Supplemental information for : “CpG islands and GC content dictate nucleosome depletion in a transcription independent manner at mammalian promoters”

Authors: Romain Fenouil[1-3,6], Pierre Cauchy[1-4,6], Frederic Koch[1-3,6,7], Nicolas Descostes[1-3], Joaquin Zacarias Cabeza[1-3], Charlène Innocenti[1-3], Pierre Ferrier[1-3], Salvatore Spicuglia[1-3], Marta Gut[5], Ivo Gut[5], & Jean-Christophe Andrau[1-3]

[1] Centre d'Immunologie de Marseille-Luminy (CIML), Aix-Marseille University, UM2, Marseille, France,

[2] Institut National de la Santé et de la Recherche Médicale (INSERM), U1104, Marseille, France,

[3] Centre National de la Recherche Scientifique (CNRS), UMR7280, Marseille, France

[4] TAGC, Case 928, 163 Avenue de Luminy, 13288 Marseille cedex 09, France.

[5] Centre Nacional D'Anàlisi Genòmica, Parc Científic de Barcelona, Baldiri i Reixac, 4 Torre I, 2a planta, 08028 Barcelona, Spain.

[6] These authors contributed equally to this work

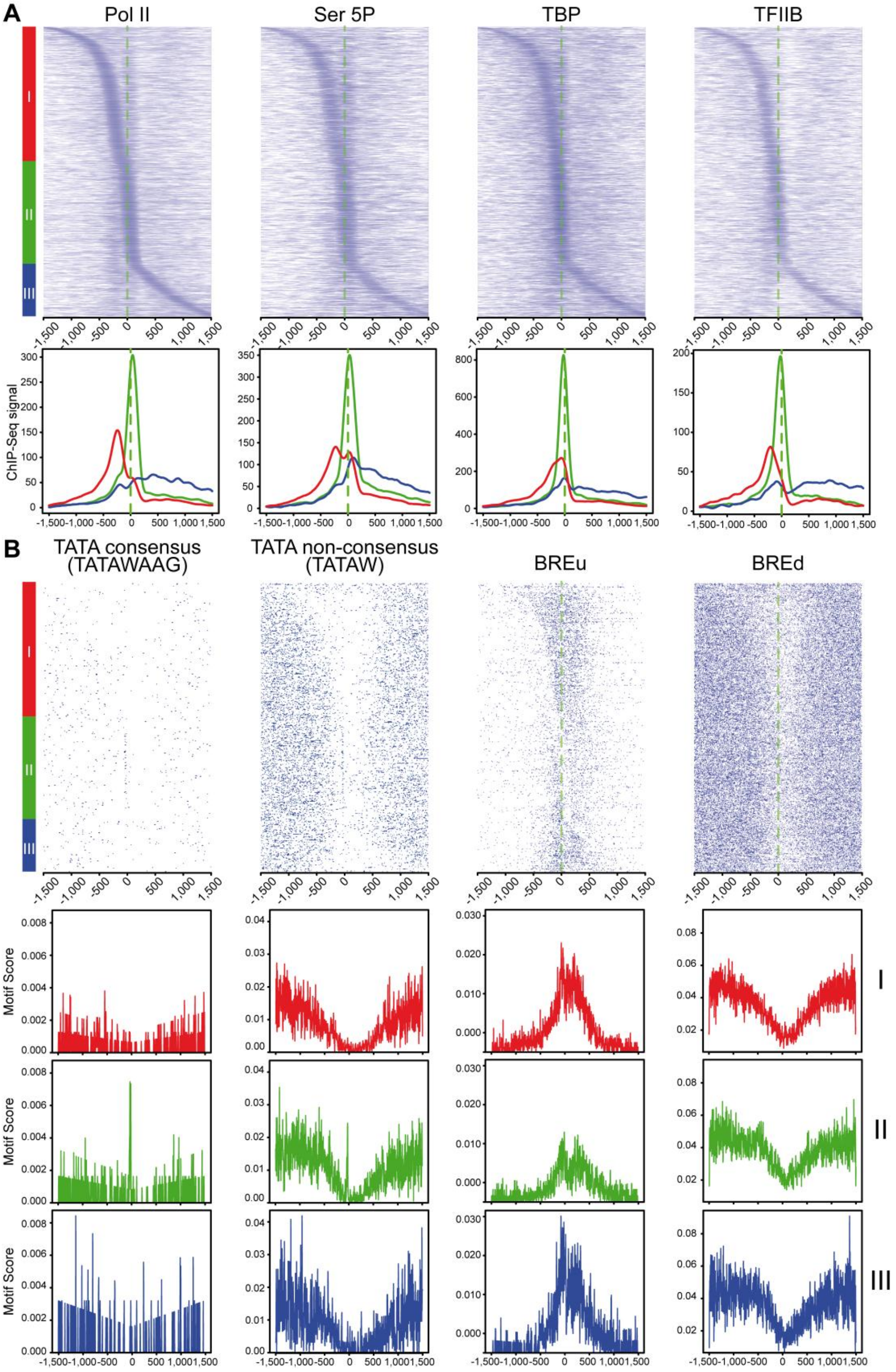
[7] Present address: Department of Developmental Genetics, Max-Planck-Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany

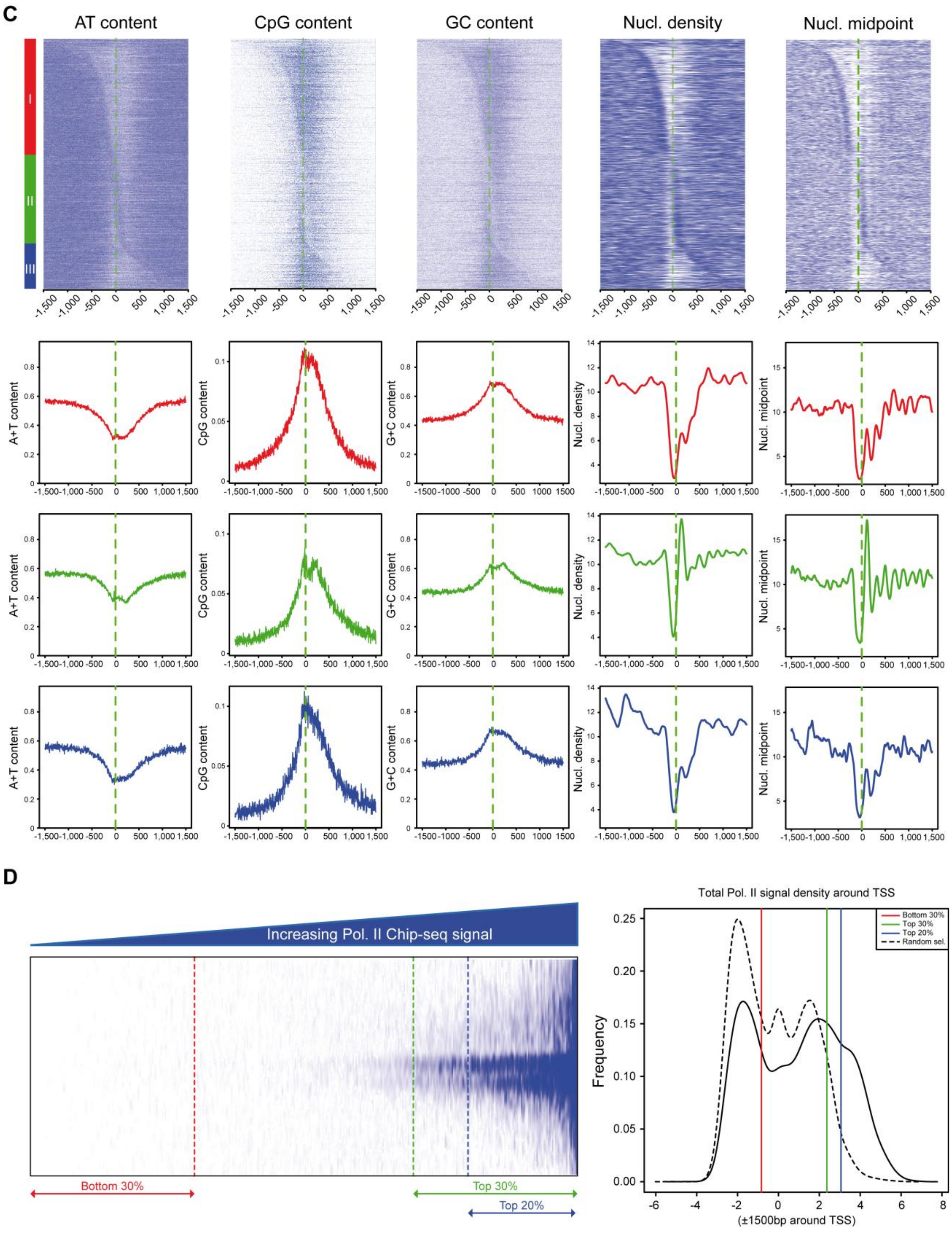
Correspondence should be addressed to J.C.A (andrau@ciml.univ-mrs.fr)

Table of Contents

I. Supplemental Figures

Supplemental Fig. 1

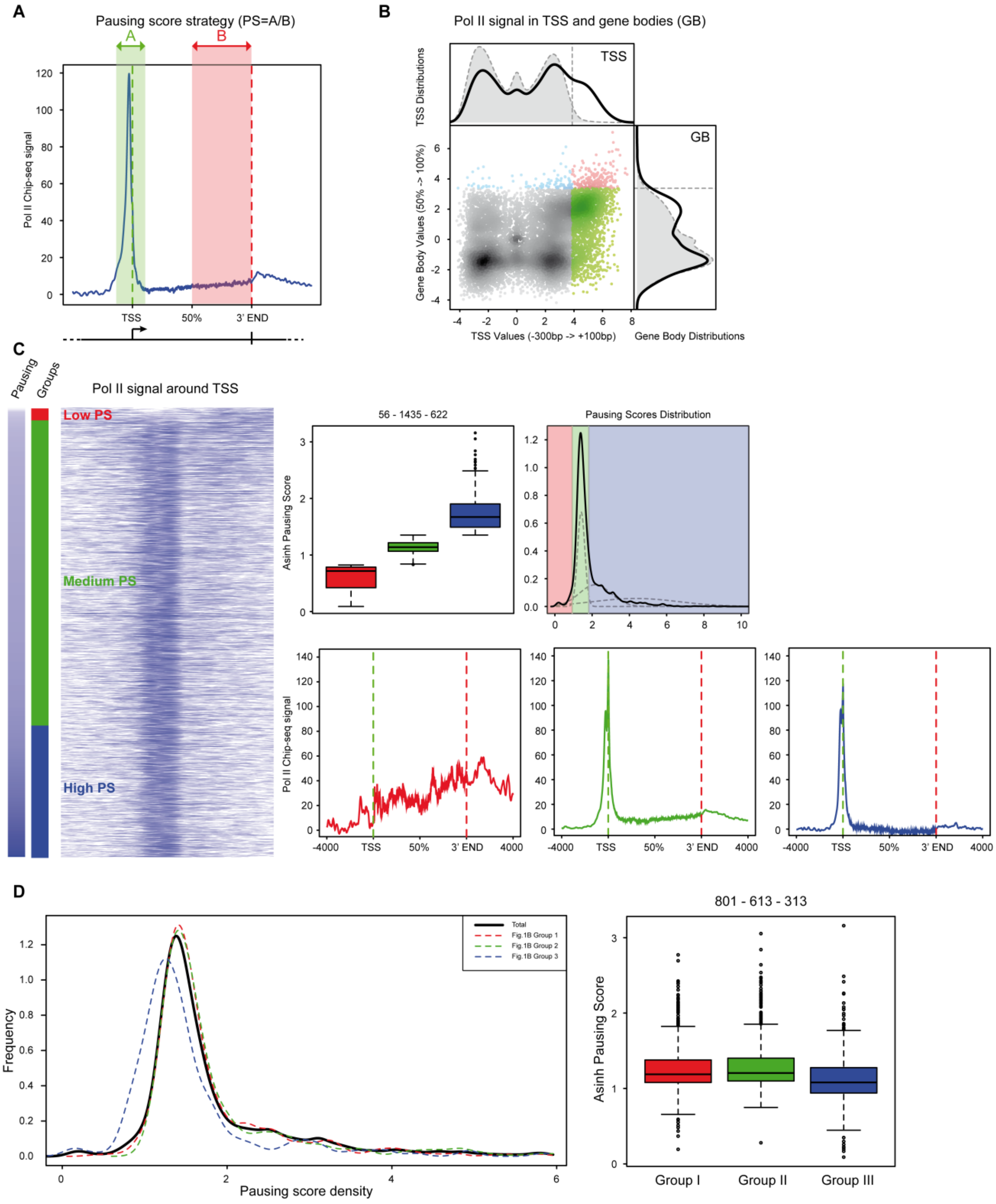



C
AT content
CpG content
GC content
Nucl. density
Nucl. midpoint
I
II
III
-1,500 -1,000 -500 0 500 1,000 1,500
-1500 -1000 -500 0 500 1000 1500
A+T content
CpG content
G+C content
Nucl. density
Nucl. midpoint
0 0.2 0.4 0.6 0.8
0 0.05 0.1
2 4 6 8 10 12 14
5 10 15
D
Increasing Pol. II Chip-seq signal
Bottom 30%
Top 30%
Top 20%
Total Pol. II signal density around TSS
Bottom 30%
Top 30%
Top 20%
Random sel.
Frequency
0.00 0.05 0.10 0.15 0.20 0.25
-6 -4 -2 0 2 4 6 8
(±1500bp around TSS)

Supplemental Fig. 1. Heat maps of intiation complex components, promoter elements and dinucleotides sorted by the main Pol II peak.

Heatmaps corresponding to Pol II peak sorting around TSSs as in Fig. 1B. **A-** Other key components of the initiation complex, including initiating Pol II (phosphorylated on Ser5P of the CTD), TBP and TFIIB are following the general Pol II profile. Average profiles of subgroups are provided underneath the clusters. **B-** Sorting of known core promoter elements motifs, including TATA-boxes, Initiator and BREs (left to right). The average profile of the sequence occurrences in each subgroup is provided underneath. Overall, TATA-boxes, when present, are found just upstream of the TSS in class II promoters, with Pol II located in close proximity. BREu motifs are more spread in classes I and III. **C-** Based on Pol II sorting, the distributions of the AT and GC content and CpG dinucleotide density are provided together with MNase signal represented as nucleosome density or positioning (midpoint). A striking exclusion between CG nucleotides and nucleosome density is observed, indicating that aNDRs are overrepresented in CG-high areas. Furthermore, the most strongly positioned nucleosomes, such as the +1 in class II, appear to be associated to AT-rich sequences instead. Whereas 89 and 82% of promoters in class I and III contain annotated CGIs respectively (as compared to 55% CGI in our starting selection), 72% of those in class II contain CGIs. The average size of CGIs in class II is of 410 bp compared to 617 and 536 bp. **D-** Heat map of Pol II ChIP-seq signal in mouse DP T-cells ranked over all TSSs (-1.5 to +1.5 kb) is shown in the left panel. Selection of Pol II-bound promoters used for analyses shown in Fig. 1, Fig. 2, Supplemental Fig. 1A-C, 2D (Top 20% signal) or Supplemental Fig. 5 (Top and bottom 30% signal) are indicated on the heat map or on the signal distribution plot (right panel). Compared to the signal of random genomic regions (8634 regions of 3 kb long), both top 20% or 30% selections are significant. The bottom 30% Pol II corresponds to the area defined by the first Gaussian of random regions and therefore a confident set of Pol II depleted regions.

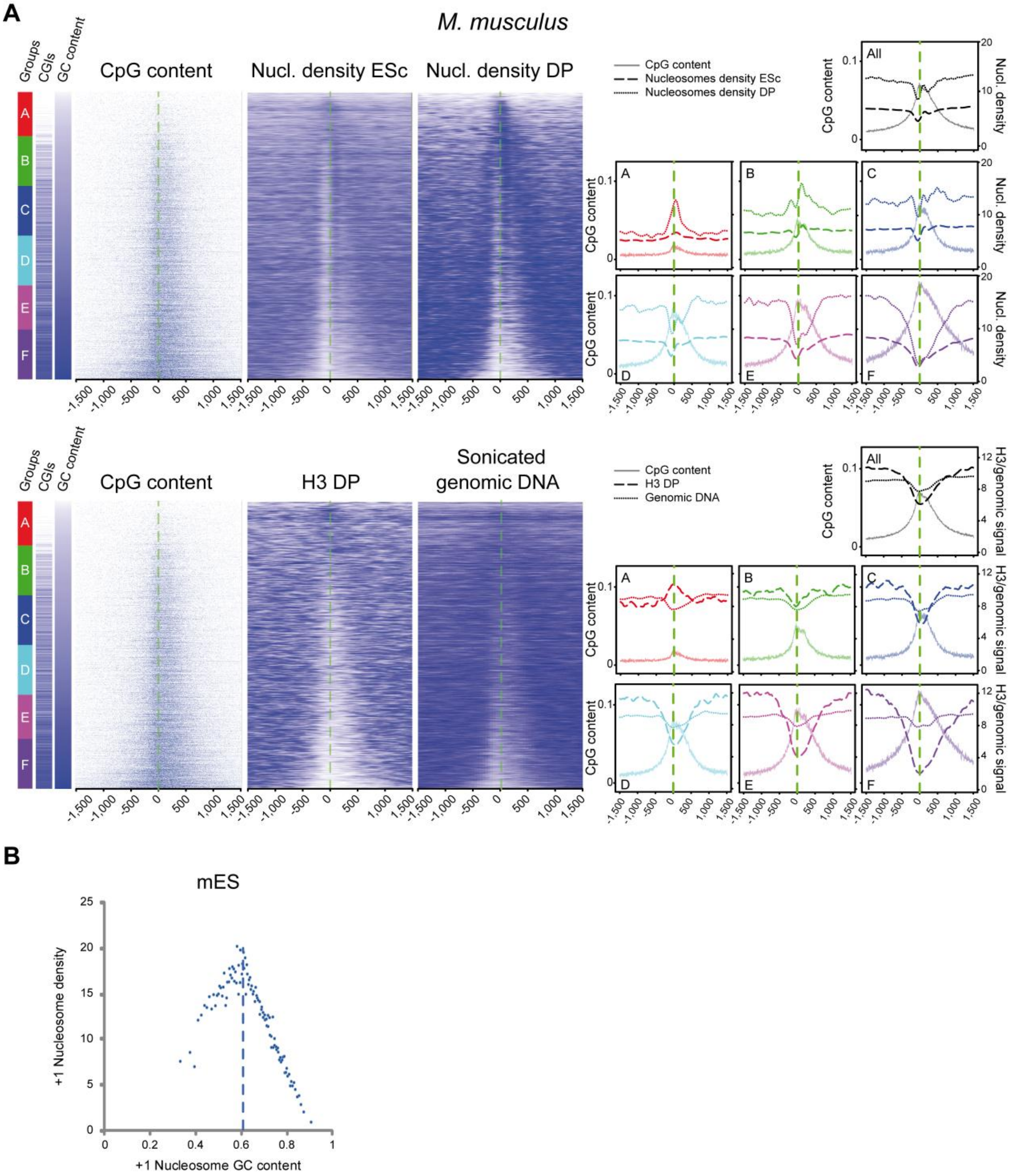
Supplemental Fig. 2



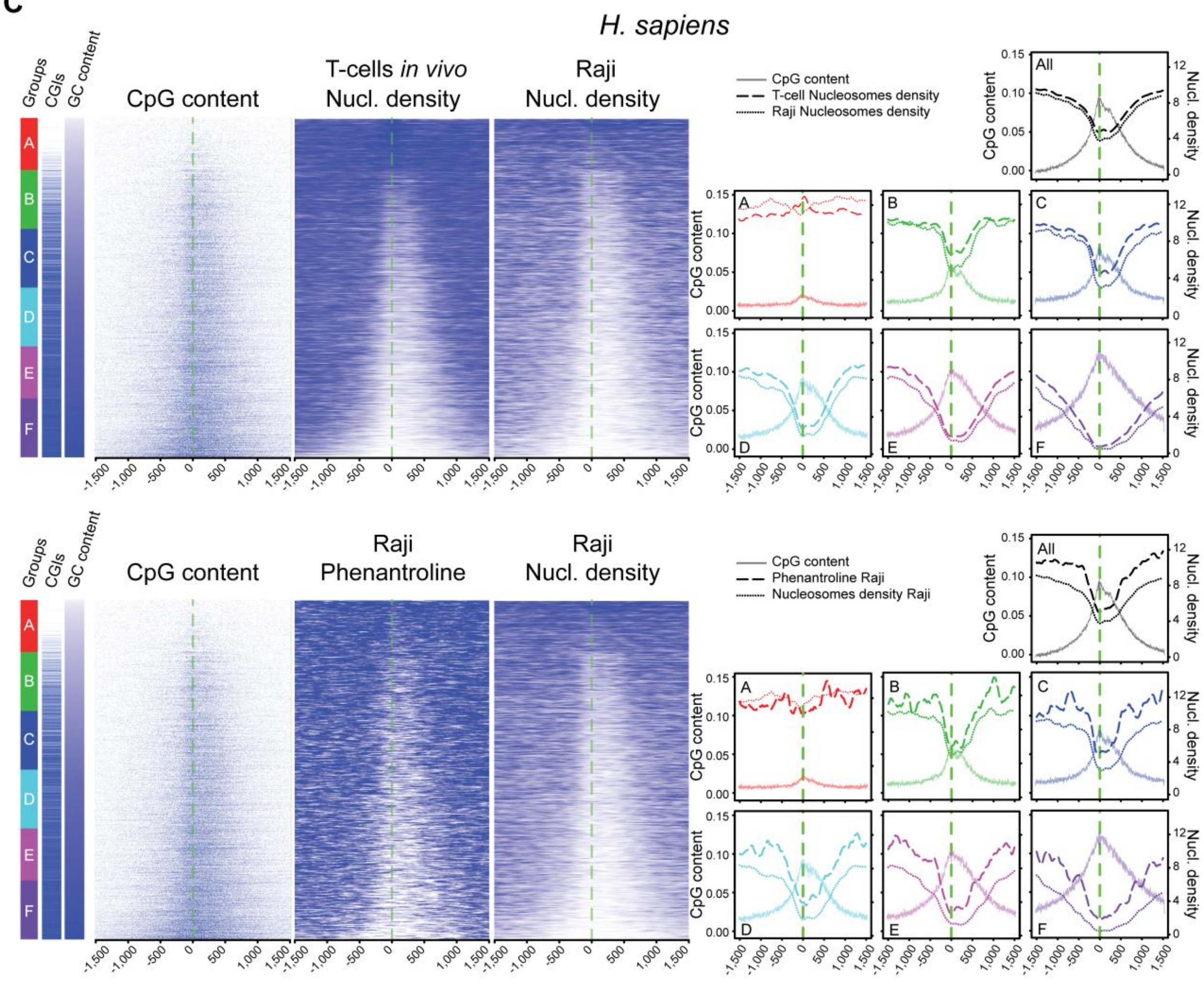
Supplemental Fig. 2. Determination of promoter pausing scores (PSs) in mouse DP T-cells and classification of group I-III from Fig. 1B

A- Principle of PS calculation. Ratio of Pol II ChIP signal over TSSs (-300+100 area) is divided by the average signal on 50-100% of the gene bodies (see Supplemental methods) on significantly bound Pol II genes. **B-** Signal distribution over TSSs (x axis) and gene bodies (Gbs) are shown in comparison with random regions distributions of equivalent sizes (dashed lines and light grey areas). We selected a highly confident set of bound TSSs (after the vertical dashed line on x axis) and GBs (on top of horizontal dashed line on the y axis), that we further merged to establish our model group of genes for PS group determination. **C-** Defining PS groups of genes was performed by plotting the PS distribution and isolating the main Gaussian using a Gaussian fit from Mclust in R (top right panel). This central group (green; 1435 genes) defines genes with medium pausing whereas the area on the left define low or no pausing (light red; 56 genes) and the area on the right (blue; 622 genes) high pausing. Box plots of these 3 groups are shown (top middle panel) as well as their average composite profile over gene region areas (bottom panels). **D-** Determination of pausing categories from promoters isolated in Fig. 1 (groups I, II, III). PSs from Groups I, II and III (isolated in Fig. 1) were calculated based on the strategy described in **A, B** and Supplemental methods. Their values are displayed as distribution curves (dashed lines in the left panel) or box plots (right panel). They allow for classification into medium/high paused for groups I and II and as low paused for group III.

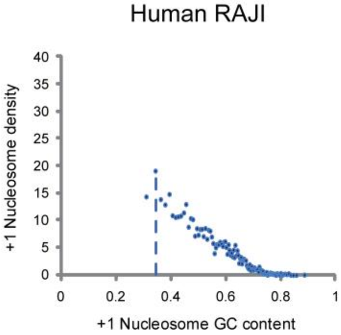
Supplemental Fig. 3



C


H. sapiens
Groups
CGIs
GC content
CpG content
T-cells in vivo Nucl. density
Raji Nucl. density
A
B
C
D
E
F
-1,500
-1,000
-500
0
500
1,000
1,500
CpG content
T-cell Nucleosomes density
Raji Nucleosomes density
All
CpG content
Nucl. density
0.00
0.05
0.10
0.15
0
4
8
12
Raji Phenantroline
Raji Nucl. density
CpG content
Phenantroline Raji
Nucleosomes density Raji

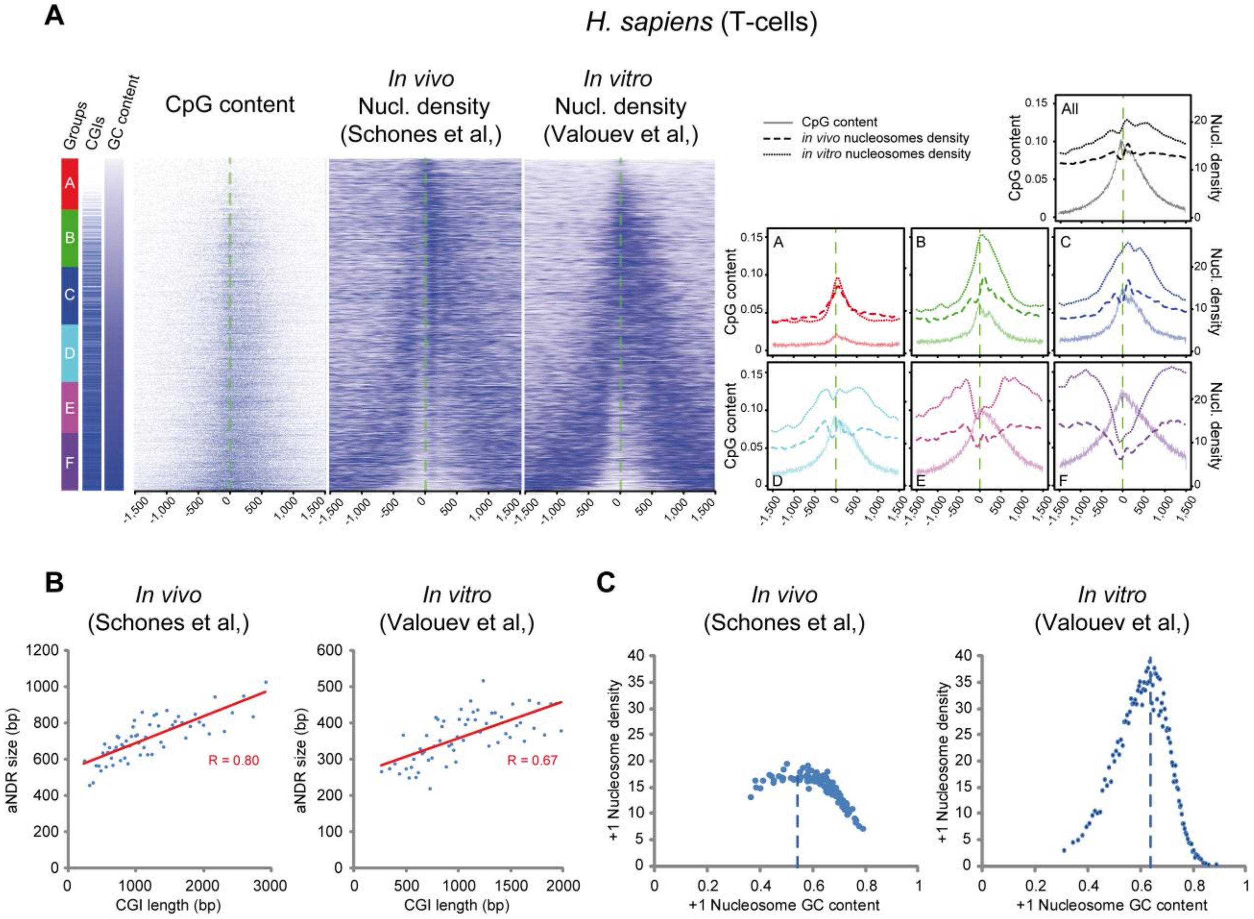
D


Human RAJI
+1 Nucleosome density
+1 Nucleosome GC content
0
5
10
15
20
25
30
35
40
0
0.2
0.4
0.6
0.8
1

Supplemental Fig. 3. Nucleosome profiling based on increasing GC content in several human and murine data sets.

In each panel, the sextile subgroup average profiles are provided to the right, together with CpG content as in Fig. 3. **A-** Comparison of CpG dinucleotide and MNase profiling in murine ES and DP T-cells and H3 ChIP-seq and genomic DNA controls. After a 0.6 GC content threshold, the level of the +1 and overall nucleosome density decreases, whereas the width of the aNDR increases. To rule out possible problems of missing fragile nucleosomes and sequencing artefacts, we performed a histone H3 ChIP on crosslinked material and sequencing of sonicated naked genomic DNA respectively. We observe the same increase in aNDR size in the case of the H3 ChIP-Seq. The genomic DNA control does indicate possible sequencing problems in the GC-richest regions that are negligible compare to specific depletion observed with MNase experiments. **B-** As for DP T-cell (Fig. 3C), increasing GC content first correlates then anti-correlates with MNase signal at the +1 nucleosome positions in mES cells. **C-** The results from panel **A** were further confirmed in the human Raji B-cell line as well as in human primary T-cells. While the overall +1 nucleosome levels appear less pronounced in human cells, the correlation of GC content and aNDR size still holds true. In order to rule out potential MNase sequence bias, we performed the experiment using a chemical cleavage reagent (1,10-phenantroline) and confirmed our results. **D-** Correlation (as in Fig. 3C and panel **B**) from data presented in panel **C.**

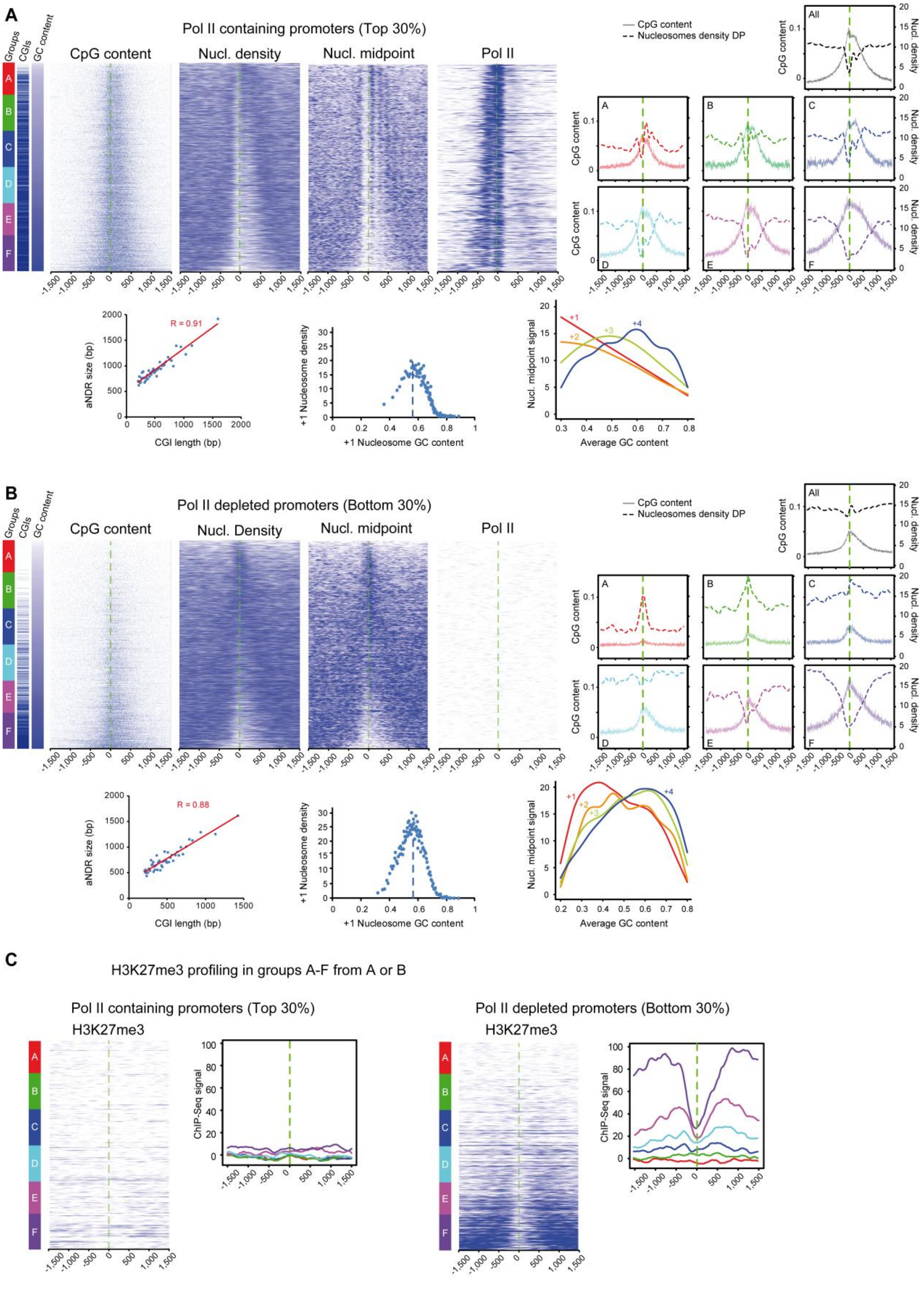
Supplemental Fig. 4



Supplemental Fig. 4. Additional comparison of nucleosome profiling from in vivo and in vitro published datasets.

Subgroup average profiles are provided to the right, together with CpG content as in Fig. 3, 4 and supplemental Fig. 3. **A-** CpG dinucleotide and MNase clustering on published datasets as in Fig. 4 but in comparison with an independent dataset (*in vivo* Schones et al.). This latter gives a more resolutive view and confirms the nucleosomal trend observed in the manuscript. **B** - Correlation of CpG length and aNDR as in Fig. 4. **C** - As for Fig. 4, increasing GC content first correlates then anti-correlates with MNase signal at the +1 nucleosome positions, this trend inversion occurring in the same GC content range (0.5 - 0.6).

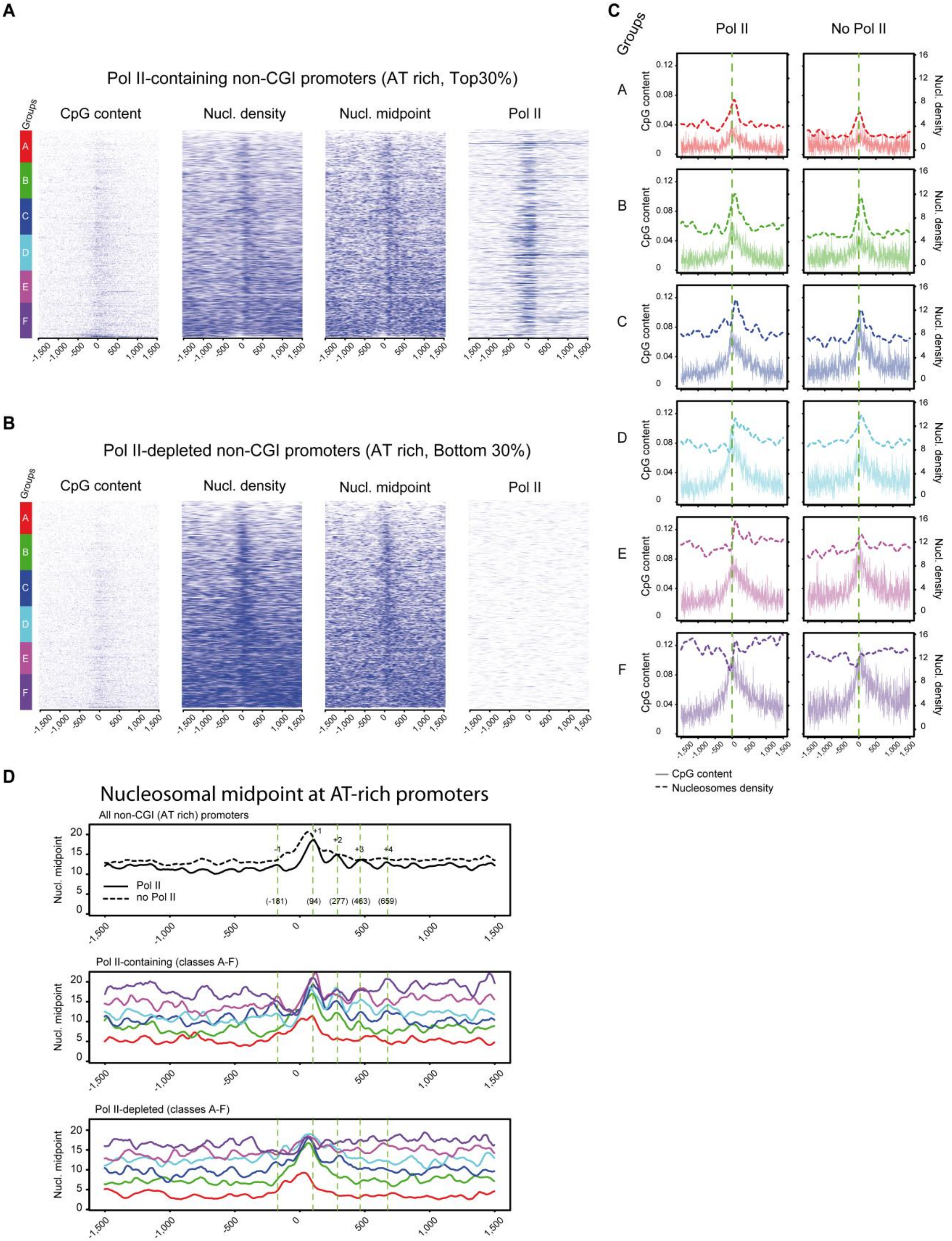
Supplemental Fig. 5



Supplemental Fig. 5. Influence of Pol II occupancy and GC content at promoters on aNDRs and nucleosome positioning.

Heat maps of CpG content and nucleosome density and their corresponding average profiles are shown from left to right on the top and correlations as described in Fig. 3B-D in the bottom for panels **A** (Pol II-containing promoters as shown in Fig. 3) and **B** (Pol II-depleted promoters as shown in Fig. 3). Gene groups are divided into sextiles by increasing GC content (red to violet boxes). Pol II-containing genes show higher GC content as well as more genes exhibiting aNDRs, as compared to Pol II-depleted genes. The number of CGI-containing promoters (blue bar, left) is higher in Pol II-containing genes than in Pol II-depleted genes. The percentages of CGI promoters in the 6 groups of Pol II^{high} are 53, 78, 86, 90, 92 and 97% and for the Pol II^{low} 0.5, 8, 24, 42, 56 and 90% from top to bottom. In Pol II^{high} genes, a transition is observed for the major nucleosome occupancy from the +1 and +2 to the +3 and +4 nucleosomes (as shown in Fig. 3D) while GC content increases. This trend is not observed in Pol II^{low} genes. Moreover, the decrease of +1 and +2 nucleosomes in the Pol II^{high} group is dramatically more marked as compared to Pol II^{low} group. **C-** Heat maps and profiling of H3K27me3 repressive mark deposited by the PcG complex in Pol II-containing and depleted groups. High H3K27me3 are associated with high GC content in Pol II^{low} promoters only. All Pol II^{high} promoters exhibit background levels of H3K27me3. Conversely, Pol II^{low} promoters are characterised by increasing H3K27me3 signal as GC content per group increases.

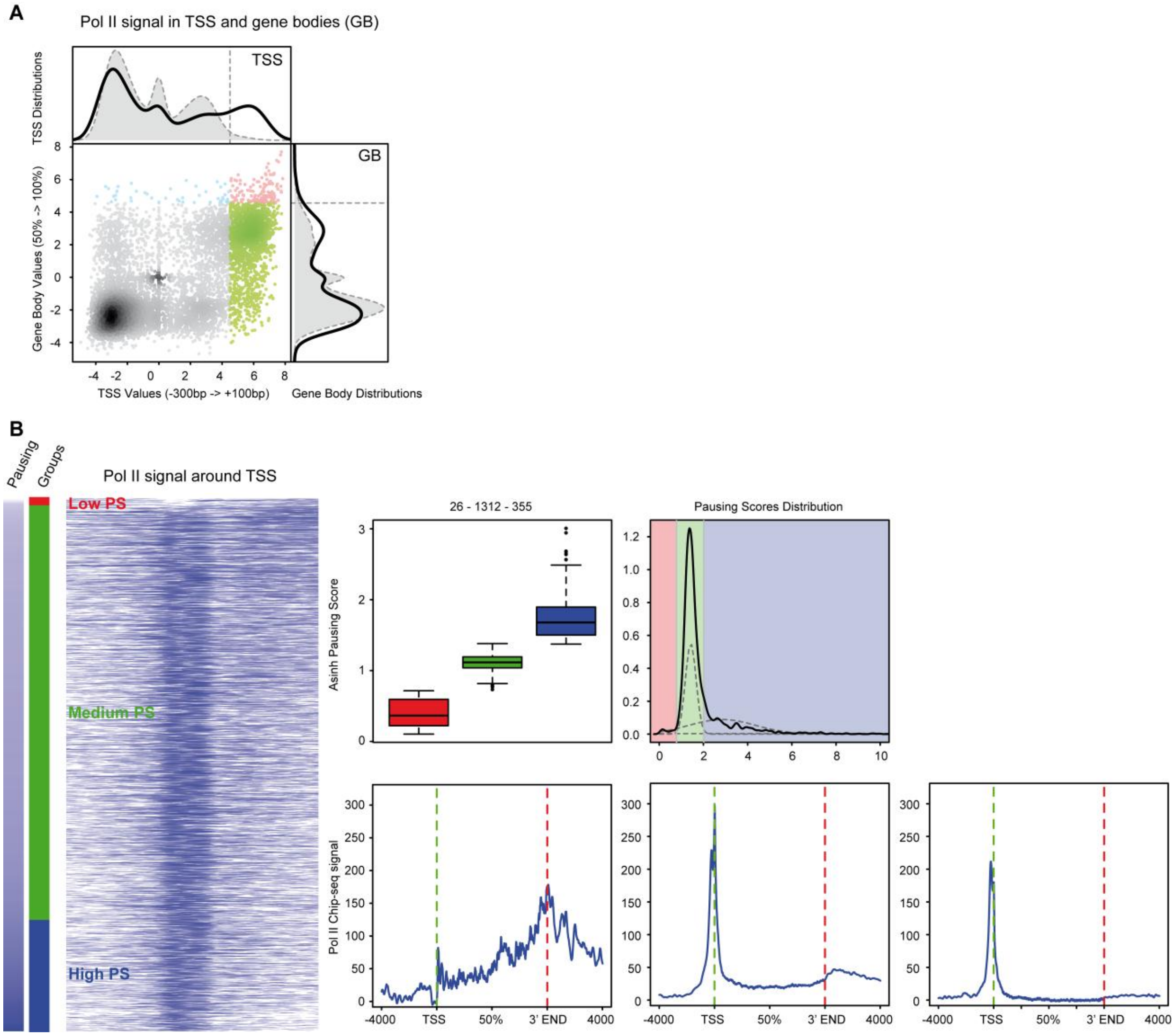
Supplemental Fig. 6



Supplemental Fig. 6. Influence of Pol II occupancy on nucleosome positioning and aNDR at non-CGI (AT-rich) promoters.

Heat maps of CpG content, nucleosome density, nucleosome midpoints ordered by increasing GC content are shown for Pol II-containing (**A**) and Pol II-depleted (**B**) non-CGI promoters only as in Fig. 5. **C-** Corresponding average profiles of subgroups A-F (genes sextiles from lowest to highest GC content) for nucleosome densities and CpG dinucleotide scores are shown. D- Nucleosome positioning as measured by midpoint analysis. Average positions of the -1 to +4 nucleosomes midpoint shown in A and B are indicated on the top panel (all non-CGI promoters). Gene group colours A-F correspond to sextiles of increasing GC content for Pol II-containing promoters (second pannel from the top) and Pol II-depleted (third pannel from the top) promoters. Periodicity in the Pol II-containing group is slightly more pronounced as compared to the Pol II-depleted although to a lesser extent than for CGI promoters from Fig. 5.

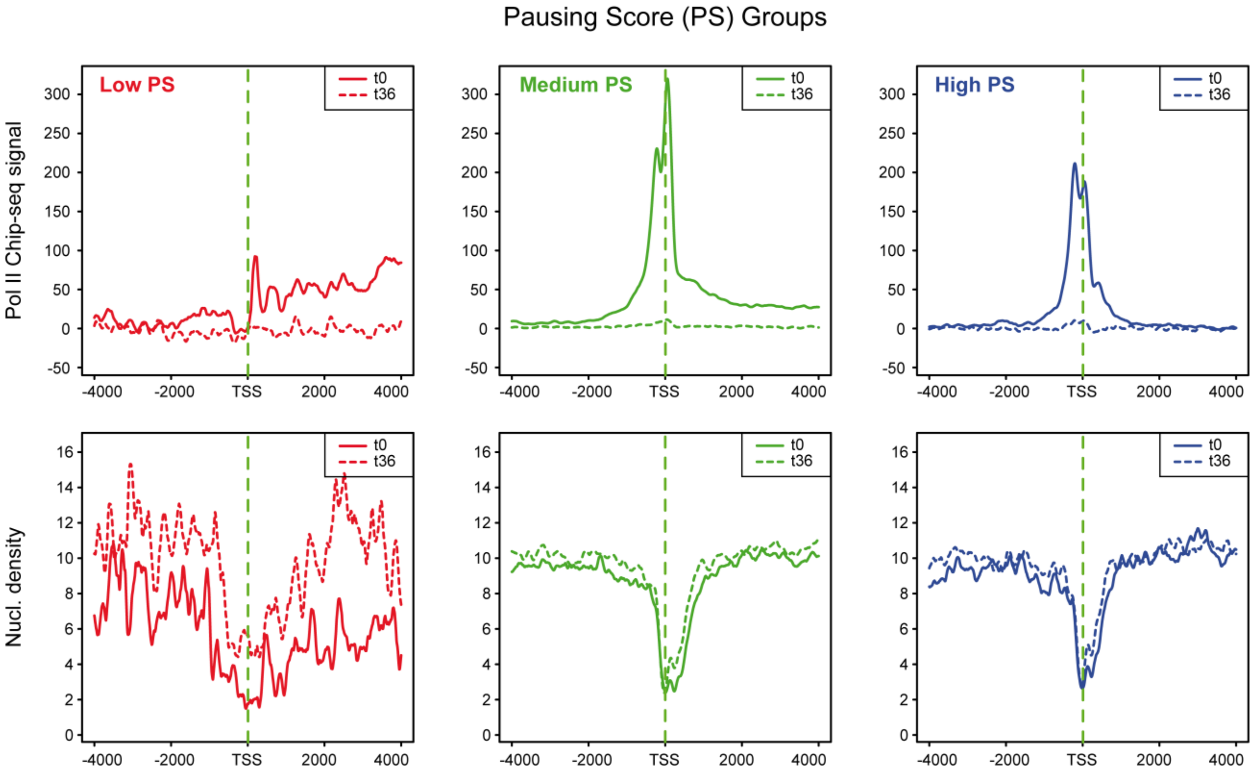
Supplemental Fig. 7



Supplemental Fig. 7. Determination of promoter pausing scores (PSs) in human Raji B-cells.

The same strategy as outlined in Supplemental Fig. 2A was used for PS determination and isolation of pausing groups for promoters used in Fig. 8A. **A-** Signal distribution over TSSs (x axis) and gene bodies (GBs) are shown in comparison with random regions distributions of equivalent sizes (dashed lines and light grey areas). We selected a highly confident set of bound TSSs (after the vertical dashed line on x axis) and GBs (on top of horizontal dashed line on the y axis), that we further merged to establish our model group of genes for PS group determination. **B-** Defining PS groups of genes was performed by plotting the PS distribution and isolating the main Gaussian using a Gaussian fit from Mclust in R (top right panel). This central group (green; 1312 genes) defines genes with medium pausing whereas the area on the left define low or pausing (light red; 26 genes) and the area on the right (blue; 355 genes) high pausing. Box plots of these 3 groups are shown (top middle panel) as well as their average composite profile over gene region areas (bottom panels).

Supplemental Fig. 8



Supplemental Fig. 8. Effect of Pol II disruption *in vivo* on aNDR for different groups of promoters based on pausing.

Effect of pausing on aNDRs after Pol II removal. The different groups of high, medium and low pausing scores determined in Supplemental Fig. 7 were investigated to assess possible differential effects of alpha-amanitin (t36 h) on aNDRs. As seen on the bottom panels (and in Fig. 7) alpha-amanitin and Pol II removal have only a moderate effect on nucleosomal disruption for all pausing groups, showing a comparable aNDR after treatment.

II. Supplemental Table

	ChIP antibodies and conditions used							
Cells	Experiment	Antibody (clone)	Origin	Reference	Cells	Antibody / beads	Washes (RIPA/TE)	Approx. DNA quantity
Mouse DP	H3K27me3	H3K27me3	mouse monoclonal	Abcam (ab6002)	5x10^{6}	2µg / 20µl	8x/1x	200ng
	total histone H3	histone H3	rabbit polyclonal	Abcam (ab1791)	5x10^{6}	2µg / 20µl	8x/1x	100ng
Mouse ES	Pol II	total (N-20)	rabbit polyclonal	Santa Cruz (sc-899x)	1x10^{8}	40µg / 400µl	8x/1x	10ng
Raji WT	Pol II	total (N-20)	rabbit polyclonal	Santa Cruz (sc-899x)	1x10^{8}	40µg / 400µl	8x/1x	22ng
Raji -amanitin (24h)	Pol II	total (N-20)	rabbit polyclonal	Santa Cruz (sc-899x)	2.5x10^{7}	5µg / 50µl	8x/1x	1ng
Raji -amanitin (36h)	Pol II	total (N-20)	rabbit polyclonal	Santa Cruz (sc-899x)	2.5x10^{7}	5µg / 50µl	8x/1x	1ng

Supplemental Table 1. Summary of ChIP conditions and DNA quantities obtained.

III. Supplemental Methods

1. Cells and growth conditions

Raji cells were grown in RPMI 1640 medium supplemented with 10% fetal calf serum (FCS), 100U/ml penicillin, 100mg/ml streptomycin and 2mM L-glutamine (Gibco/Invitrogen) at 37°C and 5% CO2. For the removal of Pol II, cells were treated with 2.5 µg/ml of α-amanitin for the indicated time points for the WT Raji cells. We checked for cell survival that was still above 90% that was still above 90% after 36h of treatment. For primary thymocytes, thymuses were extracted from wild-type C57BL/6 mice (5-6 weeks) and kept on ice in RPMI 1640 medium (Invitrogen, USA) with 10% FCS for a maximum of one hour. Animals were sacrificed in accordance with our institutional as well as European guidelines. Single cell suspensions were obtained by filtering through a 70µm nylon mesh (BD Biosciences, USA). Murine Bruce4 ES cells (Bl6) were grown as described(*1*) on a layer of mouse embryonic fibroblasts feeder cells (PMEF) in the presence of leukemia inhibitory factor (LIF) on gelatinised plates.

2. Crosslink (ChIP-Seq) and cell sorting

Raji cells were directly crosslinked in 25ml of growth medium and thymocytes were resuspended in 25ml of DPBS. ES cells were trypsynised and separated from feeders prior to crosslinking. For RNA-seq experiments, the crosslinking and quenching steps were omitted. Crosslinking was performed with the addition of 1/10th volume of crosslinking solution (11% formaldehyde, 100mM NaCl, 1mM EDTA pH 8, 0.5mM EGTA pH 8, 50mM Hepes pH 7.8) for a final formaldehyde concentration of 1% for 10 minutes at room temperature. The reaction was quenched with the addition of 250mM glycine and incubation at room temperature for 5 minutes. Cells were washed twice with cold DPBS and counted using a Vi-CELL cell counter (Beckman Coulter, France). Cells were snap-frozen in liquid nitrogen in aliquots of $5x10^7$cells/pellet.

T-cells were sorted using the AutoMACS (Miltenyi, France) system. After labelling with anti-CD8 antibodies coupled to PE (Miltenyi, France), cells were incubated with anti-PE multisort magnetic beads (Miltenyi, France) according to manufacturer’s instructions. Sorting was performed using a single positive selection. Cells were released and incubated with anti-CD4 magnetic beads (Miltenyi, France) according to manufacturer’s instructions. After a subsequent single positive selection, sorting efficiency was verified through flow cytometry analysis. Sorted CD4+/CD8+ thymocytes were counted and snap-frozen in liquid nitrogen in aliquots of $5x10^7$ cells/pellet for further processing.

3. Chromatin Preparation (ChIP)

EDTA-free protease inhibitors (Roche, France) were added to all washing buffers to a final concentration of 1x together with 0.2mM PMSF and 1µg/ml pepstatin. Aliquots of 5x10^7 cells were lysed after resuspension in 2.5ml LB1 (50mM Hepes pH 7.5, 140mM NaCl, 1mM EDTA pH 8, 10% glycerol, 0.75% NP-40, 0.25% Triton X-100) for 20 minutes on a rotating wheel at 4°C. Chromatin was collected by centrifugation for 5 minutes at 4°C and 1.350g in a tabletop centrifuge, and washed in 2.5ml LB2 (200mM NaCl, 1mM EDTA pH 8, 0.5mM EGTA pH 8, 10mM Tris pH 8) on a rotating wheel for 10 minutes at 4°C. Chromatin was collected, resuspended in 1.5ml of LB3 (1mM EDTA pH 8, 0.5mM EGTA pH 8, 10mM Tris pH 8, 100mM NaCl, 0.1% Na-Deoxycholate, 0.5% N-lauroylsarcosine) and sonicated using a Misonix 4000 (Misonix Inc, USA) sonicator for 10 cycles in the case of thymocytes, 12 cycles for ES cells and 14 cycles in the case of Raji cells (30 seconds on, 30 seconds off, amplitude 40) with the tubes submerged in ice-cold sonication solution (500mM NaCl, 20% ethanol). After the addition of Triton X-100 to a final concentration of 1%, particulates were collected through centrifugation at 20.000g and 4°C for 20 minutes. For batch processing, supernatants were combined, mixed thoroughly and subsequently aliquoted and snap-frozen in liquid nitrogen. Extracts were kept at -80°C until use and 50µl aliquots were taken to serve as input controls.

Inputs were combined with an equal volume of 2x elution buffer (100mM Tris pH 8, 20mM EDTA pH 8, 2% SDS) and incubated overnight in a water bath at 65°C for 13-15 hours. SDS was then diluted by the addition of an equal volume of TE (10mM Tris pH 8, 1mM EDTA pH 8) and RNA was digested by RNase A at a final concentration of 0.2µg/ml at 37°C for 2 hours. Samples were subsequently Proteinase K treated at 55°C for two hours at a final concentration of 0.2µg/ml. DNA was purified by two subsequent phenol:chloroform:isoamylalcohol (25:24:1, pH 8) extractions and followed by a Qiaquick purification (PCR purification columns, Qiagen, Germany). To ensure greater DNA purity, columns were washed twice with PE buffer and eluted in 50µl H2O. DNA concentration was measured using a Nanodrop 1000 (Thermo Scientific, France) and 500ng were run on 2% agarose gels to verify sonication efficiencies or alternatively 2ng using High Sensitivity DNA chips on a 2100 Bioanalyzer.

4. ChIP-Seq

All experiments were performed using Dynabeads (Invitrogen, USA) coated with Protein-G. For biological replicates, all steps were repeated using a biologically independent sample as described above. Beads were washed 3x with 1ml and subsequently resuspended in 250µl of blocking solution (0.5% BSA in 1x DPBS). After the addition of the antibody, the beads were incubated at 4°C overnight on a rotating wheel. Unbound antibodies were removed through three further washes with 1ml of blocking solution. Beads were resuspended in 100µl of blocking solution, extracts were added and the mix was incubated overnight at 4°C on a rotating wheel.

EDTA-free protease inhibitors (Roche) were added to all washing buffers to a final concentration of 1x together with 0.2mM PMSF and 1µg/ml pepstatin. Beads were washed 8 times in RIPA buffer (50mM Hepes pH 7.6, 500mM LiCl, 1mM EDTA pH 8, 1% NP-40, 0.7% Na-Deoxycholate) and once in TE+ (10mM Tris pH 8, 1mM EDTA pH 8, 50mM NaCl). Immunoprecipitated chromatin was recovered from the beads with two subsequent elution steps at 65°C for 15 and 10 minutes in 110µl and 100µl of elution buffer (50mM Tris pH 8, 10mM EDTA pH 8, 1% SDS) respectively. The two eluates were combined and incubated at 65°C overnight (13-15 hours) for crosslink reversal. DNA was purified as described for the input (see **Supplemental Table 1** for a summary of ChIP conditions for each experiment).

Prior to sequencing, ChIP DNA was quantified using the picogreen method (Invitrogen, USA) using input dilutions as standards, or by running 20% of the ChIP material on a High Sensitivity DNA chip on a 2100 Bioanalyzer. At least 1ng of ChIP or input DNA was used for library preparation according to the Illumina ChIP-seq protocol (see **Supplemental Table 1** for quantities obtained for each experiment). After end-repair and adapter ligation, fragments were size-selected (cut) on an agarose gel prior to preamplification and clustering. The size-selected and preamplified fragments were verified on a 2100 Bioanalyzer (Agilent, USA) before clustering and 36 cycle sequencing on a Genome Analyzer II (Illumina, USA) according to manufacturer’s instructions.

5. Total RNA extraction and TSS RNA-Seq

Total RNA was extracted using TRIzol (Invitrogen, USA) according to the manufacturer’s instructions with some modifications to ensure higher recovery rates of small RNAs. In brief, sorted CD4+/CD8+ cells were divided into 1x10^7 aliquots and 1ml of TRIzol was added. Cells were lysed with vigorous vortexing and pipetting. Homogenized samples were incubated at room temperature for 5 minutes and 0.2ml of chloroform was added. Samples were vigorously shaken and incubated at room temperature for an additional 5 minutes. Phase separation was carried out by centrifugation at 4°C and 12,000g in a tabletop centrifuge for 15 minutes. The aqueous phase was transferred to fresh tubes and 1.5 volumes (approximately 1ml) of isopropanol together with 10µg of linear acrylamide (Ambion, USA) were added. Samples were vortexed and incubated at room temperature for 15 minutes. The precipitated RNA was pelleted by centrifugation at 4°C and 12,000g in a tabletop centrifuge for 20 minutes. Pellets were washed with 80% ethanol, vortexed and centrifuged at 4°C and 7,500g in a tabletop centrifuge for 10 minutes. Pellets were allowed to air-dry and resuspended in nuclease-free water (Ambion, USA). DNA was digested using the rigorous Turbo DNase (Ambion, USA) treatment as per manufacturer's instructions. RNA quantity was measured on a Nanodrop 1000 and the quality was verified using RNA Nano chips on a 2100 Bioanalyzer (Agilent, USA).

Small RNAs were purified from a 10% denaturing urea polyacrylamide gel, essentially as described in the Illumina Small RNA Rev. B protocol with some minor modifications. Approximately 10µg of total RNA was run at 200V for 1 hour, until the blue front reached the bottom of the gel. We used 21G needles to puncture holes into the bottom of two 0.5ml and placed them into 2ml eppendorf tubes. RNA corresponding to 15nt-70nt was cut from the gel, diagonally cut in half and separately transferred into the 0.5ml tubes. The gel was crushed into the 2ml tubes by a two minute centrifugation at 14,000rpm. For gel elution by soaking, 0.4ml of 0.3M NaCl was added to each tube, before a 4 hour rotation at room temperature. After removal of gel particles using 0.22µm cellulose acetate filters, 10µg of linear acrylamide (Ambion, USA) and 2.5 volumes (approximately 1ml) of ice-cold absolute ethanol were added. After a 30 minute incubation at -80°C, the eluted RNA was precipitated by centrifugation at 4°C and maximum speed for 45 minutes. The pellet was washed with 1ml of room temperature 80% ethanol and resuspended in 5.7µl of water. This PAGE purification step was repeated after both the 5' and 3' adapter ligations, which were carried out according to the protocol. The resulting cDNA was purified with the QIAquick Gel Extraction kit (Qiagen, Germany) using the DNA cleanup protocol modification in order to retain DNA fragments as low as 70bp. The DNA was quantified using a Nanodrop 1000 (Thermo Scientific, France) and verified using DNA High Sensitivity 2100 Bioanalyzer chips (Agilent, USA). The library was clustered and sequenced using 76 cycles on a Genome Analyzer II (Illumina, USA) according to manufacturer's instructions.

6. MNase-Seq

For sequencing of nucleosomal DNA, $2x10^7$ cells were resuspended in 50µl Solution I (150mM sucrose, 80mM KCl, 5mM K_2HPO_4, 5mM $MgCl_2$, 0.5mM $CaCl_2$, 35mM HEPES pH 7.4) and NP40 was added to a final concentration of 0.2%. Cell membranes were permeabilized for one minute at 37°C. For nucleosomal digestion, either 10U, 20U or 40U, for thymocytes, murine ES cells and Raji cells respectively, of MNase was added with 0.5ml of Solution II (150mM sucrose, 50mM Tris pH 8, 50mM NaCl, 2mM $CaCl_2$) and incubated for 30 minutes at room temperature. The reactions were stopped with the addition of EDTA to a final concentration of 10mM. The cells were lyzed using 1.45ml of SDS Lysis Buffer (1% SDS, 10mM EDTA pH 8, 50mM Tris pH 8), with a 10 minute incubation at 4°C. A 200µl aliquot was taken for purification and the remaining extract was stored at -80°C. at -80C. An equal volume of TE (200ul) was added to the aliquot, followed by subsequent 2 hour treatments with each 0.2ug/ml final concentrations of

RNase A and Proteinase K at 37C and 55C respectively. DNA was extracted by two subsequent phenol:chloroform:isoamylalcohol (25:24:1) extractions, further purified using QIAquick PCR purifications colums (Qiagen, Germany) and eluted in 50µl of water. Nucleosomal digestion was verified by running 500ng of DNA on a 2% agarose gel as well as on DNA high-sensitivity 2100 Bioanalyzer chips (Agilent, USA). After library preparation, DNA fragments corresponding to mononucleosomes were cut from an agarose gel and subsequently clustered and sequenced using 36 cycles on a Genome Analyzer II (Illumina, USA) according to manufacturer's instructions.

7. Nucleosome preparation using Cu-10,1-phenantroline

A total of $2x10^7$ Raji cells was resuspended in Solution I and permeabilized using NP40 as described for MNase digestion. Instead of MNase, we added 80µM $CuCl_2$, 0.4mM 10,1-phenantroline (stock in absolute ethanol), 3.2mM ascorbic acid and 320µM H_2O_2 with Solution II, similar to concentrations used previously(*2*). Reactions were incubated for 2h at room temperature and stopped with the addition of EDTA to a final concentration of 10mM. Cells were then lyzed and purified as for the MNase experiments. Nucleosomal digestion was verified by running 500ng of DNA on a 2% agarose gel as well as on DNA high-sensitivity 2100 Bioanalyzer chips (Agilent, USA). After library preparation, DNA fragments corresponding to mononucleosomes were cut from an agarose gel and subsequently clustered and sequenced using 36 cycles on a Genome Analyzer II (Illumina, USA) according to manufacturer's instructions.

8. Preparation of sonicated genomic DNA for sequencing

Genomic mouse DNA was obtained commercially (Promega, France) and sonicated in TE (100ng/µl) using in a final volume of 50µl for 16 cycles (30s on/30s off) in a Bioruptor (Diagenode, Belgium) on medium power. The resulting sheared DNA displayed an average of 200bp fragments as verified on an agarose gel. The sample was used for library preparation and sequencing as described for the ChIP-Seq and MNase-Seq samples.

9. Western blot analysis of Pol II knock-down (Fig. 7 and supplemental Fig. 8)

To verify the knock-down of Pol II after α-amanitin treatments, we used $1x10^7$ cell/samples at the indicated time points of 12h, 18h, 24h, 36h and 48h. Cells were pelleted and directly resuspended in 1x Laemmli buffer (2% SDS, 80mM Tris pH 6.8, 10% glycerol, 2% β-mercaptoethanol, 0.005% bromophenol blue). Equal protein loading was checked with comassie staining of the membrances and verified by Tubulin western blotting. The chromatin was sheared by repeated snap-freezing using liquid nitrogen and boiling cycles at 95°C. Proteins corresponding to approximately 400,000 cells were separated on 5% PAGE gels and transferred to PVDF membranes using a wet-transfer system (Bio-Rad, USA). Pol II was detected using H224, an antibody against the invariant N-terminal part of Pol II (sc-9001x; Santa Cruz Biotechnology, USA).

10. Downloaded and published data

The CD4+/CD8+ ChIP-seq data sets for total Pol II, Ser5P, TBP and TFIIB were published in our previous study (GEO Accession No. GSE29362)(*3*). For human nucleosomal data, we downloaded T-cell *in vivo* and *in vitro* reconstituted datasets(*4*) (GSE25133) from the Sidow lab and re-aligned the raw tags using Bowtie(*5*) against the human genome (hg18, build 36.1). All new and downloaded data was treated equally using the in-house pipeline (see below).

11. Bioinformatic processing of sequenced tags

A. Pre-processing and pipeline

All DNA Samples were sequenced on an Illumina Genome Analyzer and aligned against the mouse 2007 (mm9, build 37) or human (hg18, build 36.1) genomes using the integrated Eland software. The Eland output files provide tag sequences, scores and coordinates. For experiments with multiple sequencing lanes (technical replicates), these files were directly merged and then processed as described. As pre-filtering steps, only non-ambiguously mapped tags were used for further processing. A threshold based on the number of sequenced tags (Nseq) has been set up as Nseq/7,000,000. All regions with a number of repeated tags (with identical sequences/coordinates) above this threshold were filtered out to remove possible sequencing and/or alignment artefacts. Remaining tags were processed using a custom R pipeline, employing the ShortRead library(*6*).

For TSS RNA-Seq, we obtained the raw sequences and detected and removed the adapter sequences using the MIRO pipeline (http://seq.crg.es/main/bin/view/Home/MiroPipeline). The obtained reads were aligned against the mouse genome using the GEM aligner (http://gemlibrary.sourceforge.net). Only unambiguously aligned tags were used for further analysis.

B. Tag extension and Wiggle file generation

In order to accurately represent and further process the ChIP-seq and input control signals, the Watson and Crick strand tags need to be merged after elongation/size extension to the gel purified fragment size. The optimal elongation size of each ChIP-seq experiment was estimated *in-silico* by employing a step-wise 10bp chromosomal sequence tag shifting and score multiplication. Tag coordinates were subsequently elongated according to this estimated DNA fragment size, corresponding to the tag shift maximizing the score. Then, a nucleotide score representing the genome-wide overlap of elongated tags was computed across both strands. Wiggle files for genome-wide scores were generated following a binning step, calculating the average score every 50bp. In order to score for the nucleosome midpoint, we performed a shift for the extreme 5' and 3' tag scores based on half of the estimated elongation size. The nucleotide score is only taking in account the first base of sequenced tags. For the RNA-Seq experiment, the elongation step was omitted and separate Wiggle files for the Watson and Crick strand were generated using 50bp bins.

C. ChIP-Seq artefact removal, normalization and input subtraction

Due to the size of the genome and relative low frequencies of binding events, we assume that more than 90% of the obtained scores from the ChIP-seq experiments represent a background (BG) level. We therefore used the genome-wide average score in each experiment to estimate the BG level. Using these average scores, it is possible to rescale the scores accordingly, acting as normalization and reducing the inter-experiment differences due to effects of different sequencing depths and/or fragment sizes.

Furthermore we employed an input subtraction step for each experiment using the normalized file for the input control. This not only allows for correction of overrepresentations of certain genomic regions due to possible (un-) favorable events during sonication and/or DNA sequencing, but also serves to reduce the signal/noise ratio especially for experiments with low enrichment values.

12. Bioinformatic analyses

A. Average binding profiles and model gene selection

To generate average binding profiles, *mm9* and *hg18* Refseq genes annotations were used to extract values from wiggle files associated with selected genes/regions. mRNA regions exhibiting a 4kb gene-free region around the boundaries were selected, excluding all genes in the vicinity of other mRNAs, snRNAs, snoRNAs or tRNAs(*7-8*) to avoid ambiguities in data interpretation and analyse a stringent set of promoters. This selection resulted in initial sets of 8634 murine and 7021 human promoters. R scripts were developed and used for retrieving bin scores inside these annotations and in a region of 4kb before and after TSSs. Based on the gene list selections, bin scores from wiggle files were used to re-center values around the TSS of all genes using linear interpolation. In total, 1000 points were interpolated for the TSS of each selected gene in all average profiles presented.

B. Promoter selection based on Pol II signal for clusters analyses

In order to select promoters based on Pol II signal, we retrieved and averaged bin scores in a region from -1500bp to +1500bp around all TSSs. Based on these scores, we selected the promoters falling in the 20 first percentiles (e.g. in the case of Fig. 1 and Supplemental Fig. 1 and Fig. 7) and the ones falling in either first or last 30 percentiles (e.g. in the case of Supplemental Fig. 5) as shown in Supplemental Fig. 1D. A similar strategy was applied for the selected CGI- or non-CGI promoters in Fig. 5 and Supplemental Fig. 6 (first or last 30 percentiles).

C. Definition of GC content, CpG content and CpG islands (CGIs)

G+C content is defined as the frequency of G or C per base-pair, looking on both strands but not redundant. G+C content densities were derived from RSATools results converted and interpolated into 3bp bin matrices, then divided by three to reflect the content per base-pair. These values were used to rank the promoters shown in Fig. 3 to Fig. 5, Supplemental Fig. 3A to Supplemental Fig. 6 although in most of the cases the CpG content (defined below) was displayed in the density profile panels (trends of GC content were however very comparable).

CpG content was defined as the frequency of CG dinucleotide content per base-pair, looking on both strands but not redundant. Two consecutive bases of a CG dinucleotide thus both exhibit a frequency of 1. CpG content densities were derived from RSATools results converted and interpolated into 3bp bin matrices, then divided by three to reflect the content per base-pair. The validity of this approach, its resulting CpG frequencies and their relation to CpG o/e were verified via CpGPlot(*9-10*) where CpG o/e is defined as ((Num of CpG/(Num of C × Num of G)) × Total number of nucleotides in the sequence).

CpG islands were defined using NCBI thresholds of at least 50% G+C content, 0.6 CpG o/e and at least 200bp in length, thus containing 16026 and 28226 CpG islands for *Mus musculus* and *Homo sapiens*, respectively.

D. Pol II peak sorting, GC content sorting and corresponding clusters

For peak distance clustering, local signal maxima were computed in [-1500bp,1500bp] windows around the TSSs of 8634 murine and 7021 human previously selected genes for RNA Polymerase II ChIP-Seq samples. Genes were subsequently ordered by increasing relative distance of ChIP-Seq local maxima to the TSS. For sorting by GC content, the GC content per base was retrieved using RSA Tools Dna-Pattern(*11*) and an average score in a region of 3Kbp around TSS was attributed to each promoter. Promoters of interest were subsequently sorted by increasing values. Corresponding samples were derived by sorting according to reference samples. Total and sextiles subgroup average profiles for reference and corresponding samples were computed using an in-house R pipeline as previously described. Heatmaps were generated viewed and color-scaled according to sample read depth using Java TreeView(*12*). For data presented in Fig. 3, the percentages of CGI-containing promoters in each group A-F from top to bottom are 4, 34, 55, 65, 78 and 93%. For data presented in Fig. 4, the percentages of CGIs-containing promoters in each group A-F from top to bottom are 6, 43, 64, 74, 82 and 92%. For Supplemental Fig. 5, the percentages of CGI-containing promoters in the 6 groups of Pol II^{high} are 53, 78, 86, 90, 92 and 97% and for the Pol II^{low} 0.5, 8, 24, 42, 56 and 90% from top to bottom.

E. CpG, dinucleotide, and promoter element clusters

Dinucleotide content was scored via RSA Tools Dna-Pattern(*11*) using a CG or AT dinecluotide search pattern. Resulting outputs were computed as frequencies every 3bp for all regions, which were subsequently converted to matrices. Similarly, we searched for TATA-box (TATAWAAG or TATAW), BREu (SSRCGCC) or BREd (RTDKKK) motifs(*13*). Clusters were obtained, viewed and identically color-scaled per sample using Java TreeView(*12*). Average profiles were performed on subgroups as described above.

F. CGI and aNDR size correlations

CGI tracks were retrieved from UCSC Genome Browser tables (http://genome.ucsc.edu/hgTables). aNDR intensity was derived by inverse MNase signal, where zero values were replaced by 0.01. Peak detection was subsequently carried out via COCAS(*14*) with 100 as the main and extension threshold, thus detecting regions completely depleted of nucleosomes. Peaks were intersected in turn with TSS coordinates of selected mouse/human genes and with CGIs. Pearson correlations were carried on CGI length and inverse MNase peak size. For correlations between CGI length and NFR area or, peak areas or maxima for inverse MNase signal were retrieved, averaged into 80 classes and sorted by increasing CGI length.

G. MNase and +1 nucleosome GC content correlation

Total MNase signal for each experiment and GC content were retrieved between positions corresponding to the +1 nucleosome (0-140bp). GC content was retrieved via RSATools(*15*). MNase signal was sorted by increasing total GC content in this region. Due to high variance in both MNase signal and corresponding GC content, both were split into 80 classes. The probability of obtaining this dataset by random permutation is infinitesimal thus negligible: for example this probability for the mouse DP data is $\frac{80}{P_{100}^{8634}}$. Total GC content against total MNase signal in [0,140bp] around the TSS was then plotted for all categories.

H. Major nucleosome position

Nucleosome consensus positions (-1, +1, +2, +3 and +4) around TSS (Fig. 3D and Supplemental Fig. 5A, B) were defined based on the average profiles of nucleosome midpoints for all promoters (-226bp, 88bp, 307bp, 484bp and 668bp respectively). Bin scores in a region of 20 bp around these coordinates were averaged and sorted according to the increasing GC content of corresponding promoters.

I. Assessment of pausing scores (PSs)

Our approach for PS determination is similar to the one described in (*16*) for traveling ratios with modifications. It essentially takes in account Pol II density on either promoter regions (TSS) or gene bodies (GB). Sequencing of short RNAs (Fig. 1) allows for determination of the product of paused transcripts (TSS RNAs) at promoters based on a size selection strategy (Seila et al, 2008). As shown in Fig. 1, Pol II density at the surrounding of TSSs exactly matches that of the TSS RNAs. This density appears optimal between -300 and +100 for both Pol II and RNAs and we therefore chose this interval to define paused Pol II density for further calculations. For genes bodies, we chose the interval 50-100% for calculation of the average GB signal as (1) more significant signal of elongating Pol II can be detected in this area and (2) it avoids to detect signal originating from the promoters (and therefore paused Pol II) for short genes or genes with exceptionally large initiation areas (Koch et al, 2011). We then plotted the distribution of the signal of both TSS and GB densities that we compared to genomic random selections of similar sizes to determine TSS and GB enrichment areas and define the set of genes with significant Pol II signal (P-val. < 0.05, Supplemental Fig. 2 and 7). We could rank this gene set by the ratio of TSS/GB that we define as the pausing index. This approach allows us to define 3 classes of genes with high, intermediate or no pausing. The term ‘pausing’ can used here as Fig. 1 indicates an excellent correlation between TSS RNA accumulation and Pol II densities at promoters. Genes with no pausing represent a minority (less than 10%) in most of our analyses in mouse or human cells.

J. Determination of CGI border lines

After sorting of CGI annotations around promoters, borders coordinates were isolated. A smoothing spline function has been applied separately for starting and ending points and the resulting line has later been overlaid on the clusters of interest.

IV. References

1. L. A. Boyer *et al.*, *Nature* **441**, 349 (May 18, 2006).
2. S. Y. Tsang, S. C. Tam, I. Bremner, M. J. Burkitt, *Biochem J* **317 (Pt 1)**, 13 (Jul 1, 1996).
3. F. Koch *et al.*, *Nature structural & molecular biology* **18**, 956 (2011).
4. A. Valouev *et al.*, *Nature*, (May 22, 2011).
5. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, *Genome biology* **10**, R25 (2009).
6. M. Morgan *et al.*, *Bioinformatics* **25**, 2607 (Oct 1, 2009).
7. P. P. Chan, T. M. Lowe, *Nucleic acids research* **37**, D93 (Jan, 2009).
8. P. A. Fujita *et al.*, *Nucleic acids research* **39**, D876 (Jan, 2011).
9. M. Gardiner-Garden, M. Frommer, *J Mol Biol* **196**, 261 (Jul 20, 1987).
10. P. Rice, I. Longden, A. Bleasby, *Trends Genet* **16**, 276 (Jun, 2000).
11. J. van Helden, B. Andre, J. Collado-Vides, *Yeast* **16**, 177 (Jan 30, 2000).
12. A. J. Saldanha, *Bioinformatics* **20**, 3246 (Nov 22, 2004).
13. T. Juven-Gershon, J. T. Kadonaga, *Dev Biol* **339**, 225 (Mar 15, 2010).
14. T. Benoukraf *et al.*, *Bioinformatics* **25**, 954 (Apr 1, 2009).
15. J. V. Turatsinze, M. Thomas-Chollier, M. Defrance, J. van Helden, *Nat Protoc* **3**, 1578 (2008).
16. P. B. Rahl *et al.*, *Cell* **141**, 432 (Apr 30, 2010).