BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 19 2013, pages 2380–2386
doi:10.1093/bioinformatics/btt404

Genome analysis

Advance Access publication July 11, 2013

# Modeling nucleosome position distributions from experimental nucleosome positioning maps

Robert Schöpflin[1], Vladimir B. Teif[2], Oliver Müller[1], Christin Weinberg[1], Karsten Rippe[2] and Gero Wedemann[1,*]

[1]Institute for Applied Computer Science, University of Applied Sciences Stralsund, Zur Schwedenschanze 15, Stralsund 18435, Germany and [2]Deutsches Krebsforschungszentrum (DKFZ) & BioQuant, Im Neuenheimer Feld 280, Heidelberg 69120, Germany

Associate Editor: Inanc Birol

**ABSTRACT**

**Motivation:** Recent experimental advancements allow determining positions of nucleosomes for complete genomes. However, the resulting nucleosome occupancy maps are averages of heterogeneous cell populations. Accordingly, they represent a snapshot of a dynamic ensemble at a single time point with an overlay of many configurations from different cells. To study the organization of nucleosomes along the genome and to understand the mechanisms of nucleosome translocation, it is necessary to retrieve features of specific conformations from the population average.

**Results:** Here, we present a method for identifying non-overlapping nucleosome configurations that combines binary-variable analysis and a Monte Carlo approach with a simulated annealing scheme. In this manner, we obtain specific nucleosome configurations and optimized solutions for the complex positioning patterns from experimental data. We apply the method to compare nucleosome positioning at transcription factor binding sites in different mouse cell types. Our method can model nucleosome translocations at regulatory genomic elements and generate configurations for simulations of the spatial folding of the nucleosome chain.

**Availability:** Source code, precompiled binaries, test data and a web-based test installation are freely available at http://bioinformatics.fh-stralsund.de/nucpos/

**Contact:** gero.wedemann@fh-stralsund.de

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online.

Received on February 13, 2013; revised on May 21, 2013; accepted on July 6, 2013

## 1 INTRODUCTION

The DNA of eukaryotic organisms is wrapped around a histone octamer, forming nucleosomes, the basic DNA packing unit of chromatin. Nucleosomes can impede the binding of transcription factors to DNA. Therefore, the positions of nucleosomes control local DNA accessibility and can directly affect the readout of DNA sequence information at regulatory genomic elements, thus regulating gene expression.

The genome-wide study of nucleosome positions has been made possible by high-throughput DNA sequencing techniques that identify the nucleosomal DNA isolated after digestion of the linker DNA between nucleosomes with micrococcal nuclease (MNase-seq) (Zhang and Pugh, 2011). However, obtaining nucleosome positions based on these MNase-seq data is fraught with difficulties: nucleosome positions have an intrinsic biological variability. Some nucleosomes, such as the +1 nucleosome at the transcription start site, are known to be positioned with high precision in yeast (Zhang *et al.*, 2011), but the positions of other nucleosomes can vary substantially (Valouev *et al.*, 2008). In addition, the results reflect the experimental methodology used. First, although MNase preferentially digests the linker DNA between the nucleosomes, the nucleosomal DNA fragments obtained in this manner display some length variation from the precise boundaries of the nucleosome. Second, the proportion of the nucleosomal DNA fragments that can be identified is affected by MNase concentration and incubation time (Allan *et al.*, 2012; Chung *et al.*, 2010; Zhang *et al.*, 2009). Third, the experiments probe a population of typically $10^5$ to $10^7$ cells. Each cell contributes its individual nucleosome configuration to this average, which is potentially incompatible with other configurations. Accordingly, experimental nucleosome occupancy maps contain overlapping and ambiguous nucleosome positions at many locations.

Several approaches and tools have been proposed to extract nucleosome positions from MNase-seq data, such as the following: peak calling on smoothed coverage data (Albert *et al.*, 2008; Flores and Orozco, 2011; Zhang *et al.*, 2008;), hidden Markov models (Lee *et al.*, 2007; Yassour *et al.*, 2008), template filtering (Weiner *et al.*, 2010), mixture models (Polishko *et al.*, 2012; Zhang *et al.*, 2012) and sliding window statistics (Becker *et al.*, 2013; Nellore *et al.*, 2012). Many of these approaches need an elaborate parameterization or cannot extract non-overlapping nucleosome configurations and decide between overlapping peaks.

The sole identification of local occupancy maxima determines the average nucleosome positions of a cell ensemble but does not cope with the sterical incompatibilities arising from averaging many, usually unsynchronized cells. If different configurations lead to overlapping occupancy peaks, it is impossible to represent the data with a single configuration of non-overlapping nucleosomes.

Here, we address these issues with two approaches going beyond peak calling. We apply a binary-variable approach that generates an ensemble of nucleosome configurations that, when combined with each other, reproduce the input data with

*To whom correspondence should be addressed.

minimal error. Ambiguous peak data can be dissected into source components, including non-overlapping nucleosome configurations. This allows quantification of the dynamics of nucleosome positioning and of the relative occlusion of DNA by nucleosomes.

For some applications, like 3D modeling of the nucleosome chain or combinatorial transcription factor binding, a single optimal configuration of non-overlapping nucleosomes is desirable, even though the data are ambiguous. We address this by applying a Metropolis Monte Carlo (MMC) algorithm that generates a dynamic population of non-overlapping nucleosomes. Our simulation algorithm yields an ensemble of nucleosome configurations that reproduces the distribution of the input data. To find an optimized placement of nucleosomes, we combine Monte Carlo simulation with a simulated annealing protocol.

Our approach compares favorably to existing peak-calling methods and it was applied to the analysis of promoter and enhancer regions in mouse embryonic stem cells (ESCs), and neural progenitor cells (NPCs) as well as embryonic fibroblasts (MEFs) derived from these (Teif *et al*., 2012). The three cell types have identical genomes but differ in their differentiation state. Comparing them by using our method allowed quantification of differences in nucleosome patterns and dynamics at binding sites for transcription factors that have important roles in cell differentiation. Furthermore, we analyzed the nucleosome density at the *Samd4* locus for the three cell types, as an example of a gene that shows changes in gene expression during differentiation.

## 2 METHODS

### 2.1 Binary-variable analysis of overlapping nucleosome populations

Positioning data from MNase-seq experiments represent an overlay of different nucleosome populations. Hence, the MNase-seq data can be imagined as a combination of different non-overlapping nucleosome configurations for the same locus. We aimed to dissect the overlay into single nucleosome configurations and quantify their occurrence such that their combination reproduces the experimental distribution with minimal error.

In a preparatory step, fragments of nucleosomal DNA from mapped paired-end reads (Fig. 1a) are transformed into occupancy data. The number of read centers is counted per base pair, smoothed with a Gaussian kernel and normalized (Fig. 1b). In the next step, we identify all peaks in the occupancy data. A peak is defined by having a greater occupancy value than the two neighboring positions. Every peak is assumed to be a potential nucleosome center. In a subsequent step, clusters of overlapping peaks are identified. A cluster is defined as a sequence of peaks that are connected by an overlap between neighboring peaks. Two peaks are considered to overlap if their peak-to-peak distance is <147 bp. A cluster comprising $n$ peaks can be represented by a vector $\vec{c} = c_1, c_2, c_3, \ldots, c_n$, with $c_i$ being the location of the $i$-th peak in the cluster. Owing to steric effects, it is not possible that all peaks of a cluster are populated by nucleosomes at the same time. Therefore, we generate different sterically possible configurations of nucleosomes for a cluster (Fig. 2). A configuration can be represented as vector $\vec{k} = k_1, k_2, k_3, \ldots, k_n \; k_i \in \{0, 1\}$ (Fig. 1c). A nucleosome at position $c_i$ is represented as $k_i = 1$, whereas $k_i = 0$ denotes a nucleosome-free position. By a recursive procedure, we create all possible configurations containing only non-overlapping nucleosomes (Fig. 2). Additionally, we apply the constraint that each configuration has to be maximally

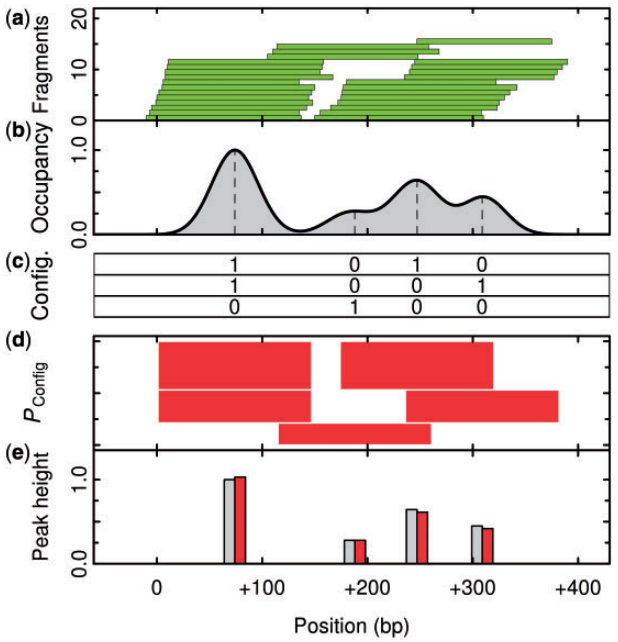


**Fig. 1.** Analysis of overlapping nucleosome populations. (**a**) Mapped paired-end fragments of nucleosomal DNA. (**b**) Reads are transferred into frequencies of nucleosomal centers, smoothed with a Gaussian kernel and normalized to probabilities (solid line) of nucleosome centers. The peaks (dashed lines) are overlapping and form a connected cluster of four potential nucleosome positions. (**c**) Configuration matrix of the cluster; value 1 marks a nucleosome; value 0 represents an empty position. (**d**) Nucleosome configurations according to the matrix in (d). Each row of rectangles represents a particular nucleosome configuration. The row height corresponds to the contribution of the configuration to the overall occupancy. The width of a rectangle is constantly 147 bp. (**e**) Peak heights from the experimental data (left bar) and reconstructed (right bar) from nucleosome configurations in (d). The difference between experimental and reconstructed peak height gives the error of the analysis

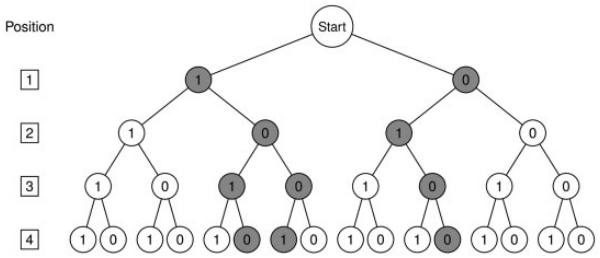


**Fig. 2.** Recursive construction of nucleosome configuration during the binary-variable analysis. The cluster shown in Figure 1 has four potential nucleosome positions. Each position has two different states: occupied by a nucleosome ($k_i = 1$) and empty ($k_i = 0$), yielding a total of $2^4$ different configurations. Only three configurations (paths along gray nodes) fulfill the constraints of having no overlapping nucleosomes and being maximally populated. The result set can be represented as a binary matrix (Fig. 1c)

populated, that is, no further nucleosome can be added without producing an overlap. The $m$ different configuration vectors can be combined to a matrix $K$ with $m$ rows and $n$ columns.

After identifying all sterically possible nucleosome configurations, we compute the proportion of the individual configurations with respect to

the original peak heights. The $n$ individual peak heights of a cluster can be represented by a vector $\vec{h} = h_1, h_2, h_3, \ldots, h_n$. Given that occupancy data are an overlay of different configurations, the contribution of the individual configurations should ideally sum to the original peak heights $\vec{h}$:

$$\vec{h} = K^T \vec{s} \quad (1)$$

where $\vec{s}$ is a vector of the size $m$, with $s_j$ as the contribution of the $j$-th configuration to the overall occupancy. The number of configurations $m$ is smaller or equal to the number of peaks $n$. When $m = n$, the equation can be solved unambiguously. When $m < n$, Equation (1) is over-determined and can, in most cases, only be solved with a certain error. We apply the Lawson–Hanson implementation of the non-negative linear least square method (Lawson and Hanson, 1995) from R-package nnls (Mullen and van Stokkum, 2012) to solve Equation (1) (Fig. 1d). Importantly, this method prevents negative coefficients, avoiding non-reasonable negative contributions. However, the value 0 is allowed, implying that a certain configuration was dismissed. After computing the relative contribution $\vec{s}_j$ of each configuration, the deviation between original and reconstructed data can be computed (Fig. 1e). The number of configurations with a value $s_j > 0$ gives a measure for the number of possible nucleosome configurations and the dynamics within the cluster.

## 2.2 Monte Carlo simulation of nucleosome populations

The binary-variable analysis of nucleosome populations presented here yields a weighted set of possible nucleosome configurations for clusters of overlapping nucleosomes. However, for some applications, a single configuration of non-overlapping nucleosomes is needed, ignoring ambiguity from experimental data. We developed a simulation approach based on an MMC protocol combined with simulated annealing to retrieve a single optimized solution from the experimental data.

Each simulation starts with an interval of nucleosome-free DNA. During the simulation, the initially nucleosome-free configuration is altered randomly by so-called Monte Carlo moves that add, delete or slide nucleosomes (Fig. 3). Each nucleosome occupies an interval of 147 bp on the DNA and excludes the binding of other nucleosomes within its boundaries.

The MMC algorithm comprises three essential steps (Binder and Heermann, 2010):

(i) Alter the nucleosome configuration by a Monte Carlo move
(ii) Compute the energy difference $\Delta E$ between new and old configuration
(iii) Accept the new configuration with the probability $\min[1, \exp(-\Delta E/k_B T)]$; step back to (1).

We derive the energy function $E$ used in step (2) from experimental data (Fig. 4). The starting point is, as described earlier in the text, fragments of nucleosomal DNA (Fig. 4a) taken from mapped paired-end reads. In the first step, input data are converted into a frequency distribution of nucleosome centers with base pair resolution (Fig. 4b). The nucleosome centers, assumed to be in the middle of a fragment, are counted for every base pair. Because the real position of the nucleosome center is unknown, a smoothing with a Gaussian kernel is then applied to the frequency distribution of nucleosome centers. A normalization step then transforms the smoothed frequency function into a probability function for nucleosome centers with values between 0 and 1. The MMC simulation is based on Boltzmann statistics. Considering a Boltzmann ensemble, the probability and the energy of a state are connected. Because we are only interested in energy differences, we derive the energy $E_i$ from the Boltzmann factor (Padinhateeri and Marko, 2011):

$$E_i = -k_B T \ln P_i \quad (2)$$

where $P_i$ is probability of finding a nucleosome center at base pair $i$, $k_B$ the Boltzmann constant and $T$ the temperature.

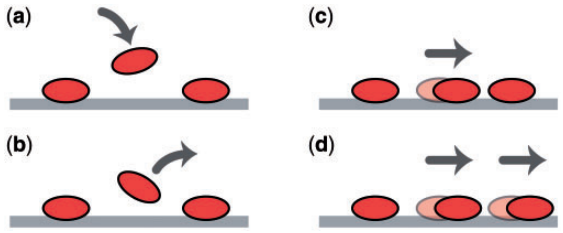


**Fig. 3.** Monte Carlo moves that are applied to alter the nucleosome configuration: (**a**) add, (**b**) delete, (**c**) slide and (**d**) pair slide

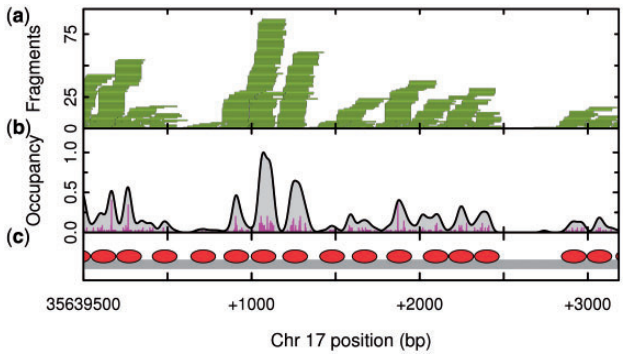


**Fig. 4.** Steps of the MMC simulation. (**a**) Intervals of nucleosomal DNA from mapped paired-end reads. (**b**) Reads are transferred into frequencies of nucleosomal centers (small spikes), smoothed with a Gaussian kernel and normalized to probabilities of nucleosome centers (solid black line). (**c**) Nucleosome (ellipses) positions after simulated annealing

The MMC criterion in step (3) ensures that genomic positions are populated with nucleosomes according to the input distribution from MNase-seq experiments (Supplementary Fig. S1). We additionally introduced a chemical potential to enforce the binding of nucleosomes:

$$E_i = -k_B T \ln P_i + \Delta N \mu \quad (3)$$

where $\Delta N$ is the change in the number of nucleosomes. The potential gives a constant energetic reward $\mu$ for every added nucleosome and an energetic penalty for every removed nucleosome. Nevertheless, genomic positions with low probability remain nucleosome-free owing to their unfavorable overall energy. The MMC algorithm produces a statistical ensemble of possible nucleosome populations. Pseudo-random numbers were generated using the Mersenne Twister algorithm (Matsumoto and Nishimura, 1998).

## 2.3 Simulated annealing for an optimized nucleosome placement

The MMC algorithm can be combined with a simulated annealing scheme to search for a single optimized solution of the nucleosome positioning problem. In this case, the simulation starts at a high temperature, causing the nucleosome population configurations to be highly dynamic. Over the simulation run, the temperature is gradually reduced with exponential decay. At low temperatures, the annealing process resembles an energy optimization because nucleosomes are allowed to move only to more favorable positions (Fig. 4c). The final nucleosome configuration of the annealing run represents the result of nucleosome positioning.

## 3 RESULTS

We analyzed experimental nucleosome maps in mouse ESCs, as well as NPCs derived from these and MEFs for the same mouse strain (Teif *et al.*, 2012). In a first step, we identified non-overlapping nucleosome configurations from the experimental data as described in Section 2.1. We weighted the contribution of each nucleosome configuration using a least squares method. Based on the weighted configurations, we determined the relative occlusion of DNA with nucleosomes and characterized the dynamics of nucleosome positioning at particular genomic features.

### 3.1 Quantifying accessibility changes at transcription factor binding sites during cell differentiation

The binary-variable analysis described in Section 2.1 yields the proportion of the overall occupancy made up from each individual nucleosome configuration. This enables quantification of the relative occlusion of DNA by nucleosomes. We focused on transcription factor binding sites in promoter and enhancer regions that were determined previously by chromatin immunoprecipitation in mouse ESCs (Chen *et al.*, 2008). For the analysis, we evaluated the center of a given transcription factor binding site and computed the percentage of weighted configurations, which occlude the binding site with a nucleosome. In contrast to the occupancy, which was normalized globally, this relative occlusion value refers to nucleosome configurations within one cluster of overlapping peaks. We found significant differences in the relative occlusion among the three cell lines (Fig. 5 and Supplementary Fig. S2). For the example shown in Figure 5, some transcription factor binding sites were found to be fully buried in a nucleosome in all configurations (Fig. 5b), whereas others were fully exposed (for example the STAT3 site, Fig. 5a). Furthermore, the accessibility of transcription factor binding sites showed large variations between the three cell types at some other sites: for example, the Tcfcp2l1 site in the enhancer region studied had a medium accessibility in ESCs and NPCs but was completely occluded in MEFs (Fig. 5b).

### 3.2 Computing stringency of nucleosome positioning to map large-scale nucleosome occupancy changes

A metric introduced by Valouev *et al.* quantifies the stringency of positioning based on the number of overlapping nucleosome positions (Valouev *et al.*, 2011). Here, we used the number of unique and incompatible nucleosome configurations as a measure of the dynamics of a particular locus. We quantified stability versus flexibility in the nucleosome positions by calculating a stringency value $S$:

$$S = 1/n_S \tag{4}$$

where $n_S$ is the number of nucleosome configurations with a value $s_i > 0$ (Section 2.1). A stringency of $S = 1$ describes a configuration with stably positioned nucleosomes, whereas a value of 0.25 means that four different nucleosome configurations coexist in a given genomic region. Thus, low stringency values indicate 'mobile nucleosome hot spots', where nucleosomes can be arranged in many different configurations. When analyzing the three different cell differentiation states in this manner, we found regions with similar stringency patterns as well as regions with significant differences (Fig. 6).

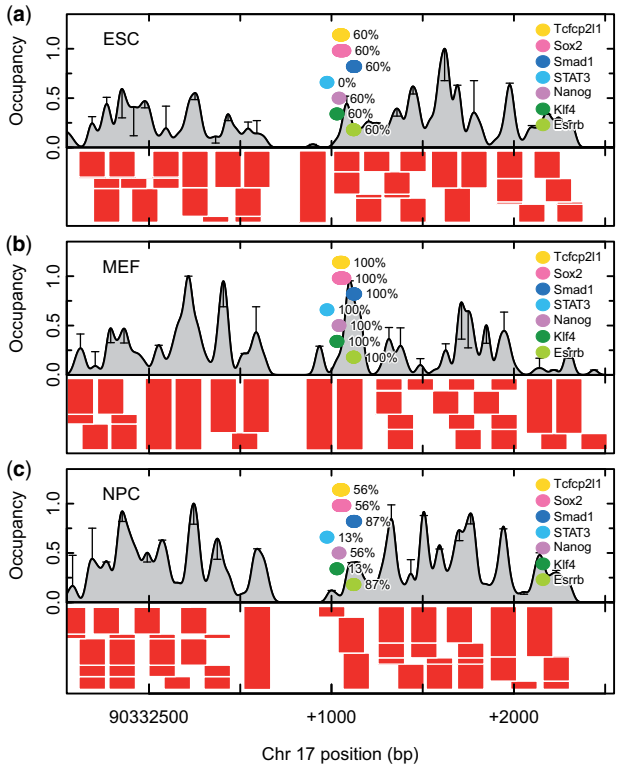


**Fig. 5.** Weighted nucleosome configurations (rectangles) for three differentiation states of mouse cells. The depicted locus contains binding sites for various transcription factors (ellipses) that are important in cell differentiation. The relative occlusion of the binding site by nucleosomes is shown as a percentage next to the transcription factor. Error bars indicate the deviation between experimental and reconstructed peak height

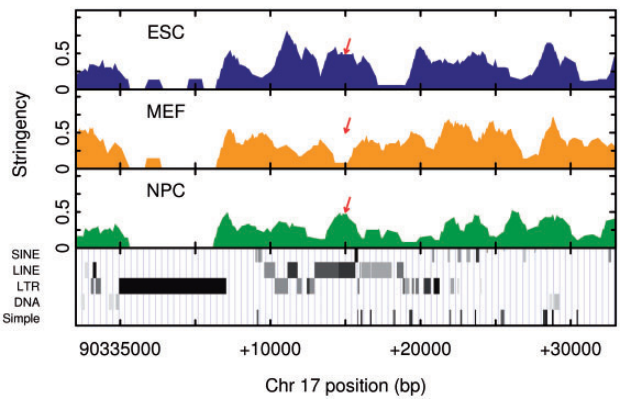


**Fig. 6.** Positioning stringency for the three differentiation states of mouse cells. Bottom panel shows the location of predicted repeats [from the UCSC genome browser (Kent *et al.*, 2002)]. Raw data were smoothed with a sliding average (window size = 1000 bp). The number of individual nucleosome configurations can vary dramatically between the cell lines. See for example the arrows in the region of LINE repeats

This analysis performed a measurement of heterogeneity regardless of its source. To subtract out variance that originates

from experimental bias, additional information is required, e.g. coverage maps from MNase-digested naked DNA or synchronized cells.

### 3.3 Applying the simulated annealing method to the Samd4 locus

The MMC approach was implemented in our open source software tool NucPosSimulator. As described in Section 2.2, it generates a dynamic nucleosome population by applying random Monte Carlo moves. The total number of positioned nucleosomes is variable and depends on the experimentally determined nucleosome occupancy, which enters into the energy function (Section 2.2). Each simulation run started with nucleosome-free DNA. Thus, in the first part of the simulation, the number of nucleosomes increased rapidly. After the equilibration phase, the frequency of adding and removing events was balanced and the number of nucleosomes fluctuated around a stable value. Nucleosomes were allowed to move around and populate all positions that were covered by reads of nucleosomal DNA. In the simulation run, the algorithm generates a statistical ensemble of possible solutions for the nucleosome positioning problem. To produce a single solution, we applied a simulated annealing scheme. During the simulations, the temperature was continuously decreased to 'freeze' the nucleosome population (Fig. 7). The annealing yielded a configuration that was optimized in terms of nucleosome density as well as fitting the experimental occupancy data. We performed a simulated annealing procedure for the nucleosomes occupancy data from the three mouse cell lines (Fig. 8) and analyzed the nucleosome maps of the *Samd4* locus, with 300 kb length including the gene locus itself and flanking DNA up- and downstream. Simulation parameters are listed in Supplementary Table S1. We found 1420, 1476 and 1513 nucleosomes for ESCs, MEFs and NPCs, respectively. Dividing by the locus length yielded a nucleosomal repeat length (NRL) of 211, 203 and 198 bp, respectively. This is slightly lower than the genome-wide global average, which yielded NRLs of 186–193 bp (Teif *et al.*, 2012). The latter is based on a region of up to ~10 nucleosomes and includes non-genic regions. Furthermore, local variations from the global average are expected so that the NRL values determined for the *Samd4* are well within the values expected from the experiments. Our approach led to a higher number of nucleosomes for a single configuration than a best-peak-first peak-calling approach. However, a high number of detected nucleosomes is only meaningful in combination with sufficient accuracy, which we validated in the next steps.

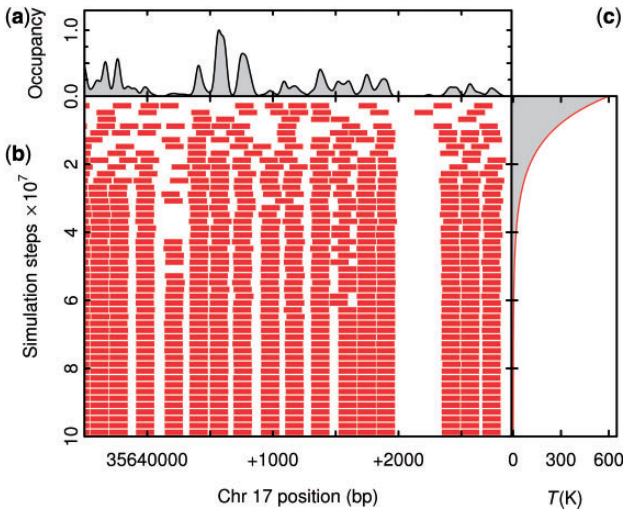


**Fig. 7.** A simulated annealing run. (**a**) The occupancy data that were used to derive the energy function. (**b**) Snapshots of the nucleosome population during a simulations run. Each row of rectangles represents one configuration. Simulation starts with naked DNA at step 0 and stops at a stable nucleosome configuration. (**c**) Annealing scheme: the temperature decreases gradually from 600 K to 0.1 K, at which point the nucleosome population freezes

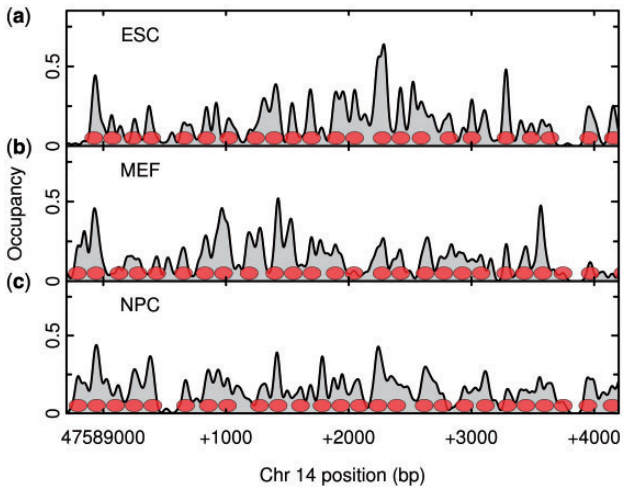


**Fig. 8.** Nucleosome positions after a simulated annealing run. Ellipses indicate nucleosome positions for (**a**) ESC, (**b**) MEF and (**c**) NPC cell lines. The nucleosomes are not always located at peak positions, but rather are located such that the overall energy of the system is optimized

### 3.4 Comparison with peak-calling approaches

As explained in the Introduction, peak-calling approaches address the problem of finding the ensemble-average nucleosome positions, which is different from the problem of finding the most probable nucleosome configuration in a single cell addressed here. Nevertheless, a comparison of our method with peak-calling approaches is instructive. As a benchmark, we compared our NucPosSimulator software with the peak-calling tool PeakPredictor from the GeneTrack project (Albert *et al.*, 2008) and the recently published positioning tools nucleR (Flores and Orozco, 2011) and NOrMAL (Polishko *et al.*, 2012). We used the *Samd4* locus as a test case and analyzed nucleosome positions from the ESC dataset (Table 1). The output of all tools is affected by their parameterization. For PeakPredictor, we used two different smoothing factors and a peak-to-peak distance of 147 bp to prevent overlapping peaks. NucPosSimulator was used with its default smoothing factor of 20 bp and a nucleosome size of 147 bp. The tool nucleR does not provide a peak prioritization strategy, and can thus return peaks from overlapping nucleosomes. However, the output contains a score for each

called peak. For comparison reasons, we used this quality value to sort the peaks and called the best peaks first while excluding less significant peaks in a neighborhood of 147 bp. The tool NOrMAL was originally developed to derive nucleosome positions from single-end reads. We transformed our paired-end data back to single-end data and performed the analysis with NOrMAL in two setups, the first allowing 35% overlap between nucleosomes and the second allowing no overlap. The analysis with NucPosSimulator yielded the largest number of positioned nucleosomes (Table 1), which is in better agreement with the experimentally determined nucleosome density (Section 3.3) than results from other tools.

In addition, we used synthetically generated nucleosome maps to assess accuracy and number of positioned nucleosomes simultaneously and compared the results with nucleR and PeakPredictor (Table 2). Synthetic nucleosome maps were sampled after addition of noise to generate more realistic data. A part of the nucleosome maps was generated with the nucleR-package (for details see Supplementary Table S2). In regularly spaced nucleosome sets with no overlap, all tools detected every nucleosome with high accuracy. In highly overlapping nucleosome configurations (e.g. the *Samd4* dataset or set 5, Table 2), the simulated annealing had clear advantages over the best-peak-first strategies, concerning the number of detected nucleosomes, while keeping a high accuracy.

Figure 9 summarizes the characteristics of the different nucleosome positioning strategies based on a synthetic nucleosome map. The binary-variable analysis recovers the source components within clusters of overlapping nucleosomes. NucPosSimulator generates an optimized non-overlapping nucleosome configuration from ambiguous experimental data. The solution is optimized to the following criteria: maximal number of nucleosomes and good compliance with the input distribution, i.e. accuracy. Best-peak-first strategies focus locally on the highest peaks but dismiss relevant peaks in the neighborhood that would contribute to an optimal overall solution.

Our simulation approach is suited for genomic regions up to several mega base pairs (Mb). In contrast to the other tools mentioned, it is currently too computationally intensive to be applied to larger genomic regions. However, in many instances it is feasible to separate large genomic areas into smaller parts (Teif and Rippe, 2012). Regions kept free of nucleosomes by, for example, the binding of CTCF (Cuddapah *et al*., 2009; Fu *et al*., 2008) act as insulators, preventing steric interactions between nucleosomes. Also, chromatin domains of ~1 Mb apparent in high-resolution microscopy and by *in situ* cross-linking are functional units (Cremer and Cremer, 2001; Müller *et al*., 2004). Thus, smaller genomic regions could be simulated independently and in parallel to reduce computation time. Furthermore, in many instances it is important to calculate the nucleosome distribution at a single enhancer/promoter.

**Table 1.** Comparison of tools for nucleosome positioning

| Positioning tool | Number of nucleosomes |
|---|---|
| NucPosSimulator (smoothing factor 20) | 1420 |
| PeakPredictor (smoothing factor 10) | 1276 |
| PeakPredictor (smoothing factor 20) | 1237 |
| nucleR (non-overlapping nucleosomes) | 1315 |
| NOrMAL (max overlap 35%) | 832 |
| NOrMAL (no overlap) | 589 |

*Note*: The number of positioned nucleosomes at the *Samd4* locus for ESCs is shown in the right column.

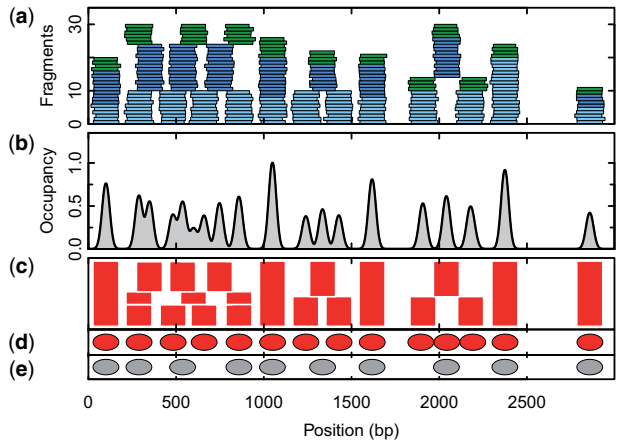


**Fig. 9.** Comparison of nucleosome positioning strategies for a synthetic nucleosome map with overlapping nucleosomes. (**a**) Short-read data comprising three different nucleosome populations. (**b**) Occupancy function derived from short-read data. (**c**) Result of a binary-variable analysis. Clusters of overlapping nucleosome peaks were dissected into their source components. (**d**) Non-overlapping nucleosome positions computed by NucPosSimulator. (**e**) Results of a best-peak-first strategy

**Table 2.** Analysis of synthetic nucleosome maps with NucPosSimulator, nucleR and PeakPredictor

| Synthetic nucleosome map (original number of nucleosomes) | NucPosSimulator | nucleR | PeakPredictor |
|---|---|---|---|
| 100 regularly spaced | 100 (0.9 bp) | 99 (1.1 bp) | 100 (0.8 bp) |
| 100 regularly spaced, 50 phase shifted | 100 (1.6 bp) | 75 (1.1 bp) | 65 (0.9 bp) |
| 150 randomly positioned | 85 (4.4 bp) | 74 (2.8 bp) | 75 (3.1 bp) |
| 100 regularly spaced, 10 removed, 50 randomly positioned | 93 (5.2 bp) | 90 (4.3 bp) | 90 (4.4 bp) |
| 100 regularly spaced, 10 removed, 100 randomly positioned | 96 (6.8 bp) | 89 (5.7 bp) | 86 (5.5 bp) |

*Note*: Columns 2–4 show the number of positioned nucleosomes and mean deviations between detected positions and original positions (value in brackets).

## 4 CONCLUSIONS

We have described a binary-variable analysis and a Monte Carlo simulation approach to analyze ambiguous nucleosome position data from MNase-seq experiments. Nucleosome maps derived from cell populations represent an overlay of nucleosome configurations from single cells that typically display some heterogeneity. The binary-variable approach analyzes the source components to dissect this heterogeneity, whereas the simulations provide an identification of distinct nucleosome configurations. By using an MC approach, we were able to derive an ensemble of nucleosome chains that can represent the experimental dataset as a sum of distinct configurations. Furthermore, we derived an optimized nucleosome placement from the MC ensemble by a simulated annealing method to derive a single non-overlapping nucleosome configuration for ambiguous MNase-seq data. Compared with previous approaches, it yields a higher number of nucleosomes, which fits better the experimentally determined nucleosome densities. The combination of these approaches offers a comprehensive view on possible nucleosome positions in cell ensembles and accounts for the incompatibility of mutually overlapping nucleosome configurations.

By applying our novel methodology to an experimental dataset, we (i) revealed different nucleosome positioning patterns between the three mouse cell types ESCs, MEFs and NPCs; (ii) derived the relative occlusion of transcription factor binding sites by nucleosomes in clusters of overlapping peaks; and (iii) introduced a stringency descriptor to quantify the flexibility of nucleosome positioning. This was applied to mobile nucleosome hot spots from regions of low stringency values that differed between the three cell lines in terms of transcription factor binding site accessibility. Thus, our method identifies biologically relevant nucleosome translocations at functional genomic elements. It can be applied to the increasing number of experimental nucleosome occupancy datasets to extract important information for the functional analysis of nucleosome positions and their translocation that cannot be obtained easily with existing techniques. In particular, we consider it as crucial to break down the ensemble average data into specific configurations that each represents the genome organization of a single cell. This is essential to further dissect nucleosome translocation changes or to evaluate the spatial folding of a given locus in meaningful manner.

## ACKNOWLEDGEMENTS

We thank Anna Sharman for critical reading of the manuscript.

*Funding*: This work was supported within project EpiGenSys by the German Federal Ministry of Education and Research (BMBF) as a partner of the ERASysBio+ initiative in the EU FP7 ERA-NET Plus program [grant number 0315712C to G.W. and 0315712A to K.R.]. Computational resources and data storage were funded by the German Research Foundation (DFG) [grant number INST 295/27-1 (to G.W.). V.T. acknowledges the support from the Heidelberg Center for Modeling and Simulation in the Biosciences (BIOMS) and a Deutsches Krebsforschungszentrum intramural grant.

*Conflict of interest*: none declared.

## REFERENCES

Albert,I. *et al.* (2008) GeneTrack—a genomic data processing and visualization framework. *Bioinformatics*, **24**, 1305–1306.

Allan,J. *et al.* (2012) Micrococcal nuclease does not substantially bias nucleosome mapping. *J. Mol. Biol.*, **417**, 152–164.

Becker,J. *et al.* (2013) NucleoFinder: a statistical approach for the detection of nucleosome positions. *Bioinformatics*, **29**, 711–716.

Binder,K. and Heermann,D.W. (2010) *Monte Carlo Simulation in Statistical Physics: An Introduction*. 5th edn. Springer, Berlin.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Chung,H.R. *et al.* (2010) The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One*, **5**, e15754.

Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.

Cuddapah,S. *et al.* (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**, 24–32.

Flores,O. and Orozco,M. (2011) nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, **27**, 2149–2150.

Fu,Y. *et al.* (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.*, **4**, e1000138.

Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Lawson,C.L. and Hanson,R.J. (1995) *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, Philadelphia.

Lee,W. *et al.* (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.

Matsumoto,M. and Nishimura,T. (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, **8**, 3–30.

Mullen,K.M. and Van Stokkum,I.H. (2012) *nnls: The Lawson–Hanson Algorithm for Non-negative Least Squares*. [R package insert] version 1.4, (NNLS). http://CRAN.R-project.org/package=nnls (31 July 2013, date last accessed).

Müller,W.G. *et al.* (2004) Generic features of tertiary chromatin structure as detected in natural chromosomes. *Mol. Cell. Biol.*, **24**, 9359–9370.

Nellore,A. *et al.* (2012) NSeq: a multithreaded Java application for finding positioned nucleosomes from sequencing data. *Front Genet.*, **3**, 320.

Padinhateeri,R. and Marko,J.F. (2011) Nucleosome positioning in a model of active chromatin remodeling enzymes. *Proc. Natl Acad. Sci. USA*, **108**, 7799–7803.

Polishko,A. *et al.* (2012) NOrMAL: accurate nucleosome positioning using a modified Gaussian mixture model. *Bioinformatics*, **28**, i242–i249.

Teif,V.B. *et al.* (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.*, **19**, 1185–1192.

Teif,V.B. and Rippe,K. (2012) Calculating transcription factor binding maps for chromatin. *Brief. Bioinform.*, **13**, 187–201.

Valouev,A. *et al.* (2008) A high-resolution, nucleosome position map of *C.elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.*, **18**, 1051–1063.

Valouev,A. *et al.* (2011) Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516–520.

Weiner,A. *et al.* (2010) High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.*, **20**, 90–100.

Yassour,M. *et al.* (2008) Nucleosome positioning from tiling microarray data. *Bioinformatics*, **24**, i139–i146.

Zhang,X. *et al.* (2012) Probabilistic inference for nucleosome positioning with MNase-Based or sonicated short-read data. *PLoS One*, **7**, e32095.

Zhang,Y. *et al.* (2008) Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics*, **9**, 537.

Zhang,Y. *et al.* (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*. *Nat. Struct. Mol. Biol.*, **16**, 847–852.

Zhang,Z. *et al.* (2011) A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science*, **332**, 977–980.

Zhang,Z. and Pugh,B.F. (2011) High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, **144**, 175–186.

# Supplemental Information: Modeling nucleosome position distributions from experimental nucleosome positioning maps

Robert Schöpflin[1], Vladimir B. Teif[2], Oliver Müller[1], Christin Weinberg[1], Karsten Rippe[2] and Gero Wedemann[1]

[1]University of Applied Sciences Stralsund, Zur Schwedenschanze 15, Stralsund 18435, Germany

[2]Deutsches Krebsforschungszentrum & BioQuant, Im Neuenheimer Feld 280, Heidelberg 69120, Germany

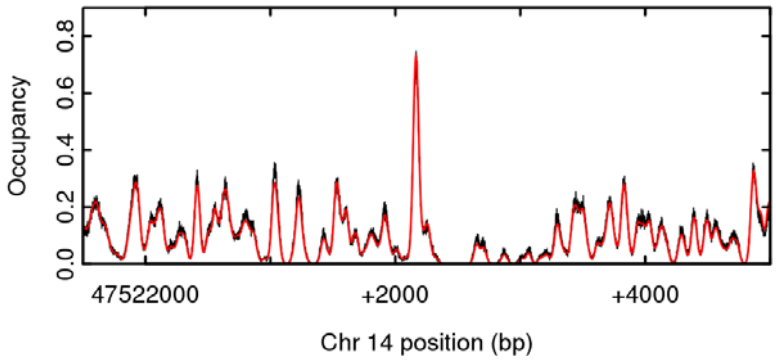


**Figure S1:** Comparison of the occupancy function from MNase-seq experiments (red curve) and simulation procedure (black curve). Simulation was performed at a constant temperature of 293 K and without a chemical potential.

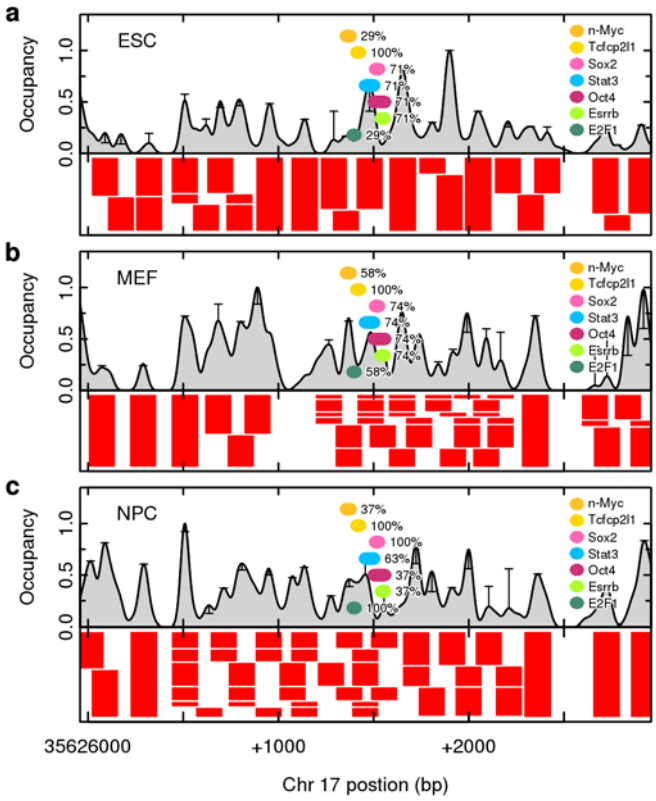


**Figure S2:** Nucleosome configurations (red filled rectangles) for three differentiation states of mouse cells, as in Fig. 4. Binding sites for different transcription factors are shown as colored ellipses. Next to the binding site the relative occlusion by nucleosomes is indicated.

| Parameter | Value |
|---|---|
| Start temperature | 600 K |
| End temperature | 0.1 K |
| Chemical potential | –8.0 |
| Nucleosome size | 147 bp |
| Simulation steps | $10^8$ |
| Annealing steps | $10^5$ |
| Smoothing (standard deviation) | 20 bp |

**Table S1:** Parameters of simulated annealing for the *Samd4* locus. A value for the chemical potential was determined empirically: we used a small test data set in which we expected to find five positioned nucleosomes at the same time, and adjusted the chemical potential iteratively such that the average number of nucleosomes in the configurations coincided with the expected value.

| Synthetic nucleosome map (original number of nucleosomes) | Parameters |
|---|---|
| 100 regularly spaced | NRL=200 bp, reads per nucleosome = 20 (± 5), read length = 147 bp (± 10), shift per read = ± 5 bp |
| 100 regularly spaced; 50 phase shifted | 100 regularly spaced nucleosomes, as above<br><br>50 additional nucleosomes were shifted by NRL/2: NRL=400 bp (± 25), reads per nucleosome = 25 (± 5), read length = 147 bp (± 10), shift per read = ± 5 bp |
| 150 randomly positioned | NRL=random, reads per nucleosome = 20 (± 5), read length = 147 bp (± 10), shift per read = ± 5 bp |
| 100 regularly spaced, 10 removed, 50 randomly positioned | nucleR-parameters: wp.num=100, wp.del=10, wp.var=20, fuz.num=50, fuz.var=50, max.cover=20, nuc.len=147, lin.len=20 |
| 100 regularly spaced 10 removed, 100 randomly positioned | nucleR-parameters: wp.num=100, wp.del=10, wp.var=20, fuz.num=100, fuz.var=50, max.cover=20, nuc.len=147, lin.len=20 |

**Table S2:** Parameters for the generation of synthetic nucleosome maps. NRL = Nucleosomal Repeat Length.