bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Retrieving the topology of chromatin domains from deep sequencing data with correlation functions

Jana Molitor[1], Jan-Philipp Mallm[1,2], Karsten Rippe*[,1] & Fabian Erdel*[,1,2]

[1] German Cancer Research Center (DKFZ) and Bioquant Center, Research Group Genome Organization & Function, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

[2] Present address: Biochemistry and Molecular Biophysics, Columbia University, New York, USA

* Address correspondence to Karsten Rippe (Karsten.Rippe@dkfz.de) or Fabian Erdel (fe2172@columbia.edu)

Running title: Correlation function analysis of chromatin

Five Key words: epigenetics, chromatin patterns, stem cell differentiation, heterochromatin, peak calling, deep sequencing analysis

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Epigenetic modifications and other chromatin features partition the genome on multiple length scales to control its biological function. Some of them like DNA methylation target single bases, whereas others such as heterochromatic histone modifications span regions of several megabases. It has now become a routine task to map chromatin marks by deep sequencing. However, the quantitative assessment and comparison of the topology of chromatin domains and their spatial relationships across data sets without *a priori* assumptions remains challenging, especially if broad domains are involved. Here, we introduce multi-scale correlation evaluation (MCORE), which uses the fluctuation spectrum of mapped sequencing reads to quantify and compare spatial patterns on multiple length scales in a model-independent manner. We used MCORE to dissect the chromatin domain topology of embryonic stem cells and neural cells by integrating sequencing data from chromatin immunoprecipitation, RNA expression, DNA methylation and chromosome interaction experiments. Further, we constructed network models that reflect the relationships among these features on different genomic scales. We anticipate that MCORE will complement current sequencing evaluation schemes and aid in the design and validation of mechanistic models for chromatin signaling.

Background

In eukaryotic cells most processes that involve interactions with the genome are controlled by the local chromatin context. Accordingly, DNA replication, DNA repair, RNA expression and RNA splicing have been found to be regulated by different combinations of DNA methylation (5mC) and histone modifications [1, 2]. The genome-wide distribution of these and other chromatin features, like binding sites of transcription factors, contact frequencies between genomic loci and transcriptional activity, can routinely be assessed by deep sequencing [1]. Recent methodological developments enable the analysis of low cell numbers or even single cells [3-5] as well as the simultaneous readout of various features [6]. Thus, sequencing data at unprecedented cellular resolution and throughput are becoming available that provide a rich source of information on molecular networks that shape the chromatin landscape. When attempting to dissect how these networks operate it becomes important to robustly identify, quantify and compare the topology of chromatin domains enriched in a given feature on multiple genomic length scales across different data sets. Currently, deep sequencing data are mostly analyzed on the basis of local enrichments of read density, with the goal to identify regions scoring positive for one or more features of interest. Most of these approaches (see Table S1 for an incomplete list) fall into two

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

categories, namely peak calling algorithms [7-9] and probabilistic network models [10-12]. While these provide lists of enriched loci or chromatin states in a straightforward manner, it remains challenging to retrieve topological quantities such as characteristic chromatin domain sizes or the spatial relationships between these domains in a reliable manner.

Identification of enriched regions is not straightforward because it requires assumptions about the properties of these regions, e.g. their width and enrichment level. This is particularly problematic for the analysis of complex patterns that involve different enrichment levels and are therefore incompatible with binarization (Fig. S1). Furthermore, the information content of the local read density at individual loci is inherently limited due to the characteristics of deep sequencing data. Complications arise from undersampling, noise and technical bias that can change the apparent pattern and introduce or mask similarities between data sets [13-15]. For example, data from chromatin immunoprecipitation sequencing (ChIP-seq) experiments of histone modifications display artificial preferences for certain genomic regions, have different sequencing depths, or vary in signal-to-noise ratio due to different levels of non-specific background [14]. Because these factors change with genomic position and affect each genomic length scale differently they are difficult to account for. Consequently, peak calling results depend on user-defined input parameters and the specific algorithm used [16, 17]. Likewise, chromatin state annotations differ with respect to state number, state identity and spatial extension of the corresponding chromatin domains [10, 11]. These uncertainties are tolerable for identifying the most enriched regions or the most prevalent chromatin states. However, they may obscure the quantitative assessment of more complex patterns such as those observed for heterochromatic regions, which contain a combination of broadly distributed histone marks, 5mC and associated proteins [18, 19]. This impedes the comparison of experimental profiles to the predictions from different mechanistic models for the formation and maintenance of heterochromatin states (e.g. [20] and references therein). Therefore it would be beneficial to identify and evaluate chromatin patterns in deep sequencing data sets independently of peak or state annotations.

Here, we introduce an approach termed multi-scale correlation evaluation (MCORE) that complements the above-mentioned repertoire of analysis methods. MCORE avoids assumptions about the shape and the amplitude of enriched regions and evaluates all mapped sequencing reads without filtering. It retrieves information from correlation functions, which are used for the discovery of patterns in noisy and possibly undersampled data sets in many fields of research [21-25]. The use of correlation functions in the context of deep sequencing has mostly been restricted to strand cross correlation for measuring fragment lengths [16, 26] and autocorrelation for comparing ChIP-seq data sets to each other [27].

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Key advantages of correlation functions are intrinsic removal of (white) noise, robust identification of characteristic spatial or temporal length scales and straightforward assessment of spatial relationships between two different features. Conveniently, correlation functions can also be used if the exact pattern geometry is unknown (Fig. S1). The length scales are encoded in the shape of the correlation function and their determination is unaffected by variations in the absolute correlation amplitude. We used MCORE to analyze the chromatin domain topology of embryonic stem cells (ESCs) and neural cells (neural progenitor/brain cells, NCs) as their differentiated counterparts, focusing on 11 different chromatin features. These data sets covered histone modifications and DNA methylation, RNA expression and genome folding, as well as binding sites of chromatin-associated proteins. For each feature we identified the associated nucleosome repeat length and the characteristic domain sizes along with their relative abundance in the genome. In a pair-wise analysis we determined the (anti-)colocalization and spatial relationship between features on different genomic scales and used the results to construct network models for chromatin signaling. We compared ESCs to NCs to retrieve information about the spatial reorganization of chromatin during differentiation and to map the global transitions that occurred at active and repressive chromatin domains. Alterations were most pronounced for H3K9me3/H3K27me3 regions that changed their size, their location within chromosome territories and their positioning relative to DNA methylation and to each other.

Results

Comparison of MCORE to other sequencing analysis workflows

The MCORE workflow in comparison to the currently most common approaches for deep sequencing analysis is illustrated in Fig. 1 and Table S1. First, all types of data sets were transformed into normalized read occupancy profiles. Similar to other methods, a normalization step was included in the MCORE analysis. This takes into account that the observed coverage of a genomic region depends not only on its actual abundance after extraction but also on multiple other factors. For ChIP-seq samples these include the propensity to be immunoprecipitated, ligated, amplified, sequenced and mapped. To correct for these multiplicative biases, sample reads were divided by input reads for immunoprecipitation (IP) experiments or by the sum of converted and unconverted reads for bisulfite sequencing (BS-seq). IP experiments such as ChIP-seq, chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) and high-throughput chromosome conformation capture (Hi-C) yielded significant background correlation due to non-specific binding of DNA and proteins to beads or bead-antibody complexes [28]. Accordingly, these

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

types of data sets were further corrected by subtraction of a weighted control IP signal obtained from an IP with non-specific antibodies (Fig. S2A). The weighting factor reflects the contribution of non-specifically precipitated DNA in each sample and removes the correlation between specific IP and control IP (Materials and Methods). As expected, the contribution of non-specific signal depended on the quality of the antibody and on the enrichment levels of the specific IP-signal. H3K9me3 ChIP-seq data, for example, were affected more strongly by this correction than H3K4me3 ChIP-seq data (Fig. S2B) because H3K4me3 domains were more distinct and exhibited larger enrichments than H3K9me3 domains.

Peak calling or dynamic network models use occupancy profiles to define peaks or chromatin states based on local enrichments (Fig. 1A). In contrast, MCORE computes correlation functions from the sequencing read occupancy without binarizing the data. To this end, normalized occupancy profiles from two different data sets were shifted with respect to each other along the genomic coordinate, and the normalized Pearson correlation coefficient for each shifting distance Δx was calculated and analyzed (Materials and Methods). In contrast to rank correlations the Pearson correlation coefficient accounts for the enrichment values within the normalized occupancy profile and therefore preserves the biologically relevant information (Fig. S3, Fig. S4). We computed three types of correlation functions with different biological meaning: (i) the correlation function between two biological replicates, yielding the domain topology for a chromatin feature (Fig. 1B), (ii) the correlation function between the same feature in two different cell types, providing information on the positional conservation of a given chromatin mark across cell types (Fig. 1C), and (iii) the correlation function between two different features in the same cell type, reflecting their genome-wide relationship such as co-localization or shifted localization (Fig. 1C). The use of at least two independent data sets (either two biological replicates or two samples interrogating different features or cell types, see Eq. 4) for the calculation of each type of correlation function suppresses spurious noise that is uncorrelated between independent experiments and does therefore not contribute to the correlation.

To compare co-localization values among differently distributed marks we normalized cross-correlation functions with respect to their replicate correlation (Materials and Methods, Eq. 5). This step was required because broadly distributed marks tended to yield smaller cross- and replicate correlation coefficients than marks forming narrow and well-positioned domains. As illustrated in Fig. 1C, positive correlation indicated co-localization at a given shift distance, whereas negative correlation reflected mutually exclusive modification or binding. Each decay length and its contribution to the correlation function encoded a domain size and its abundance, whereas superimposed oscillations reflected nucleosome spacing

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[27, 29]. Where necessary, the correlation function can be used as a starting point to identify individual regions of interest as described below.

MCORE is complementary to peak calling, which typically aims to identify enriched regions without larger gaps. As the probability to find modified regions without spurious gaps decreases with size, broad regions are prone to get lost or fragmented in such analyses. This phenomenon is more or less pronounced depending on the settings and the algorithm used as shown for H3K9me3 in Fig. S5B. Further, in peak calling it is often challenging to identify and remove false-positive/negative peaks that are caused by the inherent properties of sequencing data sets like noise, artificial overrepresentation of particular genomic regions [30, 31] or insufficient read coverage [13]. An example for H3K36me3 is shown in Fig. S5C. Finally, nested structures like broad domains that contain smaller highly enriched domains cannot be reproduced with a set of non-overlapping peaks (Fig. S1). These issues complicate the quantitative assessment of patterns and domain structures from lists of enriched regions. MCORE, however, retrieves information about patterns upstream of peak calling analyses and is relatively robust towards uncertainties at individual loci because correlation functions are calculated from the entire collection of sequencing reads in a large genomic region (see Fig. S6, S7 for the influence of read coverage).

Interpretation and quantification of correlation functions

We quantified the information contained in correlation functions by first analyzing their decay spectrum in a model-independent manner and by subsequently fitting a generic model function [25] as described in the Materials and Methods section. This is illustrated for a simulated data set in Fig. S8. As a first step, inflection points (in logarithmic representation) were numerically determined, yielding the decay lengths that are present in the correlation function. Depending on the type of correlation function these decay lengths represent domain sizes or separation distances (Fig. 1C). Furthermore, the Gardner transformation was computed, which displayed peaks at the characteristic decay lengths [32]. For multi-exponential decays the amplitudes of the Gardner spectrum are directly related to the abundance of the respective component. Both approaches were independent of input parameters or model assumptions. Finally, we fitted the correlation function to quantitatively describe the domain size spectrum (Materials and Methods). Because decay lengths and nucleosome repeat length follow from the change of the correlation coefficient with shift distance, these parameters are independent of the absolute correlation amplitude. This is beneficial for the analysis of data sets that are not properly normalized due to low sequencing depth or lack of suitable control samples.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Correlation functions can be compared to each other based on errors obtained from Fisher transformation or bootstrapping (Fig. 2, Fig. S9, Materials and Methods). These errors reflect variations of the correlation coefficient among different positions within the genomic region of interest. Whereas Fisher transformation is exact for normally distributed enrichment data, non-parametric bootstrapping is generally applicable. To validate genome-wide relationships or domain topologies we found it instructive to assess variations among different chromosomes. If more than two replicates were available, replicate correlation functions calculated for each combination of independent samples were combined to account for differences among experiments (Fig. 2, Materials and Methods). To compare cross-correlation functions in this manner at least two replicates for each interrogated feature were required. We found these errors most meaningful because the variability among biological replicates can typically not be neglected and should be used as a reference when comparing different correlation functions to each other.

We found that the shape and the amplitudes of correlation functions were well reproducible when normalized according to the workflow described above. This was also true when comparing our samples with published histone modification ChIP-seq samples from other labs (Fig. 2C, Fig. S10). Because correlation curves are series of normalized correlation coefficients, pair-wise comparison and statistical testing can be conducted at every shift distance based on the respective value and its error (Fig. S9). In general, two curves might not be globally different from each other for every shift distance but might nevertheless exhibit significant differences on a specific length scale.

In summary, MCORE yields compact genome-wide representations of chromatin features in the form of correlation functions that can be quantitatively evaluated and compared to each other. It allows to (i) determine domain topologies (Fig. 1B), (ii) assess spatial relationships (Fig. 1C), (iii) test the reproducibility of experiments, or (iv) assess variations caused by changes in experimental conditions, e.g. the use of antibodies from different suppliers (Fig. S11). In contrast to the Pearson correlation coefficient between two data sets alone, the normalized correlation function provides insight into the similarity of the data sets on a broad range of length scales. Thus, MCORE can detect changes in domain size, amplitude or relative genomic position and can be used to track the re-organization of the epigenome among different cell types as shown below.

Domain topology and nucleosome pattern of modified regions in ESCs and NCs

We used replicate correlation functions to dissect the domain structures and nucleosome patterns in ESCs and NCs throughout the genome (Fig. 3A, B, Fig. S12 and Tables S2-3). Most features studied here, such as the histone modification H3K9me3, displayed complex

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

domain size distributions with multiple characteristic length scales (Fig. 3A, B). An exception was H3K4me3, which in agreement with published data [33] formed almost exclusively distinct peaks of roughly 1900 bp or 9-10 nucleosomes in size in both ESCs and NCs. For H3K36me3 we found a typical domain size of 24-30 kb, which is of the same order of magnitude as the average gene length in the mouse genome (according to NCBI Build 37, mm9). The nucleosome repeat length varied among domains carrying different histone modifications, with 218 bp for H3K27me3 in NCs and 182 bp for H3K9me3 and H3K36me3 in NCs (Tables S2-3). This suggests that nucleosome spacing is differentially regulated and linked to the chromatin state, consistent with previous reports [27, 34].

The initial decay of most replicate correlation functions is caused by the reduced probability to find the same modification at the neighboring nucleosome and is therefore associated with a domain size of a single nucleosome. Notably, a prerequisite for this interpretation is that the occupancy profile is properly normalized and not heavily undersampled, which is validated for representative profiles in Fig. S6 and Fig. S7. Accordingly, homogenous domains that primarily contain equally modified nucleosomes produce a weaker initial decay than domains that contain a mixture of modified and non-modified or differently modified nucleosomes. Whereas the subtle initial decay for H3K4me3 in ESCs and NCs (Fig. 3A, B and Tables S2-3) is indicative of homogenous domains, the pronounced decay for H3K9me3 in NCs (Fig. 3B, Table S3) suggests that this modification forms discontinuous domains with gaps. This is corroborated by the absence of isolated nucleosomes with high H3K9me3 enrichment levels outside broader domains (Fig. S13), which could also be responsible for a steep decay in the correlation function because such nucleosomes would have unmethylated neighbors.

In summary, these observations indicate that different histone modifications form domains with different topology. Based on the domain size and frequency distribution obtained from MCORE, an assignment to specific genomic loci can be made, e.g. by evaluating the normalized occupancy profiles with a sliding window corresponding to a domain size of interest. This is illustrated in Fig. 3C and Fig. S14 for broad H3K9me3 domains, which according to MCORE prevailed in NCs.

Changes in chromatin patterns during stem cell differentiation

To identify changes of chromatin features during stem cell differentiation we conducted a comparative MCORE analysis of more than 60 deep sequencing data sets from ChIP-seq (histone modifications: H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3, binding sites of RNA polymerase II (RNAP II) and transcription factors TAF3, Oct4 and Otx2), BS-seq, RNA-sequencing (RNA-seq), Hi-C and RNAP II ChIA-PET experiments in

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ESCs and NCs (Fig. 3, 4, Fig. S15-18, Table S4). Normalized correlation amplitudes at zero shift distance were assembled into a matrix (Fig. 4A, red/blue), which reflects co-localization or mutually exclusive localization of different features. In both cell types we found more co-localizations than mutual exclusions, which suggests that the set of chromatin features analyzed here tends to localize to the same part of the genome. In general, mutual exclusions were weaker than co-localizations as judged by the absolute values of the respective normalized correlation coefficients.

In ESCs, the strongest co-localization was found among features related to actively transcribed genes (H3K4me1, H3K4me3, H3K27ac, H3K36me3, RNAP II, RNAP II ChIA-PET). Notably, H3K36me3, which is known to be associated with active genes, also co-localized with H3K9me3/H3K27me3, which are traditionally considered heterochromatin marks. This might reflect (i) the presence of repressed genes not devoid of H3K36me3 [35], (ii) the occurrence of H3K9me3 and H3K27me3 at active genes [33], and/or (iii) the presence of H3K36me3 domains outside of coding genes (Fig. S19). Mutual exclusion was found between RNAP II and the repressive marks H3K27me3 and 5mC (but not H3K9me3) in ESCs. Furthermore, inter-chromosomal contact sites were depleted around H3K27me3 in ESCs, indicating that H3K27me3 domains localized preferentially inside chromosome territories.

In NCs, co-localization among features associated with active chromatin was conserved and tended to become stronger (Fig. 4A). Most activating modifications retained their domain size structures and genomic positions on a global level (Fig. S16). In contrast, H3K9me3 and H3K27me3 redistributed during differentiation in a way that their co-localization with each other, with 5mC and with some of the activating marks like H3K4me1 increased (Fig. 4A, D, Fig. S17). In particular, the following changes are noteworthy: (i) Both the H3K9me3 and H3K27me3 modification formed broader domains in NCs compared to ESCs, which led to a stretched decay in correlation functions for NCs compared to the steeper decays in correlation functions for ESCs (Fig. 3A, B, Fig. 4B). (ii) The normalized correlation of H3K9me3 between ESCs and NCs decreased compared to the normalized correlation between replicates from the same cell type (Fig. 4B). The same tendency was observed for H3K27me3. These differences suggest partial re-location of H3K9me3/H3K27me3 during differentiation because otherwise correlation functions between ESCs and NCs would resemble the correlation function calculated for the replicates from the same cell type, and all curves in each panel would essentially be identical. (iii) The normalized correlation between H3K9me3 and H3K27me3 increased in NCs (Fig. 4C), which is indicative of stronger co-localization of both marks in NC ensembles. (iv) Correlation functions for 5mC in ESCs, NCs and between both cell types were similar (Fig. S16). Thus, global changes in the

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genome-wide 5mC pattern were minor, consistent with previous findings [33]. (v) The normalized correlation between H3K27me3 and 5mC was higher in NCs compared to ESCs (Fig. 4A, Fig. S18A), suggesting stronger co-localization of both marks. Normalized correlation between H3K9me3 and 5mC increased for large shift distances in NCs, implying that extended H3K9me3 domains formed around pre-existing 5mC sites (Fig. S18A). (vi) Substantial mutual exclusion was found between H3K9me3 and inter-chromosomal contacts in NCs but not in ESCs, which suggests that H3K9me3 was re-localized to the interior part of chromosome territories (Fig. 4C). H3K27me3 resided preferentially inside chromosome territories already in ESCs and did not change its position in NCs (Fig. 4C).

Differential relationships among chromatin features in ESCs and NCs

Next, we determined the characteristic genomic separation distance for each pair of features (Fig. 4A, green color coding). Whereas correlation functions for co-localizing features tend to decrease monotonously, correlation functions for shifted features exhibit local maxima at their characteristic separation distance (Fig. 1C). Correlation functions for features that co-localize at some places in the genome and are shifted with respect to each other at other places exhibit an initial decay that is followed by local maxima (Fig. 4C, D). This type of information is lost in evaluation schemes that exclusively assess overlap. Only for simple cases, such as H3K4me3 and H3K36me3 that localize side by side at promoters and bodies of active genes (Fig. S5), similar information can be obtained by determining distances between adjacent peaks across data sets (Fig. 4E, F). However, if at least one of the chromatin features of interest forms broad domains that become partitioned into several smaller enriched regions by peak calling, information beyond the scale of a peak or the distance between neighboring peaks is not accessible. Examples for pairs of features that are shifted with respect to each other in ESCs but overlap and co-localize in NCs are H3K4me1-H3K9me3, H3K4me3-H3K27me3 and H3K9me3-H3K27ac (Fig. 4A, D). These changes are in agreement with the global reorganization of H3K9me3 and H3K27me3 in NCs described above.

Network models for relationships among chromatin features on multiple scales

The cross-correlation functions introduced above represent the scale-dependent relationships between pairs of chromatin features. Accordingly, we used these values to construct network models that reflect the associations among all features assessed here for a particular genomic scale (Fig. 5). Features were arranged based on their associations at zero shift distance, with positively correlated features positioned close to each other (Materials and Methods). As described above, activating histone modifications such as

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

H3K4me1, H3K4me3 and H3K27ac co-localized with RNAP II and RNAP II ChIA-PET sites in both ESCs and NCs. Repressive marks including H3K9me3, H3K27me3 and 5mC were also positively associated with each other, with stronger correlations in NCs than in ESCs. H3K36me3 exhibited positive correlations with both activating marks and repressive marks. For increasing genomic distances, associations among different features changed in a characteristic manner, reflecting the action of mechanisms that establish chromatin patterns on different scales. Activating features were still associated with other activating features at the adjacent nucleosome (200 bp shift), which indicates that they form chromatin domains that extend beyond a single nucleosome. In contrast, the cross-correlation among repressive marks at neighboring nucleosomes decreased considerably compared to their correlation at the same nucleosome. This indicates the presence of nucleosomes (without an equally modified neighbor) that either carry at least two repressive marks simultaneously, switch between two different repressive marks over time, or stably carry different repressive marks in different cells. All of these scenarios would produce positive correlation in the ensemble average. At a shift distance of about ten nucleosomes (2000 bp), most associations among activating histone modifications were lost, reflecting the relatively limited spatial extension of the respective domains (Tables S2-3). In contrast, correlations between repressive marks decreased only moderately when comparing values for shift distances of one and ten nucleosomes. This finding is consistent with their occurrence in broad domains with low enrichment levels, i.e. the presence of large genomic regions that contain repressive marks at moderate density. The differential scale-dependence found for relationships among activating and among repressive marks suggests distinct topologies of the respective chromatin domains and thus fundamental differences in the mechanisms for their establishment and maintenance.

Reorganization of heterochromatin components

To further investigate the changes in heterochromatin organization during differentiation of ESCs into NCs inferred from the MCORE analysis above (Fig. 4), we dissected the core part of the network around H3K9me3. To this end we compared the distributions of the H3K9me3 mark, the histone methyltransferase SUV39H1 that sets this mark in pericentric heterochromatin, and the two heterochromatin protein 1 isoforms HP1α and HP1β to each other. Both SUV39H1 and HP1 contain chromodomains that recognize H3K9me3, but the contribution of these interactions to their genome-wide binding profiles has not been studied comprehensively. First, we asked if the two HP1 isoforms displayed cell type-specific chromatin interaction patterns. We found that the genomic distributions of HP1α and HP1β were different from each other in both ESCs (Fig. 6 A-C) and NCs (Fig. 6 D-F). In ESCs,

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HP1β formed broader domains than HP1α (Fig. 6A) that were less correlated with H3K9me3 (Fig. 6B) but rather overlapped with H3K36me3 (Fig. 6C). This somewhat surprising finding supports recent work, which showed that HP1β but not HP1α is essential for proper differentiation and maintenance of pluripotency in ESCs, where it is enriched in exons but not in pericentric heterochromatin [36]. The nuclear distribution of HP1β in ESCs might be related to its function in splicing [37]. In NCs, HP1α and HP1β displayed moderate differences in their domain structure (Fig. 6D, G), with a stronger preference of HP1αfor broad domains. In contrast to ESCs, both isoforms strongly co-localized with H3K9me3 in NCs (Fig. 6 E), in line with their well-established role as heterochromatin components in differentiated cells ([20] and references therein). Co-localization with H3K36me3 was also observed (Fig. 6F), consistent with the overlap between H3K9me3 and H3K36me3 domains in NCs found above. Next, we assessed the composition of H3K9me3 domains in NCs. H3K9me3 formed both broad and intermediately sized domains (Fig. 6D, Fig. 6G) but SUV39H1 formed only broad domains. This is apparent from the fast decay of the replicate correlation function for SUV39H1 (Fig. 6D, red). Consistently, co-localization among HP1α/β, SUV39H1 and H3K9me3 was only found within broad domains (Fig. 6E).

These findings suggest the presence of SUV39H1-independent H3K9me3 domains in NCs, which have also been described in ESCs [38], and indicate that the presence of H3K9me3 is not sufficient for stably recruiting SUV39H1 or HP1 to chromatin. The fact that different types of H3K9me3 domains within the same cell vary in molecular composition is consistent with a looping mechanism in which domains are established and maintained around high-affinity nucleation sites [20, 39] that primarily determine the genome-wide distribution of SUV39H1 (Fig. 6H). H3K9me3 domains established by other methyltransferases are expected to lack such nucleation sites and therefore to contain only transiently bound SUV39H1 molecules that are recruited via the weaker interaction between H3K9me3 and their chromodomain. Whereas a looping model does not rely on H3K9me3-dependent SUV39H1 recruitment to enable spreading, this type of recruitment is essential for feedback-driven spreading schemes. According to these, every nucleosome carrying H3K9me3 should efficiently recruit SUV39H1, no matter where it is located within the genome or how the surrounding domain was established, and H3K9me3-dependent recruitment of SUV39H1 should affect the size and stability of all H3K9me3 domains in the same way. Although further experiments are required to fully understand the underlying molecular details of heterochromatin reorganization during differentiation, these observations suggest that broad H3K9me3 domains in NCs are formed by site-specific recruitment of SUV39H1 and that H3K9me3-dependent recruitment of SUV39H1 is much weaker. These insights provide a starting point

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to uncover the pathways that are responsible for establishing differently sized heterochromatin domains with distinct molecular composition.

Model for changes of chromatin features during differentiation

The MCORE results on domain size distributions (Fig. 3), co-localizations and separation distances (Fig. 4, Fig. 5) lead us to the model for the reorganization of chromatin during differentiation of ESCs into NCs depicted in Fig. 7. H3K9me3 and H3K27me3 domains become larger, stronger co-localized (Fig. 4B, C) and rearranged around sites of preexisting 5mC during the transition from ESCs to NCs (Fig. S18A). This rearrangement leads to a number of alterations in the relationships between H3K9me3/H3K27me3/5mC and other chromatin features in NCs: (i) H3K27me3 and H3K9me3 co-localized stronger with active marks including H3K4me1, H3K4me3, H3K27ac and RNAP II as well as H3K36me3 (Fig. 4A, Fig. 5). (ii) 5mC co-localized somewhat stronger with H3K36me3 (Fig. 4A, Fig. S18A). (iii) Whereas 5mC and H3K27me3 were already depleted from the surface of the chromosome territory in ESCs (Fig. 4C, Fig. S18B), H3K9me3 moved into the interior of the territory in NCs (Fig. 4C). The positive correlations between H3K4me1-H3K27me3, H3K4me1-H3K9me3 and H3K4me1-H3K36me3 remained stronger in NCs than in ESCs also on larger genomic scales up to ten nucleosomes (Fig. 4D, Fig. 5, Fig. S17), indicating that they are caused by NC-specific broad domains. In summary, these findings suggest that the main chromatin transition during differentiation from ESCs into NCs is the rearrangement of H3K9me3/H3K27me3 domains, which in NCs extend beyond repressive heterochromatin and overlap at least to some extent with chromatin regions that carry activating histone marks. The observations in Fig. 6 suggest that site-specific nucleation sites rather than pre-existing H3K9me3 domains are responsible for recruiting the SUV39H1 methyltransferase to genomic regions covered by broad H3K9me3 domains in NCs.

Discussion

The quantitative understanding of how cells organize their genome into cell type-specific chromatin states is important for the description of all processes that require access to the genetic information. While the effects of soluble enzymes can be represented by simple rate equations, the polymeric nature of chromatin introduces a spatial relationship among nucleosome states. As a result, nucleosomes are influenced by the adjacent chromatin segments and patterns can form along the genomic coordinate. These patterns are present on different length scales and represent an extra layer of complexity, which is an essential part of the regulatory networks that control genome functions. For example, repressive

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

histone modifications form broad domains that are relatively independent from the underlying DNA sequence and can be transmitted through at least several cell divisions [20, 40-42]. Furthermore, chromosomes fold into topological domains that determine the contact frequencies between genomic loci and the proteins they are decorated with [43], thereby creating structural patterns also in three dimensions. Elucidating the mechanistic basis of these phenomena and the functional relationships among them requires techniques that can identify, quantitate and compare different topologies along the genome.

Global analysis of deep sequencing data by correlation functions

Several methods for the analysis of genome-wide sequencing data sets have been developed (see Table S1 for a non-comprehensive list). Most of them analyze deep sequencing data on the level of individual genomic positions or bins to produce lists of enriched regions. Unfortunately, this procedure is complicated by noise, bias and undersampling [13-15], and it is not straightforward to choose a threshold value for classifying enriched regions because low values lead to false-positive peaks and high values lead to false-negative results. Consequently, identifying differences in the chromatin domain topology between samples is fraught with difficulties as evident from a comparison of 14 different software tools for differential ChIP-seq analysis that yield different results [44]. These problems are especially detrimental for the analysis of broad regions with low enrichment levels that are common to heterochromatin. One popular approach to deduce information more robustly is the generation of aggregate plots around known genomic elements, such as transcription start or termination sites [45-47]. These plots sacrifice single-locus resolution in order to decrease the contribution of noise and bias and to increase statistical power. Such plots require *a priori* knowledge about the position of genomic elements to align with and are therefore of only limited applicability for the study of features that are not exclusively present in the vicinity of annotated genes (Fig. S19) or other known genomic elements.

The MCORE method introduced here uses correlation functions to find and quantify chromatin patterns. It computes Pearson correlation coefficients as underlying metrics, which is a convenient measure that has extensively been used for data comparison and statistical inference in many fields including deep sequencing analysis [16, 26, 27, 48]. When calculating correlation functions, MCORE implicitly combines multiple genomic regions to gain a correlation coefficient for each shift distance, yielding statistical robustness from a large number of reads. In this manner MCORE can quickly retrieve information on the spatial distribution of chromatin features on all length scales, while avoiding assumptions or model-dependent parameter settings like significance thresholds (Fig. S1). In contrast to

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

aggregate plots MCORE does not rely on any *a priori* knowledge about annotated genomic elements. Compared to peak calling [13], MCORE has a relatively low sensitivity to undersampling. As illustrated in Fig. S6 and Fig. S7, domain sizes obtained from normalized correlation functions for the features tested here are unaffected by a reduction of the read number down to about 12 million reads. This might be beneficial for the analysis of data sets that have low complexity, e.g. due to limitations in input material as it is the case for low input sequencing samples, or insufficient sequencing depth, which seems to be the norm for broadly distributed histone modifications [13]. Domain abundances obtained from data sets with different coverage values exhibited somewhat larger changes than domain sizes (Fig. S7). Therefore, sufficient coverage should be ensured in order to interpret these parameters, e.g. by applying MCORE to diluted data as shown in Fig. S6 and Fig. S7.

A crucial step in the MCORE workflow is correction for bias and background. Without this step artificially overrepresented regions and non-specific signal can induce similarities between data sets that are unrelated to the chromatin feature of interest. These phenomena are well known from other deep sequencing analysis methods. Because different artifacts affect the signal on different scales, their contribution and successful correction can better be assessed by multi-scale methods than by techniques that operate on a single scale. In particular, non-specific background leads to a characteristic correlation spectrum (Fig. S2B, Fig. S4), whose removal can and should be validated using the proper controls. Based on a single correlation coefficient between data sets this task is more difficult to accomplish.

Genome-wide topology of chromatin domains

MCORE extends previous techniques that assess co-localizations of chromatin features based on correlation coefficients. By evaluating entire correlation functions instead of single correlation coefficients the spatial extension of chromatin patterns on multiple genomic scales is retrieved. With this analysis we found predominantly small domain sizes of less than 2 kb for promoter/enhancer marks H3K4me1, H3K4me3, H3K27ac and RNAP II, intermediate domain sizes of 20-30 kb for H3K36me3 that marks the whole gene body including flanking regions, and domain sizes up to several megabases for H3K9me3/H3K27me3 (Fig. 3, Fig. S16, Tables S2-3). This is consistent with the size of promoters, enhancers and active genes, and with the estimates for repressive domains that were made based on visual inspection of selected genomic regions [49].

The scale-dependent relationships determined by MCORE for different histone modifications suggest that there are three types of domain topologies: (i) Short domains formed by activating marks are relatively homogenously modified, which is reflected by a large probability for finding the same or another activating modification at the next nucleosome.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Accordingly, correlation functions for activating marks such as H3K4me3 displayed only a moderate initial decay (Fig. 3), which is reflected by a low abundance of domains of the size of single nucleosomes (Tables S2-3). (ii) H3K36me3 formed domains of intermediate size that were roughly one order of magnitude broader than H3K4me3 domains. The stronger initial decay (Fig. 3) suggests the presence of short domains that are formed by single nucleosomes without an equally modified neighbor (Tables S2-3), which is consistent with the presence of more gaps in H3K36me3 domains as compared to H3K4me3 domains. (iii) Especially in NCs replicate correlation functions for H3K9me3 or H3K27me3 displayed long-range correlations that extended to shift distances of several megabases. Similar scale-dependence was also seen for correlation functions between H3K9me3 and H3K27me3 (Fig. 4C), suggesting that these domains intermingle. The respective correlation functions displayed a relatively fast decay at a shift distance of one nucleosome (Fig. 3, Fig. 4, Tables S2-3), indicating that many modified nucleosomes localize next to a non-modified or differently modified one.

The topology inferred here for H3K9me3/H3K27me3 domains (Fig. 7) fits well to the experimental observation of broad domains and low enrichment levels in the cell ensemble. In particular, the experimentally determined methylation levels that are below 50 % even for H3K9me3 in pericentric heterochromatin (see [20] and references therein) are incompatible with large stretches of adjacent fully H3K9me3-modified nucleosomes. The topology found here is consistent with a model in which methylation marks are stochastically propagated from well-positioned nucleation sites via dynamic chromatin looping [20].

Comparison of chromatin domains in ESCs versus NCs

The comparative analysis of 11 different chromatin features in ESCs and NCs conducted here shows that MCORE can efficiently identify and compare chromatin domain patterns. By integrating genome-wide data sets with very different readouts MCORE is particularly suited to assess the interplay between spatial genomic architecture and epigenetic signaling and to generate hypotheses that can be further validated in downstream applications.

The positive correlations we found among activating histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3), among repressive histone modifications (H3K9me3, H3K27me3, 5mC) and between H3K36me3 and repressive marks are in qualitative agreement with previous studies conducted with ESCs and other cell types [49-51]. Genome-wide co-localization of marks that were originally thought to affect transcription antagonistically might reflect the additional functions of these marks that are unrelated to the regulation of gene expression. For example, H3K9me3 is not restricted to heterochromatin but is also found at active genes [33, 52], H3K9me3, H3K27me3 and H3K36me3 have been

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

linked to alternative splicing [37, 53] and large portions of H3K9me3 and H3K27me3 localize to intergenic regions where they might serve completely different functions (Fig. S19, [50]). In addition, differences between a gene locus during different cell cycle stages, between alleles within the same cell, or between different cells in the ensemble can induce positive correlation between marks that do not co-localize on the same molecule, which is why conclusions drawn from sequencing data inherently refer to the average of the cell population that was analyzed. The finding that correlations among different marks are generally smaller in ESCs than in NCs fits to the model of plastic and 'hyperactive' chromatin in stem cells, which acquires distinct patterns only upon differentiation [54]. The fact that most 5mC regions persisted in ESCs and NCs (Fig. S16), were moderately depleted for inter-chromosomal contacts in both cell types (Fig. S18B), and gained H3K9me3 in NCs (Fig. S18A) suggests a model in which heterochromatic regions newly established in NCs are preferentially buried within chromosome territories (Fig. 7). H3K27me3 domains in both cell types also were found to adopt this configuration. This model fits very well to the previously reported localization of inactive domains inside chromosome territories in differentiated cells, including the H3K27me3-rich domains found in silenced Hox clusters [11, 55-57]. The observation that only a subset of H3K9me3 domains is broad and enriched for SUV39H1 suggests that heterochromatin extensions is not primarily caused by recruitment of *trans*-acting factors to preexisting H3K9me3 but rather by site-specific nucleation of SUV39H1 to domains that are to be extended during differentiation.

Conclusions

The MCORE method introduced here enables the quantitative retrieval and comparison of domain topologies and spatial relationships for different chromatin features from noisy data sets. These features make MCORE complementary to model-dependent approaches that assess the local read density at individual loci to find enriched regions. MCORE is relatively fast and yields a coarse-grained comparison of data sets that does not require user-defined input parameters, providing an unbiased starting point for in-depth analyses. We anticipate that this capability will prove to be valuable to cope with the deluge of genome-wide sequencing data sets arising from the analysis of small cell populations or even single cells both in the context of basic research and personalized medicine [4, 58].

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and Methods

Calculation of normalized occupancy profiles

Sequencing reads were mapped to the mouse mm9 assembly using Bowtie [59]. Only uniquely mapping hits without mismatches were considered and duplicates were removed. Mapped reads were processed according to the following steps: Bisulfite sequencing (BS-seq) data, which are used to map DNA methylation at single base pair resolution, are usually available as methylation scores calculated from the ratio of converted reads divided by the sum of converted and unconverted reads at a given position. These can be directly used for computing the correlation function as described below. For all other sequencing readouts the coverage was initially calculated for each chromosome by extending the reads to fragment length, yielding a histogram with the genomic coordinate on the *x*-axis and the number of reads per base pair on the *y*-axis. For Hi-C and ChIA-PET data only inter-chromosomal reads were considered. To calculate normalized occupancy profiles, samples were processed depending on the type of experiment. In general, it is important to account for fragmentation bias, library preparation bias and genome mappability. All these multiplicative biases are also included in the input sample and should cancel out in the ratio of specific signal *A* and input signal *I* (A/I). In RNA-seq experiments the input signal can be replaced by a sample of nucleosome-free, fragmented genomic DNA. For immunoprecipitation experiments, it is additionally important to account for the non-specific binding during sample preparation to obtain meaningful correlation functions (Fig. S2A). This is of increasing importance for decreasing signal-to-background ratio (Fig. S2B). The appropriate control, read coverage *C,* can be obtained from an immunoprecipitation with a non-specifically binding antibody (e.g. IgG control) or from a sample that lacks the antigen of interest (e.g. a knockout cell line). Accordingly, we devised the following strategy to compute normalized occupancy profiles that were used in the subsequent analysis. First, the normalized coverage of the control C_{norm} and of the specific immunoprecipitation sample A_{norm} were obtained by dividing by input signal according to Eq. 1:

$$C_{\mathrm{norm}} = \frac{C/I}{\langle C/I \rangle} \quad \text{and} \quad A_{\mathrm{norm}} = \frac{A/I}{\langle A/I \rangle} \tag{1}$$

Here, $\langle \ldots \rangle$ denotes averaging along the genomic coordinate. For the calculation of coverage (C/I and A/I) and average values ($\langle C/I \rangle$ and $\langle A/I \rangle$), positions with zero input coverage were neglected. Subsequently, the coverage at these positions was set to the respective average value ($\langle C/I \rangle$ or $\langle A/I \rangle$) that was calculated for the remaining positions, which eliminates fluctuations and corresponding contributions to the correlation coefficient.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the next step, non-specific background signal was removed to obtain the normalized read occupancy *O*:

$$O = A_{\text{norm}} - b \cdot C_{\text{norm}} \tag{2}$$

In Eq. 2, the parameter *b* quantifies the contribution of the control signal present as background in the sample (IP). To estimate *b*, we minimized the absolute value of the Pearson correlation coefficient r_0 at zero shift distance between the normalized occupancy *O* and the control coverage C_{norm} according to Eq. 3:

$$r_0 = \left| \frac{\sum_{i=1}^{n} \left(O_{\text{i}} - \langle O \rangle\right)\left(C_{\text{norm,i}} - \langle C_{\text{norm}} \rangle\right)}{\sqrt{\sum_{i=1}^{n} \left(O_{\text{i}} - \langle O \rangle\right)^2 \cdot \sum_{i=1}^{n} \left(C_{\text{norm,i}} - \langle C_{\text{norm}} \rangle\right)^2}} \right| \tag{3}$$

Here, *n* denotes the maximum genomic position considered for the calculation, which is typically the chromosome length. For the minimization procedure, *b* was changed between 0 and 1. Because the minimum correlation $r_0(b)$ indicates the lowest similarity between normalized occupancy profile and control, the corresponding *b* value was used for normalization according to Eq. 2.

Computation of correlation functions

The Pearson correlation coefficient *r* at shift distance Δx was calculated for the corrected data sets after shifting the two occupancy profiles with respect to each other by Δx base pairs according to Eq. 4 (equivalent to Eq. 3 but with a second shifted occupancy instead of the control coverage):

$$r(\Delta x) = \frac{\frac{1}{2} \sum_{i=1}^{n-\Delta x} \left[\left(O_{1,i} - \langle O_1 \rangle\right)\left(O_{2,i+\Delta x} - \langle O_2 \rangle\right) + \left(O_{1,i+\Delta x} - \langle O_1 \rangle\right)\left(O_{2,i} - \langle O_2 \rangle\right) \right]}{\sqrt{\sum_{i=1}^{n} \left(O_{1,i} - \langle O_1 \rangle\right)^2 \sum_{i=1}^{n} \left(O_{2,i} - \langle O_2 \rangle\right)^2}} \tag{4}$$

To sample the correlation function in a quasi-logarithmic manner [60], profiles were binned by a factor of 2 after 25 shift operations, which doubles the step size. To preserve high resolution for small shift distances, the first binning operation was carried out at a shift of Δx = 50 bp. This calculation was done for each chromosome separately because continuous domains cannot exceed chromosomal ends. Combinations of genomic positions beyond chromosome ends were neglected. Correlation functions for single chromosomes were averaged or compared to each other. In the manuscript, most correlation functions refer to chromosome 1, which is representative for all chromosomes as judged by the relatively

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

small deviations between chromosomes (Fig. 2B, Fig. 3A,B). However, correlation functions can also be calculated for smaller genomic regions (see Fig. S1 for the correlation function for a single domain).

To compare cross-correlation functions between different features normalization to the geometric mean of the two replicate correlation functions was conducted:

$$r_{\mathrm{norm}}(\Delta x) = \frac{r_{\mathrm{c}}(\Delta x)}{\sqrt{\left|r_1(0) \cdot r_2(0)\right|}} \tag{5}$$

Here, r_{c} is the cross-correlation coefficient at a given shift distance Δx, and r_1 and r_2 are the replicate correlation coefficients of the data sets used. This normalization step accounts for differences in the distributions of the features involved. For calculating the cross-correlation functions between two different features or the same feature in two different cell types at least two replicates for each sample were used. Accordingly, a cross-correlation function for each combination was computed, which results in n^2 functions for n replicates of each sample, and average and standard error were calculated based on all correlation functions.

Statistical analysis of correlation functions

Statistical analysis of data was conducted by computing standard errors and 95% confidence intervals. To assess significance and associated errors/confidence intervals for a given correlation function the following types of variations have to be considered.

Statistical error of the computed correlation function. Because correlation functions are calculated from millions of regions they have a very small statistical error. The sample size N for each shift distance Δx is given by the distance between the first and last position that is covered on the chromosome (P_{min} and P_{max}) subtracted by the shift length (Δx) with $N(\Delta x) = P_{max} - P_{min} - \Delta x$. Based on the sample size, 95% confidence intervals are obtained using the Fisher transformation [61, 62]. Because the shift distance is typically much smaller than the length of the chromosome for which the correlation coefficient is calculated, the sample size is very large and the error of the correlation coefficient is very small (Fig. 2A). As normalized occupancy values O_{i} typically follow a normal distribution reasonably well (Fig. 2D), the Fisher transformation is a good way to estimate confidence intervals for correlation coefficients. An alternative non-parametric option that is compatible with arbitrary sample distributions is bootstrapping [63]. To this end, occupancy profiles are resampled with replacement in pairs ($O_{1,\mathrm{i}}$, $O_{2,\mathrm{i}+\Delta x}$), which are subsequently used for calculation of the correlation coefficient according to Eq. 4. This procedure is repeated multiple times to obtain a distribution of correlation coefficients for every pair of resampled occupancy profiles (Fig. 2E) and every shift distance Δx. Based on the width of this

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distribution estimates for confidence intervals are obtained. For the cases tested here, bootstrapping yielded moderately larger confidence intervals than those obtained using Fisher transformation, but intervals from both methods were of the same order of magnitude (Fig. 2F).

*Variation between chromosome*s. An estimate for the error of genome-wide domain structures or relationships can be obtained by comparing correlation functions calculated for different chromosomes as shown in Fig. 2B. If the relationship is governed by the same biological mechanism on all chromosomes this variation can be used to evaluate the error.

Reproducibility of experiments. Sample preparation might introduce a global bias into a given data set. This is generally true for deep sequencing experiments irrespectively of which method is used for downstream analysis. Such variations between biological replicates might not be captured by statistical comparisons conducted on the basis of a single data set or a pair of data sets. The reproducibility of the experiment can be assessed with MCORE for data sets with at least three different biological replicates by computing the correlation function between all possible combinations of samples. For n replicates this yields $n\cdot(n-1)/2$ correlation functions. For these groups the correlation coefficients at a given shift distance can be compared. We found this approach to be particularly useful to identify variations due to different experimental conditions. For example, we evaluated the changes of ChIP-seq results after using antibodies from different companies (Fig. 2C and Fig. S11).

Comparison of two correlation functions. Correlation functions obtained by MCORE represent series of normalized Pearson correlation coefficients for different shifts Δx between occupancy profiles. Amplitudes of correlation functions calculated between different features in one cell type reflect the co-localization of the respective features. Although correlation functions are calculated here for the whole chromosome, some data sets are restricted to a fraction of the genome due to their biological structure. For example, occupancy profiles from RNA-seq display distinct gaps because only a fraction of the genome is transcribed. Thus, correlation functions between features that cover the whole genome, such as ChIP-seq data for histone modifications, have a tendency to exhibit larger amplitudes than correlation functions involving discontinuous occupancy profiles, such as those from RNA-seq experiments. This has to be considered when comparing absolute correlation amplitudes and can be accounted for to some extent by the normalization according to Eq. 5. Because domain sizes and nucleosomal spacing can be obtained from the shape of the correlation function without considering the absolute correlation amplitude, such effects only influence conclusions about (anti-)co-localization.

After correlation functions, associated errors and confidence intervals have been computed the comparison of two functions can be performed according to standard statistical tests for

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which an *R*-script is included in the supplementary material. To assess if the difference between two independent correlation functions is statistically significant a *t*-test can be conducted for each shift distance Δx individually, i.e. via pairwise comparison of correlation coefficients. Evaluating the statistical significance of a difference between two correlation coefficients r_1 and r_2 is done by Fisher transformation and testing of the null hypothesis $r_1 - r_2 = 0$ (Fig. S9). To compare domain sizes or nucleosomal spacing between two correlation functions, the positions of inflection points (in logarithmic representation) or domain sizes obtained by the fit can be compared. Notably two correlation functions might be significantly different due to different amplitudes but encode the same domain size distribution, which is independent of amplitudes.

An alternative non-parametric test to assess the difference between two series is the Kolmogorov-Smirnov test [64]. Correlation functions are considered as distributions of correlation coefficients, and the cumulative distribution is calculated. To assess the difference between two correlation functions, the supremum of the difference between both cumulative distributions is determined, which can readily be transformed into the corresponding significance [64]. Whereas this type of analysis is less sophisticated compared to pairwise comparison of correlation coefficients on all genomic scales individually, the Kolmogorov-Smirnov test is a convenient option to decide if curves are globally similar or different. Replicates can either be integrated into the analysis by comparing the suprema between different sets of correlation functions or by simply considering the average and standard error of cumulative distributions.

Quantification of MCORE correlation functions

Correlation functions obtained by MCORE provide information on the overall degree of (anti-)correlation between two deep sequencing data sets but also reflect the underlying chromatin domain structure with respect to (i) the number of chromatin domains, (ii) the relative domain abundance, (iii) the length of the respective domains, and (iv) the nucleosome repeat length. To extract the domain size distribution of a given chromatin feature, two different strategies are implemented in MCORE, which differ in the level of complexity but yield similar information. The first approach is independent of user-defined settings and computes parameters for the domain size distribution from the inflection points of the correlation function in logarithmic representation and a Gardner transformation of the correlation function. The Gardner transformation characterizes the decay spectrum of a function in a non-parametric manner [32]. This workflow represents a robust approach to evaluate genome-wide features from deep sequencing data without input parameters. In particular, inflection points are completely model-independent, whereas the Gardner

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

spectrum makes the generic assumption that the decay spectrum can be approximated by a superposition of exponential functions. The second approach can be used to quantitatively describe the domain size distribution based on a fit function. For this purpose it is crucial to avoid over-fitting of the data. Accordingly, we implemented a complementary set of four fit options that allow a robust in-depth analysis of correlation functions reporting fit parameters and their errors and thus determining domain sizes and their relative abundance. The performance of the different fit approaches is described below and in the MCORE software manual. The workflow we used in this manuscript is validated with simulated data in Fig. S8.

Least-squares spectrum fit. The exponential decay spectrum for the correlation function is optimized by conventional non-linear least squares fitting. The amplitudes for a given number of (logarithmically spaced) domains are optimized to obtain a good fit. The goal of the spectrum fitting process is to determine the length scales that are present in the decay spectrum of the curve. To this end it is not always necessary to describe the shape of the correlation function perfectly. For example, the initial decay of the function is frequently too steep to be adequately fitted with a superposition of exponential functions. Nevertheless, decay lengths are typically obtained in a robust manner. The multi-exponential fit described below typically performs equally well in identifying length scales and provides a good description of the correlation function. Thus, the least-squares spectrum fit is only recommended if the multi-exponential fit does not converge properly, i.e., if it yields length scales that are very different from those determined by inflection points.

Maximum entropy method (MEM) spectrum fit. The exponential decay spectrum is fitted similar to the least-squares method. However, the entropy of the amplitude spectrum is maximized along with the fit quality. To this end, optimization is carried out in a parameter space that is spanned by the first derivative of the entropy and the first and second derivatives of the fit quality according to the approach described previously [65]. This fit option is only recommended if the number of components obtained from the least-squares spectrum fit is much larger than the number of inflection points.

Multi-exponential fit implemented in MCORE. For multi-exponential fitting the following equation consisting of a combination of exponential functions is used:

$$F(\Delta x) = \sum_i a_i \cdot \exp\left(-\frac{\Delta x}{b_i}\right)^{n_i} \tag{6}$$

The exponential terms describe the domain structure of the correlation function, with a_i, b_i and n_i yielding the abundance, the half width and the fuzziness of the i-th domain, respectively. Small exponents n_i correspond to long-tail decays in the domain size distribution.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multi-exponential fit in R. The multi-exponential fit implemented in *R* [66] (http://www.R-project.org) uses a combination of exponential functions (see Eq. 6) multiplied with an additional oscillating term to describe the correlation function:

$$F(\Delta x) = \left(c_1 + (1 - c_1) \cdot \cos\left(\frac{\Delta x}{c_2} \pi \right) \cdot \exp\left(-\frac{\Delta x}{c_3} \right) \right) \cdot \sum_i a_i \cdot \exp\left(-\frac{\Delta x}{b_i} \right)^{n_i} \quad (7)$$

The oscillating term accounts for the nucleosomal pattern, with parameters c_1 for the strength of the nucleosomal oscillation, c_2 representing the nucleosomal repeat length and c_3 the scale on which regular nucleosomal spacing is lost. When using this approach, the minimal number of exponential terms that yielded uncorrelated fit residuals was chosen.

Peak calling

Peak calling was done using MACS [8] and SICER [9] implemented in the Genomatix software suite version v3.20715 (Genomatix, Munich, Germany). Prior to peak calling reads were preprocessed as described above including mapping to the mouse mm9 assembly by Bowtie [59], considering only uniquely mapping hits without mismatches and removing duplicates. Peak calling was done using default parameters and the input as control file. For H3K36me3 MACS mfold level 5, 10 and 30 were tested, and mfold 5 was selected. For SICER the FDR threshold was set to 0.0001, a window size of 200 bp and a gap size of 600 bp were used for H3K9me3 and H3K36me3, and a window size of 200 bp and a gap size of 200 bp were used for H3K4me3.

Network models

Graphs for network models were created and plotted using Gephi (http://gephi.github.io). Nodes were manually prearranged, and their layout was optimized using the Fruchterman-Reingold algorithm [67], which adjusts node positions based on forces that act between nodes according to the respective correlation strength.

Sample preparation for histone ChIP-seq

ESCs and neural progenitor cells from 129P2/Ola mice were cultured and differentiated as published [29]. ChIP-seq experiments and mapping of reads to the mm9 assembly of the mouse genome was conducted as described previously [20]. In brief, 10^6 cells were cross-linked with 1% PFA and cell nuclei were prepared. Chromatin was sheared by sonication to mononucleosomal fragments. ChIP was carried out with antibodies (Abcam) against H3K4me1 (ab8895), H3K4me3 (ab8580), H3K9me3 (ab8898), H3K27ac (ab4729),

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

H3K27me3 (ab6002), H3K36me3 (ab9050) or an unspecific IgG from Acris (RA073 or PP500P). For further information see Table S4. Libraries were prepared according to Illumina standard protocols with external barcodes and were sequenced with 51 bp single-end reads on an Illumina HiSeq 2000 system. After sequencing, cluster imaging and base calling were conducted with the Illumina pipeline (Illumina). 20 - 30 Mio reads were obtained for each sample. Reads were uniquely mapped without mismatches to the mm9 mouse genome using Bowtie. For RNA-seq, cells were harvested and long RNAs were isolated with the miRNeasy Mini Kit (Qiagen), DNA was digested by DNase I (Promega) for 30 min at 37°C, and libraries were prepared using the Encore Complete RNA-Seq Library Systems (NuGEN).

MCORE software

An executable Java program, including a test data set and an R script for statistical testing of the difference between two correlation functions, is available in the supplemental material and can be downloaded at http://malone.bioquant.uni-heidelberg.de/software/mcore.

Accession codes

ChIP-seq data have been deposited to the GEO database under the accession number GSE61874.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank Caroline Bauer for valuable assistance, the DKFZ Genomics and Proteomics Core Facility for technical support and expertise, and Anne Rademacher, Katharina Müller-Ott and Daniel Duzdevich for comments on the manuscript. This work was supported by grant CA146 of the Cancer Research Cooperation Program between the DKFZ and the Israel Ministry of Science and Technology (MOST) and the projects ImmunoQuant (0316170B) and PRECiSe (031L0076A) of the German Federal Ministry of Education and Research (BMBF) as well as a DKFZ intramural grant to FE.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Zhou VW, Goren A, Bernstein BE: Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 2011; 12:7-18.

2. Polo SE, Jackson SP: Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes Dev* 2011; 25:409-433.

3. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P: Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013; 502:59-64.

4. Shapiro E, Biezuner T, Linnarsson S: Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* 2013; 14:618-630.

5. Schwartzman O, Tanay A: Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 2015; 16:716-726.

6. Chabbert CD, Adjalley SH, Klaus B, Fritsch ES, Gupta I, Pelechano V, Steinmetz LM: A high-throughput ChIP-Seq for large-scale chromatin studies. *Mol Syst Biol* 2015; 11:777.

7. Barski A, Cuddapah S, Cui K, Roh T, Schones D, Wang Z, Wei G, Chepelev I, Zhao K: High-resolution profiling of histone methylations in the human genome. *Cell* 2007; 129:823-837.

8. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008; 9:R137.

9. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 2009; 25:1952-1958.

10. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al: Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 2013; 41:827-841.

11. Bickmore WA, van Steensel B: Genome architecture: domain organization of interphase chromosomes. *Cell* 2013; 152:1270-1284.

12. Zacher B, Lidschreiber M, Cramer P, Gagneur J, Tresch A: Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. *Mol Syst Biol* 2014; 10:768.

13. Jung YL, Luquette LJ, Ho JW, Ferrari F, Tolstorukov M, Minoda A, Issner R, Epstein CB, Karpen GH, Kuroda MI, Park PJ: Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res* 2014; 42:e74.

14. Meyer CA, Liu XS: Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* 2014; 15:709-721.

15. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014; 15:121-132.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

16. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012; 22:1813-1831.

17. Szalkowski AM, Schmid CD: Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief Bioinform* 2011; 12:626-633.

18. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP: H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res* 2009; 19:221-233.

19. Filion GJ, van Steensel B: Reassessing the abundance of H3K9me2 chromatin domains in embryonic stem cells. *Nat Genet* 2010; 42:4; author reply 5-6.

20. Muller-Ott K, Erdel F, Matveeva A, Mallm JP, Rademacher A, Hahn M, Bauer C, Zhang Q, Kaltofen S, Schotta G, et al: Specificity, propagation, and memory of pericentric heterochromatin. *Mol Syst Biol* 2014; 10:746.

21. Wochner P, Gutt C, Autenrieth T, Demmer T, Bugaev V, Ortiz AD, Duri A, Zontone F, Grubel G, Dosch H: X-ray cross correlation analysis uncovers hidden local symmetries in disordered matter. *Proc Natl Acad Sci U S A* 2009; 106:11511-11514.

22. Baum M, Erdel F, Wachsmuth M, Rippe K: Retrieving the intracellular topology from multi-scale protein mobility mapping in living cells. *Nat Commun* 2014; 5:4494.

23. Podobnik B, Horvatic D, Petersen AM, Stanley HE: Cross-correlations between volume change and price change. *Proceedings of the National Academy of Sciences of the United States of America* 2009; 106:22079-22084.

24. Elson EL: Fluorescence correlation spectroscopy: past, present, future. *Biophysical journal* 2011; 101:2855-2870.

25. Sengupta P, Jovanovic-Talisman T, Skoko D, Renz M, Veatch SL, Lippincott-Schwartz J: Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. *Nat Methods* 2011; 8:969-975.

26. Kharchenko PV, Tolstorukov MY, Park PJ: Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 2008; 26:1351-1359.

27. Stanton KP, Parisi F, Strino F, Rabin N, Asp P, Kluger Y: Arpeggio: harmonic compression of ChIP-seq data reveals protein-chromatin interaction signatures. *Nucleic Acids Res* 2013; 41:e161.

28. Marinov GK, Kundaje A, Park PJ, Wold BJ: Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)* 2014; 4:209-223.

29. Teif VB, Vainshtein Y, Caudron-Herger M, Mallm JP, Marth C, Hofer T, Rippe K: Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol* 2012; 19:1185-1192.

30. Jain D, Baldi S, Zabel A, Straub T, Becker PB: Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res* 2015; 43:6959-6968.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

31. Carroll TS, Liang Z, Salama R, Stark R, de Santiago I: Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front Genet* 2014; 5:75.

32. Gardner DG, Gardner JC, Meinke WW: Method for the analysis of multicomponent exponential decay curves. *J Chem Phys* 1959; 31:978-986.

33. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007; 448:553-560.

34. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A: Determinants of nucleosome organization in primary human cells. *Nature* 2011; 474:516-520.

35. Chantalat S, Depaux A, Hery P, Barral S, Thuret JY, Dimitrov S, Gerard M: Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Res* 2011; 21:1426-1437.

36. Mattout A, Aaronson Y, Sailaja BS, Raghu Ram EV, Harikumar A, Mallm JP, Sim KH, Nissim-Rafinia M, Supper E, Singh PB, et al: Heterochromatin Protein 1beta (HP1beta) has distinct functions and distinct nuclear distribution in pluripotent versus differentiated cells. *Genome Biol* 2015; 16:213.

37. Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm JP, Nissim-Rafinia M, Cohen AH, Rippe K, Meshorer E, Ast G: HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. *Cell Rep* 2015; 10:1122-1134.

38. Elsässer SJ, Noh KM, Diaz N, Allis CD, Banaszynski LA: Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells. *Nature* 2015; 522:240-244.

39. Erdel F, Muller-Ott K, Rippe K: Establishing epigenetic domains via chromatin-bound histone modifiers. *Ann N Y Acad Sci* 2013; 1305:29-43.

40. Audergon PNCB, Catania S, Kagansky A, Tong P, Shukla M, Pidoux AL, Allshire RC: Restricted epigenetic inheritance of H3K9 methylation. *Science* 2015; 348:132-135.

41. Ragunathan K, Jih G, Moazed D: Epigenetic inheritance uncoupled from sequence-specific recruitment. *Science* 2015; 348:1258699-1258699.

42. Hansen KH, Bracken AP, Pasini D, Dietrich N, Gehani SS, Monrad A, Rappsilber J, Lerdrup M, Helin K: A model for transmission of the H3K27me3 epigenetic mark. *Nat Cell Biol* 2008; 10:1291-1300.

43. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012; 485:376-380.

44. Steinhauser S, Kurzawa N, Eils R, Herrmann C: A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform* 2016.

45. ENCODE, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; 447:799-816.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

46. ENCODE, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, et al: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489:57-74.

47. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al: An atlas of active enhancers across human cell types and tissues. *Nature* 2014; 507:455-461.

48. Bardet AF, He Q, Zeitlinger J, Stark A: A computational pipeline for comparative ChIP-seq analyses. *Nat Protoc* 2012; 7:45-61.

49. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL, et al: Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* 2013; 152:642-654.

50. Xiao S, Xie D, Cao X, Yu P, Xing X, Chen C-C, Musselman M, Xie M, West FD, Lewin HA, et al: Comparative epigenomic annotation of regulatory DNA. *Cell* 2012; 149:1381-1392.

51. Lasserre J, Chung H-R, Vingron M: Finding associations among histone modifications using sparse partial correlation networks. *PLoS Computational Biology* 2013; 9:e1003168.

52. Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA: Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Molecular Cell* 2005; 17:453-462.

53. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T: Regulation of alternative splicing by histone modifications. *Science* 2010; 327:996-1000.

54. Efroni S, Duttagupta R, Cheng J, Dehghani H, Hoeppner DJ, Dash C, Bazett-Jones DP, Le Grice S, McKay RD, Buetow KH, et al: Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell* 2008; 2:437-447.

55. Cremer T, Cremer M: Chromosome territories. *Cold Spring Harb Perspect Biol* 2010; 2:a003889.

56. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L: Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* 2012; 30:90-98.

57. Morey C, Kress C, Bickmore WA: Lack of bystander activation shows that localization exterior to chromosome territories is not sufficient to up-regulate gene expression. *Genome Res* 2009; 19:1184-1194.

58. Junker JP, van Oudenaarden A: Every cell is special: genome-wide studies add a new dimension to single-cell biology. *Cell* 2014; 157:8-11.

59. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; 10:R25.

60. Schätzel K: Noise on photon correlation data: I. Autocorrelation functions. *Quantum Opt* 1990; 2:287-305.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

61. Fisher RA: Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* 1915; 10**:**507–521.

62. Fisher RA: On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* 1921; 1**:**3–32.

63. Efron B, Tibshirani RJ: *An Introduction to the Bootstrap.* Chapman & Hall; 1993.

64. Smirnov N: Table for estimating the goodness of fit of empirical distributions. *Ann Math Stat* 1948; 19**:**279–281.

65. Skilling J, Bryan RK: Maximum entropy image reconstruction: general algorithm. *Mon Not R Astr Soc* 1984; 211**:**111-124.

66. R Core Team: R: A language and environment for statistical computing. 2013**:**http://www.R-project.org/.

67. Fruchterman TMJ, Reingold EM: Graph drawing by force-directed placement. *Software: Practice and Experience* 1991; 21**:**1129-1164.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

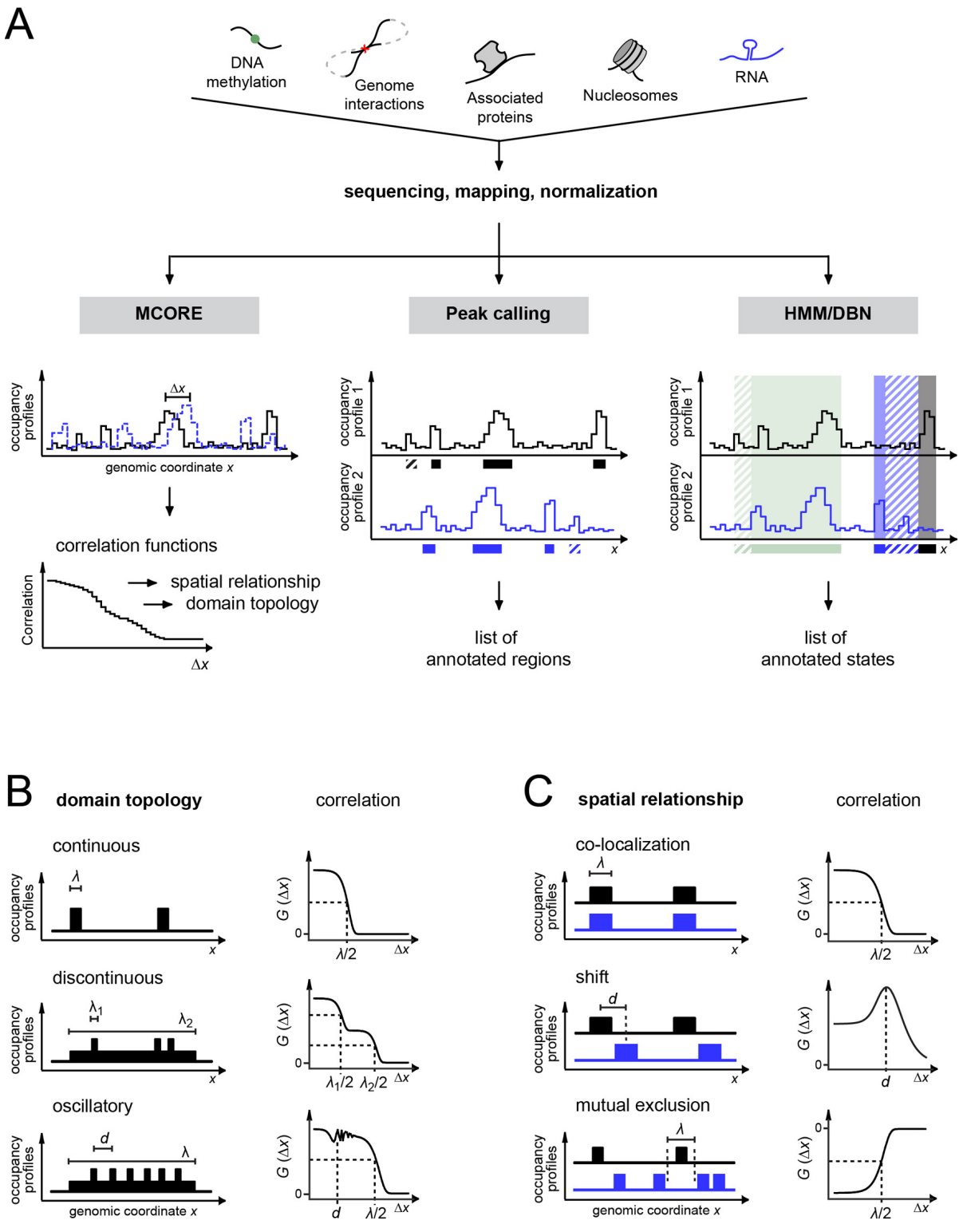


Figure 1. MCORE can identify and compare patterns in deep sequencing data sets.

(A) MCORE is suited for the analysis of deep sequencing data from various methods. Initially, mapped reads are used to compute occupancy profiles of two samples (black/blue). Subsequently, the profiles are normalized using the input sample and, if applicable, the control sample. In contrast to other methods like peak calling, hidden Markov models (HMM) or dynamic Bayesian networks (DBN), MCORE does not score enriched regions but rather shifts normalized occupancy profiles with respect to each other to compute correlation functions, which contain information about chromatin patterns as illustrated in panels B and

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

C and Fig. S1. To this end it uses all sequencing reads without filtering and avoids any assumptions about the enrichment pattern. **(B)** Correlation functions between biological replicates for the same chromatin feature contain information about its domain topology. Whereas the correlation coefficient at shift distance zero quantifies the reproducibility of the experiment, the shape of the function reflects the distribution of the feature along the genomic coordinate. Continuous domains lead to a steep decay at the shift distance that coincides with half the domain size (top), whereas broad domains containing small highly enriched regions yield multiple decay lengths (center). Arrays of equally spaced domains cause an oscillating contribution in the correlation function (bottom). Mixtures of domains with different topology yield a superposition of the respective correlation functions. **(C)** Correlation functions between two different chromatin features reflect their spatial relationship. Co-localizing features yield monotonously decaying functions (top) that resemble those between replicates discussed in the previous panel. Correlation functions for features that are shifted with respect to each other exhibit a local maximum at the shift distance (center). Mutually exclusive features are recognized by negative correlation amplitudes (bottom). Features that do not exhibit any spatial relationship with respect to each other yield no correlation for any shift distance.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

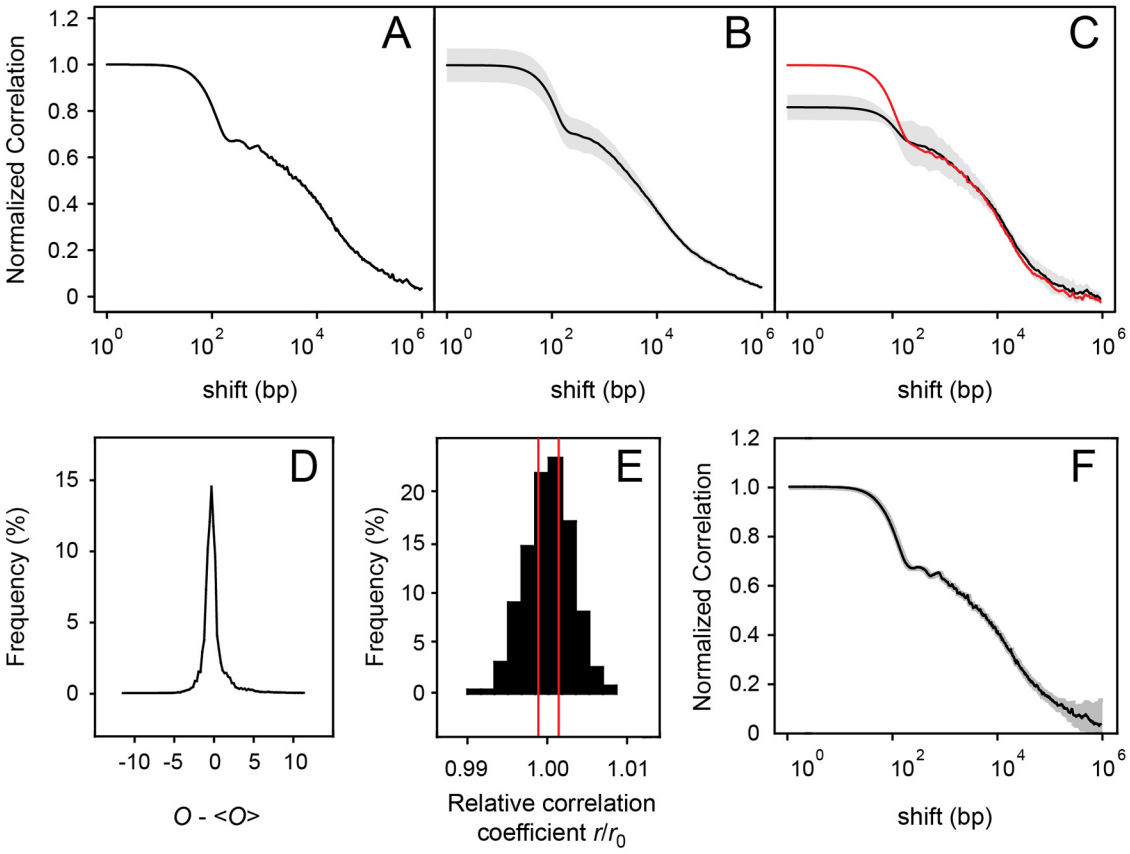


Figure 2. Confidence intervals of correlation functions. Calculations shown here were conducted for H3K36me3 in ESCs for chromosome 1. Indicated confidence intervals refer to the 95% level.

(A) Replicate correlation function (black) and its confidence interval (grey) obtained by MCORE based on the Fisher transformation (Materials and Methods). Due to the large sample size the confidence interval is smaller than 10^{-3} and within the line thickness. **(B)** Average (black) and confidence interval (grey) of correlation functions calculated for all autosomes (1-17) based on the data sets generated in this study. **(C)** Average (black) and confidence interval (grey) of three replicate correlation functions calculated from three independent biological replicates (rep1 x rep2, rep1 x rep3, rep2 x rep3), yielding information on the reproducibility of the experiment. The correlation function for ENCODE data for H3K36me3 in ESCs (red) is similar to the correlation function computed from the data sets generated in this study. The amplitude of the first domain that covers the length scale below 200 bp shift distance is different. This might be due to incomplete correction of background signal in the ENCODE data set that lacks a control IP reference, which should, however, not strongly affect the quantitation of domain sizes beyond the scale of a nucleosome. The effect of this correction is illustrated in Fig. S2. **(D)** Distribution of normalized occupancy values ($O_i - \langle O \rangle$) that were used for calculating the correlation function in panel A according to Eq. 4 (Materials and Methods). The distribution is relatively symmetric and unimodal. **(E)** Distribution of correlation coefficients obtained by bootstrapping for the correlation coefficient at zero shift distance. Each correlation coefficient was calculated after resampling the occupancy profiles with replacement as described in the Materials and Methods section. Correlation coefficients are given relative to the mean value. The 95% confidence interval obtained by this approach is roughly 3-times larger than the estimate based on Fisher transformation (shown in red). **(F)** Correlation function from panel A with non-parametric bootstrap confidence intervals for each shift distance.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

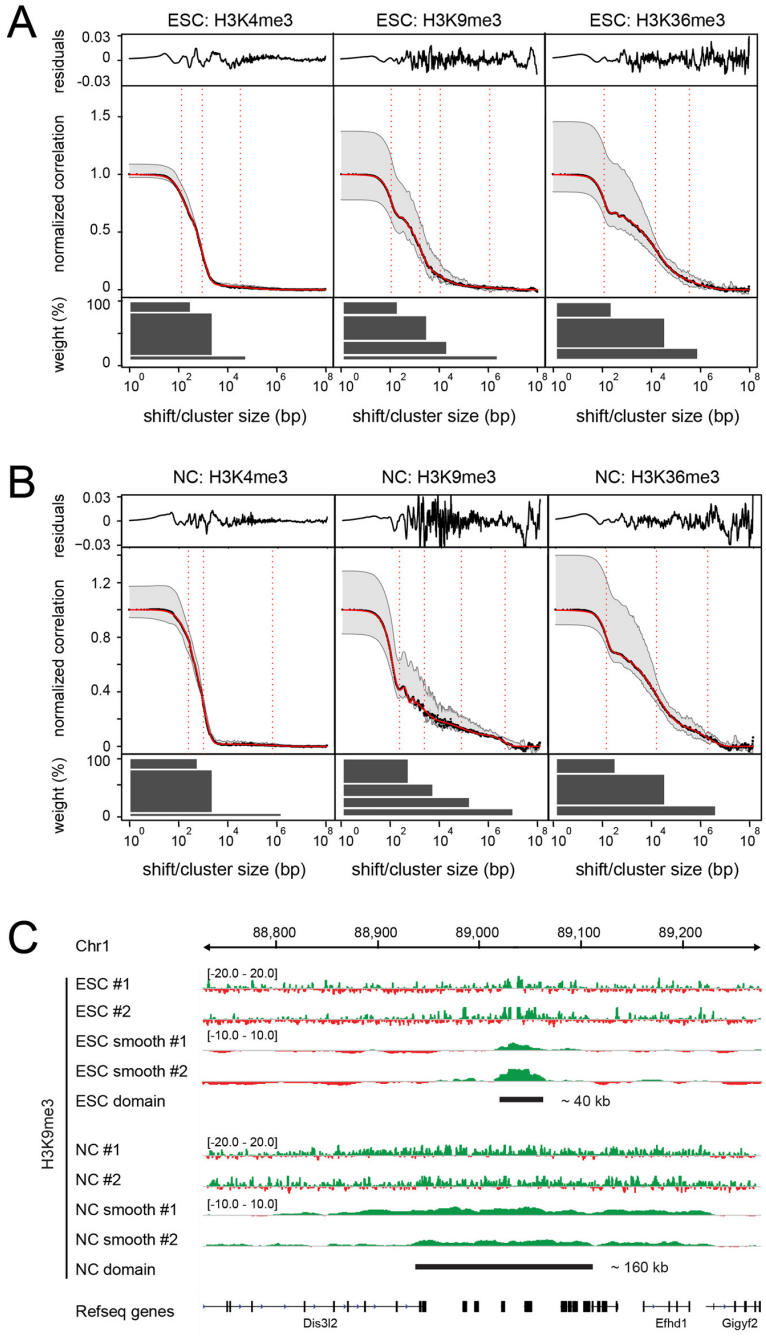


Figure 3. Quantification of domain sizes for different histone marks.

(A) Correlation functions for replicates in ESCs. Correlation functions calculated between replicates for chromosome 1 (black) and their fit functions (red) with characteristic domain sizes obtained from the fit (vertical dotted lines) are shown. Grey regions indicate maximum variation between chromosomes. Fit residuals are plotted above the correlation curves. Domain sizes and abundances calculated from the respective fit parameters are shown below the correlation curves. **(B)** Same as in panel A for NCs. **(C)** As shown in panels A and B, MCORE identified broad H3K9me3 domains spanning on average 128 kb and 7.6 Mb in NCs that were absent in ESCs. To annotate the genomic positions of these domains, read counts in a sliding window of 128 kb, which corresponded to the domain size calculated by MCORE, were evaluated. An example of a domain that became broader in NCs is shown ('#1' and '#2' denote biological replicates). For clarity, the occupancy profiles were smoothed with 0.2-times the window size ('smooth'). For window size 7.6 Mb see Fig. S14.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

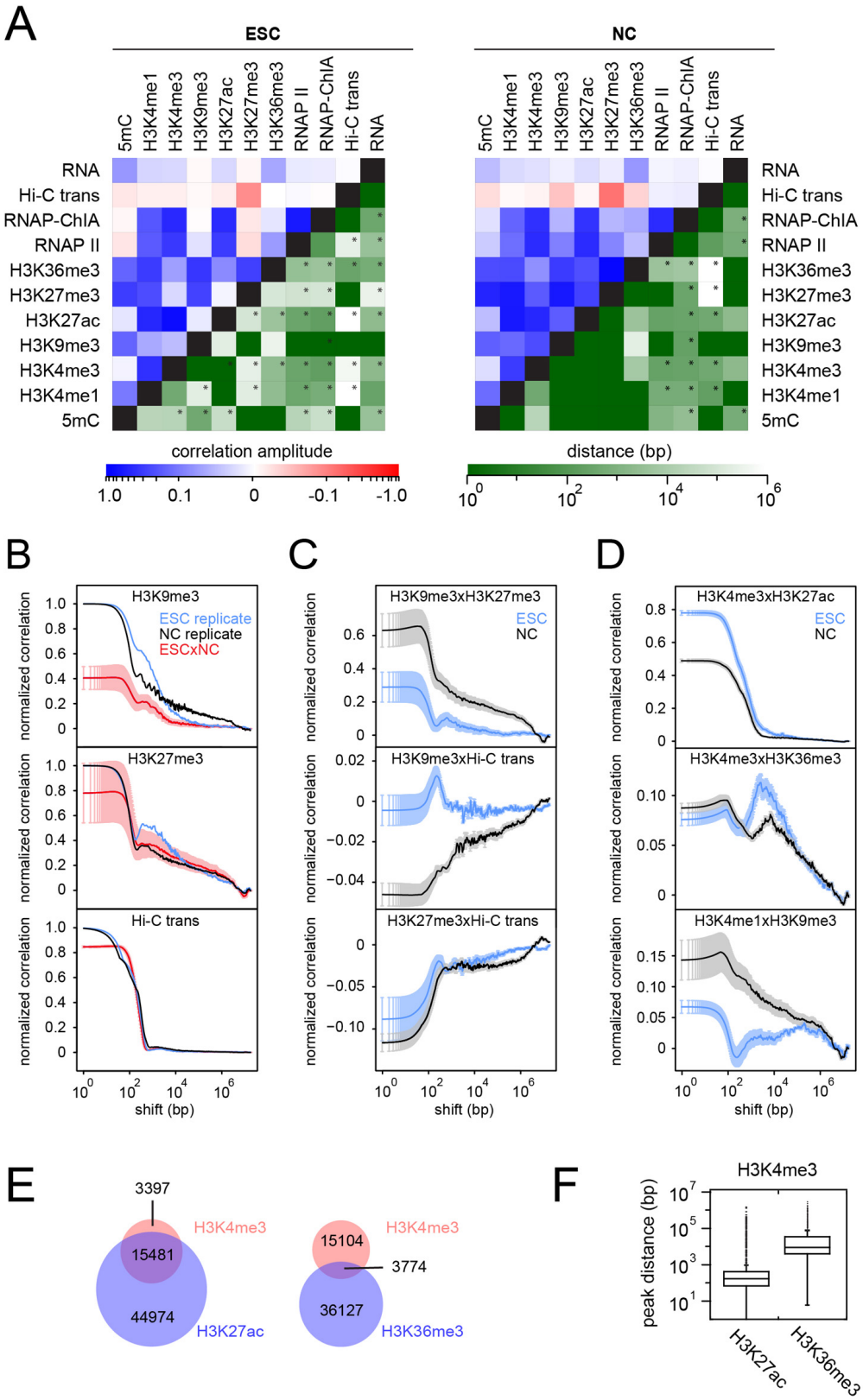


Figure 4. MCORE reveals genome-wide relationships between chromatin features.

(A) Co-localization (top, red/blue) and separation distance (distance to the largest local maximum, bottom, green) between pairs of different features in ESCs (left) and NCs (right) are illustrated. Stars indicate correlation functions for which the local maximum is also the global maximum. Hi-C trans, Hi-C inter-chromosomal contacts; RNA, RNA-seq; RNAP II-

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ChIA, RNAP II ChIA-PET. **(B)** Correlation functions for replicates of H3K9me3, H3K27me3 and inter-chromosomal contacts (Hi-C trans) in ESCs (blue) and NCs (black) show the spatial extension of these features. Average cross-correlation functions (red) between ESCs and NCs quantify the co-localization of a given feature across cell types. Averages were calculated from the four possible combinations of the two replicates for each sample (Materials and Methods). Error bars, s.e.m. **(C)** Cross-correlations between H3K9me3 and H3K27me3 (top) or H3K9me3/H3K27me3 and inter-chromosomal contact sites (Hi-C trans, center/bottom) in ESCs and NCs. Repressive domains co-localize in NCs (top) and have a tendency to be depleted for inter-chromosomal contacts (bottom). Error bars, s.e.m. **(D)** Cross-correlations between H3K4me3 and H3K27ac (top) indicate co-localization of both marks in rather small domains, whereas cross-correlations of H3K4me3 and H3K36me3 (center) reveal a relative displacement of roughly 5 kb between the two marks. In NCs, there is an additional co-localization at zero shift distance that is weaker in ESCs. Cross-correlations between H3K4me1 and H3K9me3 (bottom) show that both marks are stronger co-localized in NCs than in ESCs. The local maximum at 100 kb shift distance in ESCs suggests a separation of H3K4me1 from broad H3K9me3 domains. Error bars, s.e.m. **(E)** Peak calling in NCs as readout for co-localization. Red, peaks called by MACS for H3K4me3; blue, peaks called by SICER for H3K36me3 or by MACS for H3K27ac. The numbers of (overlapping) peaks are indicated. **(F)** Distribution of distances between called peaks. Distances were calculated from the center of the H3K4me3 peak to the center of the nearest peak in the second data set (H3K27ac or H3K36me3).

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

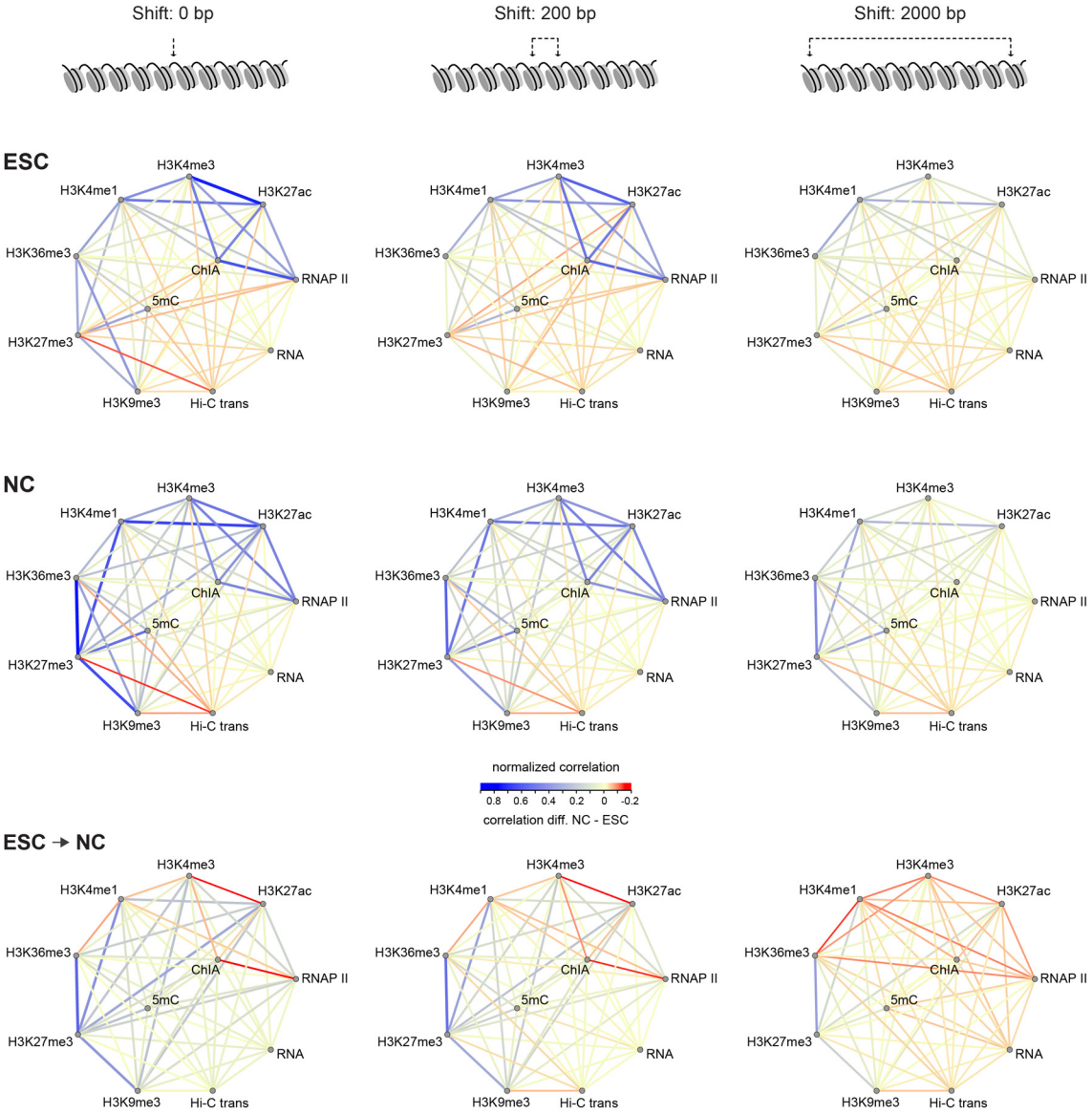


Figure 5. Network models for scale-dependent relationships among chromatin features.

(A) Network models illustrating the relationships among different chromatin features in ESCs on different scales (blue: positive correlation, red: negative correlation). Features were grouped according to their correlation at zero shift distance (left), yielding a cluster of features associated with active transcription and a cluster of marks related to gene silencing, whereas H3K36me3 co-localizes with members of both groups. The correlations among features on adjacent nucleosomes (200 bp shift distance) differ from the correlations among features at the same nucleosome (0 bp shift distance), indicating that only some features form continuous domains that extend beyond a single nucleosome. For the even larger shift distance of roughly ten nucleosomes (2000 bp), only some long-range correlations remain, which either reflect large domains of co-localizing features or features that are shifted with respect to each other. The latter two possibilities can be distinguished based on the shape of the correlation function (Fig. 1C). **(B)** Same as in panel A but for NCs. **(C)** Network models illustrating changing relationships among different chromatin features in ESCs and NCs. The difference NC-ESC is colored in blue if correlations became stronger in NCs and red if correlations became weaker in NCs. Positive correlations among repressive marks were stronger in NCs than in ESCs, particularly on larger scales. Further, inter-chromosomal contacts (Hi-C trans) were positively correlated with H3K9me3 in NCs but not in ESCs.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

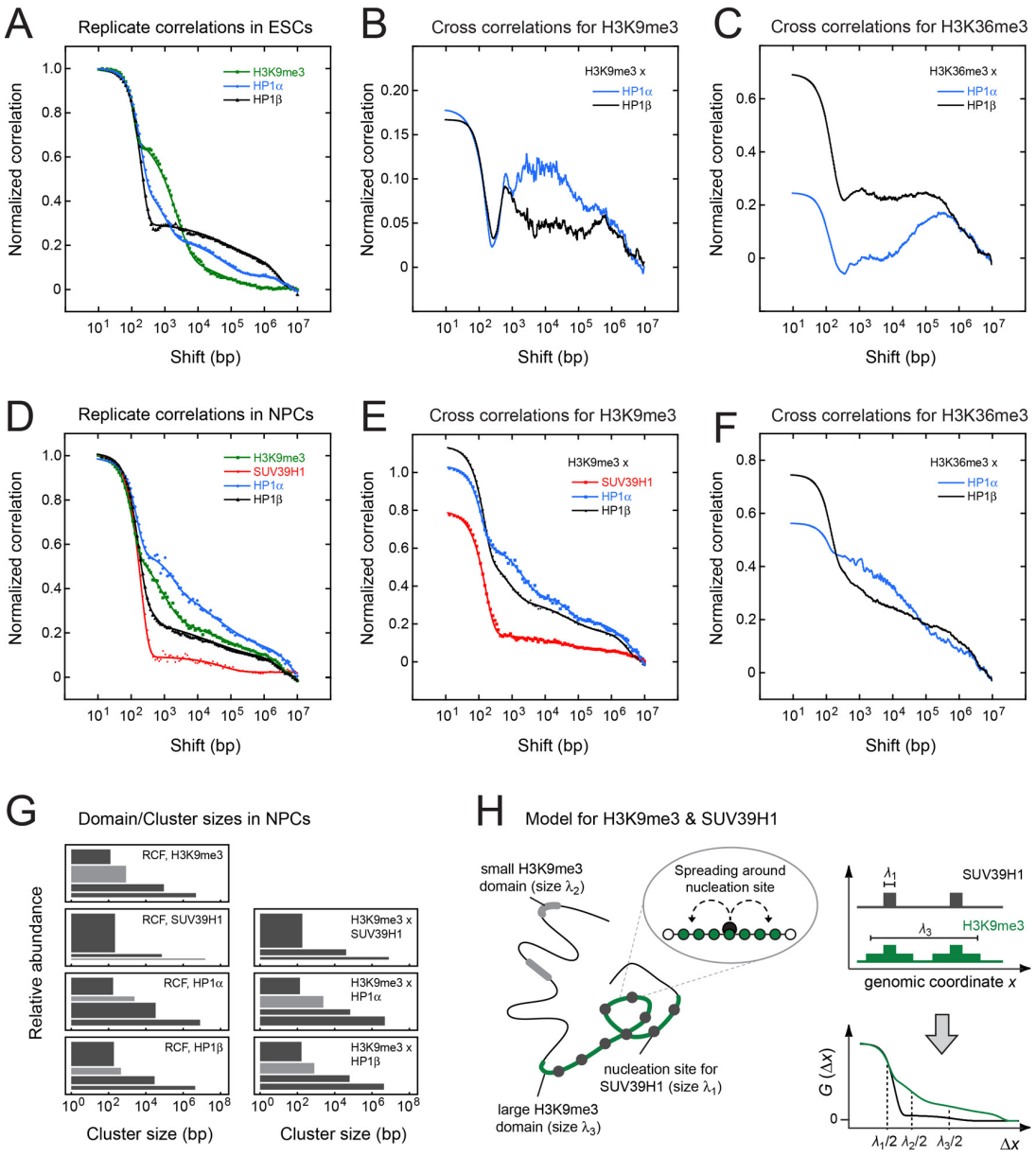


Figure 6. Interplay among H3K9me3, SUV39H1 and HP1.

(**A**) Replicate correlation functions of HP1α (blue), HP1β (black) and H3K9me3 (green) in ESCs. (**B**) Cross correlation functions of HP1α (blue) or HP1β (black) with H3K9me3 in ESCs. (**C**) Cross correlation functions of HP1α (blue) or HP1β (black) with H3K36me3 in ESCs. (**D**) Same as in panel A but for NCs and including SUV39H1 (red). H3K9me3 and HP1α/β exhibit small, intermediate and (very) broad domains. The short domain size of one nucleosome is present in the correlation functions for all marks, suggesting that domains consist of nucleation sites and gaps (as explained in the text). SUV39H1 does not form intermediately sized domains. (**E**) Same as in panel B but for NCs and including SUV39H1 (red). SUV39H1, HP1α, HP1β and H3K9me3 strongly co-localized.. Intermediate domains are not present in the cross correlation function between SUV39H1and H3K9me3, indicating that both features only co-localize in short and broad domains. In contrast, HP1α and HP1β essentially follow the H3K9me3 distribution, indicating that they do not distinguish between differently sized H3K9me3 domains. (**F**) Same as in panel C but for NCs. (**G**) Domain size distribution for correlation functions in panels D and E. (**H**) Schematic illustration of a nucleation-and-looping mechanism for the formation of SUV39H1-dependent H3K9me3 domains, which is consistent with the MCORE result.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

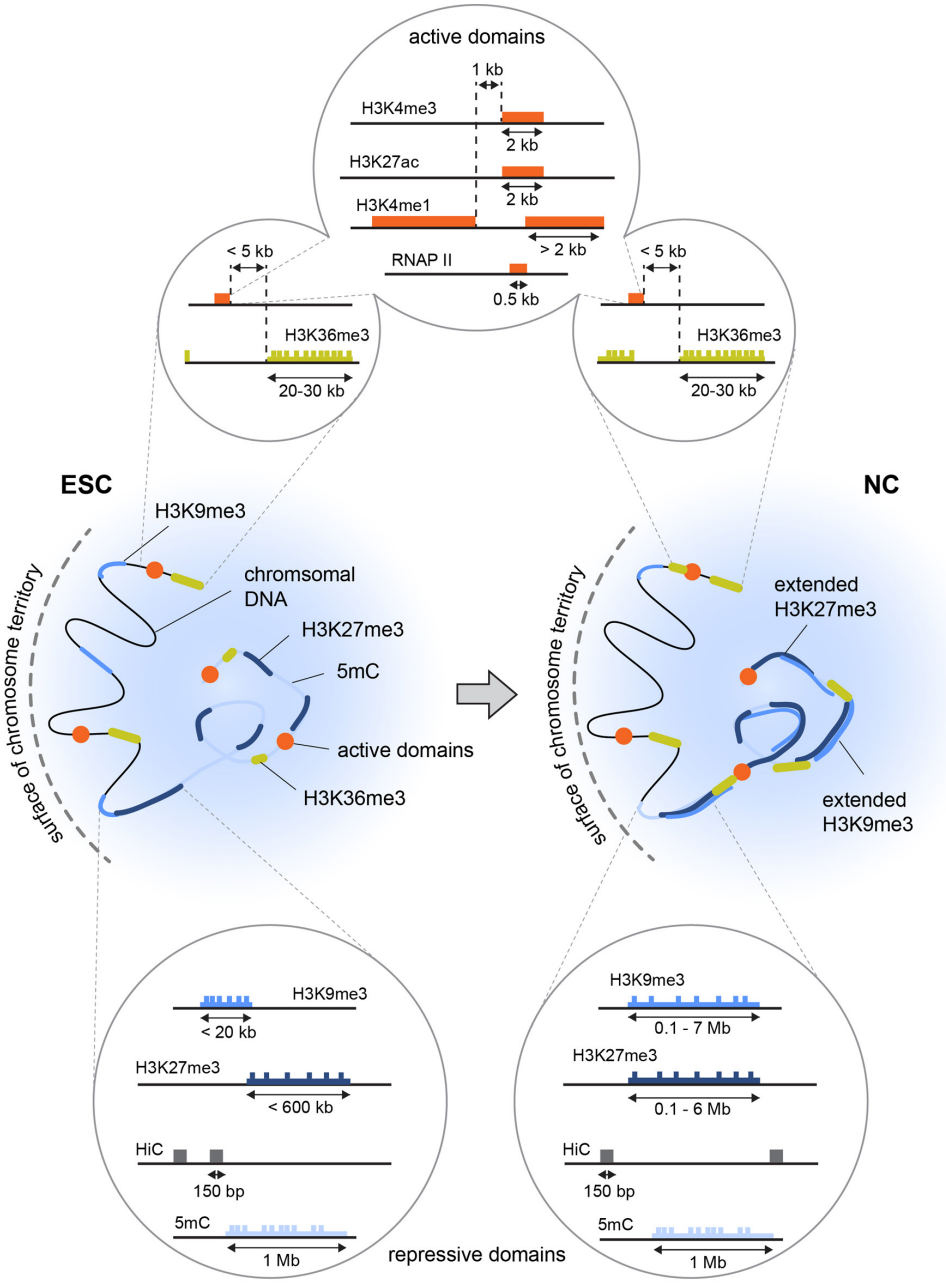


Figure 7. Alterations of chromatin features during differentiation of ESCs into NCs.

Model for the re-organization of chromatin domains during differentiation from ESCs to NCs based on the MCORE analysis of the data sets used in this study. Active domains mostly retained their organization, with H3K4me1 being partly separated from the smaller H3K4me3/H3K27ac domains in both cell types. The overlap between those marks and H3K36me3 increased in NCs, which might be due to elevated transcription of enhancers or activation of genes enriched for H3K4me1/3 or H3K27ac. Domains enriched for H3K9me3 and H3K27me3 became extended at sites of 5mC and were preferentially buried inside chromosome territories. At the same time, RNAP II re-located to the surface of the chromosome territory. The newly established H3K9me3/H3K27me3 domains were discontinuous, i.e. contained many modified nucleosomes without an equally modified neighbor. Further, they exhibited increased overlap with activating marks such as H3K4me1 and H3K4me3, which suggests that they do not exclusively contain heterochromatin but rather enclose both active and repressive chromatin domains.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Information

Retrieving the topology of chromatin domains from deep sequencing data with correlation functions

Jana Molitor, Jan-Philipp Mallm, Karsten Rippe & Fabian Erdel

Supplementary Figures

1. Strategies to retrieve information about complex patterns
2. Correction of background and multiplicative biases
3. Statistical properties of representative occupancy profiles
4. Comparison between Pearson and Spearman correlation functions
5. Peak calling for representative data sets
6. Robustness of correlation functions towards undersampling
7. Dependence of fit results on coverage
8. MCORE for simulated data sets
9. Statistical comparison of correlation functions
10. MCORE for different H3K27ac data sets
11. Quality control of ChIP-seq data
12. Fitted correlation functions for H3K27me3
13. Peak calling summary for H3K9me3
14. MCORE-directed annotation of chromatin features
15. MCORE for transcription factor binding
16. Spatial extension and co-localization of different features in ESCs versus NCs
17. Heterochromatin reorganization during differentiation
18. DNA methylation and inter-chromosomal contacts
19. Fraction of chromatin features within and near genes

Supplementary Tables

1. Comparison of MCORE with other software tools
2. Fit parameters for selected correlation functions in ESCs
3. Fit parameters for selected correlation functions in NCs
4. Summary of data sets used in this study

Supplementary References

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

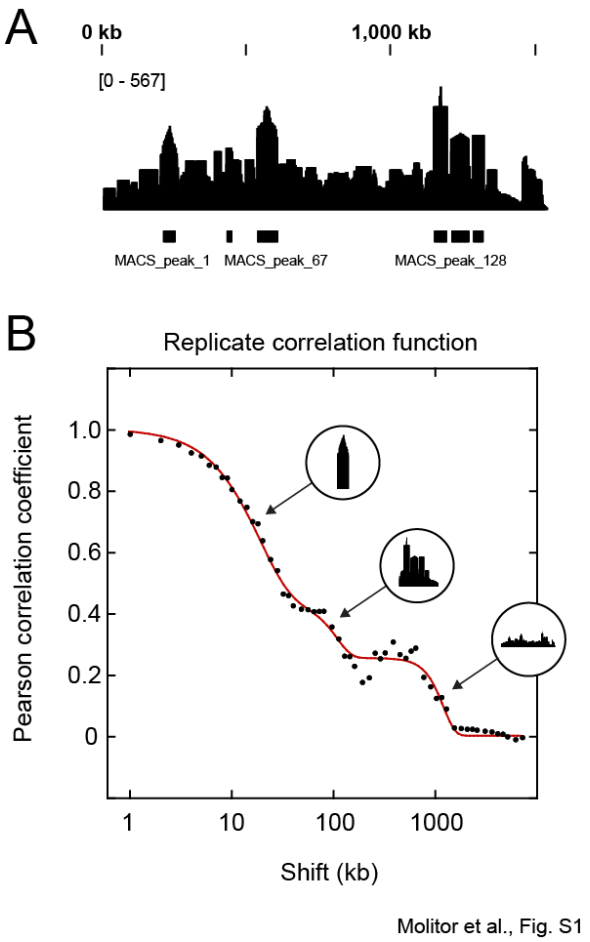


Figure S1 | **Strategies to retrieve information about complex patterns. (A)** Peak calling result for a complex domain topology involving different enrichment levels (MACS, standard settings mfold =10,30, pvalue = 1e-5). The pattern is reduced to regions that are compatible with the threshold and significance settings. Other regions are classified as background. There is no straightforward criterion to decide which threshold level should be used to separate biologically relevant enrichment from irrelevant background. **(B)** Correlation function (black dots) and multi-exponential fit according to Eq. 6 (red line) for the pattern in panel A. The correlation function yields the different length scales that are present in the pattern, including the width of highly enriched regions, the characteristic size of clusters formed by adjacent peaks, and the size of the entire enriched region. No criterion is required to decide which structures are biologically relevant and which are ignored.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

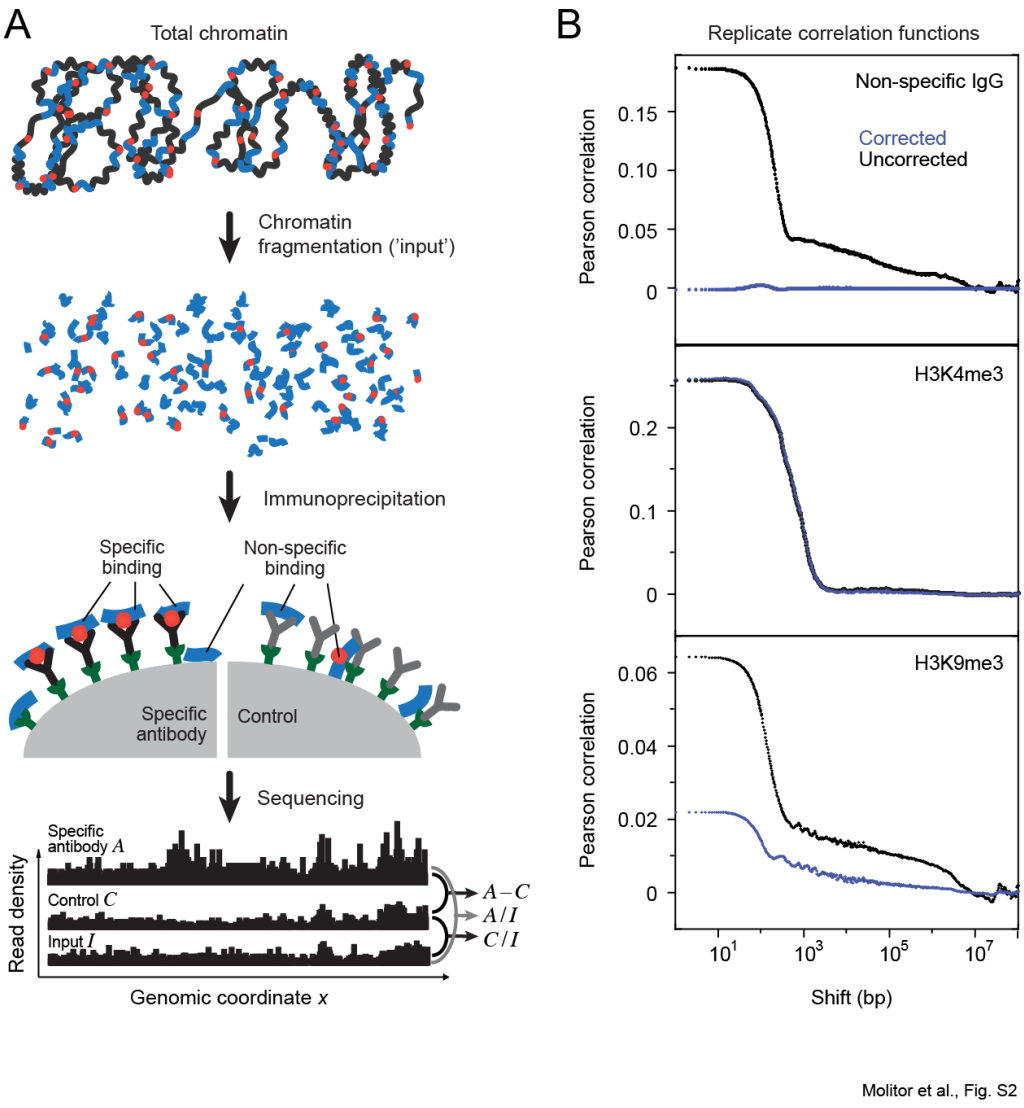


Figure S2 | **Correction of background and multiplicative biases. (A)** Fragmentation of total chromatin (black) containing a chromatin feature of interest (red) occurs with some bias and is frequently incomplete. As a result, only a fraction of chromatin (blue) is present in the input sample due to size selection during library preparation. Subsequent immunoprecipitation occurs in the presence of non-specific binding. The latter contribution can be assessed in a separate control reaction, e.g. by using an antibody that does not bind specifically to the antigen. Sequencing reads obtained from samples with the specific antibody *A*, the control *C* and the input *I* are used to calculate normalized occupancy profiles for the analysis of a given chromatin feature according to Eqs. 1-3. In brief, the coverage from the specific IP and from the control are divided by the input coverage (*A* / *I* and *C* / *I*, see Eq. 1) to account for multiplicative biases such as mappability or preferences in immunoprecipitation, ligation, amplification and sequencing. Next, the weighted control signal is subtracted from the specific antibody signal to remove additive bias caused by non-specific binding (Eqs. 2-3). Resulting profiles are used for calculating correlation functions (Eq. 4). **(B)** Correlation functions for the uncorrected (black) and corrected (blue) occupancies for control IP (IgG, top), H3K4me3 (center) and H3K9me3 (bottom) ChIP-seq replicates in neural progenitor cells. Subtraction of the weighted control IP signal removes the background correlation and thus eliminates correlation between control IP signals (top). Normalization has little effect for H3K4me3, which displays distinct peaks with considerable enrichment (Fig. S5). In contrast, it induces a significant correction for H3K9me3, which forms broad domains with moderate enrichment levels.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

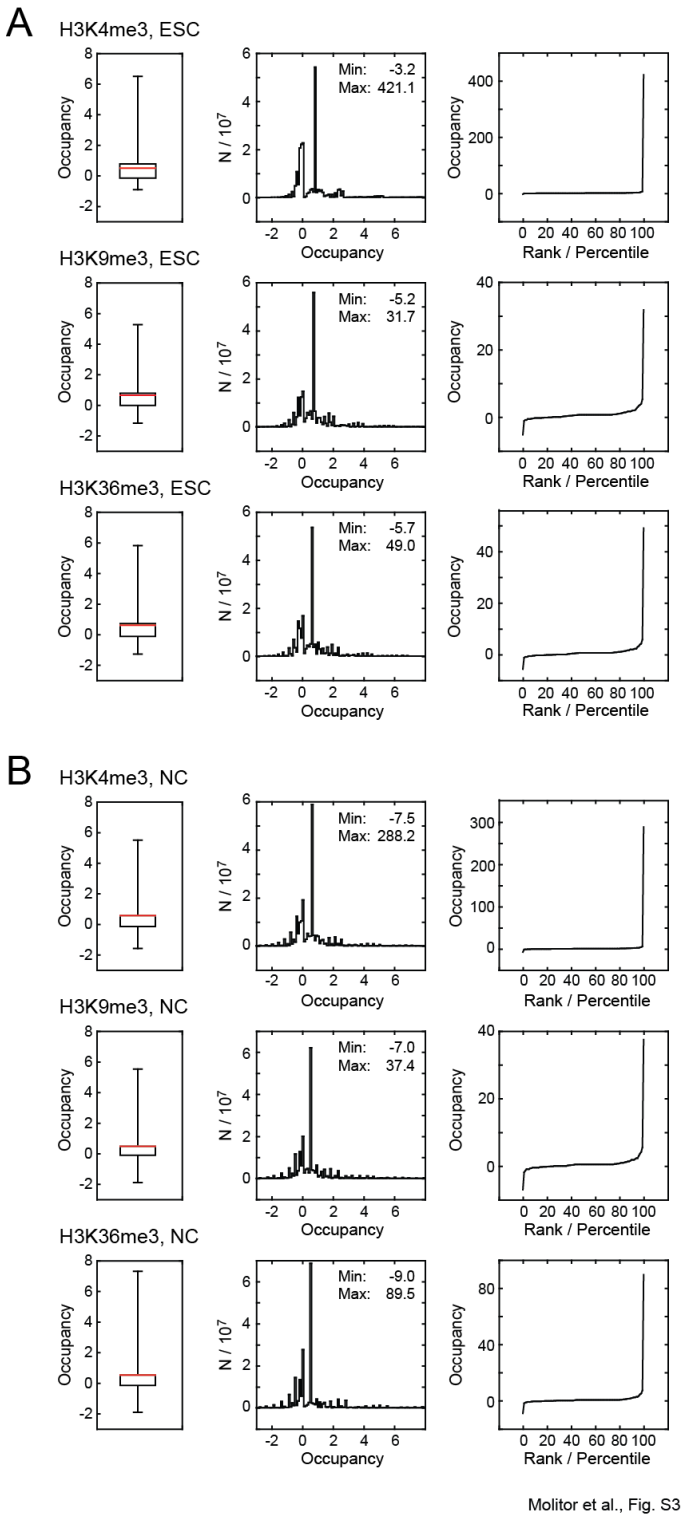


Figure S3 | Statistical properties of representative occupancy profiles. Box plots (left), histograms (center) and percentiles (right) for normalized occupancy profiles from H3K4me3, H3K9me3 and H3K36me3 ChIP-seq experiments. For box plots, the median is colored in red and the ends of the whiskers represent the 1st and 99th percentile. Minimum and maximum occupancy values are listed in the histograms. Maximum occupancies depend strongly on the feature of interest but only moderately on the cell type. The background comprises a large part of the data and its distribution is similar for all profiles (see box plots and histograms). **(A)** ESCs. **(B)** NCs.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

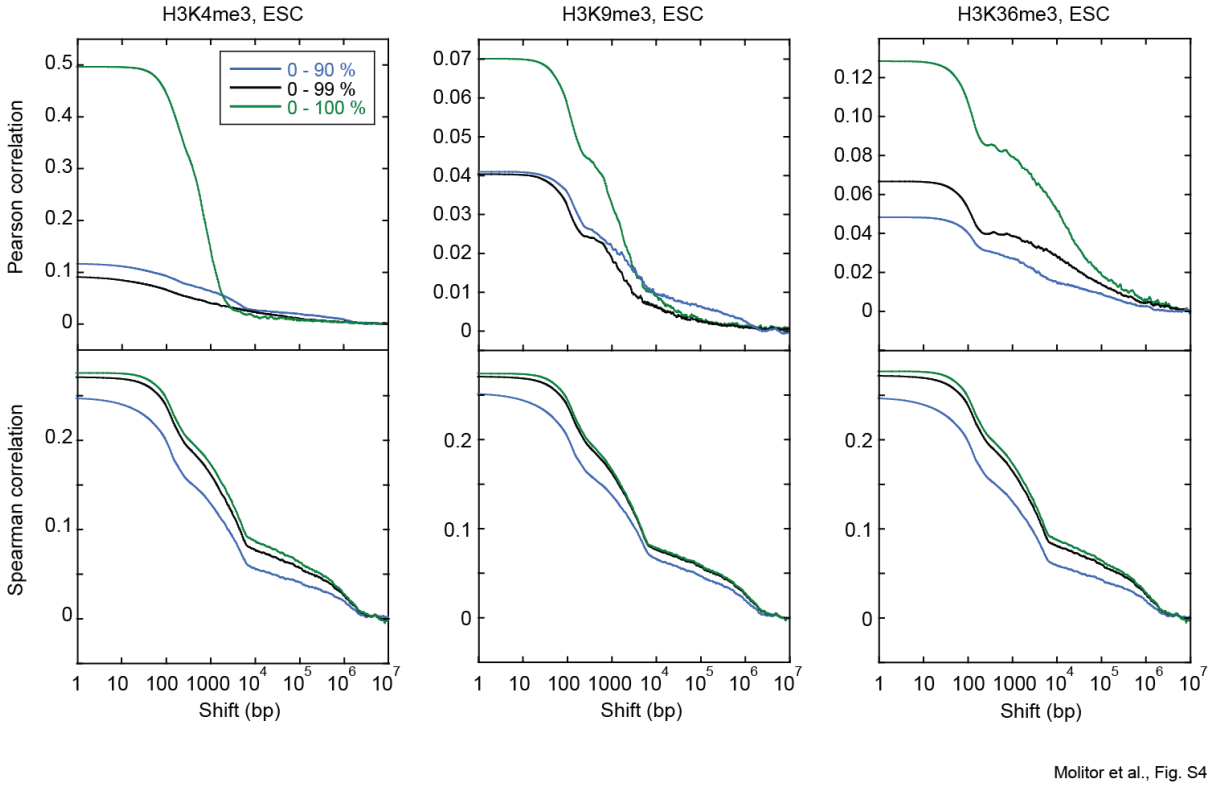


Figure S4 | Comparison between Pearson and Spearman correlation functions. Pearson (top, green) and Spearman (bottom, green) correlation functions for representative occupancy profiles in ESCs. To assess the contribution of enriched regions to the different correlation functions we replaced occupancy values above the 90th (blue) or 99th (black) percentile with the average occupancy within the rest of the genome. Spearman correlation functions exhibited only slight changes upon removal of highly enriched regions and primarily reflected the structure of the background signal that was independent of the immunoprecipitated histone mark (compare bottom left, center and right). In contrast, Pearson correlation functions reflected the properties of enriched regions, which carry the biological information, and changed their shape when these regions were omitted from the analysis. After removal of enriched regions (top, blue), Pearson correlation functions were dominated by the background signal and resembled Spearman correlation functions (bottom). The stronger background signal in Spearman correlation functions is due to the correction procedure that minimizes the background according to the Pearson metric (Eq. 3).

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

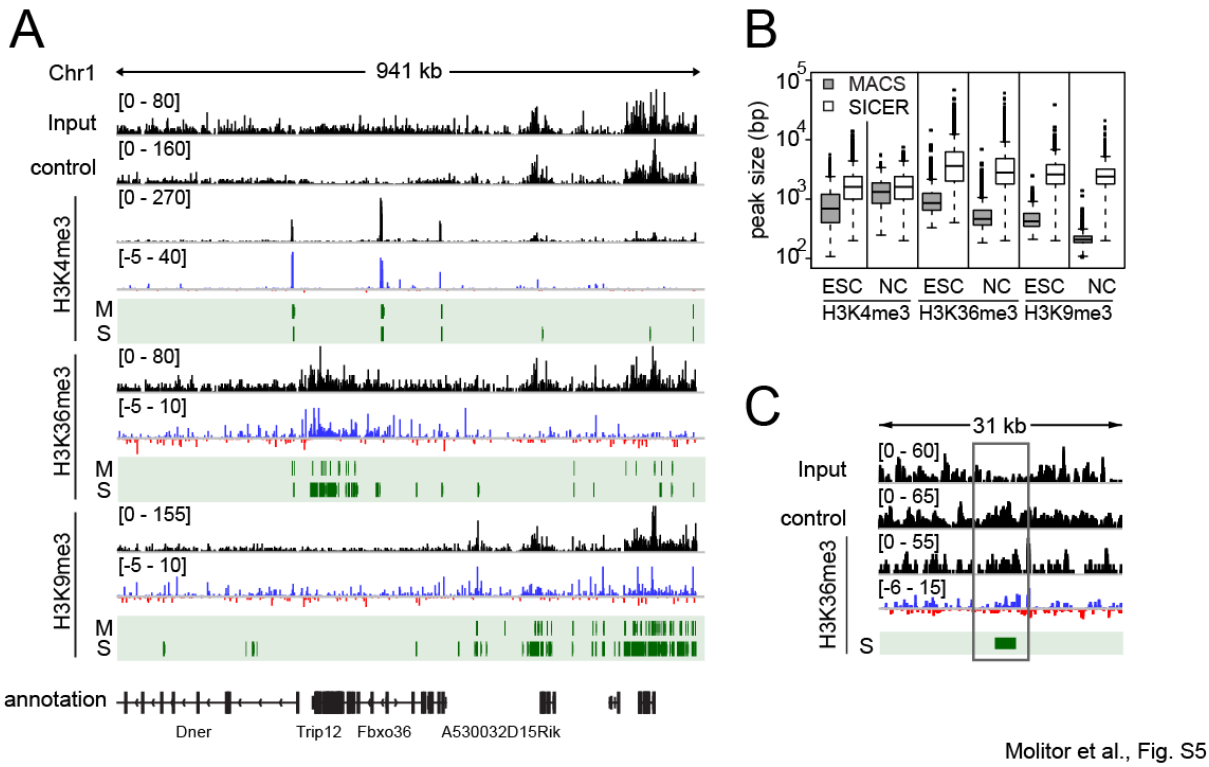


Figure S5 | Peak calling for representative data sets. (A) Read distribution (black) for sample, control (IP with a non-specific antibody) and input, normalized occupancy (red/blue), and peaks (green) called by MACS (M) and SICER (S) for H3K4me3, H3K9me3 and H3K36me3 ChIP-seq in NCs. Distinct H3K4me3 domains were reliably identified by both peak callers, results for H3K9me3 and H3K36me3 depended on the specific algorithm used (e.g. MACS and SICER). **(B)** Peak size distributions for clusters called by MACS and SICER for the ChIP-seq experiments in ESCs and NCs. Cluster sizes differ between both methods. **(C)** Example of the read distribution (black) and normalized occupancy (red/blue) for H3K36me3 ChIP-seq in NCs, including input and control. The highlighted region contains an apparent enrichment in H3K36me3 that is identified as a peak. However, similar enrichment is present in the control IP, suggesting that the signal corresponds to non-specific background. For such regions, peak calling methods that lack the possibility to simultaneously correct for biases with both an input sample and a non-specific IP give rise to false-positive peaks.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

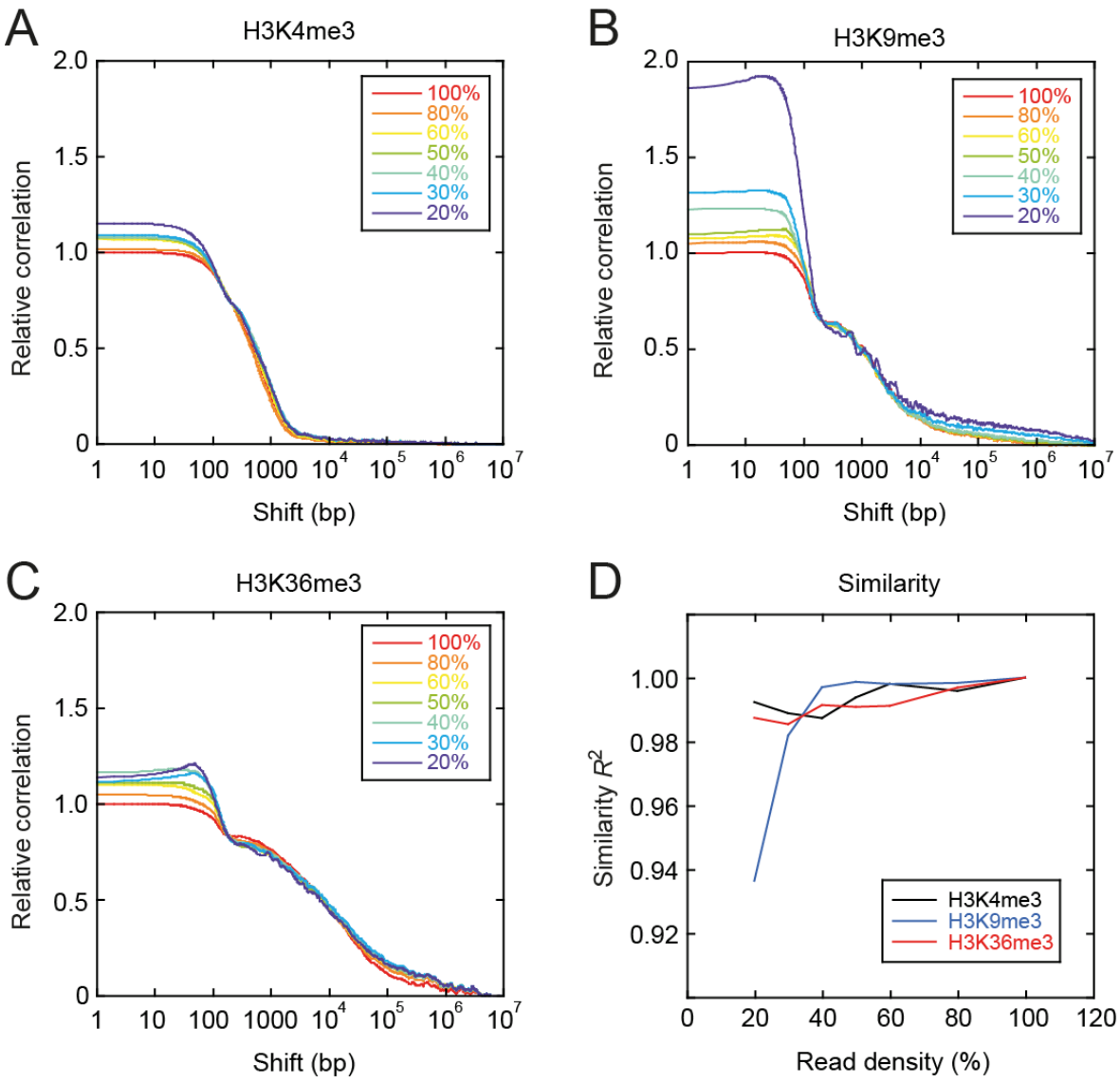


Figure S6 | Robustness of correlation functions towards undersampling. (A) Replicate correlation functions for ChIP-seq data sets of H3K4me3 in ESCs containing different numbers of reads are shown. The red curve corresponds to the entire set of reads reported in this study (100%, corresponding to 30 million reads). The other functions reflect data sets that were diluted *in silico* by randomly selecting only a fraction of reads from the entire set. Correlation functions were normalized to the 100% curve at a shift distance of one nucleosome (according to the fit parameters c2 in Table S2) because correlation coefficients for smaller shift distances do not contain information about domain structures (see Fig. S7 for domain sizes obtained by fitting). **(B)** Same as in panel A but for H3K9me3. **(C)** Same as in panel A but for H3K36me3. **(D)** Quantification of the similarity of correlation functions for diluted data sets with respect to the undiluted curve based on the coefficient of determination (R^2). Correlation functions for diluted data sets are similar to each other and to the result for the undiluted data set, with $R^2 > 0.9$. Above 40% read density, which corresponds to 12 million reads, a plateau is reached for all modifications assessed here.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

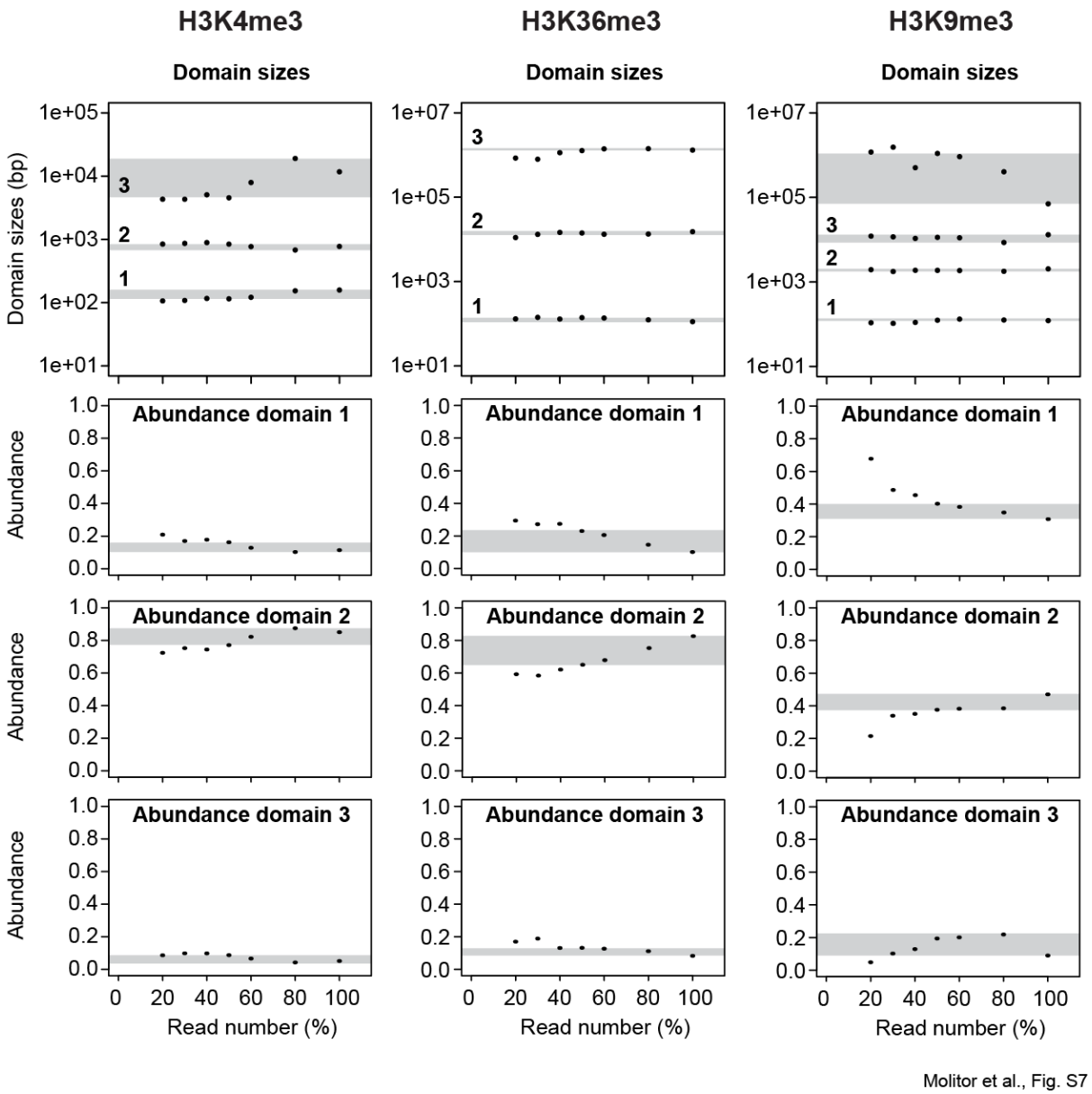


Figure S7 | Dependence of fit results on coverage. The correlation curves plotted in Fig. S6 were fitted with Eq. 6. Fit results for the domain sizes and the respective amplitudes are plotted versus coverage (domain numbers are indicated in the top panel). Grey regions show the variation of the fit results for dilution down to 50% of the reads. The most abundant domains, which represent the characteristic domain sizes for a given modification, were accurately quantified from diluted functions (top panels). Only lowly abundant large domains like the largest domain for H3K4me3 or H3K9me3 with abundance below 10% (see Table S2 for values) changed their apparent size when coverage was reduced. Due to their low abundance we did not include them in our discussion.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

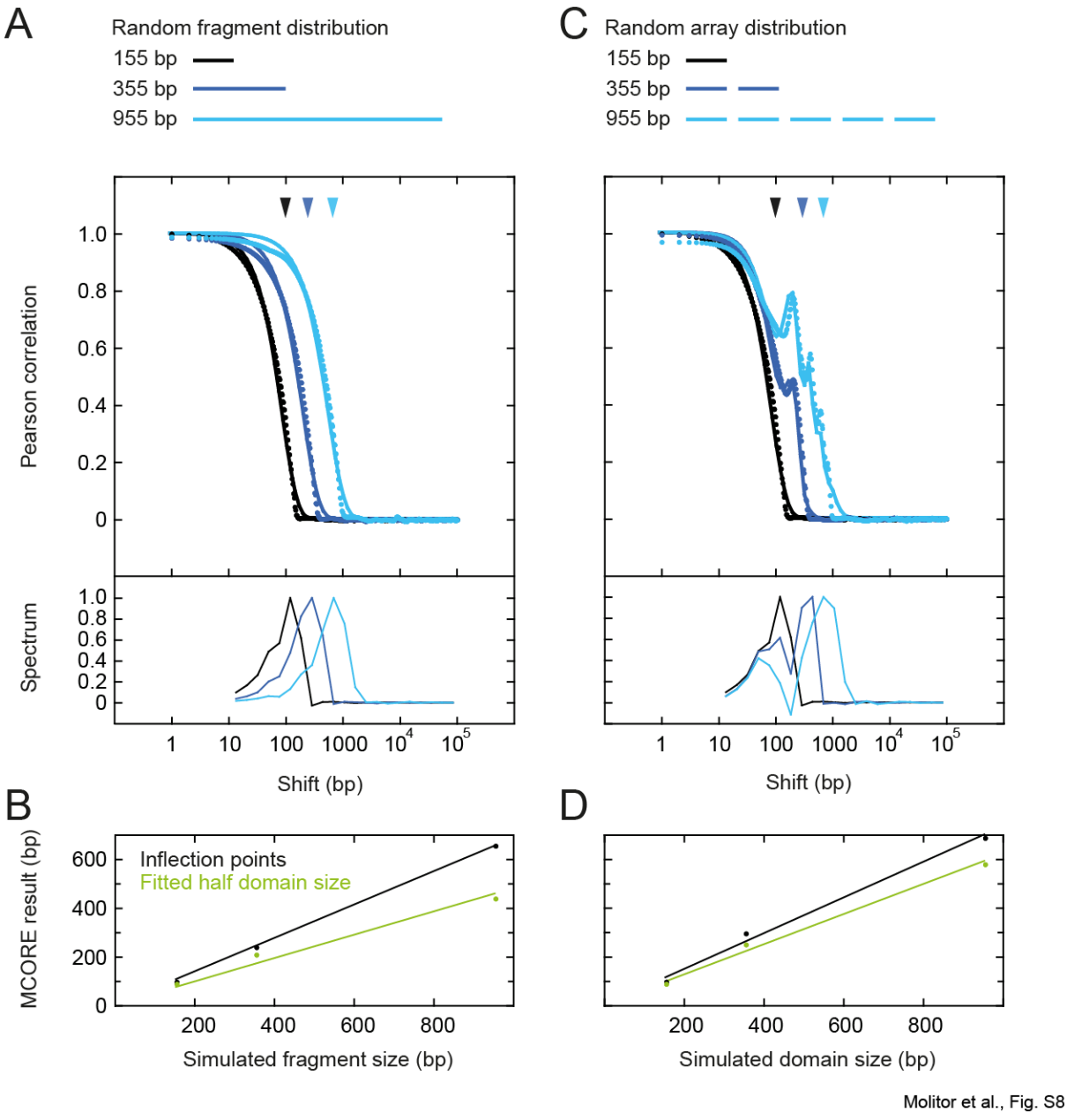


Figure S8 | MCORE for simulated data sets. (A) Correlation functions (dotted lines) for randomly distributed fragments of different size exhibit a single decay length that can be retrieved by assessing inflection points (arrowheads), by fitting the model function in Eq. 7 (solid lines) or by evaluating the decay spectrum obtained from the Gardner transformation shown below the curves. **(B)** Fit parameters obtained for the curves shown in panel A yield half domain sizes (green), whereas the positions of inflection points correspond to 0.7-times the domain sizes (black). **(C)** Correlation functions (dotted lines) for nucleosomal arrays (instead of continuous fragments as in panel A) display global decay lengths that correspond to array sizes. The decay lengths coincide with the largest inflection points depicted by the arrowheads. In addition, correlation functions exhibit an oscillatory contribution due to the nucleosomal pattern within the arrays. The nucleosome repeat length of 200 bp used for the simulation was retrieved by fitting with Eq. 7 (solid lines). **(D)** The array size in panel C is either obtained from the analysis of inflection points (black), the peaks of the decay spectrum or the fitted half domain sizes (green), with the same scaling found for continuous domains in panel B.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

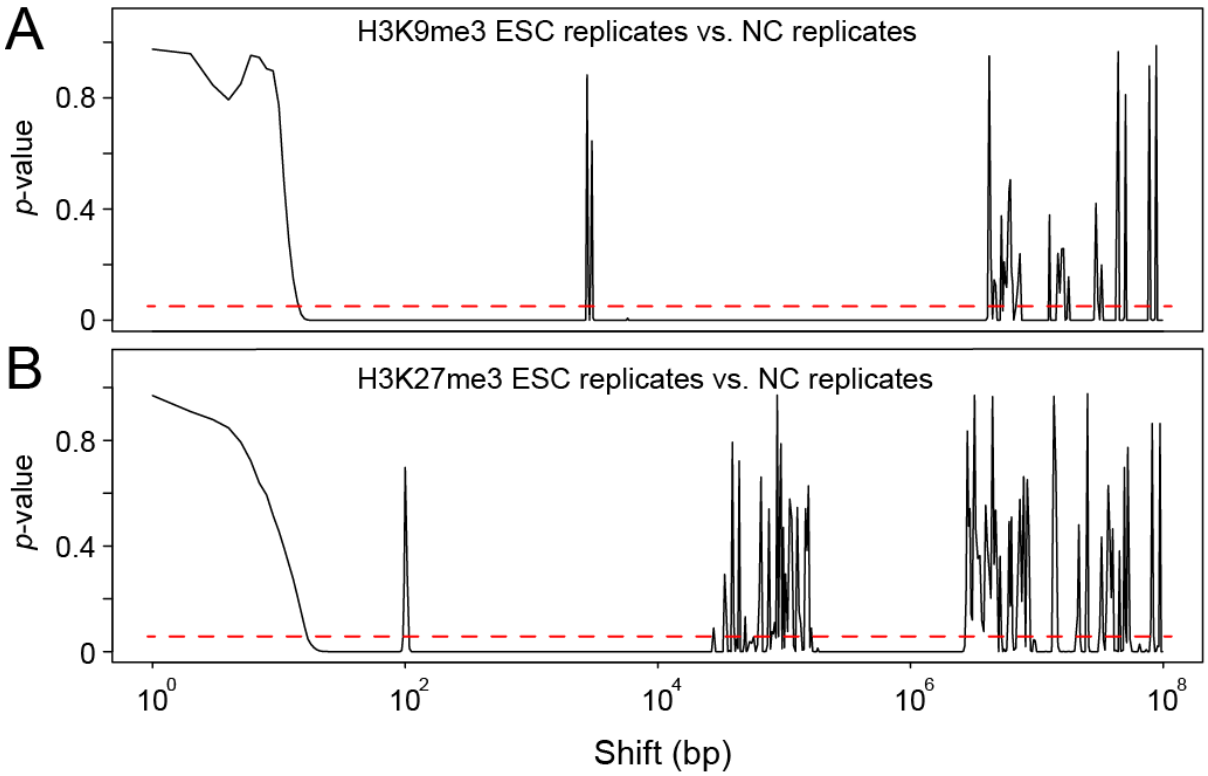


Figure S9 | Statistical comparison of correlation functions. Based on 95% confidence intervals, the statistical significance of differences between correlation functions can be assessed. A *p*-value for the difference of two functions can be obtained using a *t*-test for the correlation coefficient pair at each shift distance and the corresponding confidence intervals (Materials and Methods). Typically, the statistical error of the correlation function is very small due to the large number of genomic regions considered for the calculation of the correlation coefficient (Materials and Methods). The red dashed lines indicate a *p*-value of 0.05. **(A)** *p*-value for the difference between correlation coefficients at each shift distance are given for the replicate correlations of H3K9me3 in ESCs versus NCs. Correlation curves are shown in Fig. 4B (top, black/blue). **(B)** Same as panel A for H3K27me3. Correlation functions are shown in Fig. 4B (center, black/blue).

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

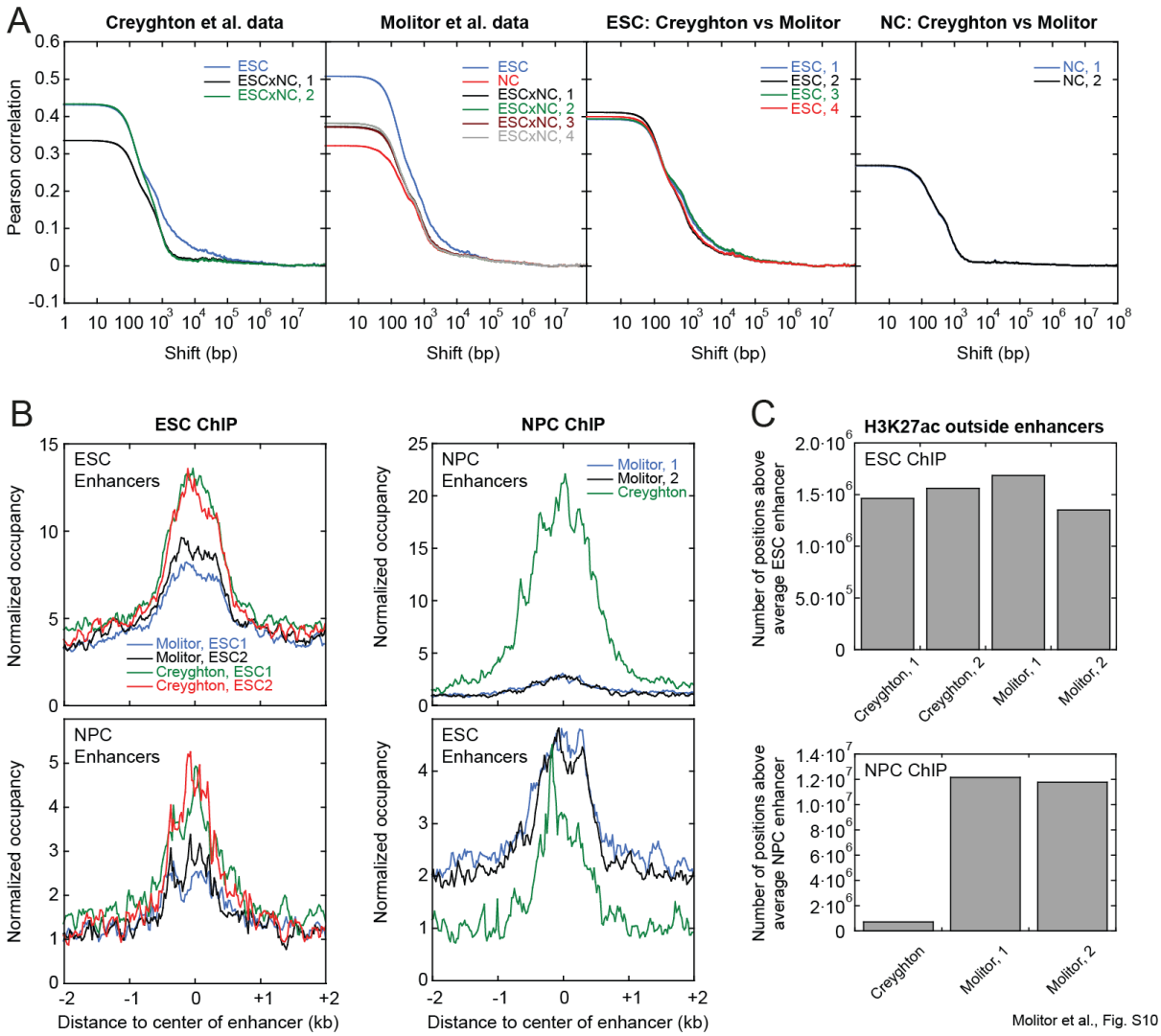


Figure S10 | MCORE for different H3K27ac data sets. (A) Correlation functions for H3K27ac data sets from this manuscript (Molitor et al. data) and from the study of Creyghton et al. [1]. Both data sets yielded similar results in the MCORE analysis. **(B)** Enrichment at the enhancers identified by Creyghton et al. was found for all data sets assessed here. **(C)** The enhancers identified by Creyghton et al. were not the only genomic regions enriched for H3K27ac. According to the MCORE results, these additional regions did not change their H3K27ac signature during development and dominated the global genome-wide distribution of the modification.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

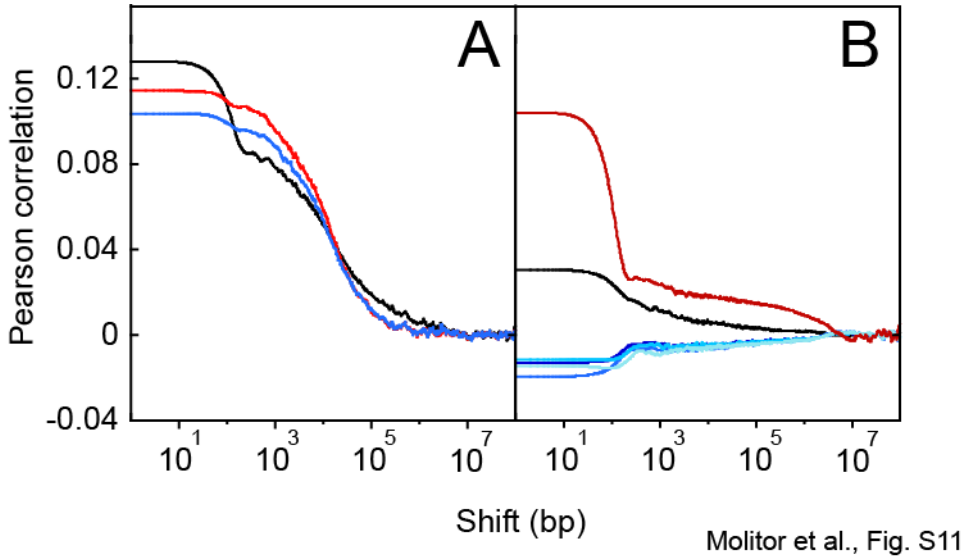


Figure S11 | Quality control of ChIP-seq data. (A) Replicate correlation functions from three ChIP-seq experiments of H3K36me3 in ESCs for all pairwise combinations, replicate 1 and 2 (black), replicate 1 and 3 (red), replicate 2 and 3 (blue). The correlation functions show variations that reflect the biological reproducibility of the experiment. **(B)** Evaluation of two different antibodies used for ChIP-seq of H3K9ac in ESCs. Two ChIP-seq experiments were conducted with polyclonal antibodies from abcam (ab4441, replicate ab1 and ab2) or Active Motif (#39137, replicates am1 and am2). Replicate correlation functions of experiments with the same antibody showed significant correlation (ab1 and ab2, red line; am1 and am2 black line) with a difference in the amplitude that indicates a higher similarity and thus a better reproducibility of ChIP-experiments conducted with the Abcam antibody. Cross-correlation functions calculated for data sets using different antibodies (blue curves for every combination of two replicates, ab1 x am1, ab1 x am2, ab2 x am1, ab2 x am2) yielded negative correlations. Thus, the two antibodies recognize different chromatin features and further validation is necessary to make conclusions on the H3K9ac distribution.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

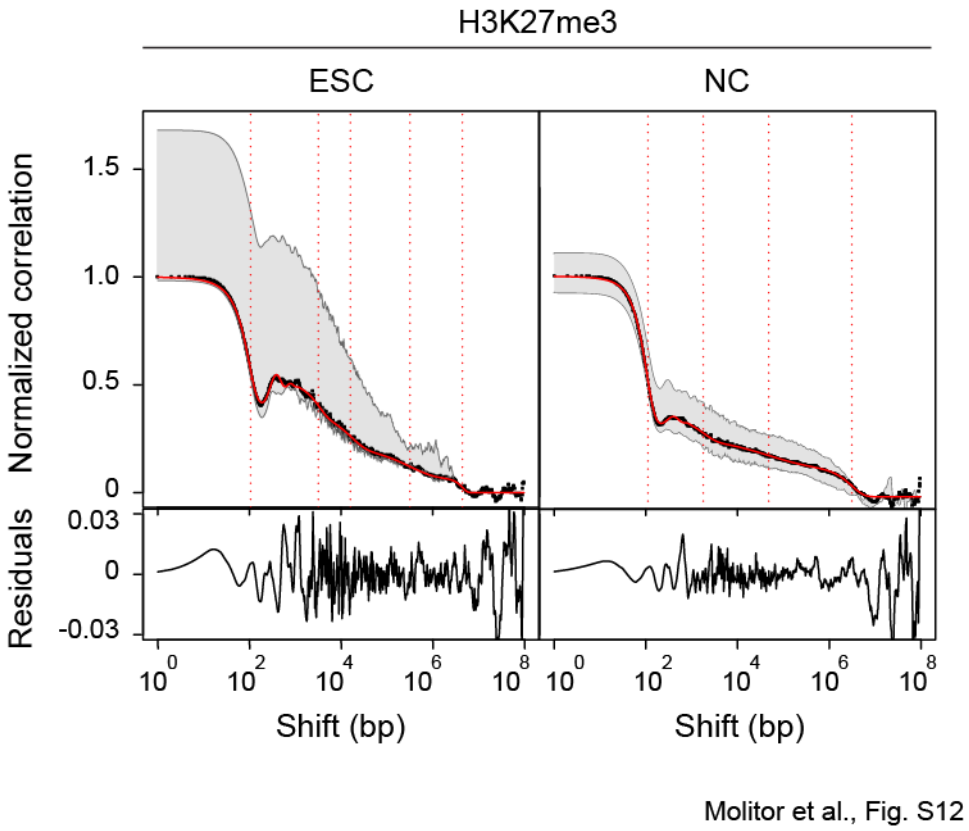


Figure S12 | Fitted correlation functions for H3K27me3. Correlation functions calculated between replicates on chromosome 1 (black) and fit functions according to Eq. 7 (red) with half domain sizes obtained from the fit (vertical dotted lines). Grey regions indicate maximum variation between chromosomes. Fit residuals for the correlation functions are shown below the curves. Fit parameters are summarized in Tables S3 and S4.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

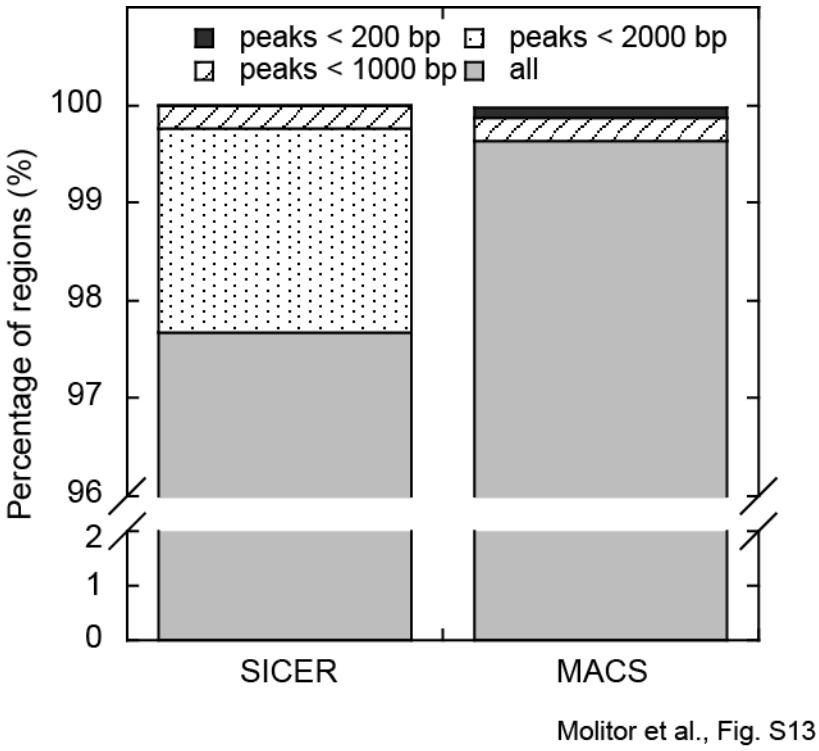
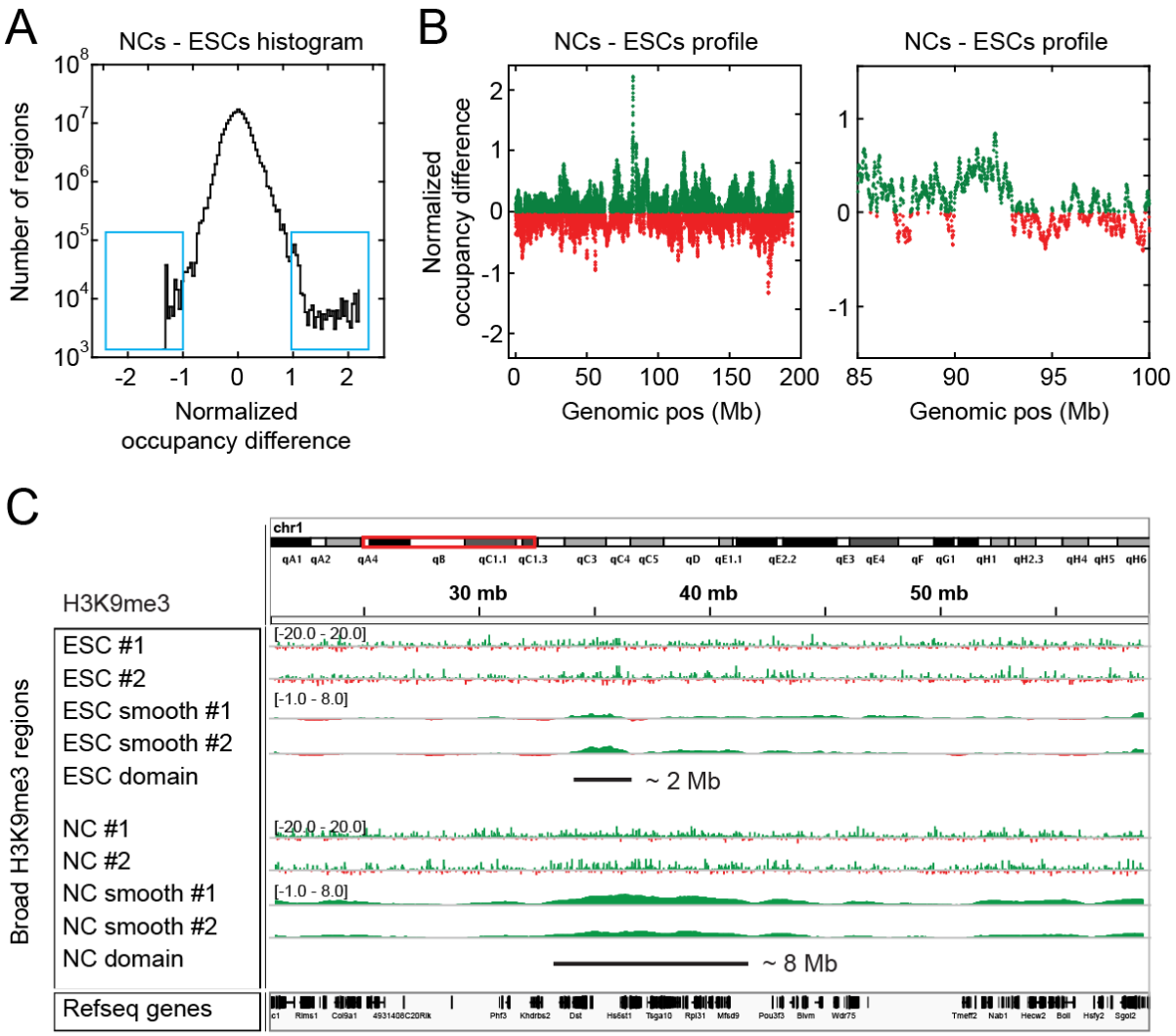


Figure S13 | Peak calling summary for H3K9me3. MACS and SICER were used to identify peaks of H3K9me3 in NCs. Parameters were used as indicated in the Material and Methods section. Numbers of peaks with different sizes are given. 100% refers to all of the peaks identified by MACS (3630 peaks containing 0.4% of all mapped reads) or SICER (35780 peaks containing 9.45% of all mapped reads).

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Molitor et al., Fig. S14

Figure S14 | MCORE-directed annotation of chromatin features. MCORE identified broad H3K9me3 domains spanning on average 128 kb and 7.6 Mb in NCs that were absent in ESCs, suggesting broadening of H3K9me3 domains during differentiation of ESCs into NCs (Fig. 3A, B, Tables S3-4). **(A)** To identify broad regions enriched for H3K9me3 in NCs but to a lesser extent in ESCs, the coverage difference for normalized occupancy profiles in ESCs and NCs was calculated in a sliding window of 128 kb in size. A histogram for the obtained values is shown. The histogram is relatively symmetric and centered at zero, indicating that most genomic regions are not differentially modified with H3K9me3 in ESCs or NCs. The tails (blue rectangles) show that the largest coverage differences are found in regions that gain H3K9me3 in NCs. **(B)** The coverage difference along chromosome 1 (left, maximum and minimum values within 10 kb bins are plotted) and a zoom-in of the genomic region in Fig. 3C (88.7 - 89.3 Mb, right) are shown. **(C)** To annotate the genomic positions of broad H3K9me3 domains, reads were counted and evaluated in a sliding window with the respective size. An example of a domain with ~7.6 Mb that became broader in NCs is shown. For clarity the occupancy profiles were smoothed with 0.2-times the window size. An example for window size 128 kb is shown in Fig. 3C.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

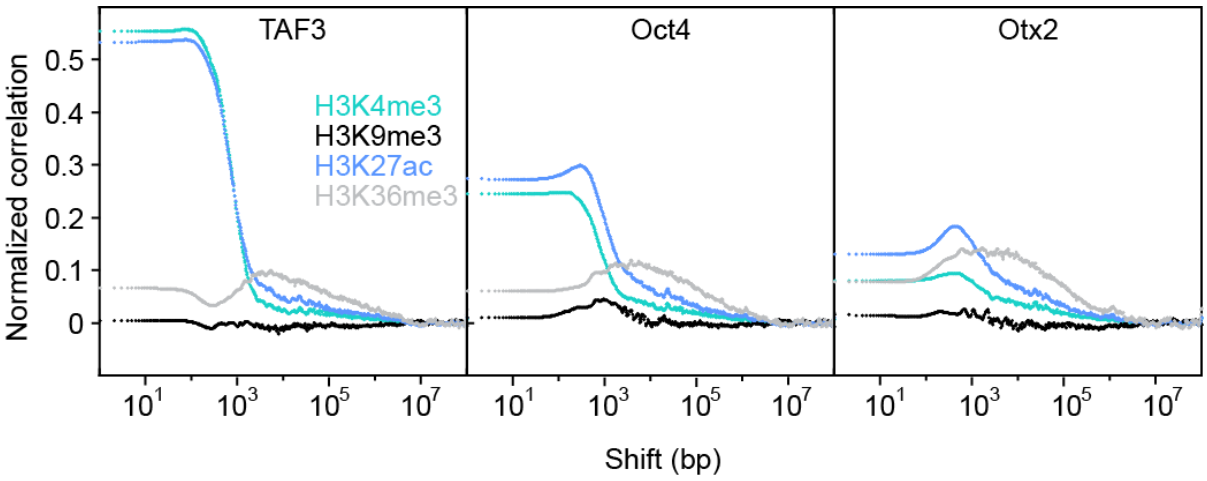


Figure S15 | MCORE for transcription factor binding. Co-localization of transcription factors with different histone modifications was studied in ESCs. Cross-correlation functions of TAF3, Oct4 or Otx2 vs. H3K4me3, H3K9me3, H3K27ac and H3K36me3 are shown. Binding of TAF3 strongly correlates with H3K4me3 and H3K27ac, which mark active promoters and enhancers in mouse ESCs [1, 2]. The binding of TAF3 to enhancers is in line with publications showing that active enhancers are transcribed by the RNA Polymerase II machinery [3] and that TAF3 mediates chromatin-looping events that regulate transcriptional activation [4]. Oct4 and Otx2 are two transcription factors that regulate pluripotency and differentiation. Their binding correlates with H3K27ac in agreement with previous reports [5]. The peaks in the correlation curves reflect the ~300 bp distance between the binding site of the transcription factor and the modified nucleosome, which was also found recently [6]. For each of the three transcription factors maximum correlation with H3K36me3 was found at shift distances around 10 kb, which is similar to the average gene length and indicates that these factors globally bind adjacent to active genes. TAF3, Oct4 and Otx2 binding is uncorrelated with H3K9me3, which is consistent with their role in active transcription.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

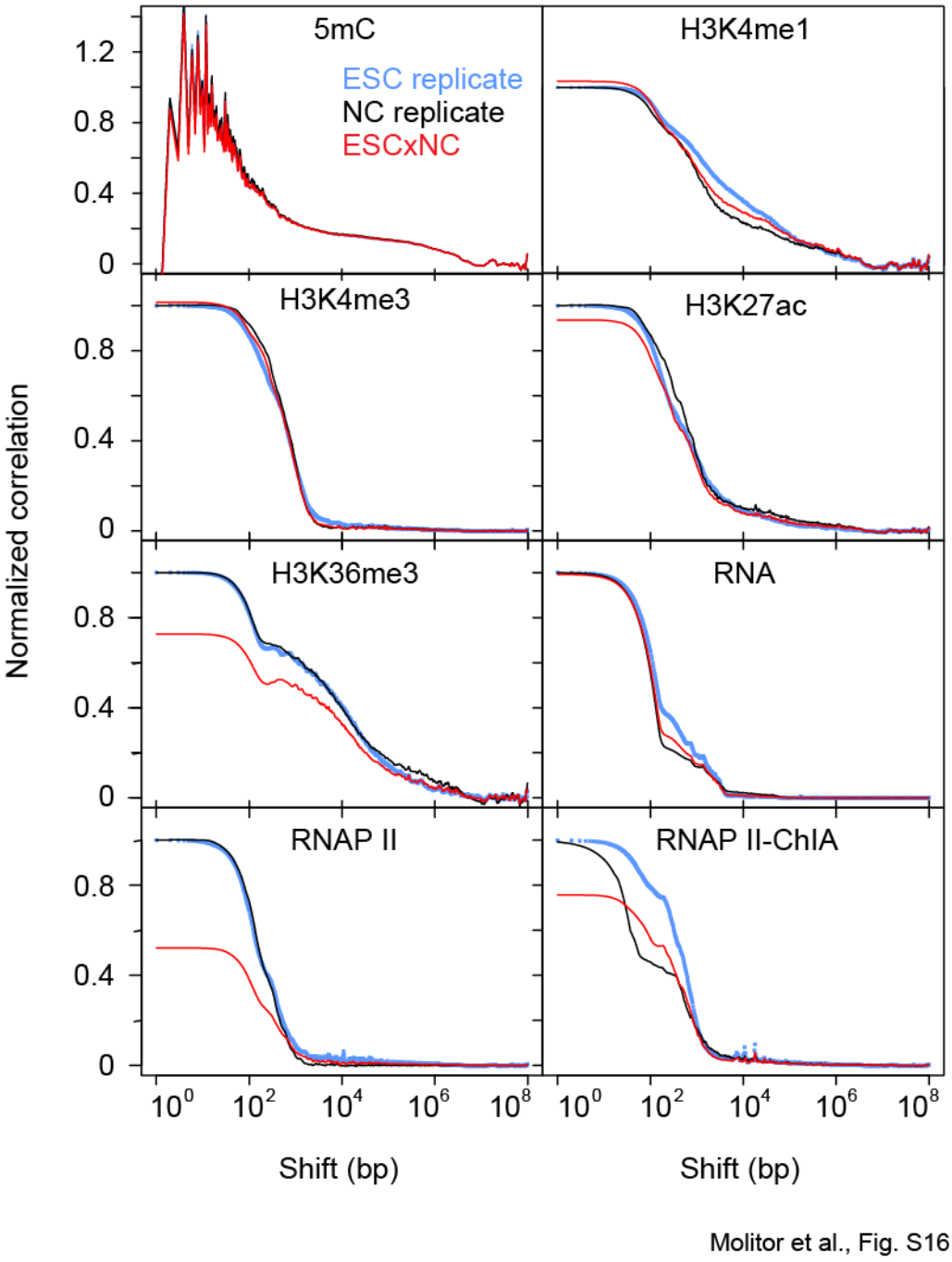


Figure S16 | Spatial extension and co-localization of different features in ESCs versus NCs. Correlation functions for replicates of H3K4me1, H3K4me3, H3K27ac, H3K36me3 and RNA Polymerase II (RNAP II) ChIP-seq, RNA-seq (RNA) and RNAP II ChIA-PET data (RNAP II-ChIA) in ESCs (blue) and NCs (black) reflect the domain topologies of the respective features. Cross-correlation functions (red) between the same feature in ESCs and NCs quantify the co-localization of the feature in both cell types. Most features depicted here do not drastically change their global distribution during differentiation because cross- and replicate correlation functions are similar to each other.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

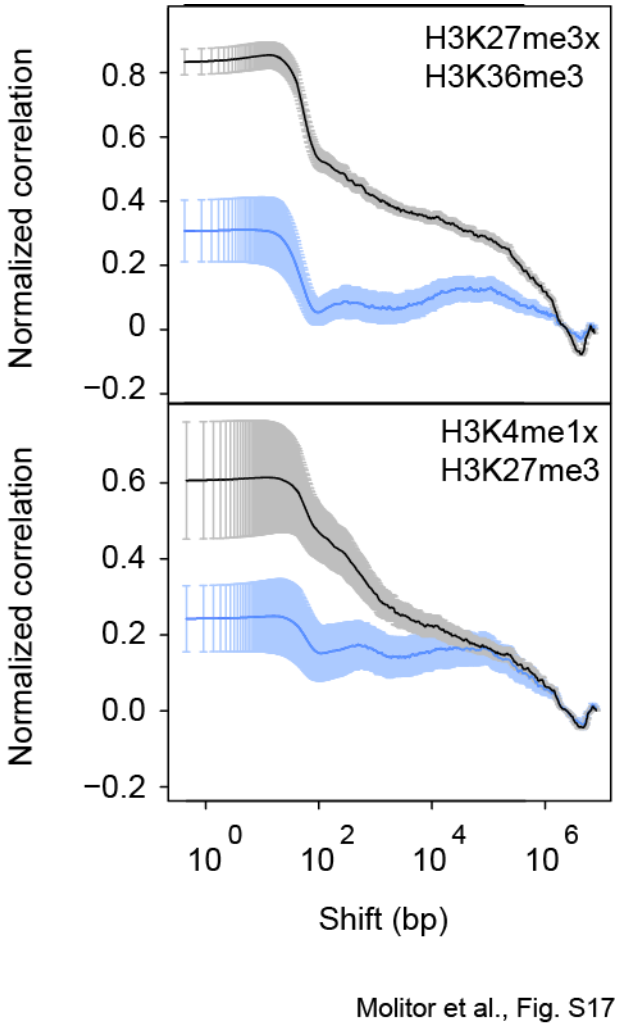


Figure S17 | Heterochromatin reorganization during differentiation. Cross correlation functions between H3K27me3 and H3K4me1/H3K36me3 in ESCs (blue) or NCs (black) are shown. H3K27me3 exhibited increased co-localization with activating marks in NCs. Error bars indicate s.e.m. among replicates.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

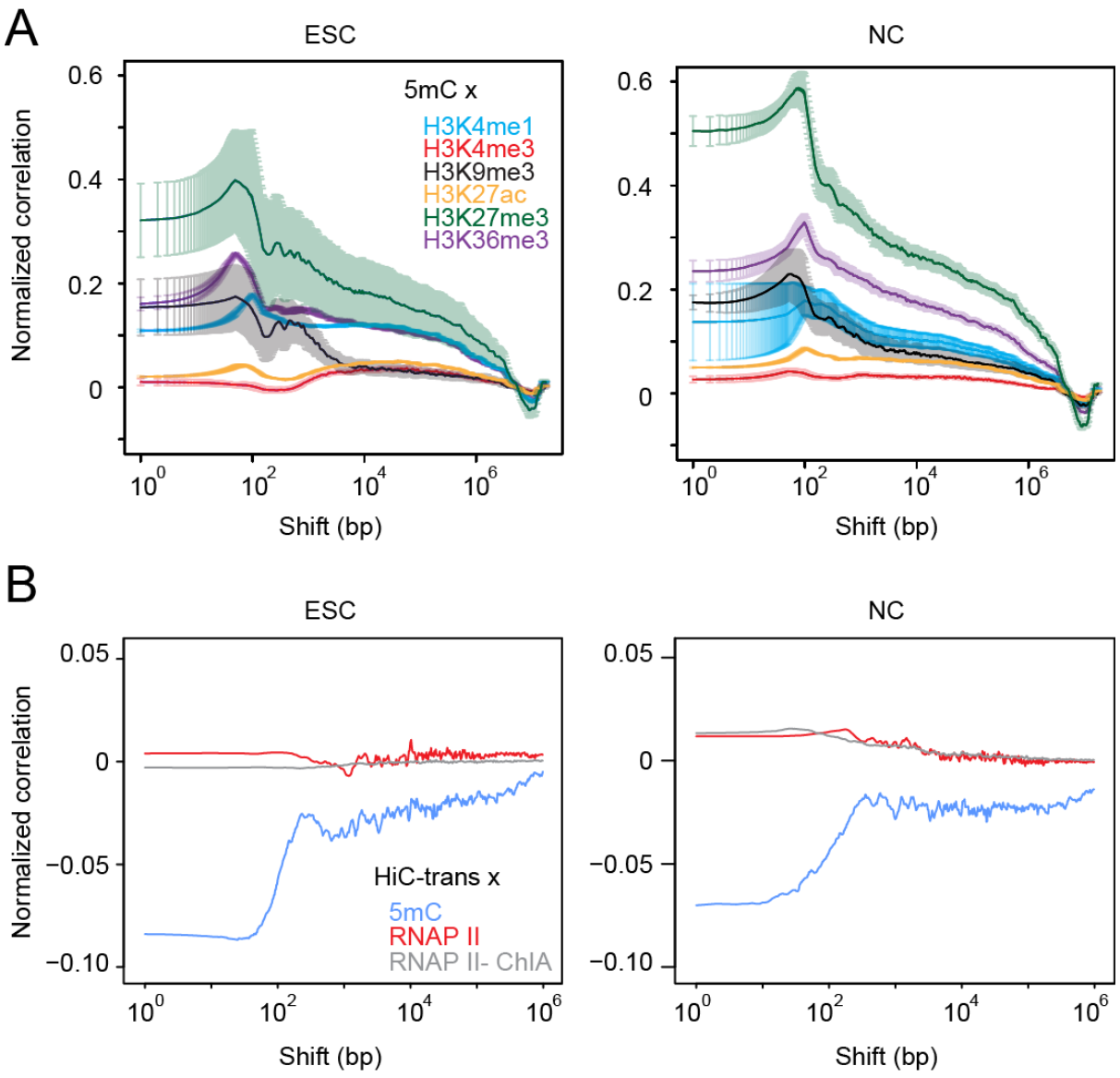


Figure S18 | DNA methylation and inter-chromosomal contacts. (A) Cross correlation functions for DNA methylation and different histone modifications in ESCs (left) and NCs (right) are shown. Error bars indicate s.e.m. among replicates. **(B)** Cross-correlation functions for inter-chromosomal contact sites (Hi-C trans) and DNA methylation (5mC), RNA Polymerase II (RNAP II) and RNAP II ChIA-PET (RNAP II-ChIA) in ESCs (left) and NCs (right) are shown. RNAP II and RNAP II contact sites became moderately enriched at the surface of the chromosome territory in NCs, whereas 5mC tended to localize inside chromosome territories in both cell types.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

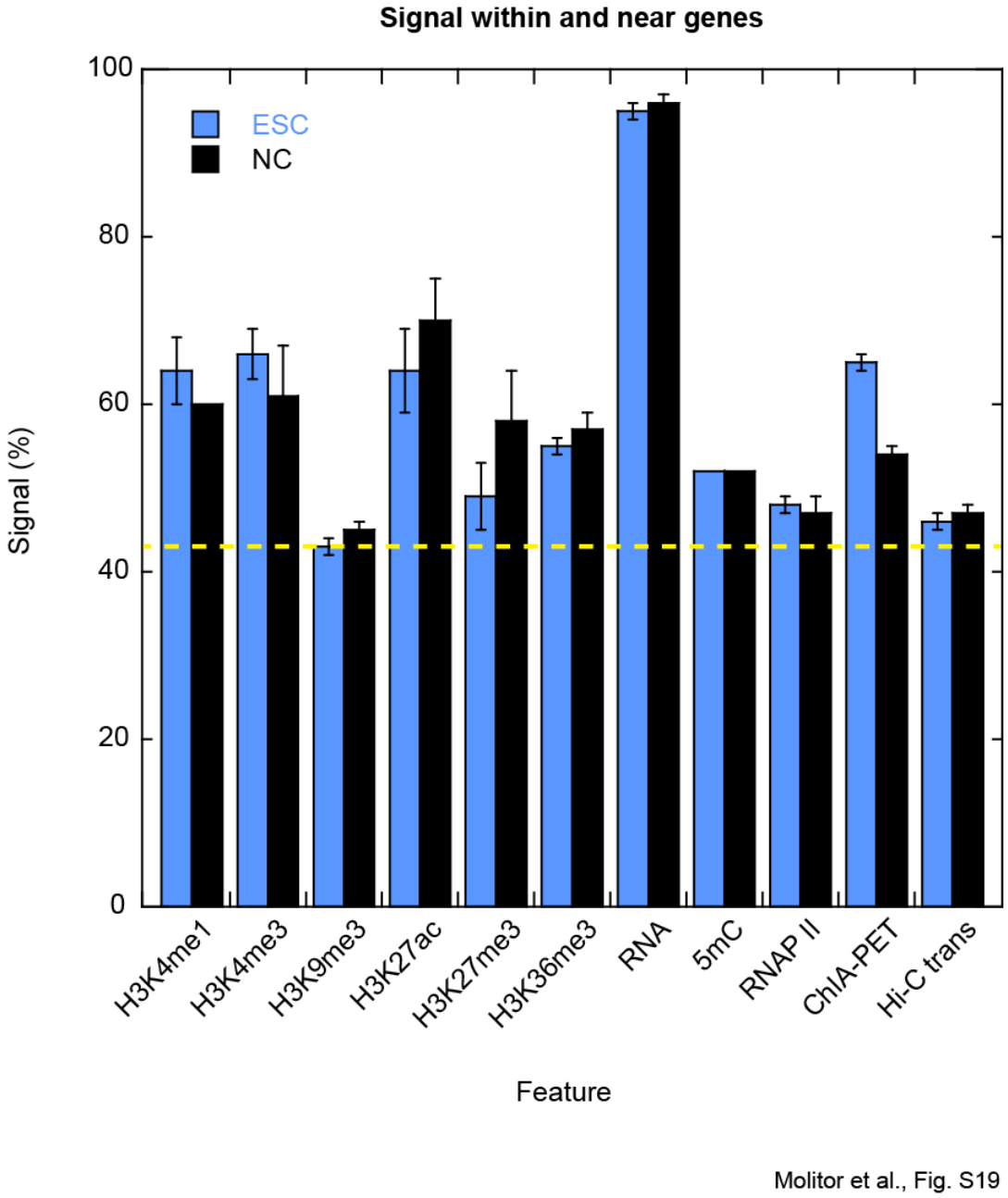


Figure S19 | Fraction of chromatin features within and near genes. For each chromatin feature the fraction of signal within or near genes (gene-proximal region, between 5 kb upstream of the transcription start site and 5 kb downstream of the transcription termination site) was calculated from normalized occupancy profiles for chromosome 1. In contrast to the RNA signal that exhibits strong preference for genes, histone modifications typically associated with active genes like H3K36me3 were also located distant from genes. The yellow dashed line marks equal partitioning between gene-distal and gene-proximal regions because the latter one spans 43% of the uniquely mappable portion of chromosome 1.

	Correlation function	Sliding window binning [a]	Peak [a] calling	Multi-scale representation	Probabilistic network models	Deconvolved correlation	Strand-specific correlation
Tool(s)	MCORE	cisGenome, SiSSRs, SPP	MACS, SICER	MSR	ChromHMM, Segway	Arpeggio	SPP
Platform	Java	various	Python	Matlab script or compiler runtime	Java, Python	Java	R-script
Sequencing data type	Unrestricted	Unrestricted	ChIP-seq	Unrestricted	Unrestricted	ChIP-seq	ChIP-seq
Mixed data type analysis implemented	Yes	No	No	No [b]	Yes	No	No
Applications	Quality control, domain features, spatial relations	Local feature enrichment	Local feature enrichment	Multi-scale feature enrichment	Segmentation into feature states	Comparison of data sets, local structure	Quality control for sequencing data
Correction [c]	Input and/or control	Input or control	Input or control	Mappability, GC content, input or control	Input or control	Input or control	None
Detected feature scale	1 bp – 1 chromosome	1 bp – 1 chromosome	< 10 kb (MACS) variable (SICER)	1 bp – 1 chromosome	1 bp – 1 chromosome	40 bp – 8 kb	Fixed window size
Information on shifted relationships	Yes	No	No	Limited [b]	No	No	No
Required input parameters	None	Window size	MACS: p-value threshold, tag length/shift SICER: size of gap & window, FDR	Resolution, scale number, *p*-value threshold	State number, *p*-value threshold	None	None
Number of data sets	2	1	1	1 [b]	>1	1	1
Noise sensitivity	Low [d]	Low	High [d]	n.d.	n.d.	Low	Low

Genome locus annotation	No	Yes	Yes	Yes	Yes	No	Yes
Output [e]	Domain sizes, nucleosome spacing, spatial relationships, normalized occupancy	Enrichment over average	Local enrichment	Scale dependent enrichment	Length distribution, abundance of chromatin states	Feature profile, nucleosome spacing	Peak separation distance
Operating system [f]	All	All	All	Unix, Windows	All (ChromHMM) Linux (Segway)	Unix, Mac OS X	All
Comment	Low sensitivity to noise, bias and undersampling	Read counting in a window of predefined size	Restricted scale-range	Can be applied as a peak caller with pruning.	Predefined number and type of states.	Removes large-scale structures by filtering	Recommended analysis prior to peak calling
Reference	This study	[10-12]	[13-15]	[16]	[17-20]	[21]	[10, 22]

Table S1 | Comparison of MCORE with other software tools

The table represents a non-comprehensive list of representative tools that are used to extract information about chromatin features from deep sequencing data sets.

[a] Exemplary tools are mentioned. For other programs see compilations in ref. [7, 8].

[b] MSR can be applied to identify region of simultaneous enrichments for two different ChIP-seq data sets by computing a matrix of segments, but this analysis is not part of the current implementation. In some cases differential correlation of the matrix indicates the presence of shifted correlations.

[c] Control reactions depend on the type of sequencing data and could involve for example a ChIP-seq reaction without the specific antibody.

[d] See ref. [9] for peak calling and Fig. S3 for MCORE

[e] The “enrichment” analysis of a given feature would also provide the information about its depletion with respect to a given average signal.

[f] All operating systems refers to Unix, Windows and Mac OS X.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ESC	H3K4me3		H3K9me3		H3K27me3		H3K36me3	
number of domains	3		4		5		3	
	value	SE	value	SE	value	SE	value	SE
a1 (%)	18.0	0.5	27.6	0.6	25.3	1.2	26.9	<0.5
a2 (%)	75.7	0.6	46.4	2.4	20.0	4.3	51.4	3.0
a3 (%)	6.3	0.3	21.0	3.0	23.3	6.0	21.7	1.8
a4 (%)	-	-	5.0	3.9	22.1	3.5	-	-
a5 (%)	-	-	-	-	9.3	8.3	-	-
b1 (bp)	132	2	107	2	106	3	119	2
b2 (bp)	926	6	1586	18	3198	173	14803	296
b3 (kb)	33	6	11	2	16	2	356	105
b4 (kb)	-	-	1121	704	322	46	-	-
b5 (kb)	-	-	-	-	4481	195	-	-
c1 (%)	99	fixed	98	<0.05	69	1	97	1
c2 (bp)	173	fixed	182	9	207	5	182	5
c3 (bp)	1000	fixed	654	340	219	9	802	303
n1	1.97	0.05	2.20	0.10	3.31	0.27	2.30	0.50
n2	1.25	0.01	1.11	0.00	1.96	0.25	0.62	0.01
n3	0.38	0.02	0.64	0.10	1.28	0.33	0.45	0.04
n4	-	-	0.39	0.10	0.79	0.17	-	-
n5	-	-	-	-	3.96	0.97	-	-

Table S2 | Fit parameters for selected correlation functions in ESCs. Correlation functions calculated for replicates of H3K4me3, H3K9me3, H3K27me3 and H3K36me3 (Fig. 3A, Fig. S12) were fitted with Eq. 7 (Materials and Methods), yielding the indicated fit parameters and corresponding standard errors (SE). The minimum number of domains required to yield uncorrelated fit residuals was chosen. The amplitudes a1-a5 represent the relative domain abundance, the decay length parameters b1-b5 represent half of the respective domain sizes, and the value of c2 reflects nucleosome spacing. See text and Materials and Methods for further details.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NC	H3K4me3		H3K9me3		H3K27me3		H3K36me3	
number of domains	3		4		4		3	
	value	**SE**	**value**	**SE**	**value**	**SE**	**value**	**SE**
a1 (%)	18.2	1.5	47.5	3.1	54.7	1.2	25.7	0.4
a2 (%)	79.9	1.5	23.2	3.6	11.5	1.9	57.3	0.9
a3 (%)	1.9	1.9	17.7	2.9	17.5	2.4	17.0	0.9
a4 (%)	-	-	11.6	5.5	16.3	3.2	-	-
b1 (bp)	243	4	202	19	111	2	111	1
b2 (bp)	985	14	2036	142	1791	91	11848	287
b3 (kb)	617	106	64	13	48	9	1412	85
b4 (kb)	-	-	3771	256	3132	131	-	-
c1 (%)	98	<0.5	74	4	82	2	99	<0.5
c2 (bp)	134	3	175	4	218	15	182	6
c3 (bp)	3017	3017[a]	367	41	224	27	11505	11505[a]
n1	2.01	0.12	2.31	0.60	2.67	0.09	2.47	0.07
n2	1.43	0.03	1.11	0.15	1.54	0.24	0.59	0.01
n3	0.53	0.07	0.62	0.13	0.52	0.09	0.79	0.05
n4	-	-	1.56	0.24	1.64	0.15	-	-

Table S3 | Fit parameters for selected correlation functions in NCs. Correlation functions calculated for replicates of H3K4me3, H3K9me3, H3K27me3 and H3K36me3 (Fig. 3B and Fig. S12) were fitted with Eq. 7, yielding the indicated fit parameters and corresponding standard errors (SE) as described in the Materials and Methods section and the legend to Table S2.

[a] Fit error truncated since it exceeded the allowed parameter range

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

target	cell type	accession replicate1	accession replicate2	reference
Input	ESC	GSM1516068	GSM1516069	This study
Input	ESC	SRX499123	SRX499124	[5]
IgG	ESC	GSM1516070	GSM1516071	This study (RA073)
IgG	ESC	GSM1516072	GSM1516073	This study (PP500P)
IgG	ESC	SRR331056	SRR331057	[4]
5mC	ESC	SRX080191		[23]
H3K27ac	ESC	GSM1516076	GSM1516077	This study (ab4729)
H3K27me3	ESC	GSM1516074	GSM1516075	This study (ab6002))
H3K36me3	ESC	GSM1516082	GSM1516083	This study (ab9050)
H3K4me1	ESC	GSM1516080	GSM1516081	This study (ab8895)
H3K4me3	ESC	GSM1516086	GSM1516087	This study (ab8580)
H3K9me3	ESC	GSM1516084	GSM1516085	This study (ab8898)
Hi-C	ESC	SRX116341	SRX116342	[24]
Input	ESC	SRR317225	SRR317226	ENCODE
Oct4	ESC	SRX499114	SRX499115	[5]
Otx2	ESC	SRX499116	SRX499117	[5]
RNA	ESC	GSM1516088 GSM1516089	GSM1516090 GSM1516091	This study
RNAP II	ESC	SRR489721	SRR489722	ENCODE
RNAP II-ChIA	ESC	SRX243706	SRX243707	[25]
TAF3	ESC	SRR331054	SRR331055	[4]
Input	NPC	SRX604258	SRX604259	[26]
IgG	NPC	GSM1516092	GSM1516093	This study (RA073)
5mC	NPC	SRX080193-5		[23]
H3K27ac	NPC	GSM1516096	GSM1516097	This study (ab4729)
H3K27me3	NPC	GSM1516094	GSM1516095	This study (ab6002))
H3K36me3	NPC	SRX604262	SRX604263	[26]
H3K4me1	NPC	GSM1516100	GSM1516101	This study (ab8895)
H3K4me3	NPC	GSM1516102	GSM1516103	This study (ab8580)
H3K9me3	NPC	SRX604260	SRX604261	[26]
Hi-C	Cortex	SRX128219	SRX128220	[24]
Input	Brain E14.5	SRR489727	SRR578284	ENCODE
RNA	NPC	GSM1516104 GSM1516105	GSM1516106 GSM1516107	This study
RNAP II	Brain E14.5	SRR578272	SRR578273	ENCODE
RNAP II-ChIA	NPC	SRX243710		[25]

Table S4 | Summary of data sets used in this study.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary References

1. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al: Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* 2010; 107:21931-21936.

2. Zentner GE, Tesar PJ, Scacheri PC: Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* 2011; 21:1273-1283.

3. Natoli G, Andrau JC: Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet* 2012; 46:1-19.

4. Liu Z, Scannell DR, Eisen MB, Tjian R: Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* 2011; 146:720-731.

5. Buecker C, Srinivasan R, Wu Z, Calo E, Acampora D, Faial T, Simeone A, Tan M, Swigut T, Wysocka J: Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* 2014; 14:838-853.

6. Yang SH, Kalkan T, Morissroe C, Marks H, Stunnenberg H, Smith A, Sharrocks AD: Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency. *Cell Rep* 2014; 7:1968-1981.

7. Pepke S, Wold B, Mortazavi A: Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 2009; 6:S22-32.

8. Park PJ: ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009; 10:669-680.

9. Jung YL, Luquette LJ, Ho JW, Ferrari F, Tolstorukov M, Minoda A, Issner R, Epstein CB, Karpen GH, Kuroda MI, Park PJ: Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res* 2014; 42:e74.

10. Kharchenko PV, Tolstorukov MY, Park PJ: Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 2008; 26:1351-1359.

11. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 2008; 26:1293-1300.

12. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K: Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 2008; 36:5221-5231.

13. Barski A, Cuddapah S, Cui K, Roh T, Schones D, Wang Z, Wei G, Chepelev I, Zhao K: High-resolution profiling of histone methylations in the human genome. *Cell* 2007; 129:823-837.

14. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008; 9:R137.

bioRxiv preprint doi: https://doi.org/10.1101/054049; this version posted May 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

15. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 2009; 25:1952-1958.

16. Knijnenburg TA, Ramsey SA, Berman BP, Kennedy KA, Smit AF, Wessels LF, Laird PW, Aderem A, Shmulevich I: Multiscale representation of genomic signals. *Nat Methods* 2014; 11:689-694.

17. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B: Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* 2010; 143:212-224.

18. Ernst J, Kellis M: ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012; 9:215-216.

19. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS: Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 2012; 9:473-476.

20. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al: Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 2013; 41:827-841.

21. Stanton KP, Parisi F, Strino F, Rabin N, Asp P, Kluger Y: Arpeggio: harmonic compression of ChIP-seq data reveals protein-chromatin interaction signatures. *Nucleic Acids Res* 2013; 41:e161.

22. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012; 22:1813-1831.

23. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al: DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 2011; 480:490-495.

24. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012; 485:376-380.

25. Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, Lim J, Tai E, Poh HM, Wong E, et al: Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* 2013; 504:306-310.

26. Muller-Ott K, Erdel F, Matveeva A, Mallm JP, Rademacher A, Hahn M, Bauer C, Zhang Q, Kaltofen S, Schotta G, et al: Specificity, propagation, and memory of pericentric heterochromatin. *Mol Syst Biol* 2014; 10:746.