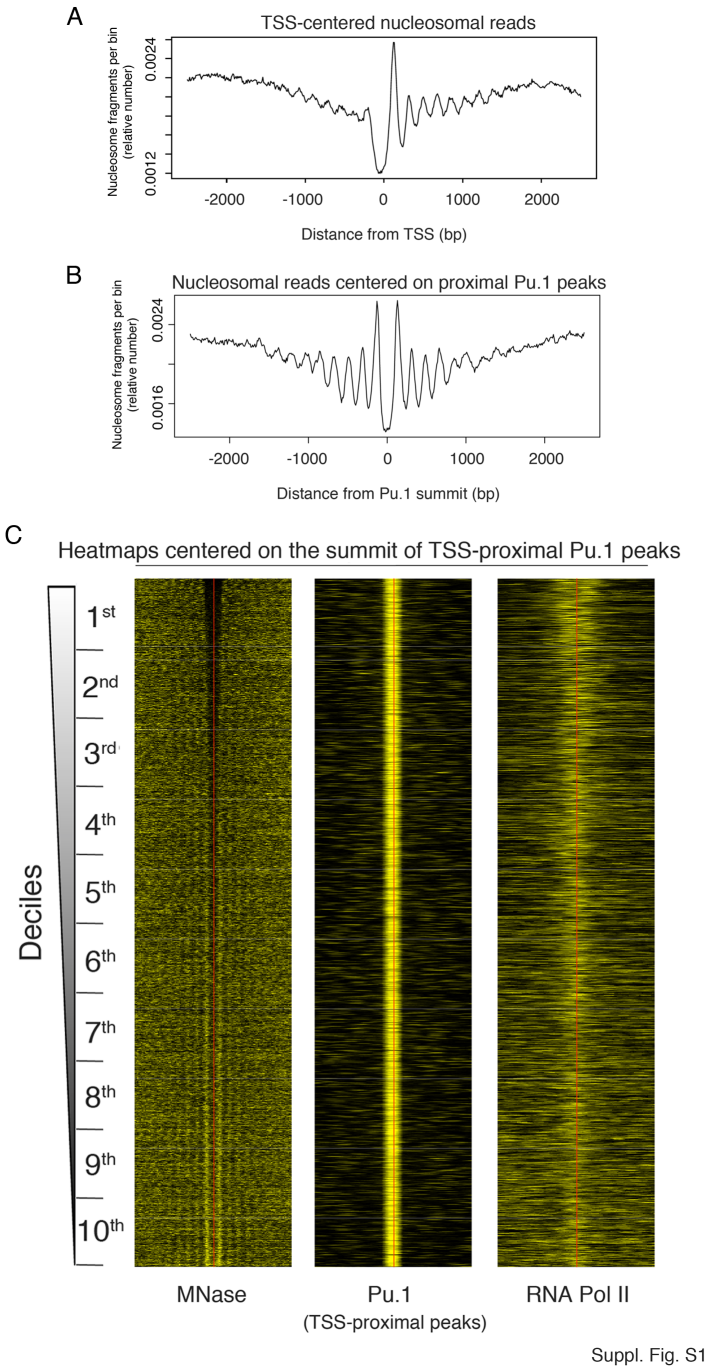
Molecular Cell, Volume *54*

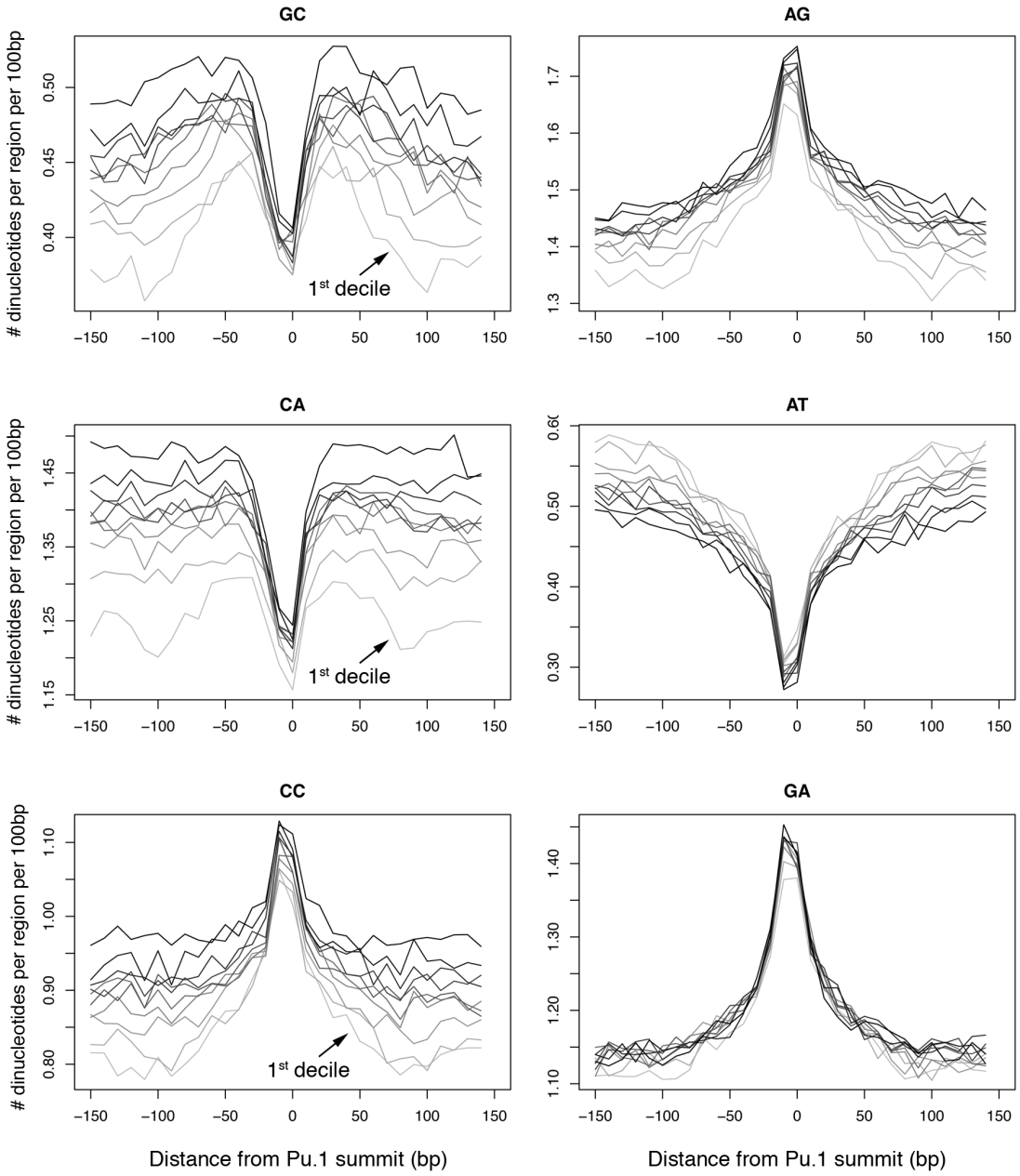
Supplemental Information

Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers

Iros Barozzi, Marta Simonatto, Silvia Bonifacio, Lin Yang, Remo Rohs, Serena Ghisletti, and Gioacchino Natoli

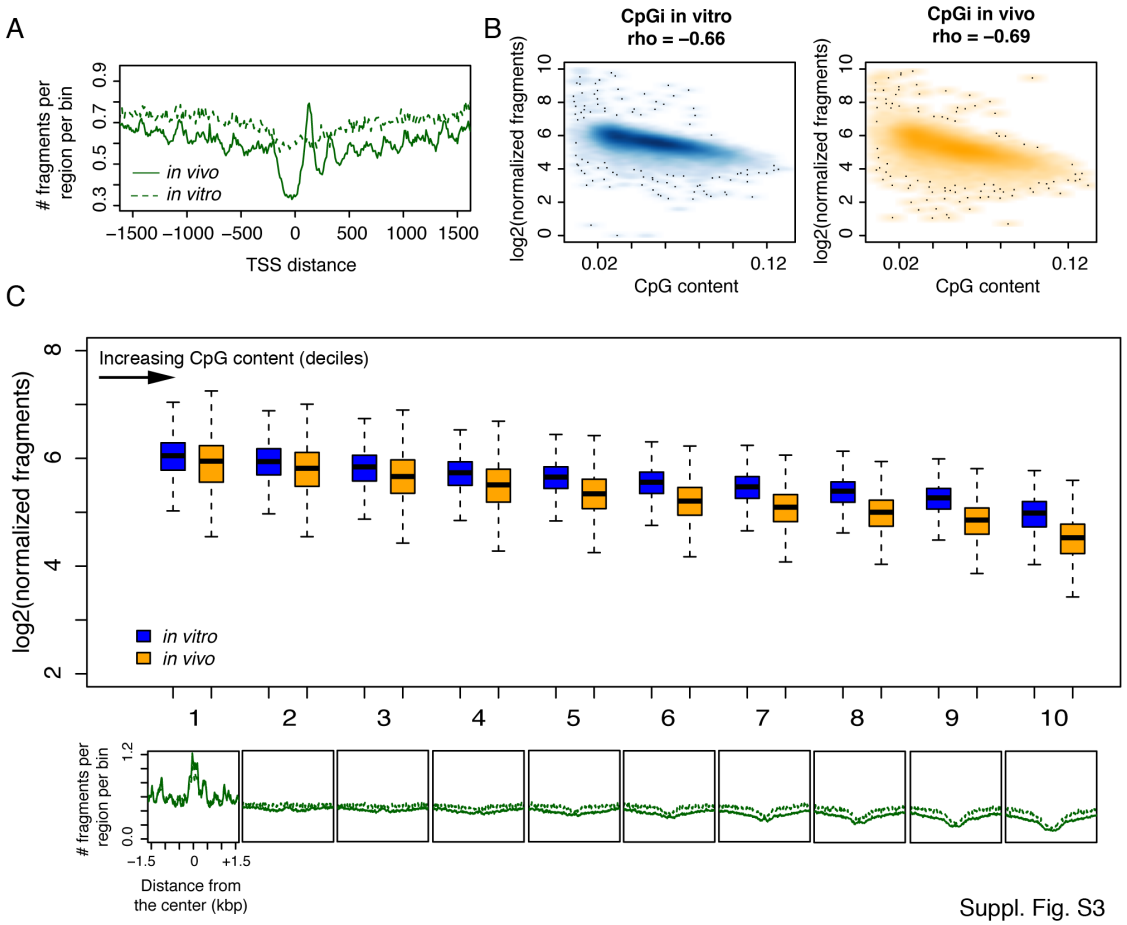


Supplemental Figure S1 (related to Fig. 1). *Nucleosomal features at TSSs and around TSS-proximal Pu.1 sites*. Midpoints of nucleosomal fragments (**A**) around annotated TSSs or (**B**) around the summit of TSS-proximal Pu.1 sites were quantified in 10 bp bins. The number of fragments in each bin was normalized by the total number of fragments in the area. The same information in **B**) is shown in **C**) as heatmap (first from the left), sorted from top to bottom based on decreasing occupancy of the NDR. Heatmaps of RNA Pol II and Pu.1 are also shown on the right side of the MNase data.



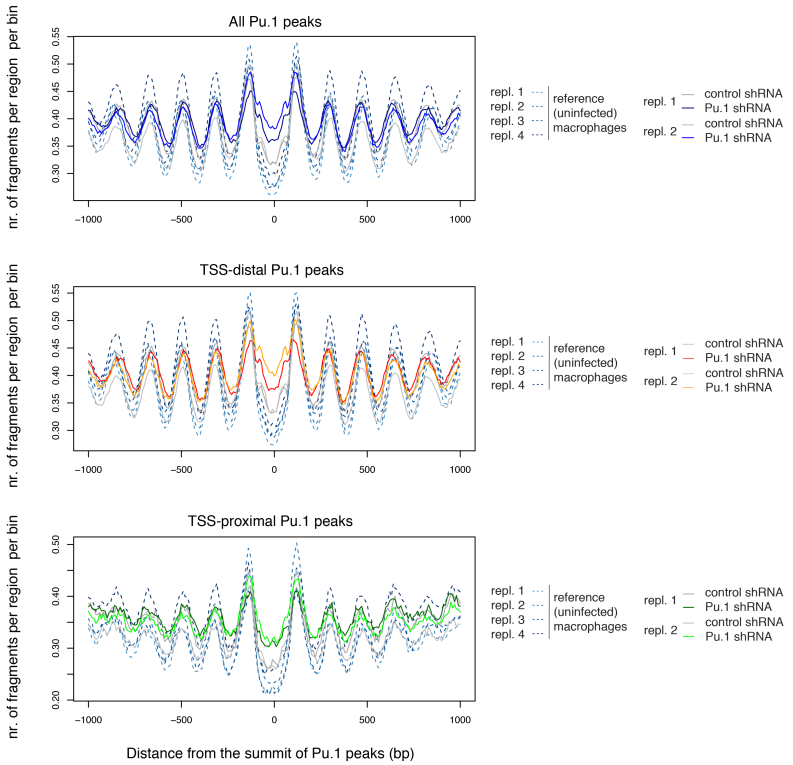
Suppl. Fig. S2

Supplemental Figure S2 (related to Fig. 2). *Dinucleotide composition of Pu.1-bound distal regions*. Regions were divided in deciles according to the NDR width (as shown in Fig 1). Left side: selected dinucleotides showing differences among deciles (indicated by a grayscale; light gray: 1st decile; black: 10th decile). Right side: dinucleotides showing similar frequencies around Pu.1 sites in different deciles. It should be noticed that AG and GA are obligatory dinucleotides of the Pu.1 binding site (5'-AGAGGAAGTG-3') and are thus strongly enriched around the summit of Pu.1 sites.

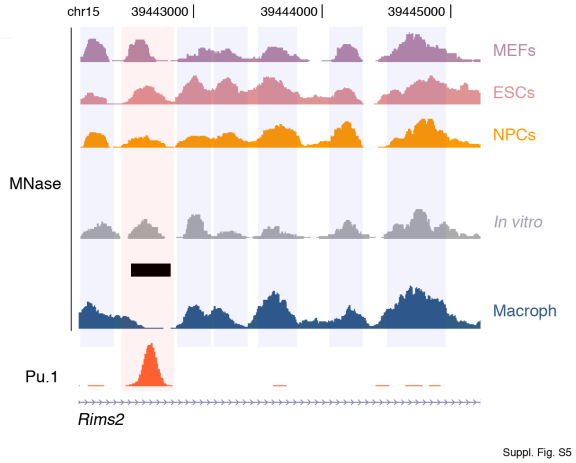


Suppl. Fig. S3

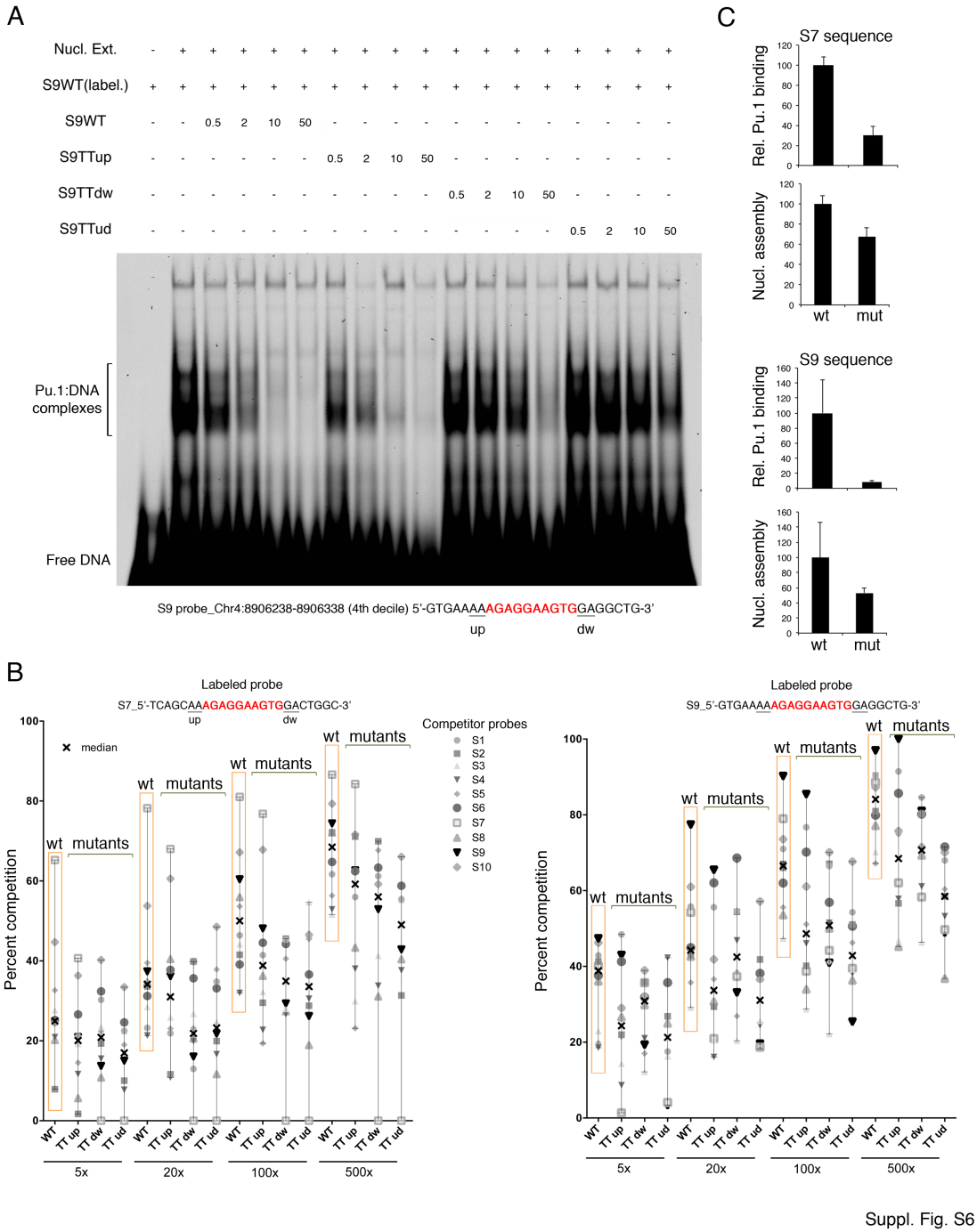
Supplemental Figure S3 (related to Fig. 3). *Nucleosome depletion at TSS and CpG islands in vitro and in vivo*. **A**) Cumulative distributions of the midpoints of the nucleosomal fragments (in 10 bp bins) centered on the TSSs of RefSeq genes for *in vitro* (dashed line) and *in vivo* (solid line) datasets. **B**) Scatter plots showing the relationship between CpG content (x-axis) and nucleosome occupancy (y-axis) at annotated CpG islands (Bird at al. 2010) *in vivo* (orange) and *in vitro* (blue). **C**) CpG islands were further divided into deciles (according to CpG content) and nucleosome occupancy was quantified for each one of them. Boxplots show the average number of the midpoints of the nucleosomal fragments per island (normalized to kbp and sequencing depth) for each decile. Cumulative distributions show the same information aligned to the centers of the CpG island for *in vivo* (solid line) and *in vitro* (dashed line) datasets, highlighting the relative nucleosome enrichment (1st decile) or depletion (upper deciles) of the island compared to the flanking regions.



Supplemental Figure S4 (related to Fig. 4). *Impact of Pu.1 depletion on nucleosomal patterns*. Cumulative distribution plots of midpoints of nucleosomal fragments are shown for all Pu.1 peaks or selectively for TSS-distal and TSS-proximal peaks.



Supplemental Figure S5 (related to Fig. 5). A representative snapshot of *in vivo* and *in vitro* nucleosome patterns.



Suppl. Fig. S6

Supplemental Figure S6 (mainly related to Fig. 6). *Testing the effects of DNA shape features on Pu.1 binding and nucleosome assembly.* **A)** A representative competitive EMSA with an infrared dye-labeled 24 nt probe (S9) and ten unlabeled competitors (indicated as S1 to S10 and corresponding to mouse genomic sequences bound by Pu.1 and bearing high-affinity Pu.1 sites). Competitors were either unmutated (wt), bearing mutations in the nucleotides upstream (up, sequence underlined) or downstream (dw), or both mutations upstream and downstream of the core Pu.1 binding site (shown in red) (ud). Numbers indicate picomoles of competitor DNA. ud: sequences mutated in both the two nucleotides upstream and downstream of the Pu.1 sites. **B**)

Competitive EMSAs with two infrared dye-labeled probes (S7, left; S9, right) and a panel of unmutated or mutated competitors. Data were quantified by LI-COR. **C**) *In vitro* nucleosome assembly and Pu.1 ChIP using two 150 nt sequences (S7 and S9) in which the Pu.1 site flanks were left unchanged (wt) or both mutated (mut) as above to alter DNA shape.

Supplemental Table legends.

Suppl. Table S1. *Sequencing statistics of the MNase-seq samples*. The total numer of high quality, uniquely aligned and properly paired reads for each MNase-seq sample is provided. The indicated numbers of reads represent occurrences after filtering for PCR duplicates. Datasets from ESCs, NPCs and MEFs were downloaded from the literature (Teif et al., 2012).

Suppl. Table S2. *List of the Pu.1 ChIP-seq datasets collected from the literature.* GEO accession numbers (Barrett et al., 2013) for each IP is provided along with the description of the cell type, the mouse strain, the antibody used and the GEO accession number of the control sample (Input/IgG).

Suppl. Table S3. *SVM performances and lists of the selected features along ten train/test randomizations.* Accuracy, sensitivity and positive predictive value over the training and test datasets are provided for ten distinct train/test randomizations (Performances datasheet). The corresponding sets of selected features are shown (Features_over_multiple_runs datasheet). For each feature that has been selected at least once, a 0/1 flag indicates if the feature has been retained in that particular random initialization. The total number of times the feature has been selected (#) is also indicated.

Suppl. Table S4. *EMSA oligonucleotides.*

Supplemental Experimental Procedures.

Cell culture and retroviral infection. Animal experiments were performed in accordance with the Italian Laws (D.L.vo 116/92 and following additions), which enforce the EU 86/609 Directive. Macrophage cultures from bone marrows of C57/BL6 mice (Harlan) were generated as described (Ghisletti et al., 2010). The hairpin used in this study to deplete Pu.1 was selected among five designed using a publicly available software (http://katahdin.mssm.edu/siRNA). The sequence is available upon request. The shPU.1 sequence was cloned in a modified version of TtRMPVIR inducible retroviral vector (Genbank HQ456318)(Zuber et al., 2011) in which the puromycin resistance gene was inserted. The empty vector, containing an sh-Renilla sequence was used as control. At day 0 bone marrow cells were isolated and 4×10^6 cells/plate were seeded in 10 cm dishes in TET-free BM medium. Cells were infected twice (in two consecutive days after plating) using supernatants from transfected Phoenix-ECO packaging cells. Puromycin selection (3 µg/ml) started on day 3. At day 5, shPU.1 expression was induced for 48 hours using doxycycline (0.5 µg/ml).

Antibodies. The anti-Pu.1 rabbit polyclonal antibody was generated in-house against the N-terminus of mouse Pu.1 (aa. 1-100; NP_035485.1) and affinity purified. Normal Rabbit IgG (Santa Cruz, SC2027) were used as control in ChIP and immunodepletion experiments. Anti-vinculin antibody (Sigma V9131) was used as loading control in western blots. Secondary IRDye antibodies were from Li-Cor (#926-68021 and 926-32210); Odyssey scanner and software (Li-Cor) were used for infrared fluorescence acquisition and quantification.

MNase digestion. MNase digestion was performed starting from 8-12 x 10^6 cells. Cell pellets were resuspended in a 15 mM NaCl, 15 mM Tris-HCl [pH 7.6], 60 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 0.3 M sucrose buffer (0.5 mM PMSF, 1 mM DTT, 0.2 mM spermine, 1 mM spermidine) and lysed upon addition of 0.4% NP40. Nuclei were washed with a 15 mM NaCl, 15 mM Tris-HCl [pH 7.6], 60 mM KCl, 0.3 M sucrose buffer (0.5 mM PMSF, 1 mM DTT, 0.2 mM spermine, 1 mM spermidine). Digestion was performed with 1.3 units of MNase (Roche 10107921001) in a 20 mM Tris-HCl [pH 7.6], 5 mM $CaCl_2$ digestion buffer, for 100 minutes at 37°C. Nucleosomal DNA was isolated by diluting nucleosomes in digestion solution to a final concentration of 5 mM $MgCl_2$, 5 mM $CaCl_2$, 70 mM KCl and 10 mM HEPES [pH 7.9]. Digestion with 5 units of MNase was let proceed for 100 minutes at 37°C and stopped by adding EDTA to a final concentration of 50mM. DNA was purified from octamer proteins with the Qiagen PCR purification kit. Purified DNA was then run in a 1% agarose gel and the mononucleosomal band cut and purified first with Millipore DNA Gel Extraction Kit and then with the Qiagen PCR purification kit. Mononucleosomal DNA was prepared for HiSeq2000 sequencing using the Illumina standard protocol.

In vitro nucleosome assembly. Naked genomic DNA was purified from mouse macrophages by three consecutive phenol/chloroform extractions. DNA was sonicated to obtain fragments smaller than 2 kb, and fragments ranging from 600 to 2,000 bp were purified with Solid-Phase Reversible Immobilization (SPRI) beads (Agencourt AMPure XP, Beckman Coulter). DNA was combined with recombinant histones (EpiMark™ Nucleosome Assembly Kit, NEB E5350) to generate nucleosomes by salt dialysis (Luger et al., 1999). DNA molecules were considered as multiple of 150 bp nucleosome-assembling units. Assembly reaction was performed mixing octamers and nucleosome-assembling units in a 1:2 molar ratio so that DNA was not limiting and octamer would assemble according to the sequence preference. For the experiment shown in Fig. 5E chimeric DNAs were produced by gene synthesis with T3 and T7 sequences at the 5' and 3' respectively, and then amplified by PCR. PCR products were mixed with genomic DNA (used as competitor) in a ratio 1:10 and then assembled with octamers.

In vitro ChIP. In vitro nucleosomes were partially digested with MNase (5U for 2 minutes in the digestion buffer described above) to obtain mainly di- and tri-nucleosomes and to eliminate any residual unwrapped DNA. They were then incubated with macrophage-derived nuclear extracts. Nuclear extracts were prepared from 20 x 10^6 cells. Cells were first lysed with hypotonic buffer (10 mM Tris-HCl, 1 mM KCl, 1.5 mM $MgCl_2$), then nuclei were lysed with a high-salt buffer (50 mM Tris-HCl, 200 mM NaCl, 10% glycerol, 0.2% NP40) and diluted 1:2 with a dilution buffer (10 mM Tris-HCl, 2 mM EDTA). Nuclear extracts were subjected twice (2 hours and overnight) to immunodepletion with 8 μg of Pu.1 antibody or normal rabbit IgG as control. Incubation of nuclear extracts and *in vitro* nucleosomes was performed at 4°C for 2 hours, then 5 μg of anti-Pu.1 antibody were added for 1 hour and DNA-protein complexes recovered with G protein-coupled magnetic beads. Beads were washed 6 times with wash buffer (30 mM Tris-HCl, 200 mM NaCl, 10% glycerol, 0.1% NP40, 1 mM EDTA) and twice with TE. DNA was eluted in TE-2% SDS. DNA

was then purified by Qiaquick PCR purification kit and quantified with PicoGreen (Invitrogen). ChIP DNA was prepared for HiSeq2000 sequencing following standard protocols. For the experiment shown in Fig. 5E, *in vitro* assembled nucleosomes were partially digested with MNase (5U for 1 min at 37°) to digest free DNA. Nuclear extracts were prepared as above from 293T cells transfected 36 h befor) with mock vector or pcDNA3-Flag_Pu.1. QPCR with T3 (aattaaccctcactaaagg) and T7 (cctatagtgagtcgtatta) primers was used for all the probes.

ChIP-Sequencing. Cross-linked nuclear lysates were sonicated and then immunoprecipitated with 5 μg of Pu.1 antibody or normal rabbit IgG. Antibodies were pre-bound to G protein-coupled paramagnetic beads (Dynabeads) in PBS-0.5% BSA and incubated with lysates overnight at 4°C. Beads were washed six times in a modified RIPA buffer (50 mM HEPES [pH 7.6], 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0.7% Na-deoxycholate) and once in TE containing 50 mM NaCl. DNA was eluted in TE-2% SDS and crosslinks reversed by incubation overnight at 65°C. DNA was then purified by QIAquick PCR purification kit (Qiagen) and quantified with PicoGreen (Invitrogen). ChIP DNA was prepared for HiSeq2000 sequencing following standard protocols.

Electrophoretic mobility shift assays and probe design. After macrophage lysis with hypotonic buffer (10 mM Tris-HCl, 1 mM KCl, 1.5 mM $MgCl_2$), nuclei were lysed in nuclear extract buffer (250mM KCl, 10% glycerol, 25mM Hepes) by freeze/thawing. 10 mg of nuclear extract were incubated with 2 mg of poly dI:dC and increasing amount of non-labeled competitors (0.5, 2, 10 and 50 pmol). After a 10 min pre-incubation with non-specific and specific competitor, 0.1 pmol of the labeled probe were added to each sample. The mix was incubated at room temperature for 20 min and then loaded in a 5% polyacrylamide non-denaturing gel. The gel was run in TBE 2,5 x for 2 hours at 100 V. The gel was scanned with the Li-Cor Odissey Infrared Imaging System and shifted bands were quantified. EMSA competitor sequences with mutations predicted to affect 3D DNA shape (**Suppl. Table S4**) were designed as follows. We randomly selected the same number (800) of bound and unbound regions showing the exact known Pu.1 core consensus site (AGAGGAAGTG). Among these sequences, DNA shape seemed to be favorable in sequences with a relative enrichment of the AA dinucleotide upstream of the consensus and a relative depletion of the TT dinucleotide downstream the consensus binding site. Following this observation, we selected a number of Pu.1-bound regions in intermediate deciles showing favorable DNA flanks and binding in the *in vitro* ChIP. Three different templates to assess the impact of mutated flanks on DNA shape were designed for each one of the selected sequences, as follows: i) the two upstream nucleotides (usually AA) were changed to TT; ii) the two downstream nucleotides (AG/GG/AA) were changed to TT; iii) i and ii were combined.

Computational Methods

ChIP-seq data analysis. After quality filtering, 51 nt long reads were aligned onto the mm9 release of the murine genome using Bowtie v0.12.7 (Langmead, 2010). Only unique alignments were retained, allowing up to two mismatches compared to the reference genome (options -m 1 -v 2). Peak calling was performed using MACS v1.4 (Zhang et al., 2008) with a bandwidth (bw parameter) of 100 (bp). Cell-type specific input was used as control. A golden set was defined by filtering peaks with a *p*-value lower than or equal to 1e-10. This set was annotated over Ensembl genes (Flicek et al., 2013) using GIN (Cesaroni et al., 2008) (*priority* set to “gene” and *promoter definition* to “-20000”). The coordinates of the genes were downloaded from the UCSC genome browser (Fujita et al., 2011) on 2011, July 7th. Peaks within +/- 2.5 kbp from TSSs were

considered as TSS-proximal while all the others were defined as TSS-distal. In order to visualize the raw profiles on the Genome Browser (Flicek et al., 2013), wiggle files were generated with MACS v1.4 and converted to bigWig (Fujita et al., 2011).

De novo motif discovery. Considering all the sequences (+/- 50 nucleotides from the Pu.1 summit) in each Pu.1-bound TSS-distal decile separately, *de novo* motif discovery was performed using mCUDA-MEME (Liu et al., 2010), an ultrafast scalable motif discovery algorithm based on MEME (version 4.4.0) (Bailey et al., 1994) for mutliple GPUs. The following parameters were used: -dna -mod zoops -evt 0.01 -nmotifs 20 -minw 6 -maxw 16 -revcomp -maxsize 1000000.

TFBSs over-representation analysis. Considering the sequences in each Pu.1-bound TSS-distal decile separately, we used Pscan (Zambelli et al., 2009, PMID 19487240) to detect statistically significant over-represented DNA motifs. Given a dataset of position weight matrices (PWMs), representing experimentally determined binding preferences for known transcription factors (TFs), Pscan scans each input sequence for the best match to each one of these PWMs. It then uses these values to build a distribution and compare it with that obtained applying the same procedure to a background set. Pscan returns a *p*-value for each PWM so that significantly over-represented binding motifs can be identified. We used as PWMs the set described in the section "*Measuring features in DNA strings*", the entire set of Pu.1-bound TSS-distal as background and the DNA sequences +/- 50 nucleotides from the Pu.1 summits as regions to scan. We then collapsed the information of TF-specific PWMs to their structural family, as defined in TFClass (Wingender et al., 2013). In total, 83 families were examined. Considering a given family, the lowest *p*-value among those obtained for the PWMs of the TFs in the family was chosen as representative. We then kept only those TF families showing a *p*-value equal or lower than 1e-5 in at least one out of ten deciles. –log10(*p*-value) were then used as input for the hierarchical clustering of the TF families, using 1 minus Spearman's Rank Correlation Coefficient along the deciles as measure of distance, and complete linkage. For visualization purpose, any –log10(*p*-value) exceeding 20 was set to this value. The cluster analysis and the heatmaps were done in R.

In vitro Pu.1 ChIP-seq data analysis. Analyses were performed as described in the previous paragraph but considering a lower statistical threshold for calling the peaks ($p <= 1e\text{-}5$). Nucleosomal occupancy over the sites was calculated as the number of paired-end fragments spanning the experimentally determined Pu.1 summits.

Defining a collective Pu.1 cistrome. Every murine ChIP-seq dataset of sufficient quality available in the literature was downloaded from the Gene Expression Omnibus (Barrett et al., 2011) (**Suppl. Table 2**) and analyzed as described in the previous paragraphs. Cell-type specific inputs were used as control (**Suppl. Table 2**). Genomic tracks were generated using MACS (Zhang et al., 2008) and normalized to a fixed sequencing depth for visualization. All the ChIPs considered were carried out with the same antibody (Santa Cruz sc-352), with the exception of the BMDMs-derived dataset generated in our lab, which was also included in the analysis.

In order to define regions bound by Pu.1 in at least one of the seven cell types binding events from distinct experiments were matched. First of all, the enriched regions were further split into their components (dense homotypic clusters, which often span up to few kilobases, are recognized by MACS as a single highly-enriched region). To achieve this aim, PeakSplitter (Salmon-Divon et al., 2010) was run on the individual ChIP-seq profiles, considering only

enriched regions with a *p*-value <= 1e-5 (as determined by MACS) and using the following parameters: -c 5 -f -v 0.7. Only subpeaks with 20 or more reads spanning the summit in at least one ChIP-seq profile were considered for further analysis. Irrespective of their cell type of origin, coordinates of Pu.1-bound regions from different cell types were merged if their summits were found within 250 bp from each other. These regions were then annotated as TSS-proximal or TSS-distal as described in the previous paragraphs.

A genome-wide map of regions containing high-affinity canonical Pu.1 sites. The motif-scanning tool FIMO (Grant et al., 2011) (MEME version 4.6.1) was used to identify DNA stretches that could be potentially bound by Pu.1 (these sequences will be referred to as canonical binding sites or bound/w sites). FIMO was run at a *p*-value threshold of 1e-4, with default parameters except that no *q*-value was calculated. A published PWM for the Spi1 DNA-binding domain (DBD)(Wei et al., 2010) was used as representative of Pu.1 binding preferences.

High-throughput sequencing could have missed some of the regions identified because of mappability issues. Intuitively, the longer the read, the lower the probability to map to more than one region in the genome. Since the shortest reads in the datasets under investigation are 36 nt long, mappability scores computed for 36 nt reads were used. Scores were extracted from bigWig tracks (Derrien et al., 2012) downloaded from the UCSC Genome Browser. For any given region (considered as 50 bp upstream and downstream of each canonical binding site identified), mappability scores ware retrieved using custom scripts. Any region showing at least one bp with a mappability score of 1 were further considered.

Using this procedure, 613,210 putative Pu.1-binding sites were identified. Among those, 41,472 overlap a bound site of the cell-type a-specific Pu.1 cistrome, meaning that 571,738 (93.2%) of the sites were never contacted by Pu.1 *in vivo*. On the other hand, among the bound sites, 41,472 showed a canonical high affinity Pu.1 binding site (Wei et al., 2010) within 50 bp from the peak summit (see Suppl. Fig. S5, the threshold was set to 50 bp by elbow method), accounting for 42.9% of the total (the bound sequences without a canonical binding sites are referred to as bound/wo).

Measuring features in DNA strings. Features were assessed in a 300 bp window (unless differently indicated) centered on the summit of the ChIP-seq peaks in case of bound regions, and to the invariant GGAA core of the Pu.1 binding site in case of the unbound ones.

These features can be divided into five broad categories, namely PWMs, *k*-mers, repetitive elements, DNA shape, and nucleosome theoretical occupancy. Each group is described in details:

- PWMs provide quantitative descriptions of the known binding sites for a TF (Stormo, 2000). They can be used to assess putative binding in any DNA string. PWMs were collected from the literature (the total number of PWMs gathered from each publication is reported in brackets):
  - Portales-Casamar et al., 2010 (Portales-Casamar et al., 2010)(146)
  - Jolma et al., 2013 (Jolma et al., 2013) (843)
  - Jolma et al., 2010 (Jolma et al., 2010) (26)
  - Hallikas et al., 2006 (Hallikas et al., 2006) (4)
  - Badis et al., 2009 (Badis et al., 2009) (104)
  - Berger et al., 2008 (Berger et al., 2008) (177)
  - Wei et al., 2010 (Wei et al., 2010) (27)
  - Kulakovskiy et al., 2013 (Kulakovskiy et al., 2013) (481).

FIMO (Grant et al., 2011) (included in MEME 4.6.1) scans an input region of DNA for occurrences of a PWM. It computes a log-likelihood ratio score with respect to each sequence position and converts these scores to *p*-values. FIMO was run on the regions of interest (using a 300 bp as well as a 100 bp window) and the corresponding *p*-values were transformed according to the formula $-\log_{10}(p\text{-value})$. Only *p*-values equal or lower than 1e-4 were considered, otherwise a *p*-value of 1 was assigned to the region. In case of multiple matches to the same region, only the match with the best *p*-value was considered. In this way each region was described with a single value for each one of the PWMs.

Since the set of PWMs gathered from the literature was highly redundant, motifs were grouped according to their DNA-binding domain. A straightforward approach to group motifs would be to cluster them based on sequence similarity. Nevertheless, a familial binding profile ignores the flanking positions of PWMs that are not aligned but which may be important in discriminating false positives. A recent paper (Oh et al., 2012) suggested an alternative approach, i.e. to consider overlapping binding sites predicted by redundant PWMs representing TFs from the same family. In line with this approach, PWMs were grouped according to their classification in families and subfamilies in TFClass (Wingender et al., 2013). A total number of 83 families and 263 subfamilies were considered. The lowest FIMO *p*-value among those obtained for the PWMs in a given family or subfamily was chosen as representative for each one of them. This approach also has the advantage of reducing the initial number of features to be considered in supervised feature selection.

Furthermore, the sum of families and subfamilies that show at least a significant occurrence for one PWM was used as a proxy for cooperative binding at the region.

- The sum of G+C content and the individual *k*-mers (with *k* equal to 2 or 4) counts were calculated.
- Repetitive elements in the mm9 genome were retrieved from the RepeatMasker (Smit and Riggs, 1996) track of the UCSC genome browser. A BED file for each class of repetitive elements was generated and superimposed with the regions of interest.
- The three-dimensional DNA shape features (Rohs et al., 2009) were predicted for the local sequence context around the 10 bp within the ETS core motif and for additional 15 bp on each flank (Gordan et al., 2013). The four DNA shape features used in this study to describe Pu.1 binding sites included minor groove width (MGW), Roll, propeller twist (ProT), and helix twist (HelT) (as implemented in Yang et al., 2014). These four structural features have been previously found to affect protein-DNA binding specificity (Rohs et al., 2010). Measurements for these features were obtained using a method for high-throughput prediction of DNA shape (Zhou et al., 2013) based on data from all-atom Monte-Carlo simulations.
- Nucleosome theoretical occupancy was calculated using a published algorithm (Kaplan et al., 2009). Calculations were performed using a sliding window of 147 bp. The average value among all the sliding windows was used as a proxy for the region.

Supervised learning using Support Vector Machines. Support vector machines (SVMs)(Cortes and Vapnik, 1995) are supervised learning algorithms used to discover patterns useful for classification and regression analysis (Drucker et al., 1997). Given a set of training

examples (each belonging to one, and only one, category) an SVM training algorithm builds a model that can be used to categorize new examples. SVMs are mainly used for binary classification, while multi-class implementations are available (Chang and Lin, 2011). Although SVMs are linear classifiers, they can perform non-linear classification using what is called the "kernel trick". This trick implies that SVMs are still performing linear classification, but input data points are mapped into high-dimensional feature spaces by a so-called kernel function.

SVMs were applied to classify the Pu.1-bound regions with a canonical binding site from unbound sites. Considering 41,472 Pu.1-bound regions, the same number of regions was randomly selected among the unbound sites. LibSVM (Chang and Lin, 2011) was used to train and test two-class SVMs. The initial features were distributed as such (for each one of the categories considered, the number of features is reported in brackets):

- PWMs (reduced to the families and subfamilies of non-redundant TFs: 686);
- *k*-mers (G+C content, $k=2$ and $k=4$: 147);
- DNA shape features in core binding site and 15 bp flanks (MGW, Roll, ProT, and HelT: 146);
- Repetitive elements (15);
- Theoretical nucleosome occupancy (Kaplan et al., 2009) (1).

Given the large amount of features, a feature selection procedure (Guyon et al., 2003) to identify the smallest set with the highest predictive power was devised. Using 20% of the total instances, ten forward feature selection runs were performed randomizing training and validation datasets (50% each). The features selected in at least one out of ten randomizations were then pooled and used to train the machine on the entire 20% and test on the remaining 80%. Training and test datasets were also randomized ten times. For each round of randomization, uninformative features, namely those showing no variance across the examples, were discarded. Features were then properly scaled (range 0-1) and ranked according to the value of absolute Spearman's rank correlation calculated among the values and the class of the training examples. Only those with a value ≥ 0.04 were retained (threshold was estimated by elbow method). Forward selection consisted in adding features one by one (according to the described ranking) and keeping only those whose inclusion improved the accuracy on the validation set of at least 0.1%. This entire routine was wrapped into Python and R code.

A grid search was performed in order to choose the set of parameters that yield the best performances on the validation dataset. In practice, for each round of feature selection, an exhaustive search through a manually specified subset of parameters was performed, and the set of parameters with the highest improvement of performance was retained. SVM with no kernel (linear SVM) or with radial basis function (RBF) kernel were tested. In both cases, hyper-parameter C was chosen among the interval {0.01, 0.1, 1, 101, 100, 1000}. In case of RBF, parameter g was chosen among the interval {0.0001, 0.001, 0.01, 0.1, 1, 10}. All the possible combinations were tested.

When the feature selection routine was allowed to selected between linear SVM or using the RBF as kernel and an exhaustive search for parameters was performed (grid search), the RBF kernel was systematically preferred over the linear SVM. Nevertheless, while performances on the validation dataset increased, those on the test dataset dropped to values lower than those obtained using the linear SVM.

Performances were assessed using three indexes (bound are the positive set, unbound are the negative set, TP = true positive, FP = false positive, TN = true negative, FN = false negative):

- Overall accuracy, defining the fraction of instances correctly predicted, calculated as (TP+TN) / (TP+FP+TN+FN);

- Sensitivity, calculated as TP / (TP + FN), high values indicate that the machine performs very well in recognizing positive (bound) examples;
- Positive Predictive Value (PPV), calculated as TP / (TP + FP), high values correspond to a higher number of negative examples (unbound) predicted as positive (bound).

MNase-seq data analysis. Paired-end 101 nt long reads were quality filtered and mapped to the mouse genome (mm9, NCBI Build 37) using Bowtie v0.12.7 (Langmead, 2010). The following parameters were used: -v 3 -m 1 -S -l 0 -X 250. In this way, the paired-end fragments with a unique match to the genome that showed three or fewer mismatches were retained. Duplicated fragments, which are likely to arise from selective PCR amplification, were discarded (see Suppl. Table S1). Namely, given multiple fragments with both ends mapping to the same genomic coordinates, all fragments but one were discarded. Wiggle files at single bp resolution were generated with BedTools (Quinlan and Hall, 2010). In order to extract nucleosomal positions from this population-averaged profile PeakSplitter (Salmon-Divon et al., 2010) was run genome-wide on the wiggle file (with options -c 5 -f -v 0.7). For each one of the resulting regions the total number of fragments spanning the putative nucleosome dyad (namely the coordinate with the highest number of overlapping fragments) was calculated. This figure was used as proxy for occupancy. The dispersion of the midpoints of these fragments around the putative dyad (measured as standard deviation) was instead used as proxy for positioning. A custom C++ script performed these calculations.

Paired-end fragments for ESCs, NPCs and MEFs (Teif et al., 2012), aligned to the mm9 reference genome, were downloaded from GEO. Alignments were processed as described in the previous paragraph. Final numbers of reads are summarized in Suppl. Table S1. Unless differently specified, all the heatmaps, the cumulative distributions and the nucleosome density plots were computed using a 10 bp binning and the midpoint of each sequenced fragment as a proxy for the nucleosome dyad (hereinafter referred to as midpoint analysis).

In order to sort the regions based on the size of the central nucleosome-depleted region (NDR) we first calculated the number of nucleosome midpoints falling into the central 300 bp ($\pm$ 150 bp) of each region. These numbers were then used to indicate the overall occupancy of the area (the lower the number, the higher the depletion).

Support Vector Regressors. Support Vector Regressors (SVRs) are a variant of SVMs that can be applied to address regression problems (Drucker et al., 1997). It was used to assess the fraction of variability in the nucleosomal occupancy pattern at Pu.1-bound and unbound sites in cells where Pu.1 is not expressed or in *in vitro* chromatin reconstitution experiments that can be explained by the same features selected by the SVM. The theoretical nucleosomes occupancy(Kaplan et al., 2009) was excluded and an SVR was in parallel trained and tested using this feature alone.

The log2-transformed number of fragments spanning the center of each region was used as a proxy for the nucleosome occupancy at bound and unbound sites (corresponding to the Pu.1 ChIP-seq summit for the bound and to the GGAA core in case of the unbound site). These numbers were calculated for the ESCs, NPCs, MEFs and the *in vitro* datasets.

The entire set of bound and unbound sites was split into 90% training and 10% test datasets. The following procedure was run using the set of features coming from each one of the ten randomizations of the training and test datasets and separately for each condition (ESCs, NPCs, MEFs and *in vitro*). Features were scaled to range 0-1. The training dataset was used to fit

the experimentally determined counts of fragments according to sequence features. The model obtained was then used to predict the nucleosome counts over the test set. Performances were evaluated through the coefficient of determination (R^2), calculated as the squared Pearson Correlation Coefficient among the predicted and the observed counts. This R^2 can be interpreted as the percentage of variation in the data that is explained by the model (i.e. the variation in the nucleosome occupancy that is explained by the features in the sequence). As mentioned, an independent SVR was fed with the theoretical nucleosomes occupancy alone, and its performances compared to those obtained by the model trained on all the remaining features.

The SVR implementation in the R package e1071 with RBF kernel was used.

Pu.1 depletion experiments. BMDMs were infected with a retroviral vector either containing a short hairpin targeting the mRNA of Renilla (hereafter referred to as "empty") or Pu.1 (hereinafter referred to as "shPu.1"). ChIP-seq data from both samples were analyzed for enrichment *vs.* the input DNA. All these peaks identified in the "empty" (using a *p*-value threshold of 1e-10) were retained only if also present in the untreated Pu.1 sample obtained in "wild-type" conditions (see previous paragraphs describing ChIP-seq analyses).

In order to get a quantitative picture of the effect of the Pu.1 depletion, the entire set of peaks was sorted based on the ratio of the reads in the "empty" *vs.* the "shPu.1". Reads were counted in a window of 200 bp around the Pu.1 summit. After adding a pseudo-count of 1 and normalizing for sequencing depth, ratios were calculated and used to split the dataset into quartiles (the 1st quartile corresponds to lower ratios, namely peaks that are not affected by the depletion, while the 4th quartile encompasses the peaks with the lowest occupancy in the Pu.1-depleted cells compared to the control). Bulk differences in nucleosomal occupancy at these sites were evaluated summing up the nucleosomal fragments whose midpoint mapped into the 160 bp centered on the peak summit (corresponding to the area that would ideally be occupied by a nucleosome if Pu.1 is not bound). The difference among the resulting distributions was tested using a Wilcoxon signed-rank test (which is a paired, non parametric test).

Statistics and plots. All plots were drawn and statistics was performed using the R package.

Supplemental References

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science 324, 1720-1723.

Bailey, T.L., Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., et al. (2011). NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic acids research 39, D1005-1010.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 133, 1266-1276.

Cesaroni, M., Cittaro, D., Brozzi, A., Pelicci, P.G., and Luzi, L. (2008). CARPET: a web-based package for the analysis of ChIP-chip and expression tiling data. Bioinformatics 24, 2918-2920.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2, 1-27.
Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach Learn 20, 273-297.
Derrien, T., Estelle, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigo, R., and Ribeca, P. (2012). Fast computation and applications of genome mappability. PloS one 7, e30377.
Drucker, H., Burges, C.J., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. Advances in neural information processing systems, 155-161.
Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. Nucleic acids research 41, D48-55.
Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. Nucleic acids research 39, D876-882.
Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C.L., et al. (2010). Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity 32, 317-328.
Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell reports 3, 1093-1104.
Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017-1018.
Guyon, I., Andr, #233, and Elisseeff (2003). An introduction to variable and feature selection. J Mach Learn Res 3, 1157-1182.
Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 124, 47-59.
Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome research 20, 861-873.
Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell 152, 327-339.
Kaplan, N., Moore, I., Fondufe-Mittendorf, Y., Gossett, A., Tillo, D., Field, Y., LeProust, E., Hughes, T., Lieb, J., Widom, J., et al. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458, 362-366.
Kulakovskiy, I.V., Medvedeva, Y.A., Schaefer, U., Kasianov, A.S., Vorontsov, I.E., Bajic, V.B., and Makeev, V.J. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic acids research 41, D195-202.
Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11, Unit 11 17.
Liu, L., Schmidt, B., Liu, W., Maskell, D.L. (2010). CUDA-MEME: accelerating motif discovery in biological sequences using CUDA-enabled graphics processing units". Pattern Recognition Letters 31(14): 2170 – 2177.
Luger, K., Rechsteiner, T.J., and Richmond, T.J. (1999). Preparation of nucleosome core particle from recombinant histones. Methods Enzymol 304, 3-19.
Oh, Y.M., Kim, J.K., Choi, S., and Yoo, J.Y. (2012). Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices. Nucleic acids research 40, e38.
Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic acids research 38, D105-110.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu Rev Biochem 79, 233-269.
Rohs, R., West, S., Sosinsky, A., Liu, P., Mann, R., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature 461, 1248-1253.
Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. (2010). PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC bioinformatics 11, 415.
Smit, A.F., and Riggs, A.D. (1996). Tiggers and DNA transposon fossils in the human genome. Proceedings of the National Academy of Sciences of the United States of America 93, 1443-1448.
Stormo,G.D. (2000) DNA binding sites: representation and discovery. Bioinformatics, 16, 16–23.
Teif, V., Vainshtein, Y., Caudron-Herger, M.Ø., Mallm, J.-P., Marth, C., H√∂fer, T., and Rippe, K. (2012). Genome-wide nucleosome positioning during embryonic stem cell development. Nature structural & molecular biology 19, 1185-1192.
Wei, G.H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J 29, 2147-2160.
Wingender, E., Schoeps, T., and Donitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. Nucleic acids research 41, D165-170.
Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W.W., Gordân, R., and Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. 42, D148-155.
Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic acids research *37*, W247–52.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137.
Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic acids research 41, W56-62.
Zuber, J., McJunkin, K., Fellmann, C., Dow, L.E., Taylor, M.J., Hannon, G.J., and Lowe, S.W. (2011). Toolkit for evaluating genes required for proliferation and survival using tetracycline-regulated RNAi. Nat Biotechnol 29, 79-83.