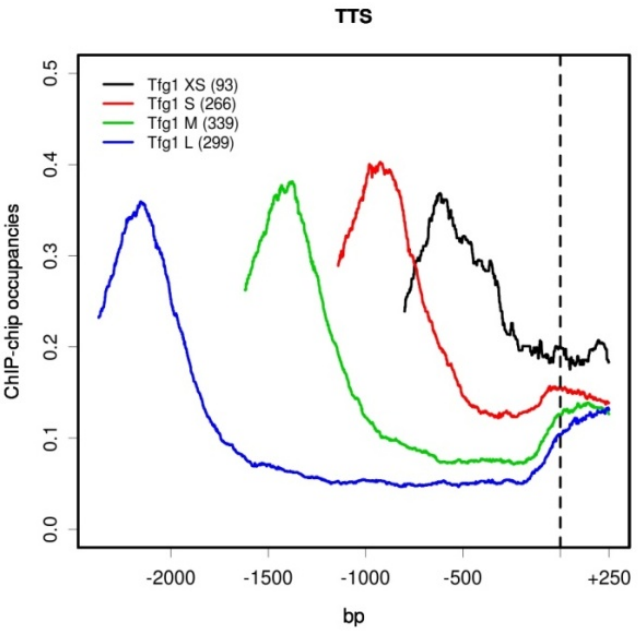

TSS
TFIIB XS (93)
TFIIB S (266)
TFIIB M (339)
TFIIB L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
0.5
-250
+500
+1000
+1500
+2000
bp

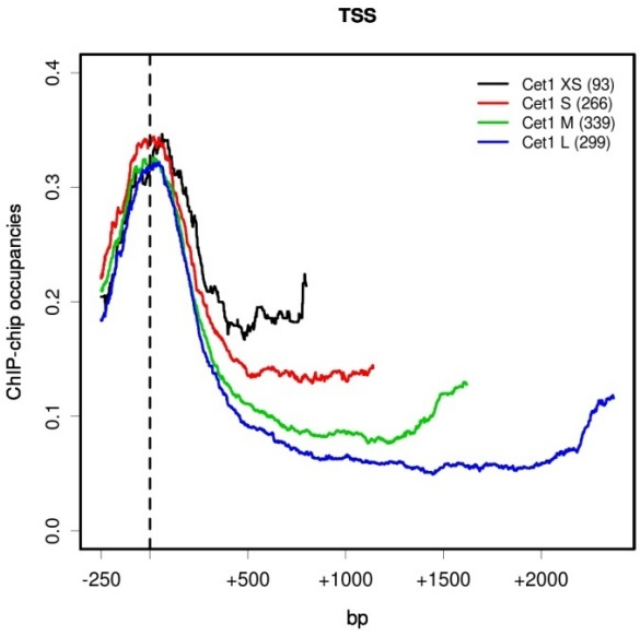

TTS
TFIIB XS (93)
TFIIB S (266)
TFIIB M (339)
TFIIB L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
0.5
-2000
-1500
-1000
-500
+250
bp

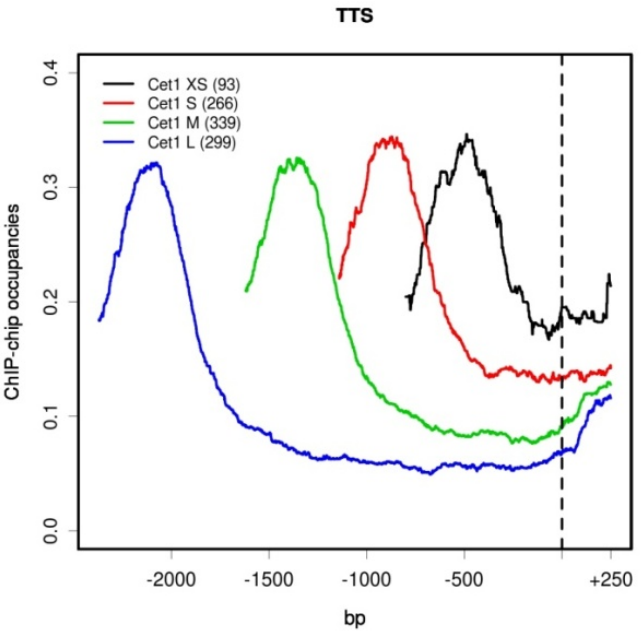

TSS
Kin28 XS (93)
Kin28 S (266)
Kin28 M (339)
Kin28 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
0.5
-250
+500
+1000
+1500
+2000
bp

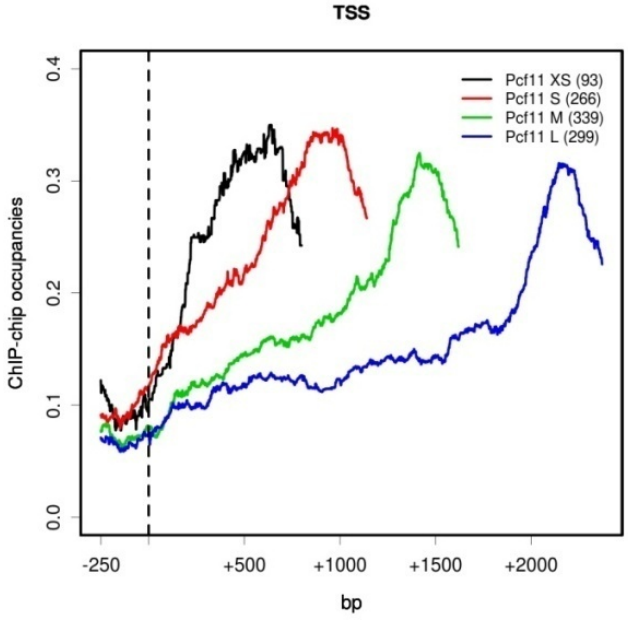

TTS
Kin28 XS (93)
Kin28 S (266)
Kin28 M (339)
Kin28 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
0.5
-2000
-1500
-1000
-500
+250
bp

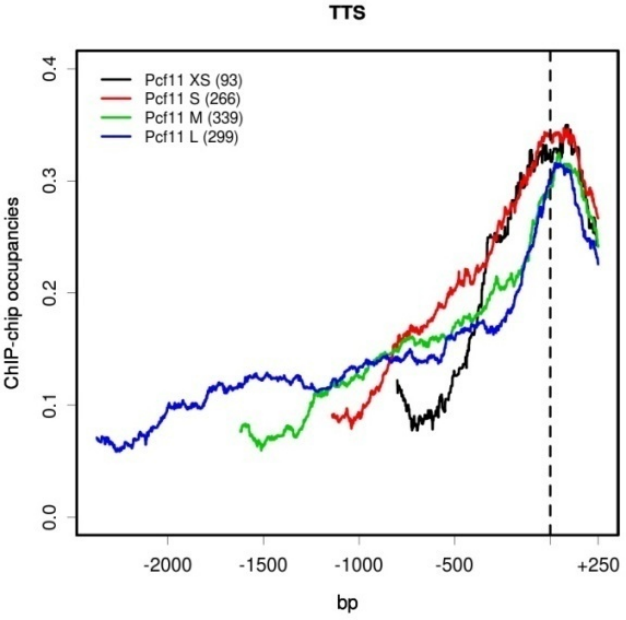

TSS
Tfg1 XS (93)
Tfg1 S (266)
Tfg1 M (339)
Tfg1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
0.5
-250
+500
+1000
+1500
+2000
bp

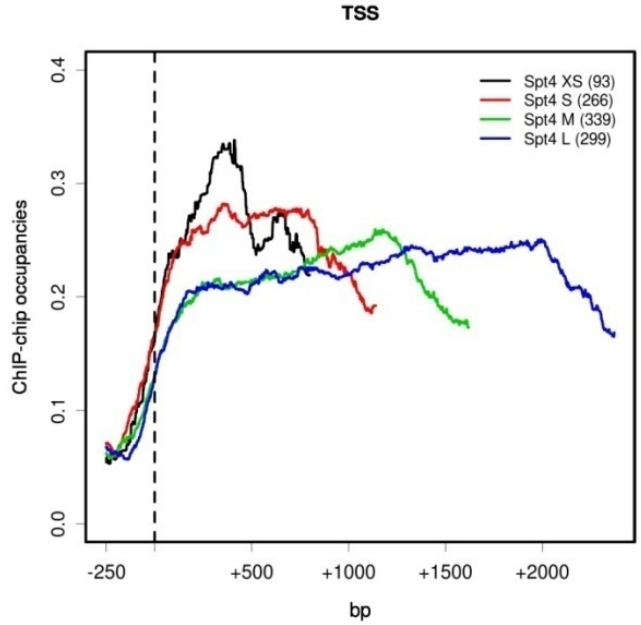

TTS
Tfg1 XS (93)
Tfg1 S (266)
Tfg1 M (339)
Tfg1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
0.5
-2000
-1500
-1000
-500
+250
bp

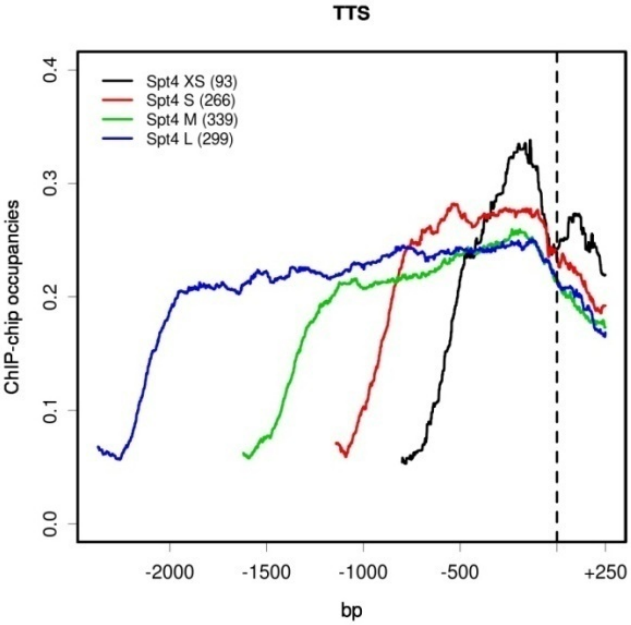

TSS
Cet1 XS (93)
Cet1 S (266)
Cet1 M (339)
Cet1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-250
+500
+1000
+1500
+2000
bp

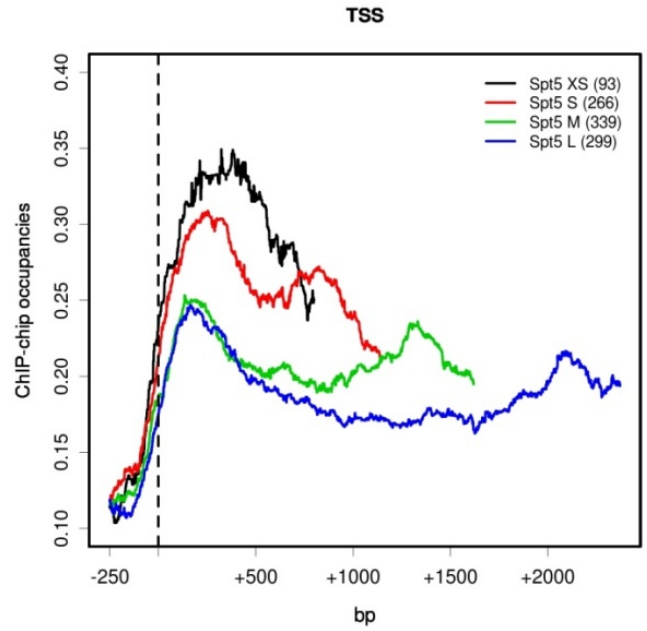

TTS
Cet1 XS (93)
Cet1 S (266)
Cet1 M (339)
Cet1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-2000
-1500
-1000
-500
+250
bp

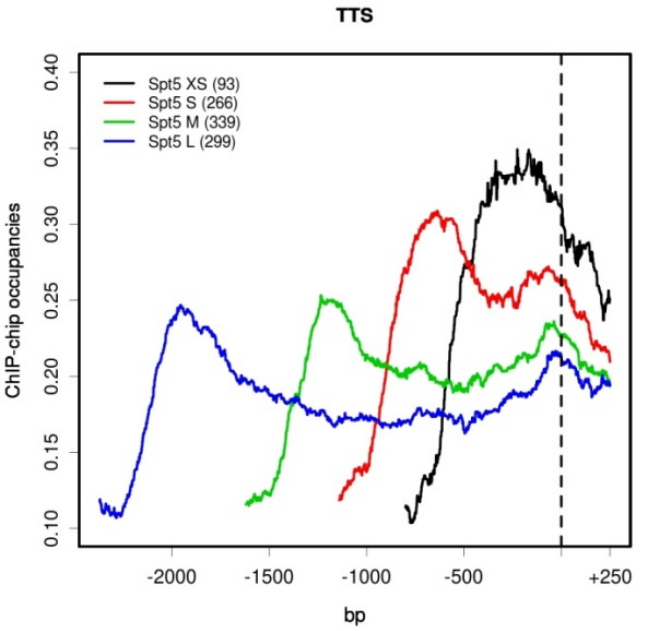

TSS
Pcf11 XS (93)
Pcf11 S (266)
Pcf11 M (339)
Pcf11 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-250
+500
+1000
+1500
+2000
bp

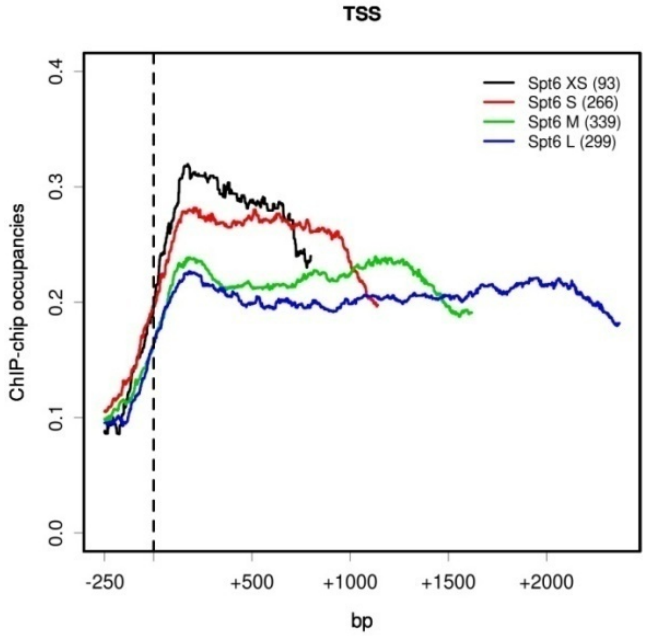

TTS
Pcf11 XS (93)
Pcf11 S (266)
Pcf11 M (339)
Pcf11 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-2000
-1500
-1000
-500
+250
bp

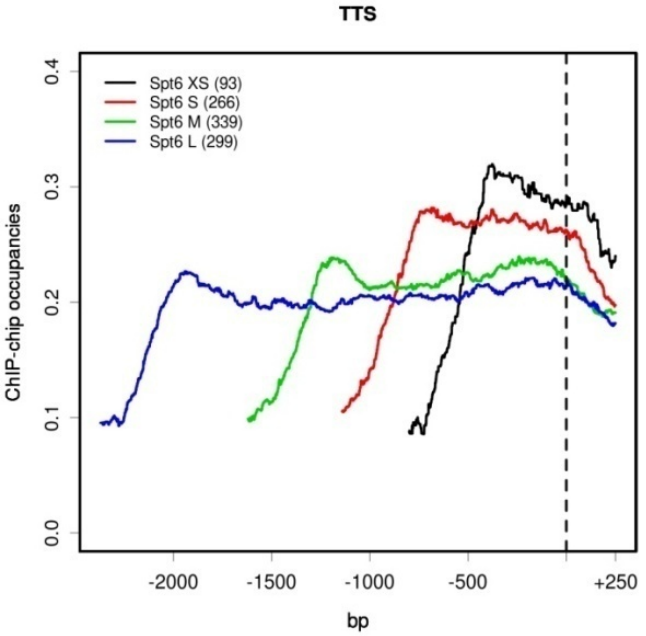

TSS
Spt4 XS (93)
Spt4 S (266)
Spt4 M (339)
Spt4 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-250
+500
+1000
+1500
+2000
bp

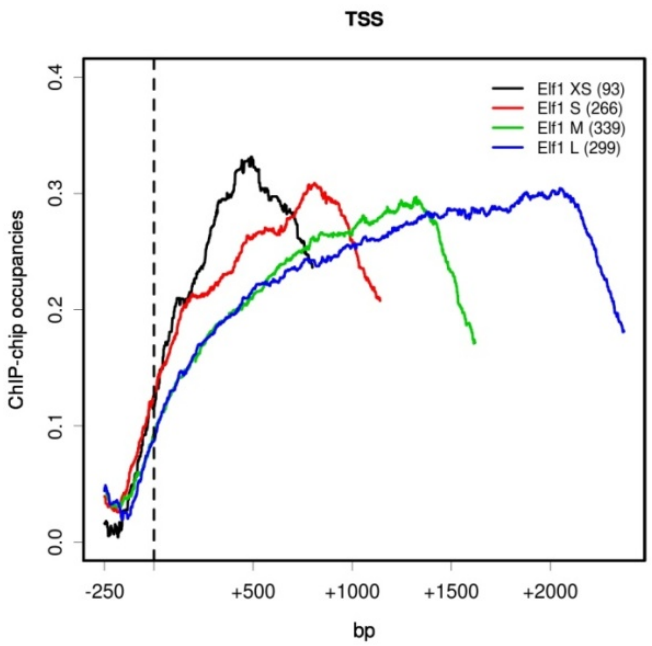

TTS
Spt4 XS (93)
Spt4 S (266)
Spt4 M (339)
Spt4 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-2000
-1500
-1000
-500
+250
bp

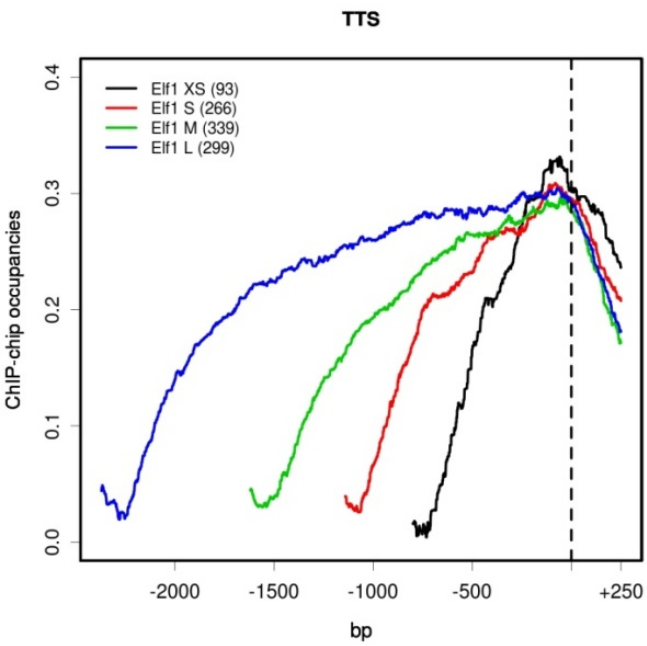

TSS
Spt5 XS (93)
Spt5 S (266)
Spt5 M (339)
Spt5 L (299)
ChIP-chip occupancies
0.10
0.15
0.20
0.25
0.30
0.35
0.40
-250
+500
+1000
+1500
+2000
bp

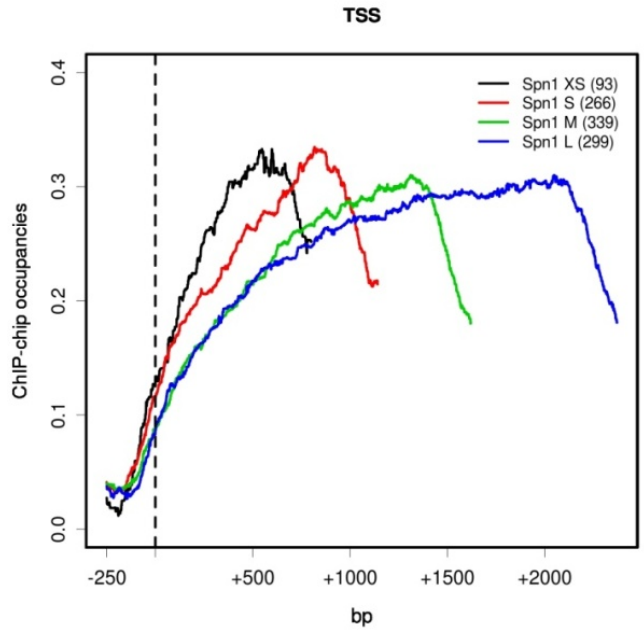

TTS
Spt5 XS (93)
Spt5 S (266)
Spt5 M (339)
Spt5 L (299)
ChIP-chip occupancies
0.10
0.15
0.20
0.25
0.30
0.35
0.40
-2000
-1500
-1000
-500
+250
bp

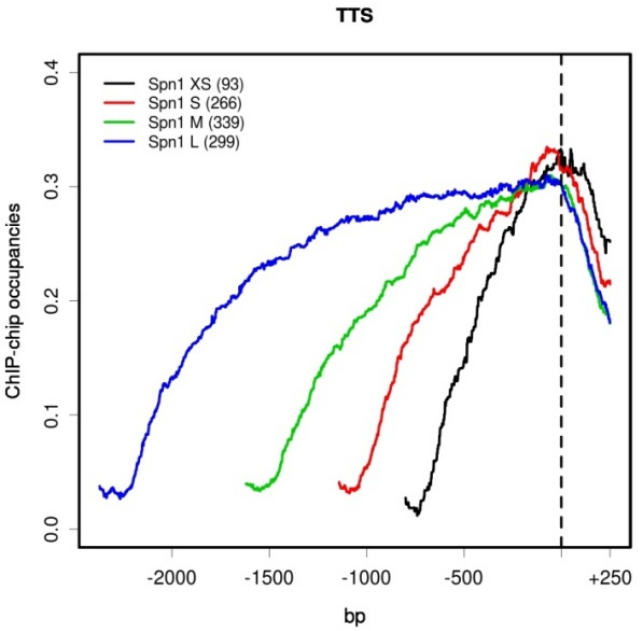

TSS
Spt6 XS (93)
Spt6 S (266)
Spt6 M (339)
Spt6 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-250
+500
+1000
+1500
+2000
bp

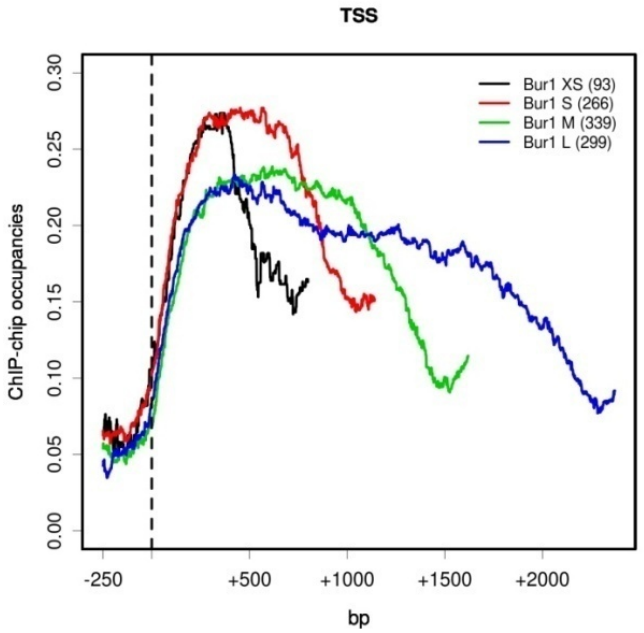

TTS
Spt6 XS (93)
Spt6 S (266)
Spt6 M (339)
Spt6 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-2000
-1500
-1000
-500
+250
bp

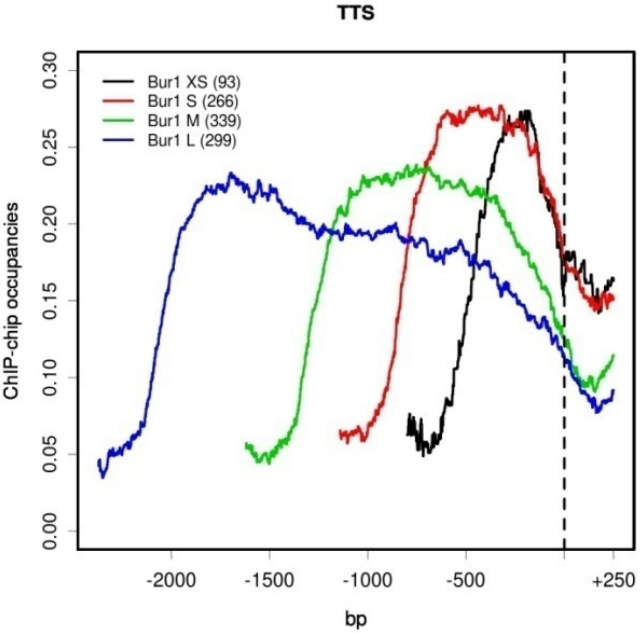

TSS
Elf1 XS (93)
Elf1 S (266)
Elf1 M (339)
Elf1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-250
+500
+1000
+1500
+2000
bp

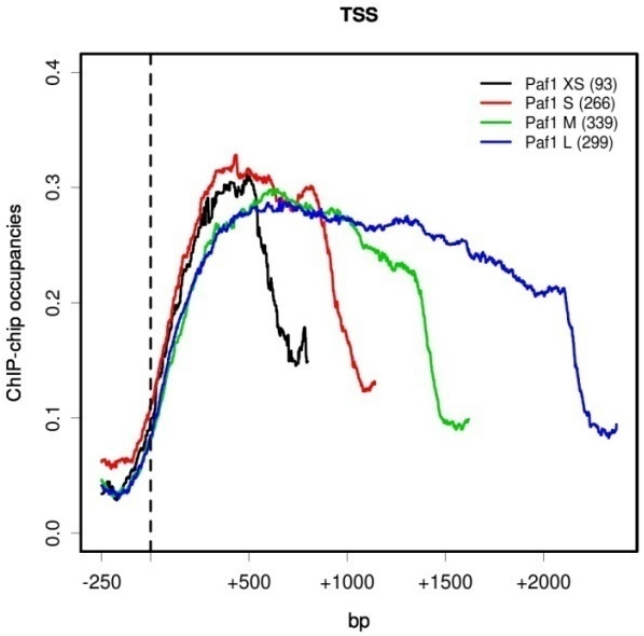

TTS
Elf1 XS (93)
Elf1 S (266)
Elf1 M (339)
Elf1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-2000
-1500
-1000
-500
+250
bp

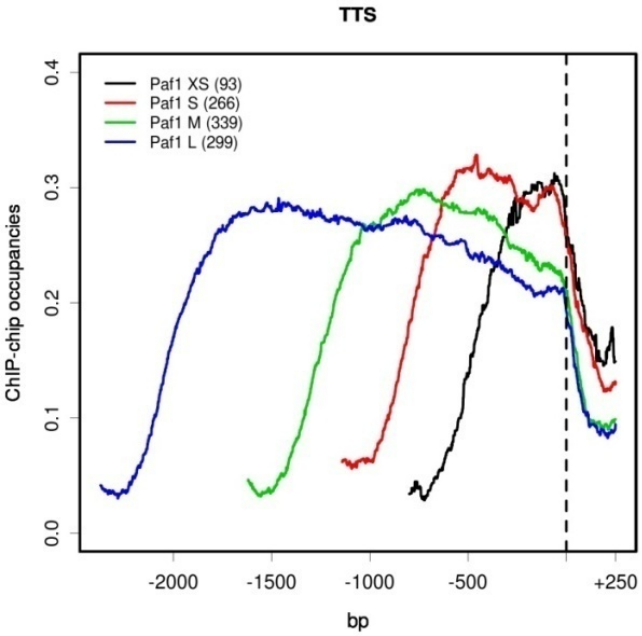

TSS
Spn1 XS (93)
Spn1 S (266)
Spn1 M (339)
Spn1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-250
+500
+1000
+1500
+2000
bp

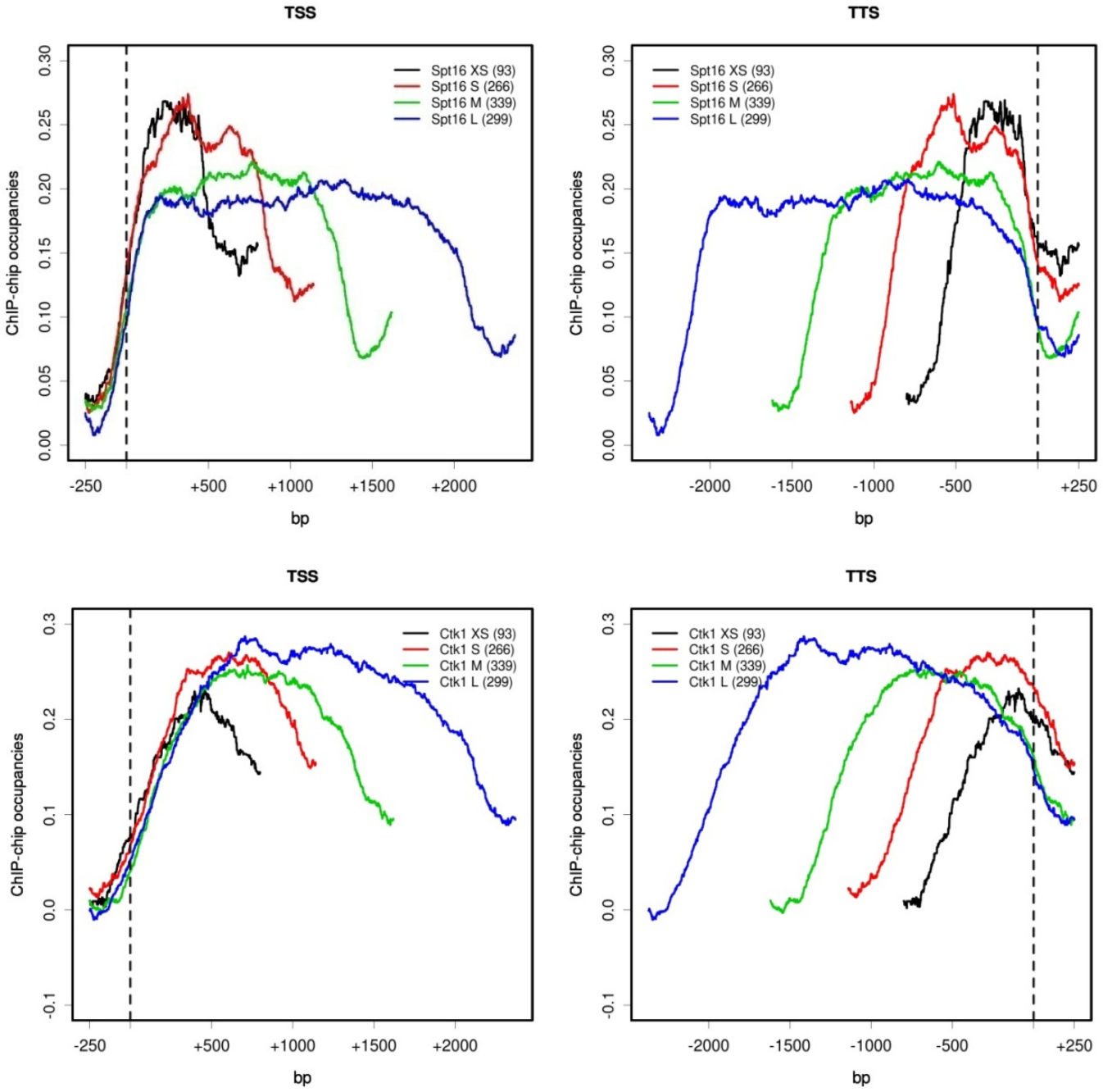

TTS
Spn1 XS (93)
Spn1 S (266)
Spn1 M (339)
Spn1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-2000
-1500
-1000
-500
+250
bp


TSS
Bur1 XS (93)
Bur1 S (266)
Bur1 M (339)
Bur1 L (299)
ChIP-chip occupancies
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-250
+500
+1000
+1500
+2000
bp

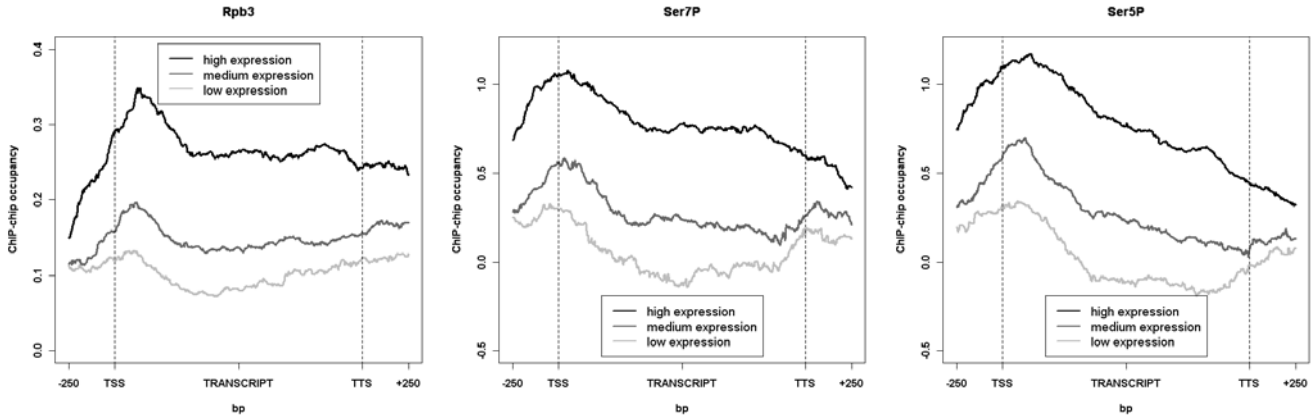

TTS
Bur1 XS (93)
Bur1 S (266)
Bur1 M (339)
Bur1 L (299)
ChIP-chip occupancies
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-2000
-1500
-1000
-500
+250
bp

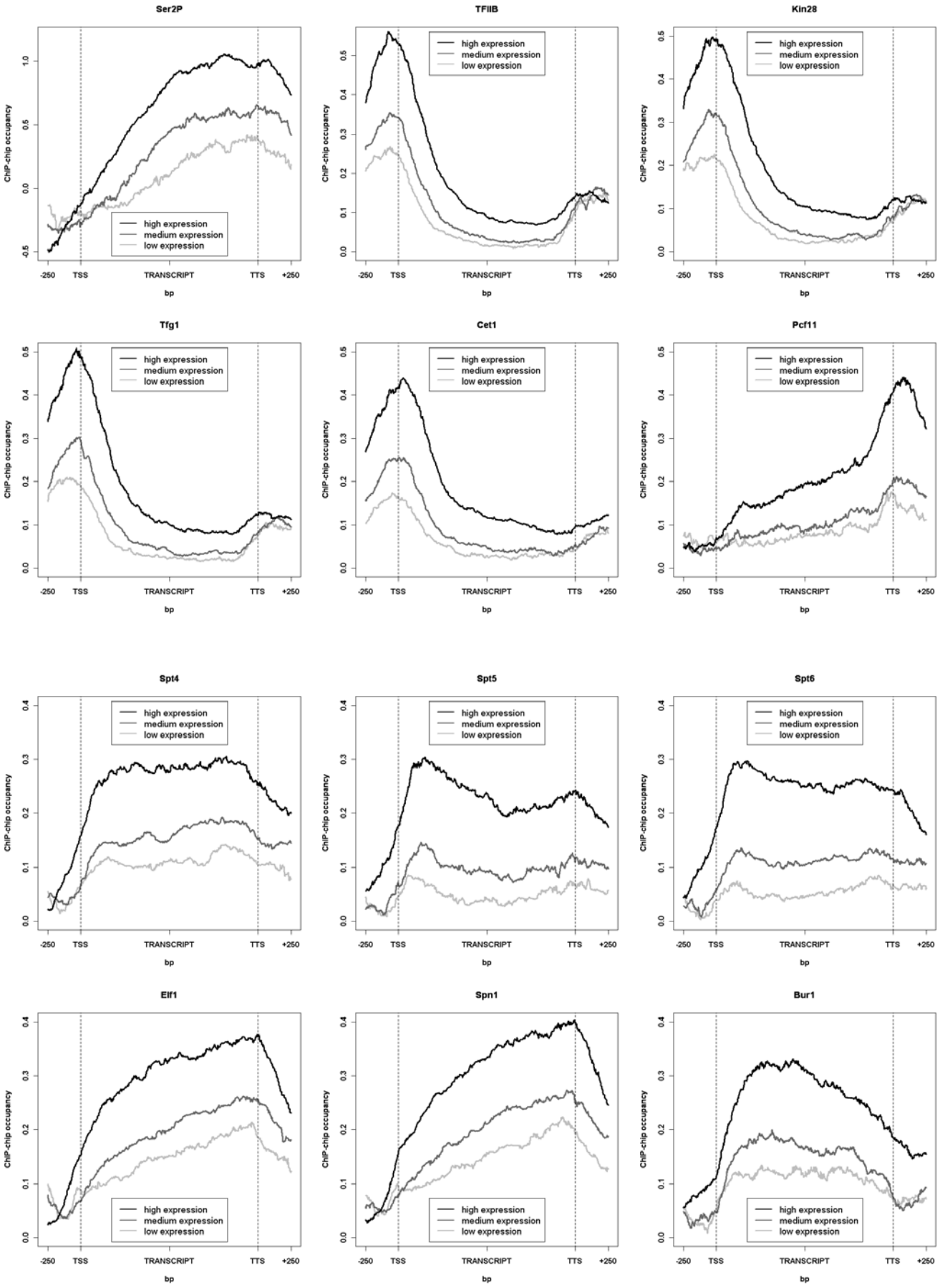

TSS
Paf1 XS (93)
Paf1 S (266)
Paf1 M (339)
Paf1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-250
+500
+1000
+1500
+2000
bp

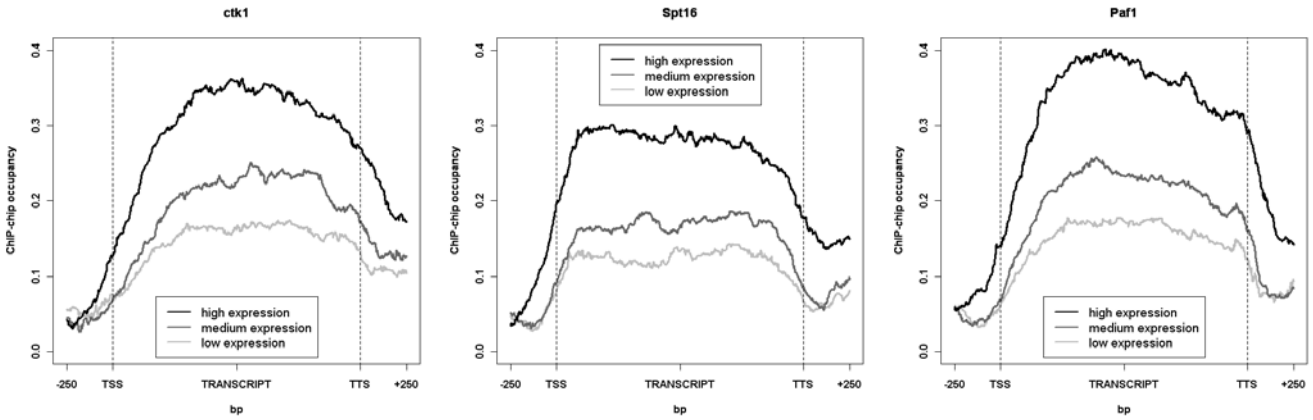

TTS
Paf1 XS (93)
Paf1 S (266)
Paf1 M (339)
Paf1 L (299)
ChIP-chip occupancies
0.0
0.1
0.2
0.3
0.4
-2000
-1500
-1000
-500
+250
bp


TSS
TTS
Spt16 XS (93)
Spt16 S (266)
Spt16 M (339)
Spt16 L (299)
ChIP-chip occupancies
bp
Ctk1 XS (93)
Ctk1 S (266)
Ctk1 M (339)
Ctk1 L (299)

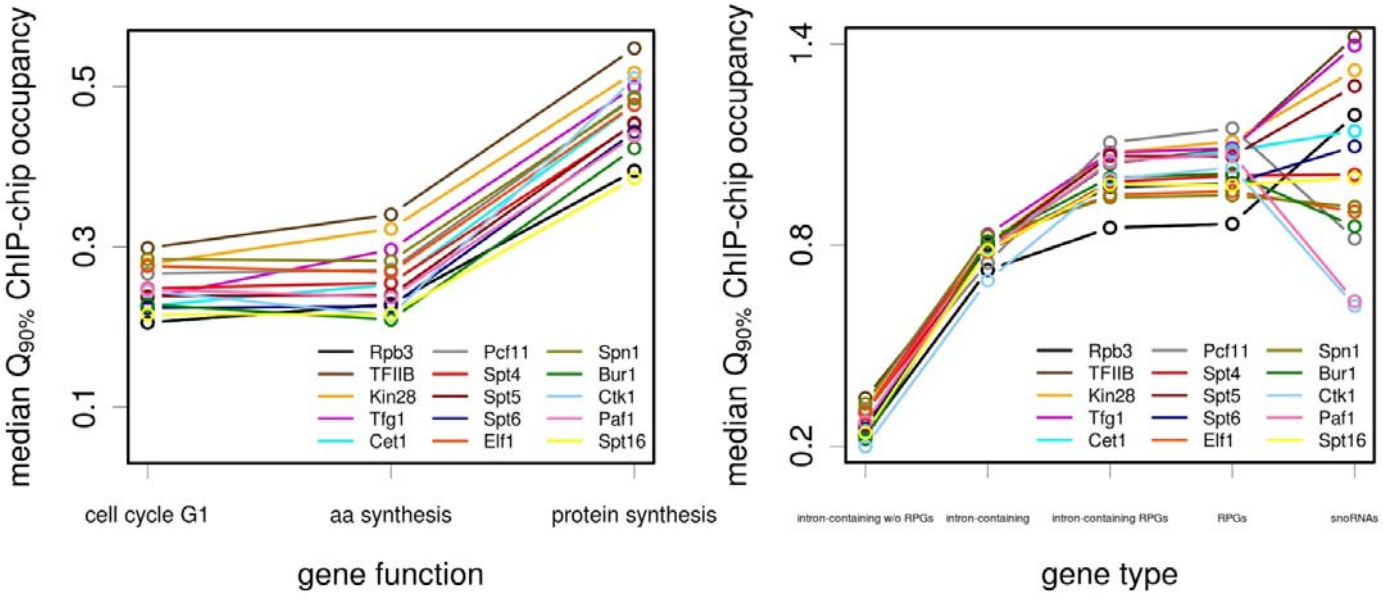
e


Rpb3
Ser7P
Ser5P
high expression
medium expression
low expression
ChIP-chip occupancy
TSS
TTS
TRANSCRIPT
bp

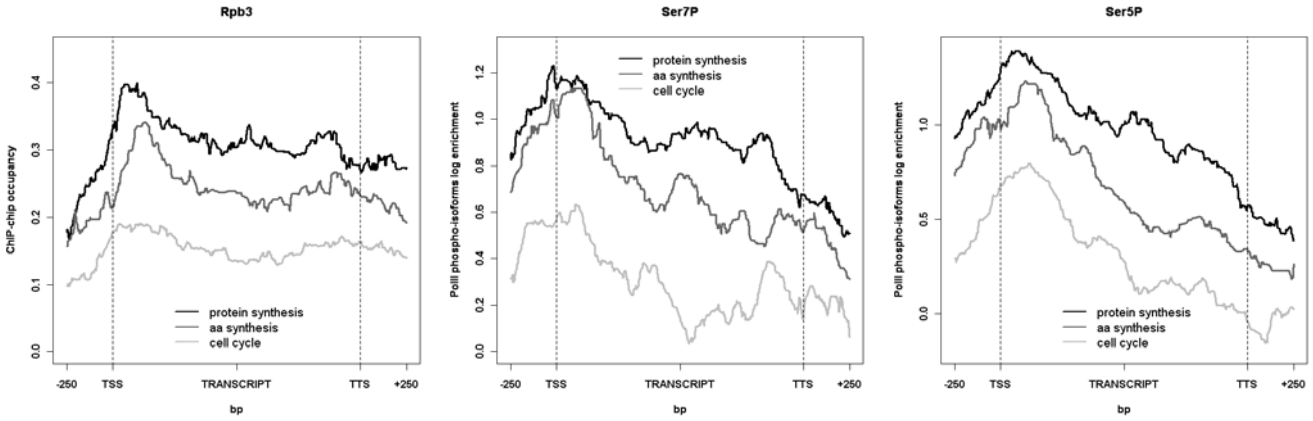

Ser2P
TFIIB
Kin28
Tfg1
Cet1
Pcf11
Spt4
Spt5
Spt6
Elf1
Spn1
Bur1
ChIP-chip occupancy
high expression
medium expression
low expression
-250
TSS
TRANSCRIPT
TTS
+250
bp

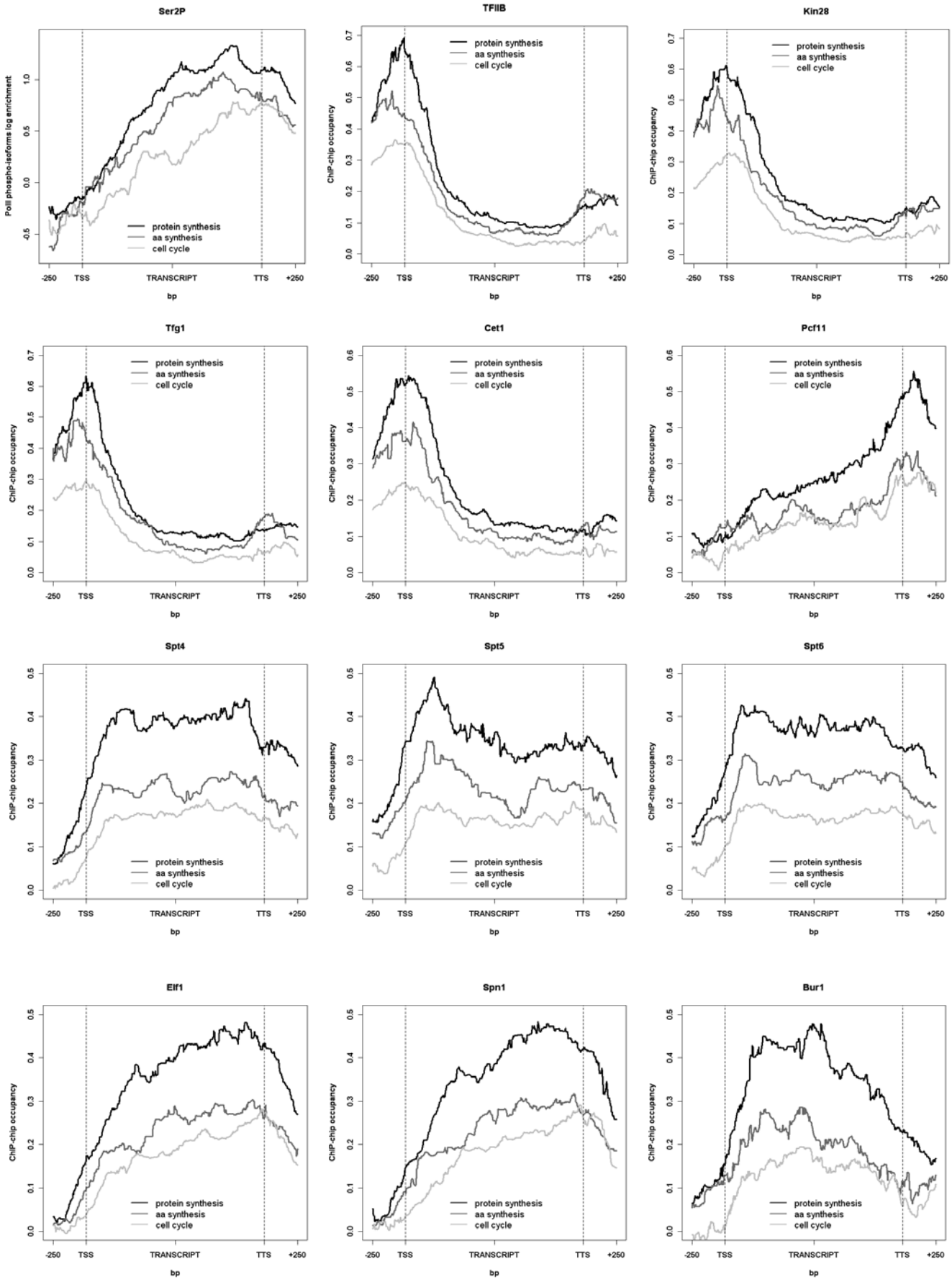

ctk1
Spt16
Paf1
ChIP-chip occupancy
high expression
medium expression
low expression
-250
TSS
TRANSCRIPT
TTS
+250
bp

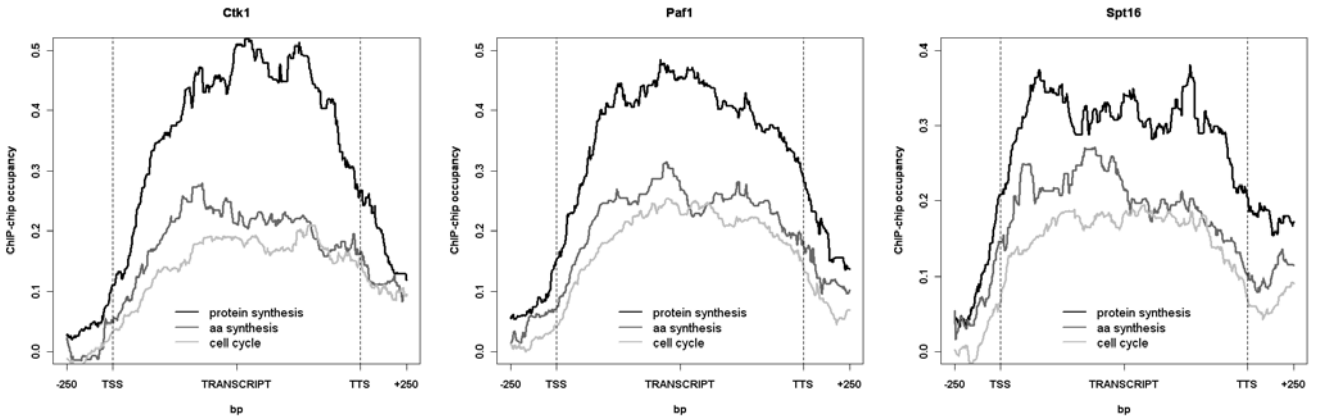
f


median Q90% ChIP-chip occupancy
Rpb3
TFIIB
Kin28
Tfg1
Cet1
Pcf11
Spt4
Spt5
Spt6
Elf1
Spn1
Bur1
Ctk1
Paf1
Spt16
cell cycle G1
aa synthesis
protein synthesis
gene function
intron-containing w/o RPGs
intron-containing
intron-containing RPGs
RPGs
snoRNAs
gene type

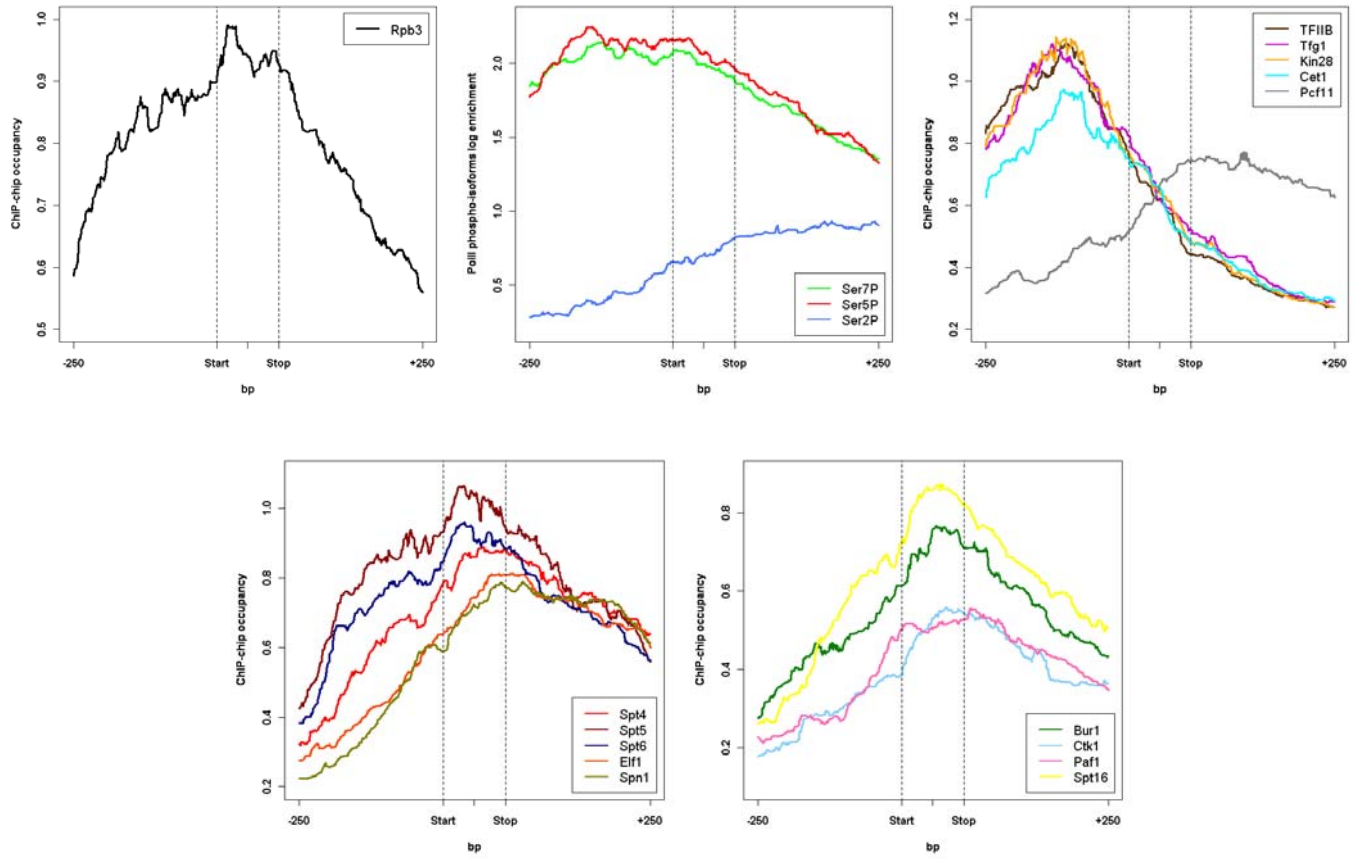
g


Rpb3
Ser7P
Ser5P
ChIP-chip occupancy
Pol II phospho-isoforms log enrichment
protein synthesis
aa synthesis
cell cycle
-250
TSS
TRANSCRIPT
TTS
+250
bp

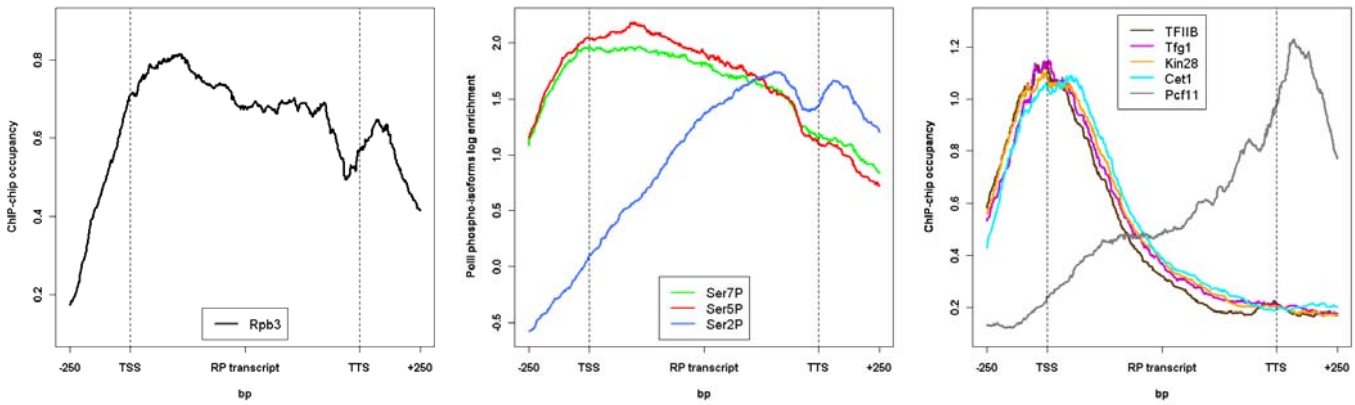

Ser2P
TFIIB
Kin28
Tfg1
Cet1
Pcf11
Spt4
Spt5
Spt6
Elf1
Spn1
Bur1
Pol II phospho-isoforms log enrichment
ChIP-chip occupancy
protein synthesis
aa synthesis
cell cycle
-250
TSS
TRANSCRIPT
TTS
+250
bp



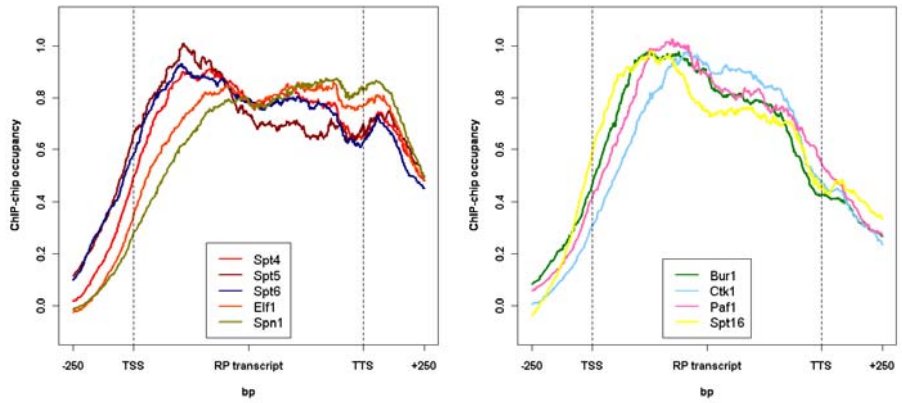
h



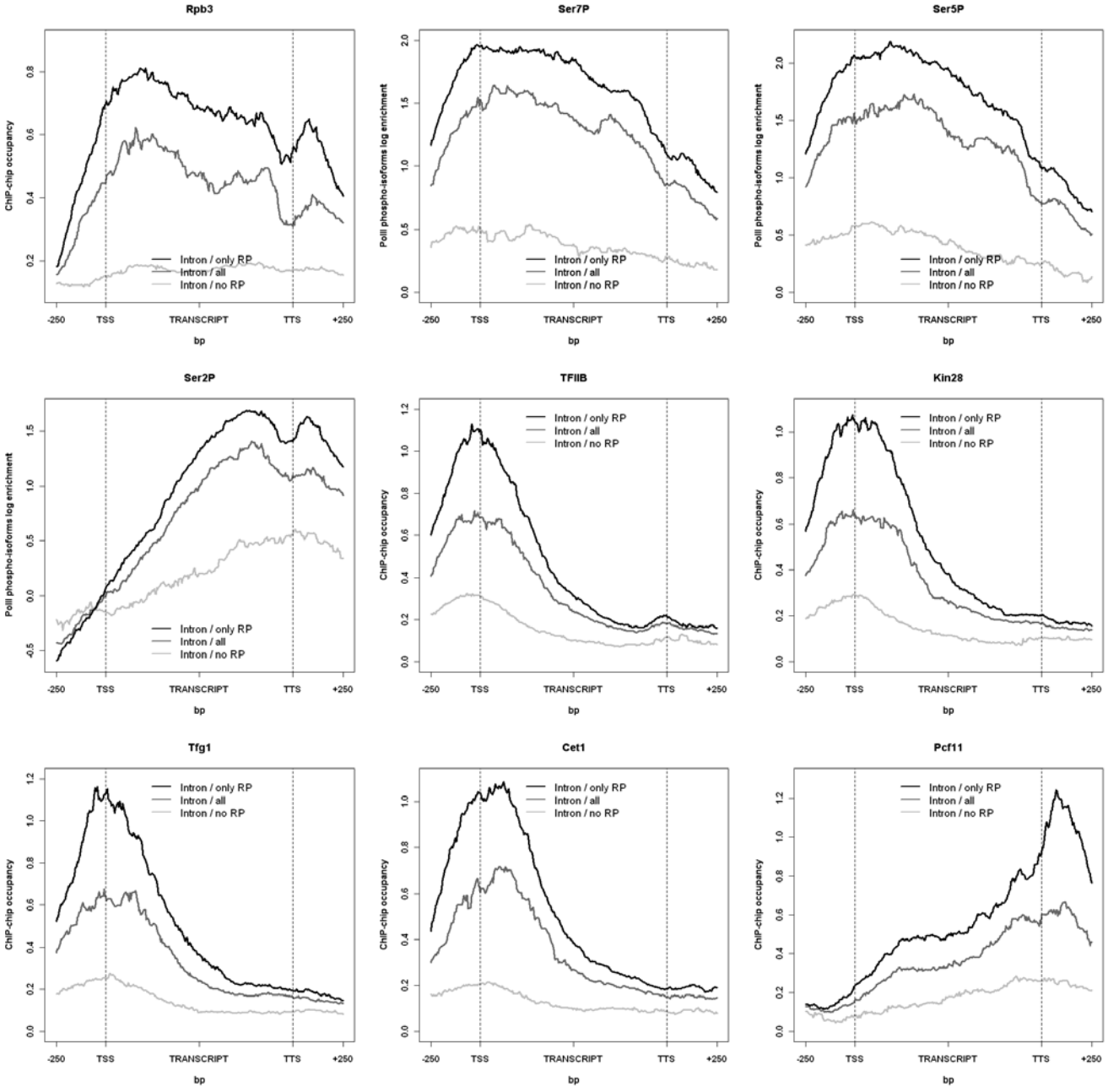
i

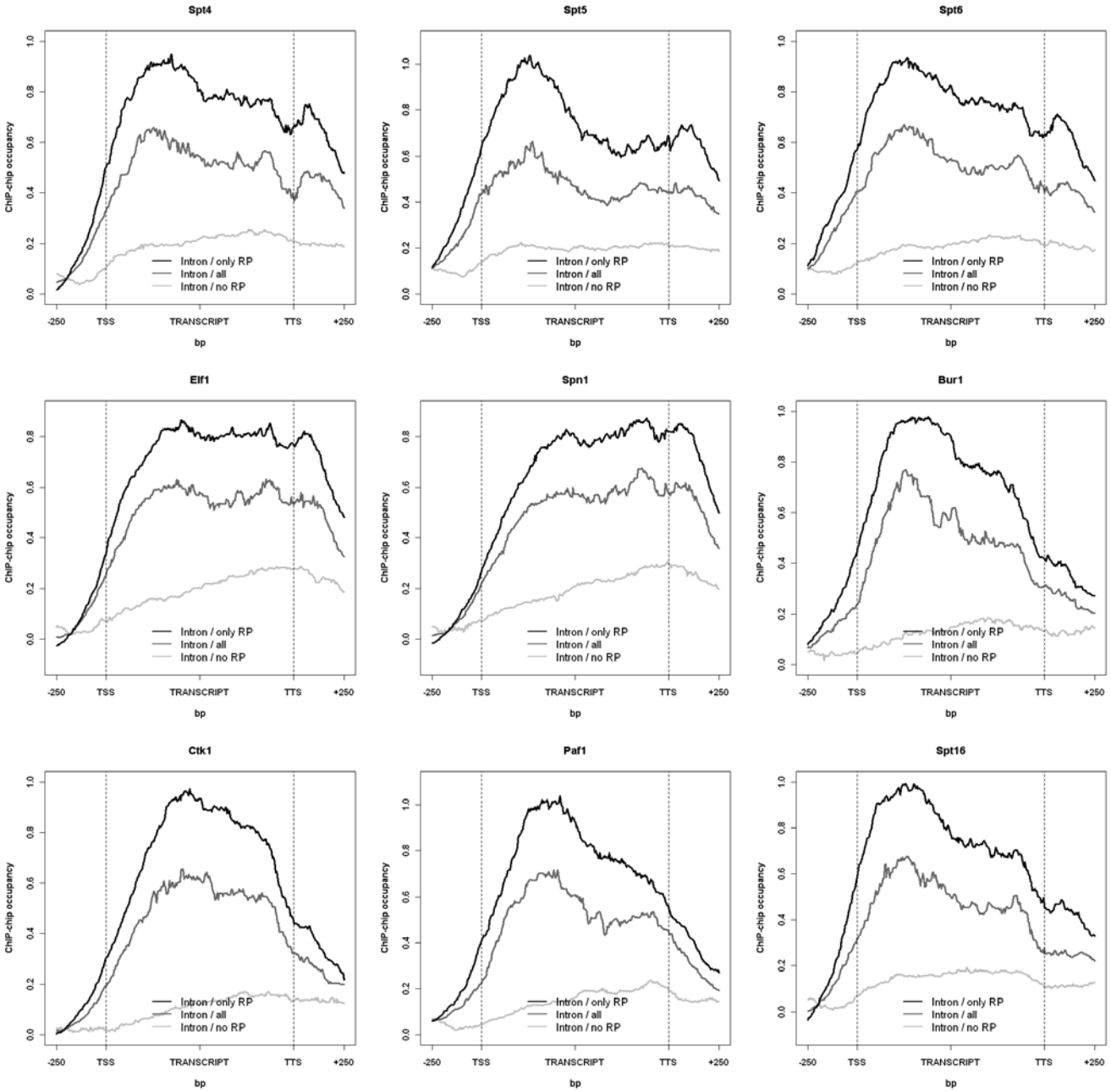


Nature Structural & Molecular Biology: doi:10.1038/nsmb.1903

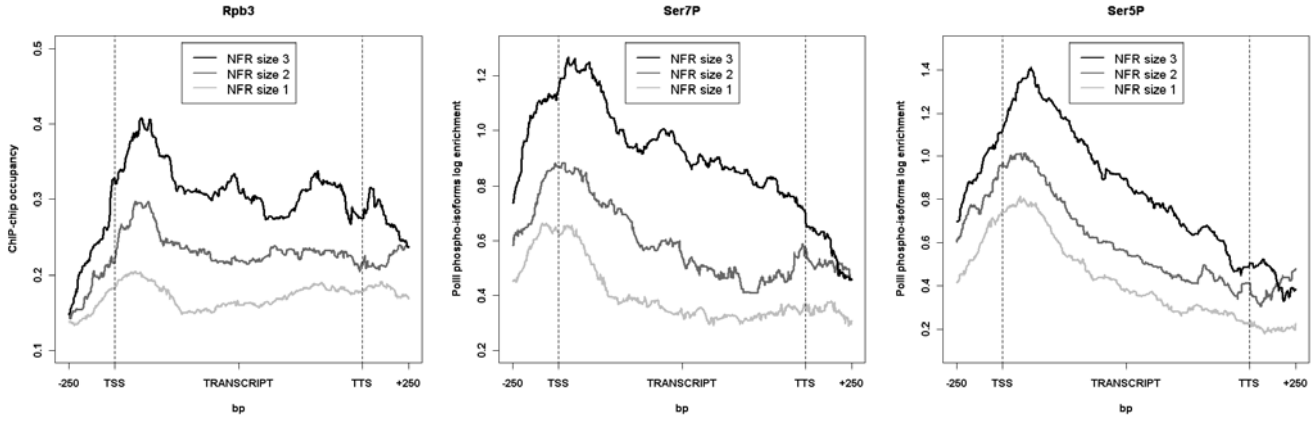

ChIP-chip occupancy
Spt4
Spt5
Spt6
Elf1
Spn1
Bur1
Ctk1
Paf1
Spt16
-250
TSS
RP transcript
TTS
+250
bp

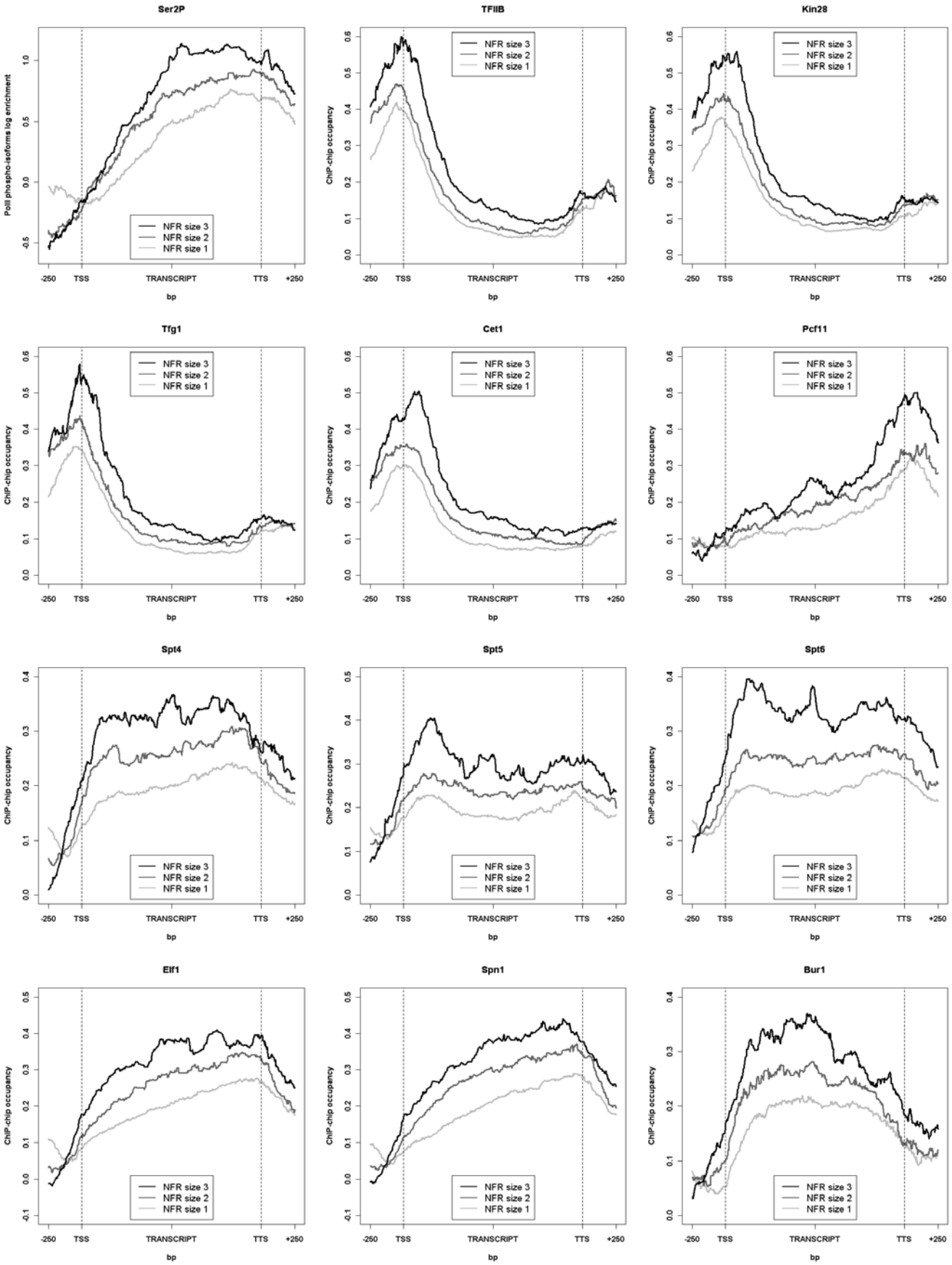
j

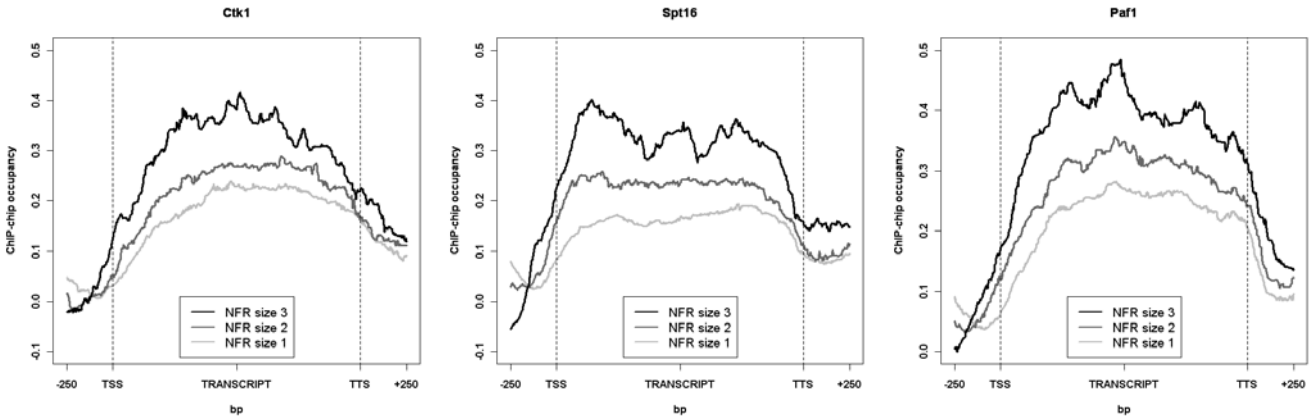

Rpb3
Ser7P
Ser5P
Ser2P
TFIIB
Kin28
Tfg1
Cet1
Pcf11
ChIP-chip occupancy
PolII phospho-isoforms log enrichment
Intron / only RP
Intron / all
Intron / no RP
-250
TSS
TRANSCRIPT
TTS
+250
bp


Spt4
Spt5
Spt6
Elf1
Spn1
Bur1
Ctk1
Paf1
Spt16
ChIP-chip occupancy
Intron / only RP
Intron / all
Intron / no RP
-250
TSS
TRANSCRIPT
TTS
+250
bp

k

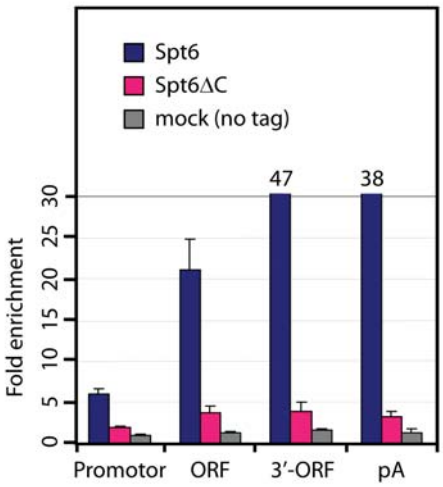

Rpb3
Ser7P
Ser5P
ChIP-chip occupancy
PolII phospho-isoforms log enrichment
NFR size 3
NFR size 2
NFR size 1
-250
TSS
TRANSCRIPT
TTS
+250
bp


Ser2P
TFIIB
Kin28
Tfg1
Cet1
Pcf11
Spt4
Spt5
Spt6
Elf1
Spn1
Bur1
PolII phospho-isoforms log enrichment
ChIP-chip occupancy
NFR size 3
NFR size 2
NFR size 1
-250
TSS
TRANSCRIPT
TTS
+250
bp



Supplementary Figure 9 Gene-averaged median profiles for different gene classes. (a) Division of yeast genes into gene length classes. The ALL set (1140 genes) was grouped into four ORF length classes (see Supplementary Methods): Xtremely Short (XS), ranging from 256 to 511 bp, Short (S) ranging from 512 to 937 bp, Medium (M) 938 to 1537 bp, and Long (L) 1538 to 2895 bp, comprising 93, 266, 339, and 299 genes, respectively. (b) Gene-averaged median profiles for length class Short (S). (c) Gene-averaged median profiles for length class Long (L). (d) Gene-averaged median profiles of four gene length classes aligned at TSS and pA, respectively. Ser7/5/2P occupancies correspond to S7/5/2P $\log_2$ enrichments. (e) Gene-averaged median profiles for genes in three different expression level classes. The ALL set of 1140 quality-filtered genes (see Supplementary Methods) was partitioned into three groups: low (25%-50% quantile), medium (50%-75% quantile), and high (>75% quantile) expression, according to data from Dengl *et al.*[1]. From these, genes with ORF lengths between 938 and 1538 were selected. Ser7/5/2P occupancies correspond to S7/5/2P $\log_2$ enrichments. (f) The composition of the transcription complex is independent of gene type and function. Analogous to Fig. 4 for gene function (left) and gene function (right). Gene functions include 213 G1 cell cycle genes, 206 amino acid (aa) synthesis genes, and 421 protein synthesis genes, obtained as transcription modules from Ihmels *et al.*[19]. Gene types comprise 63 intron-containing genes excluding Ribosomal Protein Genes (RPGs), 139 intron-containing genes, 76 intron-containing RPGs, 106 RPGs, and 68 snoRNAs. (g) Gene-averaged median profiles for different functional classes. Gene-averaged profiles of genes belonging to the functional/transcriptional modules protein synthesis, amino acid biosynthesis, and cell cycle G1[19]. Only genes of length class M were considered (65, 27, and 20 genes, respectively). (h) Averaged profiles of 66 yeast snoRNAs with length <300 bp (i.e. 66 out of 77 snoRNAs in yeast). All factors are recruited to snoRNA genes. These genes also showed CTD phosphorylation. (i) Averaged profiles of 106 yeast ribosomal protein genes with length 1000 $\pm$ 300 bp (i.e. 106 out of 137 ribosomal protein genes in yeast). (j) Averaged profiles of 139 intron-containing genes with known TSS and pA and length 900 $\pm$ 300 bp ("Intron / all"), and subgroups of these containing only ribosomal protein genes ("Intron / only RP", 76 genes), or containing only non-ribosomal protein genes ("Intron / no RP", 63 genes). Due to the high occupancy of ribosomal protein genes and snoRNAs, genes were analyzed without distance filtering. The sharp dips just upstream of the pA in (h) and (i) coincide with the very AT-rich nucleosome free regions. They are likely to be caused by nonlinear GC content-dependence of the probe hybridization efficiency, which is not completely corrected out by the reference normalization for the most highly occupied genes. (k) Gene-averaged median profiles for three classes with different nucleosome free region (NFR) size. The M set of genes was partitioned into three groups with NFR sizes of 1, 2 and 3 nucleosomes length (based on three main clusters of genes classified according to their NFR organization[20]).

Note: For details on gene-averaged profiles see Fig. 1 and Supplementary Methods. TTS = transcript termination site (corresponds to the polyadenylation (pA) site).



Supplementary Figure 10 Deletion of the Spt6 C-terminal CTD-binding domain results in a reduction of fold enrichment at the gene encoding ADH1. A TAP-tagged Spt6 wild-type strain and a Spt6 variant strain lacking the 202 C-terminal residues (Spt6ΔC) as well as a BY4741 strain (without tag) were analyzed by ChIP of the ADH1 gene. Precipitated DNA was used for quantitative PCR amplification with primers directed against the promoter, central and 3' regions of the ORF, and against the pA site. The fold enrichments for the different positions of the ADH1 gene over an ORF-free region on chromosome V are given on the y-axis. Error bars show standard deviations for three independent experiments.

Supplementary Tables

Supplementary Table 1 Replicate correlations.

	Rpb3	S7P	S5P	S2P	TFIIB	Tfg1	Kin28	Cet1	Spt4	Spt5
Pearson *R*[a]	0.920	0.918	0.886	0.973	0.964	0.917	0.930	0.843	0.905	0.975

	Spt6	Spt6ΔC	Elf1	Spn1	Spt16	Paf1	Bur1	Ctk1	Pcf11
Pearson *R*	0.939	0.566	0.946	0.975	0.681	0.921	0.733	0.911	0.727

[a] Pearson correlations between replicate measurements were calculated after division of the factor signal intensities by input reference intensities and smoothing using the sliding window smoothing procedure (window half size of 75 bp) implemented in the R package *Ringo*[13].

Supplementary Table 2 Peaks and transitions in gene-averaged occupancy profiles.

Feature	Peaks around TSS[a]		5′-transition (position)[b]	Peaks around pA site[a]		3′-transition (% of max)[c]
Genes[d]	M	ALL	M	M	ALL	M
RNA polymerase II						
Rpb3	149 ± 10	131 ± 6	-20 ± 21			
S7P	26 ± 44	18 ± 21				
S5P	128 ± 16	117 ± 10				
S2P				-252 ± 36	-204 ± 51	
Initiation and 5′-capping						
TFIIB	-52 ± 13	-45 ± 6				
TFIIH (Kin28)	-41 ± 10	-25 ± 6				
TFIIF (Tfg1)	-34 ± 10	-36 ± 6				
Cet1	6 ± 18	14 ± 8				
Elongation and 3′-processing						
Spt5	195 ± 12	197 ± 9	15 ± 19			71 ± 6
Spt4			20 ± 13	-204 ± 16	-211 ± 12	67 ± 4
Spt6	202 ± 12	204 ± 9	25 ± 16			75 ± 4
Elf1			26 ± 15	-73 ± 28	-74 ± 9	83 ± 3
Spn1			42 ± 15	-52 ± 28	-52 ±9	83 ± 3
Spt16			-13 ± 12			21 ± 4
Paf1			36 ± 10			29 ± 3
Bur1			62 ± 12			26 ± 4
Ctk1			66 ± 9			46 ± 3
Pcf11				72 ± 13	58 ± 6	

[a] Peak positions were determined separately for the ALL gene set (1140 genes) and length class M (339 genes). Median profiles were calculated in a region ± 250 bp around the TSS or the pA site, and smoothed with cubic splines (R package: stats, function: smooth.spline, parameter: spar = 0.9). The maximum value of the smoothed curve was selected as peak position. To estimate the uncertainty of the peak position, we drew 1000 bootstrap samples from the set of genes, recalculated peak positions

for each bootstrap sample as described, and estimated their scatter using the Median Absolute Deviation (MAD) measure.

[b] 5′-transitions were determined for length class M (339 genes). Median profiles were calculated in a region ± 150 bp around the TSS, and smoothed with cubic splines (R package: stats, function: smooth.spline, parameter: spar = 0.9). The position at 11.5% occupancy of the smoothed curve (corresponding to the point after which the elongation factor profiles start to change) was defined as 5′-transition point. To estimate the uncertainty of the 5′ transition, we drew 1000 bootstrap samples from the set of genes, recalculated 5′-transitions for each bootstrap sample as described, and estimated their scatter using the Median Absolute Deviation (MAD) measure.

[c] 3′-transitions were determined for length class M (339 genes) as the percentage of the maximum occupancy at 100 nt downstream of the pA site. Median profiles were calculated in a region -250 nt upstream of the TSS and 250 nt downstream of the pA site, and smoothed with cubic splines (R package: stats, function: smooth.spline, parameter: spar = 0.9). To estimate the uncertainty of the 3′-transition, we drew 1000 bootstrap samples from the set of genes, recalculated 3′-transitions for each bootstrap sample as described, and estimated their scatter using the Median Absolute Deviation (MAD) measure.

[d] Gene selection to calculate gene-averaged profiles described in Supplementary Methods.

Supplementary Methods

Yeast strains and epitope tagging

Saccharomyces cerevisiae (*S. cerevisiae*) BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) strains containing C-terminally tandem affinity purification (TAP) tagged versions of target proteins were obtained from Open Biosystems. The Spt6ΔC strain lacking the 202 C-terminal residues was generated as described[1]. BY4741 untagged wild-type strain (Open Biosystems) was used for chromatin immunoprecipiation (ChIP) coupled to tiling microarray (ChIP-chip) analysis of the different Pol II phospho-isoforms, for mock IP in ChIP-chip experiments, and for control in TAP purifications. All epitope-tagged strains were validated. First, gene-specific PCR was performed to confirm that the TAP tag was at the correct genomic position. Second, Western blotting with anti-TAP (PAP, Sigma) antibodies was performed to verify if the tagged protein of interest was properly expressed. Third, the growth of the various tagged strains compared to the non-tagged wild-type strain was monitored to rule out any influence of the epitope tag on yeast growth. This was done by serial dilutions of the various yeast strains on YPD plates at 30°C for two days (Supplementary Fig. 1).

Chromatin immunoprecipitation with TAP-tagged yeast strains

For the yeast TAP-tagged strains, ChIP was performed as described[2], with modifications. Briefly, yeast strains containing TAP-tagged versions of the proteins as well as an untagged wild-type strain (for mock IP) were grown in 600 ml YPD medium to mid-log phase (OD_{600} ~ 0.8). For the Spt6ΔC mutant the cell number was doubled (1.2 L culture) since the cell lysis efficiency was reduced 2-fold. Yeast cultures were treated with formaldehyde (1%, Sigma F1635) for 20 min at room temperature. Cross-linking was quenched with 75 ml of 3 M glycine for 30 min at room temperature. All subsequent steps were performed at 4°C with pre-cooled buffers and in the presence of a fresh protease-inhibitor mix (1 mM Leupetin, 2 mM Pepstatin A, 100 mM Phenylmethylsulfonyl fluoride, 280 mM Benzamidine). Cells were collected by centrifugation at 4000 rpm (Sorvall SLA-1500 rotor, Sorvall Evolution RC centrifuge) for 5 min, washed twice with 1 × TBS (20 mM Tris-HCL at pH 7.5, 150 mM NaCl) and twice with FA lysis buffer (50 mM HEPES-KOH at pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na deoxycholate, 0.1% SDS, 1 × protease inhibitor mix). Cell pellets were flash-frozen in liquid nitrogen and stored at -80°C. Cell pellets were thawed on ice and resuspended in 1 ml FA lysis buffer. Cells were disrupted by vortexing (neoLab 7-2020) in the presence of 1 ml silica-zirconia beads (Roth) for 3 min at full speed at 4°C, followed by an incubation of the sample for 2 min on ice. This was repeated 12 times. The success of the cell lysis was monitored by photometric measurements and the cell lysis efficiency was usually >80%. The chromatin was washed twice with FA lysis buffer and sonicated by application of a Bioruptor™ UCD-200 (Diagenode Inc.) to yield an average DNA fragment size of 250 bp as determined by agarose gel electrophoresis (Supplementary Fig. 1). This was achieved by sonifying the sample 35 min at the “high” intensity setting with alternating sessions of 30 sec of sonication

followed by 30 sec of resting. 30 µl and 100 µl of the washed and fragmented chromatin samples were saved as input materials and for control of the average chromatin fragment size (see below), respectively. 800 µl of the remaining chromatin sample was immunoprecipitated with 20 µl IgG Sepharose™ 6 Fast Flow beads (GE Healthcare) at 4°C for 4 h on a turning wheel. Immunoprecipitated chromatin was washed 3 times with FA lysis buffer, twice with high-salt FA lysis buffer (500 mM instead of 150 mM NaCl), twice with ChIP wash buffer (10 mM Tris-HCl at pH 8.0, 0.25 M LiCl, 1 mM EDTA, 0.5% NP-40, 0.5% Na deoxycholate) and one time with TE buffer (10 mM Tris-HCl at pH 7.4, 1 mM EDTA). Immunoprecipitated chromatin was eluted for 1 h at 65°C in the presence of the ChIP elution buffer (50 mM Tris-HCl at pH 7.5, 10 mM EDTA, 1% SDS). Eluted immunoprecipitated chromatin as well as input material and material for control of the average chromatin fragment size were subjected to Proteinase K (20 µl of 20 mg/ml Proteinase K from *Engyodontium album*, Sigma P4850) digestion at 37°C for 2 h and reversal of crosslinks (at 65°C over-night). Samples used for determining the average chromatin fragment size were phenol-extracted twice and ethanol-precipitated over-night. The pellet was resuspended in 20 µl TE buffer (10 mM Tris-HCl at pH 7.4, 1 mM EDTA at pH 8.0) and incubated with 10 µl RNase A/T1 Mix (2 mg/ml RNase A, 5000 U/ml RNase T1; Fermentas) at 37°C for 1 h. The resulting DNA sample was electrophoretically separated on a 1.5% agarose gel. DNA of the IP, mock IP and input samples was purified with the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions, except that the final elution was performed with 100 µl DNAse-free water. RNA was digested by adding 5 µl of RNAse A (10 mg/ml, Sigma) at 37°C for 20 min. DNA was again purified with the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions. In case of the IP sample, the eluate was concentrated via vacuum manifold to a final volume of 10 µl. The total volume was used for DNA amplification (see below).

Chromatin immunoprecipitation of phosphorylated Pol II isoforms

For ChIP analysis of the Pol II phospho-isoforms, chromatin preparation was performed as above, except that it was conducted in the presence of a combination of phosphatase inhibitors (1 mM NaN_3, 1 mM NaF, 0.4 mM Na_3VO_4). For chromatin immunoprecipitation a set of monoclonal antibodies with strong specificity and affinity for phosphorylated serine residues S5P (3E8), S2P (3E10) and S7P (4E12) were applied. These antibodies were generated by Chapman et al.[3] and were a generous gift from Dirk Eick (Helmholtz Zentrum München). It was reported that the amount of antibody used influences the occupancy profiles obtained for Pol II phospho-isoforms[4]. We therefore carried out ChIP experiments with different amounts of antibodies before the ChIP-chip analyses. The results of these antibody titration experiments are shown for three different regions of the two housekeeping genes ADH1 and PMA1 (Supplementary Fig. 1d). Briefly, for the 3E10 antibody detecting S2-phosphorylated Pol II, the occupancy behaviour remained nearly the same for the different amounts of antibody tested (5 to 200 µl) with peak levels towards the 3′ end of the gene. With increasing amount of that antibody, S2P levels showed only a marginal rise in the 5′ regions of ADH1 and PMA1. Acceptable fold enrichments could be reached with 25 µl 3E10 antibody. The 3E8 antibody directed against S5-phosphorylated Pol II showed the

strongest reactivity at the 5′ end of genes, especially when lower amounts were used (5 to 20 µl). With increasing amounts of antibody (100 to 200 µl) the signal clearly persists throughout the transcribed region with no clear peak level at 5′ end of ADH1 and PMA1. Since best fold enrichments could be observed with 20 µl of 3E8 antibody, this amount was used in further ChIP experiments. With respect to 4E12 antibody detecting S7-phosphorylated Pol II a change in the occupancy behaviour could be observed for different amounts tested (5 to 200 µl). With lower amounts of antibody (5 µl), the signal increased towards the 3′-end of genes. When more than 25 µl were applied, the strongest signal could be detected for the 5′ ends of genes. This trend intensified with increasing amounts of antibody resembling the occupancy behaviour of S5P. Acceptable fold enrichments could be reached with 50 µl 4E12 antibody and this amount was used in further ChIP experiments. 30 µl and 100 µl of the washed and fragmented chromatin samples were saved as input materials and as control of the average chromatin fragment size, respectively. The remaining 800 µl of sheared chromatin solution was immunoprecipitated with 20 µl, 25 µl or 50 µl of 3E8, 3E10 and 4E12 rat monoclonal antibodies (cell culture supernatant) at 4°C overnight on a rotating wheel, respectively. 25 µl of Protein A and Protein G Sepharose were added and incubated at 4°C for 1.5 h on a rotating wheel. Immunoprecipitated chromatin was treated as described above.

Quantitative real-time PCR (qPCR)

For ChIP experiments, input and immunoprecipitated (IP) samples were assayed by qPCR to assess the extent of protein occupancy at different genomic regions. Primer pairs directed against promoter, coding and terminator regions of the housekeeping genes *ADH1*, *ACT1* and *PMA1* as well as against a heterochromatic control region of chromosome V (chrV) were designed and the corresponding PCR efficiencies determined. All primer pairs used in this study had PCR efficiencies in the range of 95-100%. PCR reactions contained 1 µl DNA template, 2 µl of 10 µM primer pairs and 12.5 µl iTaq SYBR Green Supermix (Bio-Rad). Quantitative PCR was performed on a Bio-Rad CFX96 Real-Time System (Bio-Rad Laboratories, Inc.) using a 3 min denaturing step at 95°C, followed by 49 cycles of 30 s at 95°C, 30 s at 61°C and 15 s at 72°C. Threshold cycle (Ct) values were determined by application of the corresponding Bio-Rad CFX Manager software version 1.1 using the Ct determination mode “Regression”. Fold enrichment of any given region over an open reading frame (ORF)-free heterochromatic region on chromosome V was determined and calculated essentially as described[5]. Sequence information of primer pairs used in this study is available upon request.

DNA labeling and microarray handling

DNA samples were amplified and re-amplified with GenomePlex© Complete Whole Genome Amplification 2 (WGA2) Kit using the Farnham Lab WGA Protocol for ChIP-chip[6] (http://www.genomecenter.ucdavis.edu/farnham/protocol.html). DNA quantity and quality control was performed with a ND-1000 Spectrophotometer (NanoDrop Technologies) and was usually larger than 1 µg. In addition, DNA quality was monitored

by agarose gel electrophoresis. The re-amplification was performed in the presence of 0.4 mM dUTP (Promega U1191) to allow later enzymatic fragmentation. The enzymatic fragmentation, labeling, hybridization and array scanning were done according to the manufacturer's instructions (Affymetrix Chromatin Immunoprecipitation Assay Protocol P/N 702238). Enzymatic fragmentation and terminal labeling were performed by application of the GeneChip WT Double-Stranded DNA Terminal Labeling Kit (P/N 900812, Affymetrix). Briefly, re-amplified DNA was fragmented in the presence of 1.5 µl uracil-DNA-glycosylase (10 U/µl) and 2.25 µl APE1 (100 U/µl) at 30°C for 1 h 15 min. The average fragment size was in the range of 50-70 bases as determined by automated gel electrophoresis on an Experion system (Bio-Rad Laboratories, Inc.) that allowed the analysis of small amounts of DNA. The fragmented DNA was then labeled at the 3′-end by adding 2 µl and 1 µl of terminal nucleotidyl transferase (TdT, 30 U/µl) and GeneChip DNA Labeling Reagent (5 mM), respectively. 5.5 µg of fragmented and labeled DNA were hybridized to a high-density custom-made Affymetrix tiling array[7] (PN 520055) at 45°C for 16 h with constant rotational mixing at 60 rpm in a GeneChip Hybridization Oven 640 (Affymetrix, SantaClara, CA). Washing and staining of the tiling arrays were performed using the FS450_0001 script of the Affymetrix GeneChip Fluidics Station 450. The arrays were scanned using an Affymetrix GeneChip Scanner 3000 7G. The resulting raw data image files (.DAT) were inspected for any impairment. The CEL intensity files were used for bioinformatics analysis.

Protein purification and identification

TAP-tagged proteins were purified by the TAP method essentially as described[8]. Proteins associated with the purified TAP-tagged proteins were identified either by mass spectrometry (Zentrallabor für Proteinanalytik, Ludwig-Maximilians-Universität München) or by western blot analysis using monoclonal antibodies directed against the HA epitope (3F10, Roche Applied Science), Rpb1 (8WG16, Santa Cruz Biotechnology, Inc.) and Rpb3 (1Y26, NeoClone Biotechnology).

Replicate measurements, reference samples, and data quality control

At least two independent biological replicates were analyzed for each factor (replicate correlations are shown in Supplementary Table 1). Mock IP and input (genomic background) measurements were used for normalization (see below). Two biological replicates were used for the mock IP (R = 0.65). Due to the very high reproducibility/correlation between input samples of different factors (comparable to factor replicate correlations), we took input samples of three factors, Rpb3, Spt4, and Spt6, and used them as triplicate measurements to normalize all factors except the Pol II phospho-isoforms and Spt6ΔC. The latter were normalized by dividing them by their matched input data. All data normalization procedures were performed using R[9] and Bioconductor[10]. For data import of the Affymetrix CEL files and the conversion into the basic Bioconductor object class for microarray data *ExpressionSet*, we used the R package *Starr*[11]. Quality assessment of each measured array was done by inspection of

raw image files, density-plots, scatter-plots, and MA-plots, in order to avoid processing flawed arrays.

Normalization with the use of mock IP and input reference data

Our normalization thus consists of three steps. First, we performed quantile normalization between replicate measurements (not between non-replicate measurements). Second, for each condition (including the reference measurements) we averaged the signal for each probe by calculating the geometric average over the replicate intensities. Third, data from all factors were normalized using a combined mock IP plus input reference normalization. The rationale for the last step is explained in detail below.

In ChIP-chip experiments, it is important to correct for sequence- and genomic region-specific biases in the efficiency of the various biochemical and biophysical steps. Fragmentation of the chromatin, PCR amplification, immunoprecipitation, labeling, and hybridization to the array can produce strong biases. Although reference-free normalization procedures have been suggested that can reduce these biases[12], the cleanest and most efficient method consists in measuring a reference signal and dividing the true signal intensities obtained from the immunoprecipitated protein by the reference intensities. This is in our experience the most important step in data normalization when using arrays with short probes such as the Affymetrix arrays used here, which exhibit strong GC-content bias. As reference signal one can use the intensities obtained by hybridizing the input (genomic background) fraction to a tiling array or, alternatively, a mock IP. Both procedures are popular and it probably depends on individual experimental conditions and array platforms which of the two perform better in correcting for systematic biases without unduly increasing statistical noise.

We developed a simple mathematical model to describe the fragmentation, amplification, labeling, and hybridization biases and the bias through unspecific binding in the ChIP-chip measurements in order to understand the effects of different normalization procedures. We found that, using either the mock IP or the matched input as reference signal, one can only correct for some of these effects. Based on our model, we derive a combined normalization using both mock IP and input signal that should correct for all these sources of bias. Crucial to our method is that our mock IP employs the same antibody as the factor IP, but using a wild type yeast strain with untagged protein. In this way, we measure unspecific binding of the TAP tag-directed antibody that can be used for subtracting the unspecific binding component in the factor IP.

Let x be the genomic coordinate, $B(x)$, $M(x)$, and $S(x)$ the array signals obtained from hybridizing the input, the mock IP, and the factor IP, respectively, and $p(x)$ the occupancy profile due to the specific binding of the antibody to the factor of interest. We can model these array signals by

$$B(x) = a_B \times b(x),$$

$$M(x) = a_M \times b(x) \times u(x),$$

$$S(x) = a_S \times b(x) \times (u(x) + p(x)),$$

with a_B, a_M, and a_S unknown scaling constants for the input, mock IP, and factor IP array measurements, respectively, $b(x)$ the input intensity profile describing the fragmentation, amplification, labeling, and hybridization biases, and $u(x)$ the profile describing the effect of unspecific binding of the antibody to other DNA-bound proteins and protein complexes. We seek to obtain the occupancy profile $p(x)$. Using normalization with the input, we would get

$$S(x) / B(x) = (a_S / a_B) \times (p(x) + u(x)),$$

showing that the unspecific binding of the antibody will lead to distortions of the ChIP enrichment signal which will be the more serious the less specific the antibody binds to the factor and the less sequence specificity the factor has in binding to the genomic DNA. Using normalization with the mock IP, we would obtain

$$S(x) / M(x) = (a_S / a_M) \times (1 + p(x) / u(x)),$$

which should work better than the previous version in cases where unspecific binding dominates the signal from the specific binding, but which introduces a bias through sequence-specific effects of the unspecific binding (e.g. through a nucleosome density-mediated GC bias of the unspecific binding signal).

We therefore introduce a combined normalization using both the input and mock IP signals:

$$(S(x) - (a_S / a_M) \times M(x)) / B(x) = (a_S / a_B) \times p(x).$$

If we assume that X% of the genomic regions do not bind the factor of interest, we can estimate the factor a_S/a_M as the global (X/2)% quantile of $S(x)/M(x)$,

$$a_S / a_M = Q_{(X/2)\%}[S(x) / M(x)],$$

since with this choice (X/2)% of the values of the normalized signal ($S(x)$ - (a_S/a_M) × $M(x)$)/$B(x)$ will be below zero. Note that, in practice, the value of a_S/a_M depends only weakly on the value of X and can be estimated to much better than a factor 2. Only when severely overestimating it (by more than a factor of 2) will the correction of unspecific binding be detrimental.

To be able to give absolute occupancy values on a scale between 0% and 100%, we need to estimate the factor a_S/a_B. For this purpose, we assumed that the highest occupancies of our measured factors correspond to 100% occupancy. We estimate the highest genome-wide occupancies using the Y% quantile instead of the genome-wide maximum probe signal to obtain an estimation that is robust against statistical noise. In this study, we chose a quantile of Y=99.8% for all factors, which corresponds to the

highest-bound ~6000 probes on our tiling arrays. Then, a_S/a_B is estimated as the Y% quantile of the factor occupancy:

$$a_S / a_B = Q_{Y\%}[(S(x) - (a_S / a_M) \times M(x)) / B(x)]$$

Our normalization method should improve on the normalization procedures commonly used by correcting the signal for unspecific binding. Our method cannot correct for sequence dependent effects of cross-linking efficiency. These effects represent, however, an inherent limitation of ChIP-chip and ChIP-seq techniques.

We used the combined mock IP plus input normalization method to calculate occupancy profiles for all factors except the CTD phospho-isoforms (S2P, S5P, S7P). Since for these no mock IP measurements with the same antibody as used for the IP could be done, they were normalized simply by dividing through genomic input intensities.

Calculation of transcript-wise occupancy profiles

In order to calculate occupancy profiles over genes or other genomic features the normalized occupancy signal at each nucleotide of the region was calculated as the median signal of all probes overlapping this position (6.5 probes on average). Individual probe intensities were further smoothed using the sliding window smoothing procedure (window half size of 75 bp) implemented in the R package *Ringo*[13].

Gene selection to calculate gene-averaged median profiles

We start with all nuclear Saccharomyces cerevisiae S288C protein-coding genes classified as 'verified' or 'uncharacterized' by the Saccharomyces Genome Database[14] (5769 genes). To align gene profiles across entire transcripts, only genes with available TSS and pA assignments from RNA-seq experiments[15] were taken into account (4366 genes). Genes with TSS (pA) measurements downstream (upstream) of the annotated ATG (Stop) codon were excluded. To remove possible wrongly annotated TSSs and pAs, we only included genes with TSS (pA) annotations showing a distance less than 200 bp to the corresponding downstream (upstream) ATG (Stop) codon (3448 genes). As a result of the limited ChIP-chip resolution and the compactness of the yeast genome with its short intergenic regions (median inter-ORF length: 368 bp, median inter-transcript length: 259 bp), a gene's factor occupancy profile can have spurious contributions from flanking genes. To minimize these “spill-over” effects, we focused on genes exhibiting a minimal ORF and transcript distance to flanking genes of 250 bp and 200 bp, respectively (1786 genes). Furthermore, we restricted our analysis to the 50% highest expressed nuclear protein-coding genes according to measurements by Dengl et al.[1] (1140 genes, ALL set). We grouped genes into four ORF length classes: Xtremely Short (XS) ranging from 256 to 511 bp, Short (S) 512 to 937 bp, Medium (M) 938 to 1537 bp, and Long (L) 1538 to 2895 bp, comprising 93, 266, 339, and 299 genes, respectively (Supplementary Fig. 9a). Profiles within these groups were scaled to median gene length. We calculated gene-averaged profiles by taking the median over

gene factor profiles. For facilitating the comparison of elongation factor occupancies and in particular their slope in the region near the TSS, we shifted the traces in Fig. 1e and f by up to 0.1 on the occupancy scale such that they overlapped in the region [TSS-250,TSS].

Pairwise profile correlations and correlation network

Analyses were done using 4366 genes with available TSS and pA annotations[15]. Pairwise Pearson correlations over factor occupancy profiles were calculated between concatenated gene profiles ranging each from TSS-250 bp to pA+250 bp. The correlation-based network was calculated using the GraphViz's Neato algorithm[16] employing an edge-weighted, spring-embedded layout procedure attempting to minimize a global energy function, which is equivalent to statistical multi-dimensional scaling.

Singular value decomposition

For each of the nine elongation factors *f* and for each of the 4366 genes *g* for which we had TSS and pA annotations[15], we calculated 90%-quantiles of occupancies within a region [TSS-250bp, pA+250bp] as a robust proxy for peak occupancies. This resulted in a 9×4366 matrix. From each matrix element, we subtracted the average over its row, i.e., over its factor. The resulting matrix X_{fg} was subjected to SVD, yielding singular values $\sigma_1 \geq ... \geq \sigma_9 \geq 0$ and unit-length, orthogonal, singular vectors $u_1,..,u_9$, $v_1,..,v_9$, such that $X_{fg} = \Sigma_{i=1,..,9}\ \sigma_i \times u_{if} \times v_{ig}$. The *k*'th term in this sum, $\sigma_k \times u_{kf} \times v_{kg}$, can explain a fraction $\sigma_k^2 / \Sigma_i\ \sigma_i^2$ of the data variance. For Fig. 2c (right), we repeated the SVD analysis with all 15 factors (3 initiation factors, Cet1, Rpb3, 9 elongation factors, and Pcf11) and, for Fig. 2d, we used the set of Cet1, Rpb3, S7P, S5P, S2P, 9 elongation factors, Pcf11.

To reveal correlations contained in the 14.4% of the total variance that was not contributed by the first term in the SVD (Fig. 2d), we subtracted from the data matrix the values from the first term of the SVD, i.e., $X_{fg} - \sigma_1 \times u_{1f} \times v_{1g}$. This resulted in the matrix of residual correlations shown in Fig. 2d. To ensure that the residual correlations were not caused by spill-over effects among neighboring genes, we used a very stringently filtered set of 97 spatially well-separated genes. First, we demanded that the distances to the nearest 'verified' or 'uncharacterized' nuclear ORF, snoRNA, snRNA, ncRNA, CUT or SUT (according to [14] and [17]) be at least 500bp. Second, we used only genes whose neighboring genomic transcripts were both annotated to be transcribed from the same strand. In this way, we made sure that TSSs and pAs of neighboring transcripts were well separated. The resulting matrix of residual correlation is very similar to the one shown in Fig. 2d (Supplementary Fig. 8b bottom), confirming the validity of the analysis on the full set of 4366 genes. We determined the standard errors of each of the correlation coefficients by taking bootstrap samples from the columns of matrix $X_{fg} - \sigma_1 \times u_{1f} \times v_{1g}$ and obtained errors of ±0.047 and below. Hence, the residual correlations are not caused by statistical noise but rather mirror actual physical and functional associations.

Supplementary References

1. Dengl, S., Mayer, A., Sun, M. & Cramer, P. Structure and in Vivo Requirement of the Yeast Spt6 SH2 Domain. *Journal of Molecular Biology* **389**, 211-225 (2009).
2. Aparicio, O. et al. Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo. in *Current Protocols in Molecular Biology* (John Wiley & Sons, Inc., 2005).
3. Chapman, R.D. et al. Transcribing RNA Polymerase II Is Phosphorylated at CTD Residue Serine-7. *Science* **318**, 1780-1782 (2007).
4. Kim, M., Suh, H., Cho, E.-J. & Buratowski, S. Phosphorylation of the Yeast Rpb1 C-terminal Domain at Serines 2, 5, and 7. *Journal of Biological Chemistry* **284**, 26421-26426 (2009).
5. Fan, X., Lamarre-Vincent, N., Wang, Q. & Struhl, K. Extensive chromatin fragmentation improves enrichment of protein binding sites in chromatin immunoprecipitation experiments. *Nucl. Acids Res.* **36**, e125- (2008).
6. O'Geen, H., Nicolet, C., Blahnik, K., Green, R. & Farnham, P. Comparison of sample preparation methods for ChIP-chip assays. *Biotechniques* **41**, 577-580 (2006).
7. David, L. et al. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5320-5325 (2006).
8. Puig, O. et al. The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification. *Methods* **24**, 218-229 (2001).
9. R Development Core Team. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, Vienna, Austria, 2009).
10. Gentleman, R. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80 (2004).
11. Zacher, B. & Tresch, A. Starr: Simple Tiling ARRay analysis of Affymetrix ChIP-chip data. *Accepted by BMC Bioinformatics* (2010).
12. Johnson, W.E. et al. Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences* **103**, 12457-12462 (2006).
13. Toedling, J., Sklyar, O. & Huber, W. Ringo - an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* **8**, 221 (2007).
14. SGD project. Saccharomyces Genome Database.
15. Nagalakshmi, U. et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344-1349 (2008).
16. Gansner, E. & SC, N. Improved force-directed layouts. Vol. 1547 364-373 (Springer, 1998).
17. Xu, Z. et al. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033-1037 (2009).
18. Longtine, M.S. et al. Additional modules for versatile and economical PCR-based gene deletion and modification in <I>Saccharomyces cerevisiae</I>. *Yeast* **14**, 953-961 (1998).
19. Ihmels, J. et al. Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**, 370-377 (2002).
20. Hesselberth, J.R. et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Meth* **6**, 283-289 (2009).

21. Ghavi-Helm, Y. et al. Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription. *Genes & Development* **22**, 1934-1947 (2008).
22. Raha, D. et al. Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proceedings of the National Academy of Sciences* **107**, 3639-3644 (2010).
23. Jasiak, A.J. et al. Genome-associated RNA Polymerase II Includes the Dissociable Rpb4/7 Subcomplex. *Journal of Biological Chemistry* **283**, 26423-26427 (2008).