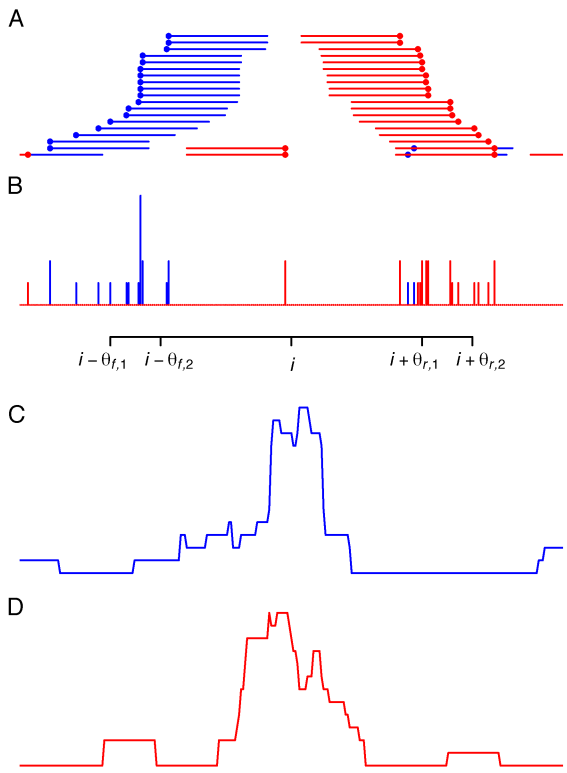
Supplementary Material

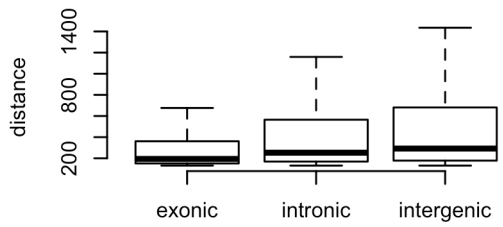
Contents:

Supplementary Figures
Supplementary Tables 1,2 and 4
Supplementary Methods

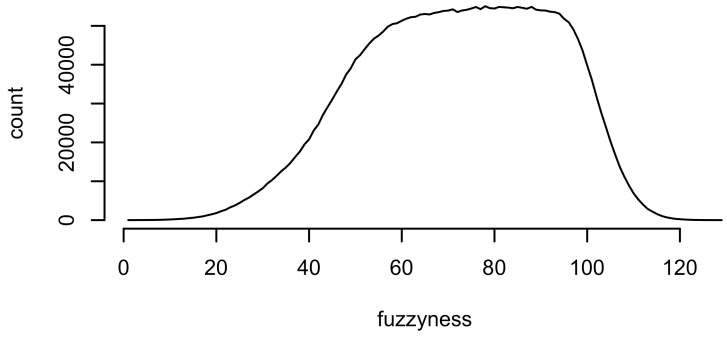
Supplementary Figures



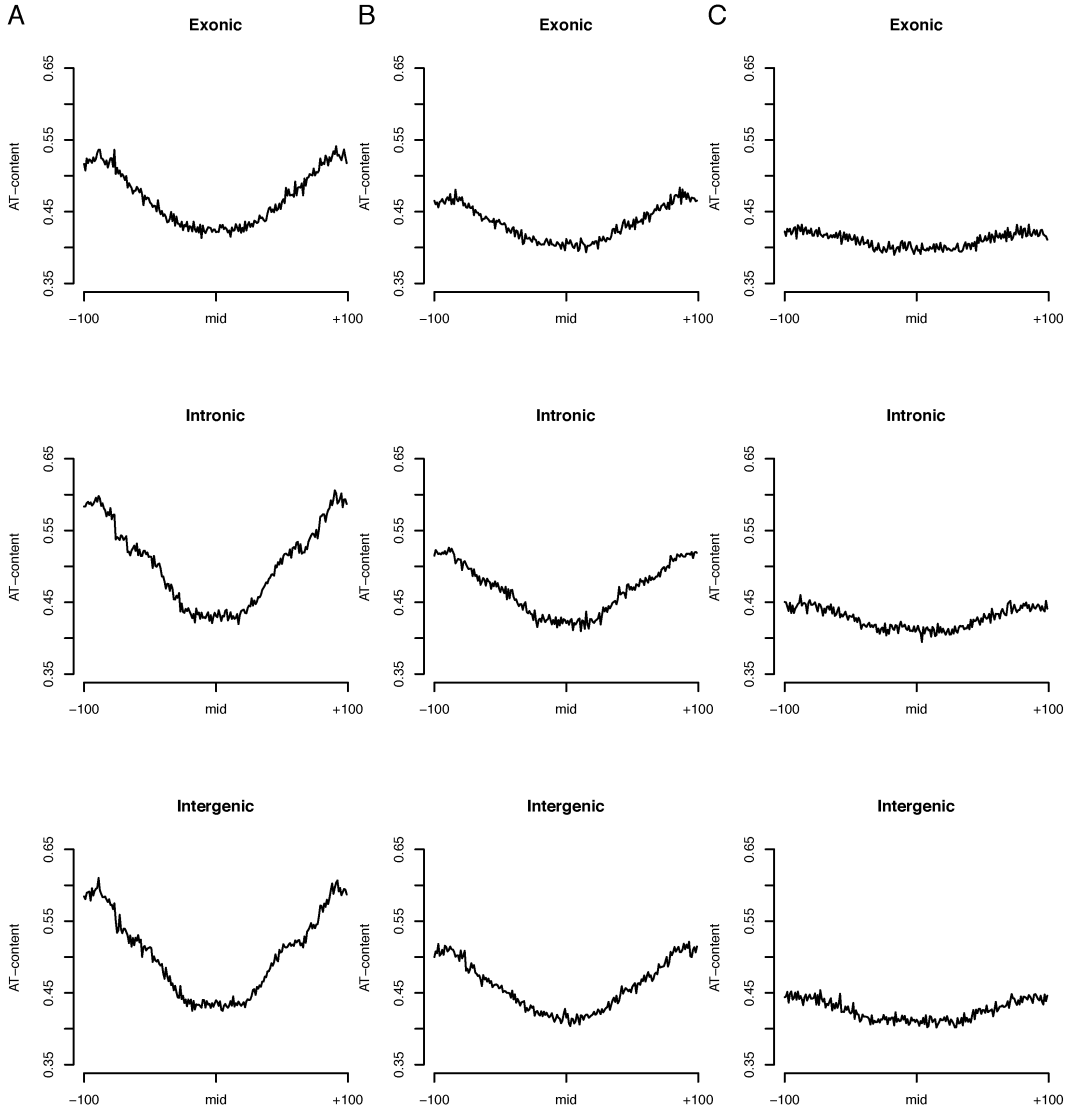
Supplementary Figure 1: The SuMMIt approach for determining support of nucleosome mid-positions. (A) Sense (blue) and antisense (red) reads and (B) the counts of their start positions per bp as indicated by the dots in (A). Sense and antisense reads supporting a nucleosome mid-position at position i are those that are located within windows at $[i\text{-}\theta_{f,1}, i\text{-}\theta_{f,2}]$ and $[i+\theta_{r,1}, i+\theta_{r,2}]$, respectively, were $\theta_{f,1}$, $\theta_{f,2}$, $\theta_{r,1}$, and $\theta_{r,2}$ are determined by the size range of sequenced DNA fragments (see below for details). Summations over sense read start positions (C) and antisense read start positions (D) over such windows flanking each bp in the genome are used by SuMMIt for modeling and prediction of nucleosome mid-positions.



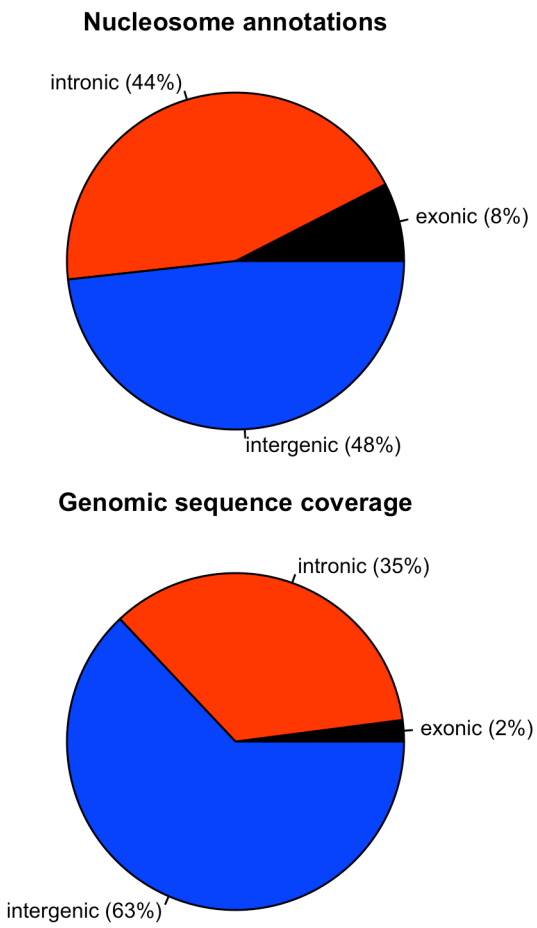
Supplementary Figure 2: Box-and-whisker plots of interregional distances between mid-positions of non-conflicting nucleosome interior regions in HepG2 TGFB- cells in exonic, intronic and intergenic regions. Boxes depict the interquartile ranges (IQR) of the data while each extreme whisker depicts an existing value no more than 1.5 times the interquartile range from the box.



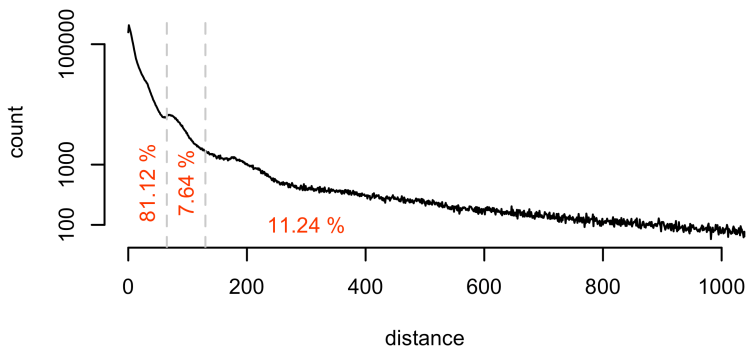
Supplementary Figure 3: Frequency of fuzzyness scores for nucleosomes in TGFB-cells.



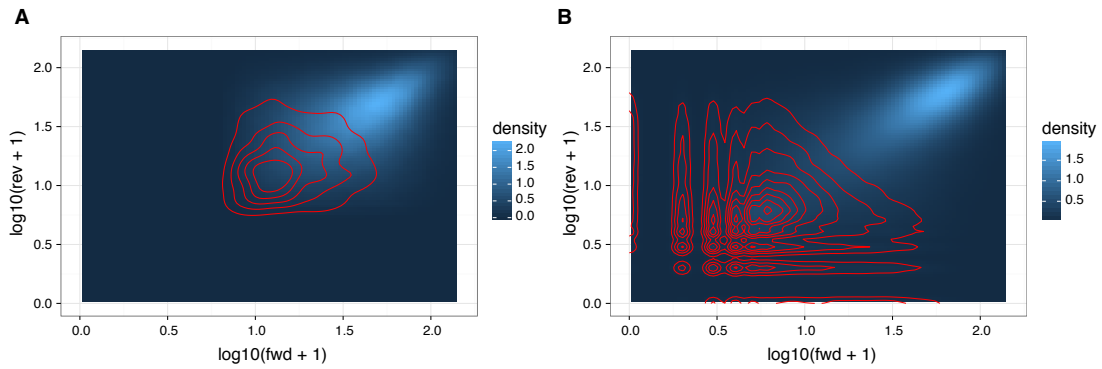
Supplementary Figure 4: Average AT-contents of DNA sequences of 201 bp in length centered at nucleosomal mid-positions in exonic (top row), intronic (middle row) and intergenic (bottom row) regions. Columns group nucleosomes into phased (A), intermediate (B) and fuzzy (C).



Supplementary Figure 5: Annotations of nucleosome interior regions in HepG2 TGFB+ cells. Top pie chart shows the distribution of nucleosomes in exonic, intronic and intergenic regions. For comparison (bottom pie chart), the genomic sequence coverages of these regions are shown.



Supplementary Figure 6: Frequency of the minimal intersample distances between nucleosome mid-positions in TGFB- and TGFB+ cells. Dashed vertical lines depict distances of 65 bp and 130 bp.



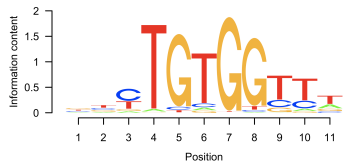
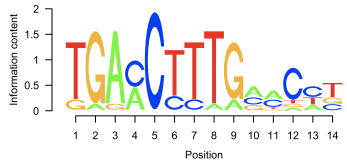
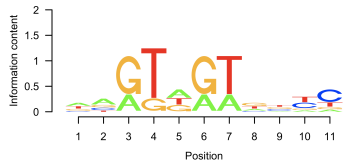
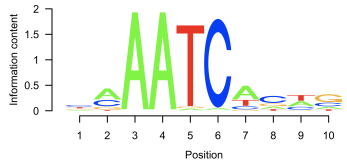
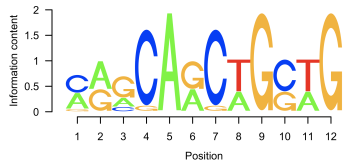
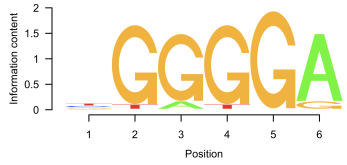
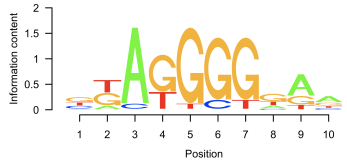
Supplementary Figure 7: (A) Density of read counts on forward (fwd) and reverse (rev) strands for SuMMIt nucleosome calls. Red 2D contour depict density of nucleosome calls made by SuMMIt that was not predicted by PING. (B) Density of read counts on forward (fwd) and reverse (rev) strands for PING nucleosome calls. Red 2D contour depict density of nucleosome calls made by PING that was not predicted by SuMMIt. The plots clearly show that nucleosome calls unique to PING have a more unbalanced ratio between forward and reverse strand tags.

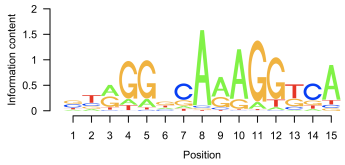
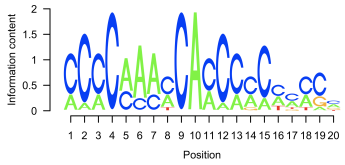
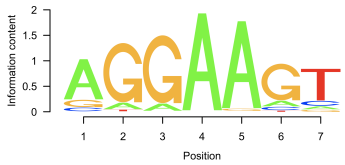
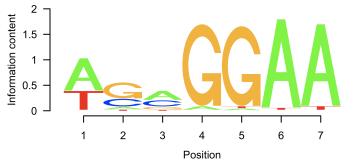
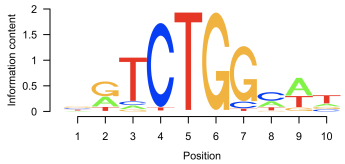
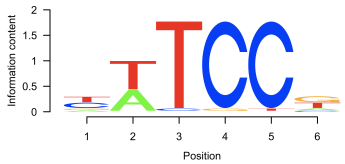
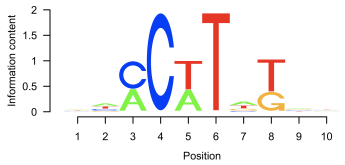
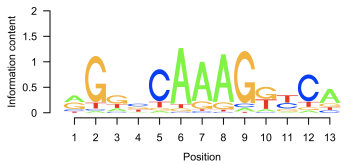
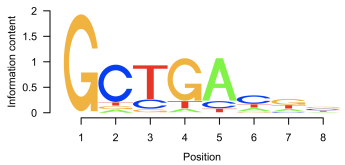
Supplementary tables

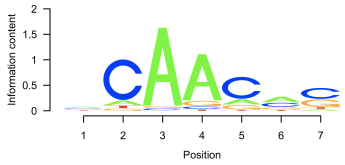
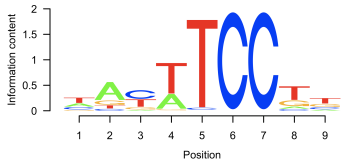
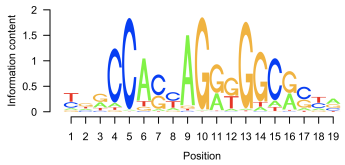
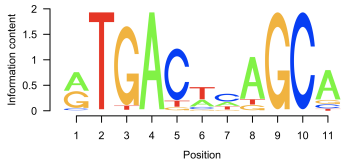
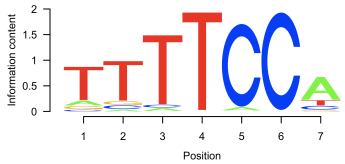
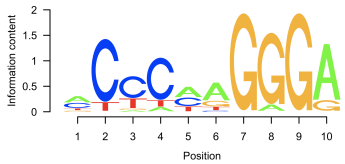
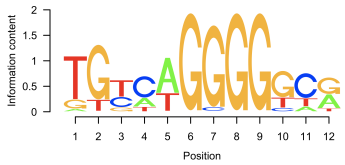
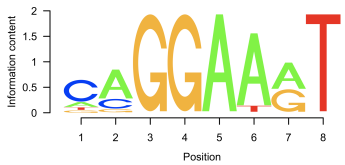
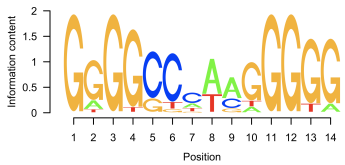
Supplementary Table 1: Summary of SOLiD read placement.

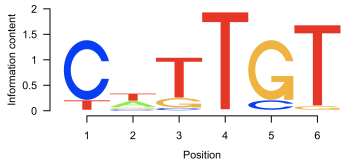
Sample	Nr. of reads	Nr. of placed reads	Nr. of uniquely placed reads	Read coverage	Nucl. coverage
TGFB-	1,321,302,648	581,068,643	396,028,207	6.4	18.8
TGFB+	1,238,104,193	457,625,710	301,461,448	4.9	14.3

Supplementary Table 2: Transcription factors with motifs that were overrepresented in sequences around loci of nucleosome depletion in TGFB+ cells.

JASPAR model	JASPAR logo
MA0002.2 RUNX1 Ig-fold	
MA0017.1 NR2F1 Zinc-coordinating	
MA0027.1 En1 Helix-Turn-Helix	
MA0038.1 Gfi Zinc-coordinating	
MA0055.1 Myf Zipper-Type	
MA0056.1 MZF1_1-4 Zinc-coordinating	
MA0057.1 MZF1_5-13 Zinc-coordinating	

MA0065.2 PPARG::RXRA Zinc-coordinating	
MA0073.1 RREB1 Zinc-coordinating	
MA0080.2 SPI1 Winged Helix-Turn-Helix	
MA0081.1 SPIB Winged Helix-Turn-Helix	
MA0092.1 Hand1::Tcfe2a Zipper-Type	
MA0098.1 ETS1 Winged Helix-Turn-Helix	
MA0109.1 Hltf Zinc-coordinating	
MA0114.1 HNF4A Zinc-coordinating	
MA0117.1 Mafb Zipper-Type	

MA0133.1 BRCA1 Other	
MA0136.1 ELF5 Winged Helix-Turn-Helix	
MA0139.1 CTCF Zinc-coordinating	
MA0150.1 NFE2L2 Zipper-Type	
MA0152.1 NFATC2 Ig-fold	
MA0154.1 EBF1 Zipper-Type	
MA0155.1 INSM1 Zinc-coordinating	
MA0156.1 FEV Winged Helix-Turn-Helix	
MA0163.1 PLAG1 Zinc-coordinating	

MA0442.1 SOX10 Other Alpha-Helix	

Supplementary Table 4. Primer locations for the HNF-alpha qPCR-validations.

Primer Name	Forward primer (5'-3')	Reverse primer (5'-3')	Primer coordinate (hg18)
Chr13:48101749	GAAGGCGTTGCAGTTTGAG	CACAGGCCCATAGTTACGC	chr13:48101648-48101749
chr11:35370495	ATTGCTGAGCTTTCGCTGAT	AGCCCTACTTTTGCTGGATG	chr11:35370400-35370522
Chr10:10273763	ACCCAGTCCGTTGTGGAG	CCCTCTCCTCGCTGCTTAG	chr10:102737131-102737234
chr9:136396847	AGCGTGGACTTTGGCATC	AAGGGGAGGGACCCACA	chr9:136396524-136396629
chr7:75487514	GGGACAAAAGTCCAGACAGC	AATCAGAACTGCACCTGTGG	chr7:75487508-75487623
chr7:28962120	GAGCCACAGGGCAGACAC	AAGTTCCACCGCTTCGTCTA	chr7:28962168-28962294
chr5:172976292	CCATCTCCGGCGTTCTTTAT	GCGCCTACAGTTCCCATATT	chr5:172976337-172976449
Chr3:188945813	TTCTCGGAATTTGAGCTTCG	GCTTTGAGGGCTTTTGGTT	chr3:188945970-188946078
chr1:35098085	CTGTGACCCTGCCCACTG	TAGCTTCCTTCCGTCCTTTG	chr1:35097963-35098043
chr1:31883148	CCAGGGGAGCTAAGTGATTG	CCCAACCTGACCACCTCTT	chr1:31883230-31883340
ISLR	TTGTTGCTGCAGAGAAGCAG	GTGCCTGGAACCATCCACT	chr15:72253126-72253226
Chr14des	TGGGATCCAACATATAGCACA	GTGGTTTGGGGTTTGAAAAA	chr14:53279280-53279379
CD36_3	CAGTCAATATTCATTAAAGGGCAGT	TCTCTCCCAGCTCCTTTGAG	chr7:79836930-79837020

Supplementary Table 5. Genomic distribution of inferred loci with nucleosomal depletion in TGFβ unstimulated cells with associated overrepresented TF binding motifs in defined categories according to distance from exons and genes.

	Exonic	**Intronic proximal**	**Intronic distal**	**Intergenic proximal**	**Intergenic distal**
Total loci	635	2185	1231	870	3674
Nr. of loci with motif	139	1012	500	293	2808
% of Total loci	21.9%	46.3%	40.6%	33.7%	76.4%
TFs	Myf	MZF1_1-4, Klf4, Prrx2, Pdx1, SPI1	Prrx2, Pdx1, NHLH1, SOX10	NFATC2, Klf4, MZF1_1-4, CTCF	FEV, Pdx1, Prrx2, NR2F1, PPARG::RXRA, Myf, NHLH1, EBF1, INSM1, SPI1, SPIB, NF-kappaB, ELF5, ETS1, MZF1_1-4, SOX10, CTCF, NFATC2, Hltf
Nr. of associated genes	133	1051	449	327	
Nr. of associated exons	177	4393			

Supplementary Table 6. Comparison of performance of nucleosome positioning methods.

	Nr. Reads	**Total Time[1]**	**Max Memory**	**Nr Predicted Features**
SuMMIt	**37.3M**	**< 1h**	**< 1Gb**	**462'977**
PING 2.0	**37.3M**	**< 15h**	**8.8 Gb**	**365'238**
NORMAL	**37.3M**	**> 408h[2]**	**~ 2 Gb**	
SuMMIt	**8M[3]**			**88'754[3]**
PING 2.0	**8M[3]**			**60'274[3]**
NORMAL	**8M**	**~ 1.3h**	**~ 2 Gb**	**42'295**

[1] For SuMMIt, this includes preprocessing of the bed-files with SICTIN. The other programs directly accept the bed-format.
[2] The run was terminated after 17 days of execution.
[3] The results were extracted from the 37.3M run.

Supplementary Table 7. Comparison of results of nucleosome positioning methods.

chr1, 37.3M		SuMMIt	PING 2.0	NORMAL
	SuMMIt		2'509	
	PING 2.0	95'545		

Supplementary methods

Modeling of nucleosome mid-positions

For positioning of nucleosomes, we considered counts of start positions for sense reads (X) and antisense reads (Y), i.e.

$$X = [x_i]_{i=1}^{N}\ ,\ Y = [y_i]_{i=1}^{N}\ ,\ \text{where } x_i, y_i \geq 0.$$

The distance between sense and antisense read start positions defining a nucleosome is expected to be 147 bp (1). Therefore, in theory, the mid-position of a nucleosome is determined by sense and antisense reads at (147-1) / 2 = 73 bp upstream and downstream defining the start and end of a nucleosome, respectively. In practice, the data is heterogeneous due to, for instance, variability of nucleosome positioning between cells, biased DNA sequence directed MNaseI cleavage and alignment problems. Hence, it is wise to consider window-counts of sense and antisense reads when defining nucleosome mid-positions. In addition, the lengths of selected fragments subjected to sequencing may be known and should therefore be considered. If the lengths of selected DNA fragments after MNaseI cleavage are within a given range, [min_d, max_d], we can assume that the start positions of matching sense and antisense reads should fall within flanking windows [i-$\theta_{f,1}$, i-$\theta_{f,2}$] and [i+$\theta_{r,1}$, i+$\theta_{r,2}$] of a nucleosome mid-position i defining the start and end of nucleosomal DNA, respectively, where

$$\theta_{f,1} = \theta_{r,2} = \left\lceil \frac{max_d - 1}{2} \right\rceil \text{ and}$$

$$\theta_{f,2} = \theta_{r,1} = \theta_{f,1} - w\ ,\ \text{where}$$

$$\mathrm{w} = \left\lceil \frac{max_d - min_d}{2} \right\rceil .$$

We define W^+ and W^- to cover window-counts of sense (X) and antisense (Y) read data, respectively, in relation to a putative nucleosome mid-position:

$$W^+ = \left[w_i^+ \right]_{i=\theta_{f,1}+1}^{N-\theta_{r,2}} = \left[\sum_{j=i-\theta_{f,1}}^{i-\theta_{f,2}} x_j \right]_{i=\theta_{f,1}+1}^{N-\theta_{r,2}},$$

$$W^- = \left[w_i^- \right]_{i=\theta_{f,1}+1}^{N-\theta_{r,2}} = \left[\sum_{j=i+\theta_{r,1}}^{i+\theta_{r,2}} y_j \right]_{i=\theta_{f,1}+1}^{N-\theta_{r,2}}.$$

Since a large proportion of windows will not represent nucleosome mid-positions, we considered W^+ and W^- to be distributed as mixtures of Poisson distributions (2), defining true interaction sites and background noise, i.e

$$W^+ \sim p_s Pois(\lambda_s) + p_{\neg s} Pois(\lambda_{\neg s}),$$

$$W^- \sim p_e Pois(\lambda_e) + p_{\neg e} Pois(\lambda_{\neg e}),$$

where λ_s and λ_e denote the parameters of Poisson distributions defining counts of reads supporting the starts and ends of nucleosomal DNA, respectively. $\lambda_{\neg s}$ and $\lambda_{\neg e}$ denote the parameters of Poisson distributions considering background noise. The p's denote the mixture proportions. We required that $p_s = p_e$, since for every start, there should be an end. Thus, $p_{\neg s} = p_{\neg e} = (1 - p_s)$. Since the total coverage of nucleosomal DNA in the genome was unknown, we used non-informative priors for the p's using Dirichlet distributions $p_s, p_{\neg s} \sim Dir(\delta_s, \delta_{\neg s})$, where $\delta_s = \delta_{\neg s} = 1$. The λ's were also unknown and needed to be estimated. Vague Gamma priors were used for the λ's:

$\lambda_s \sim Ga(\alpha_s, \beta_s), \quad \lambda_{\neg s} \sim Ga(\alpha_{\neg s}, \beta_{\neg s}),$
$\lambda_e \sim Ga(\alpha_e, \beta_e), \quad \lambda_{\neg e} \sim Ga(\alpha_{\neg e}, \beta_{\neg e})$, where
$\alpha_s = \alpha_{\neg s}$, $\alpha_e = \alpha_{\neg e}$, $\beta_s = \beta_{\neg s}$ and $\beta_e = \beta_{\neg e}$.

One characteristic of Poisson distributed data is that the variance equals the mean. This rarely happens with real-life data. To handle over-dispersion (variance greater than mean) we included appropriate measures in the hyper parameters of the Gamma distributions (2, chapter 9):

$$\alpha_1 = \alpha_s = \alpha_{\neg s} = \frac{\overline{w^+}^2}{s^2_{w^+} - \overline{w^+}^2}, \quad \beta_1 = \beta_s = \beta_{\neg s} = \frac{\alpha_s}{\overline{w^+}},$$

$$\alpha_2 = \alpha_e = \alpha_{\neg e} = \frac{\overline{w^-}^2}{s^2_{w^-} - \overline{w^-}^2}, \quad \beta_2 = \beta_e = \beta_{\neg e} = \frac{\alpha_e}{\overline{w^-}}, \text{ where}$$

$\overline{w^+}$ ($\overline{w^-}$) and $s^2_{w^+}$ ($s^2_{w^-}$) denote the mean and variance of W^+ (W^-), respectively.

The following parameters were unknown and estimated from the data through Gibbs sampling:

$$\Omega_1 = \{p_s, p_{\neg s}, \lambda_s, \lambda_{\neg s}\},$$
$$\Omega_2 = \{p_e, p_{\neg e}, \lambda_e, \lambda_{\neg e}\}.$$

In each iteration m we updated the parameters conditional on the allocations of the previous iteration (m-1) and updated the allocations conditional on the parameters of the current iteration (2, chapter 3.5.2, algorithm 3.3) (see below for details). The current implementation of SuMMIt allows for inferring one genome-wide model or multiple chromosome-wise models. In the present study, the genome-wide approach was used.

Nucleosome predictions

Having estimated the parameters of the Poisson mixtures, we predicted nucleosome mid-positions guided by log-odds (LO) of the posterior for nucleosome mid-position against background noise for W^+ and W^- data separately. A nucleosome mid-position was called whenever support were given from both sense data ($LO^+ > 0$) and antisense data ($LO^- > 0$), where

$$LO_i^+ = \log\left(\frac{p(\text{Nucleosome start flanking } i \mid W^+, \Omega_1)}{p(\text{No nucleosome start flanking } i \mid W^+, \Omega_1)}\right)$$
$$= \log\left(\frac{p(w_i^+ \mid \lambda_s) p_s}{p(w_i^+ \mid \lambda_{\neg s}) p_{\neg s}}\right),$$

$$LO_i^- = \log\left(\frac{p(\text{Nucleosome end flanking } i \mid W^-, \Omega_2)}{p(\text{No nucleosome end flanking } i \mid W^-, \Omega_2)}\right)$$
$$= \log\left(\frac{p(w_i^- \mid \lambda_e) p_e}{p(w_i^- \mid \lambda_{\neg e}) p_{\neg e}}\right).$$

Parameter estimation of Ω's via Gibbs sampling

For convenience, we formulate the Poisson mixtures in a hierarchical manner using the variable $Z = \{Z_1, Z_2\}$ representing the allocation of observations to the components:

$p(w_i^+ \mid \lambda, z_{1,i} = j) = Pois(w_i^+ \mid \lambda_j)$,

$p(w_i^- \mid \lambda, z_{2,i} = k) = Pois(w_i^- \mid \lambda_k)$,

where $p(z_{1,i} = j) = p_j$ and $p(z_{2,i} = k) = p_k$,

with $z_{1,i} \in \{s, \neg s\}$ and $z_{2,i} \in \{e, \neg e\}$.

The Gibbs sampling procedure follows.

Start with some initial allocations $Z_1^{(0)}, Z_2^{(0)}$.

1. Update of parameters $\Omega_1^{(m)}, \Omega_2^{(m)}$ (conditional on $Z_1^{(m-1)}, Z_2^{(m-1)}$):
   a. Update of the mixing proportions:

      Sample $\mathrm{p}_{\mathrm{s}}^{(\mathrm{m})}$, $\mathrm{p}_{\neg\mathrm{s}}^{(\mathrm{m})}$ from $Dir\left(\delta_s + n_{1,s}\left(Z_1^{(m-1)}\right), \delta_{\neg s} + n_{1,\neg s}\left(Z_1^{(m-1)}\right)\right)$,

      where $n_{1,j}\left(Z_1^{(m-1)}\right) = \left|\left\{i \mid z_{1,i}^{(m-1)} = j\right\}\right|$.

      Set $\mathrm{p}_{\mathrm{s}}^{(\mathrm{m})} = \dfrac{\mathrm{p}_{\mathrm{s}}^{(\mathrm{m})}}{(\mathrm{p}_{\mathrm{s}}^{(\mathrm{m})} + \mathrm{p}_{\neg\mathrm{s}}^{(\mathrm{m})})}, \mathrm{p}_{\neg\mathrm{s}}^{(\mathrm{m})} = \dfrac{\mathrm{p}_{\neg\mathrm{s}}^{(\mathrm{m})}}{(\mathrm{p}_{\mathrm{s}}^{(\mathrm{m})} + \mathrm{p}_{\neg\mathrm{s}}^{(\mathrm{m})})}$.

      Set $\mathrm{p}_{\mathrm{e}}^{(\mathrm{m})} = \mathrm{p}_{\mathrm{s}}^{(\mathrm{m})}$ and $\mathrm{p}_{\neg\mathrm{e}}^{(\mathrm{m})} = \mathrm{p}_{\neg\mathrm{s}}^{(\mathrm{m})}$.

   b. Update of the λ's:

      While $\lambda_s^{(m)} \leq \lambda_{\neg s}^{(m)}$:

      Sample $\lambda_s^{(m)}$ from $Ga\left(\alpha_1 + \sum_{i \mid z_{1,i}^{(m-1)} = s} w_i^+, \beta_1 + n_{1,s}\left(Z_1^{(m-1)}\right)\right)$,

      Sample $\lambda_{\neg s}^{(m)}$ from $Ga\left(\alpha_1 + \sum_{i \mid z_{1,i}^{(m-1)} = \neg s} w_i^+, \beta_1 + n_{1,\neg s}\left(Z_1^{(m-1)}\right)\right)$

      Similarly for $\lambda_e^{(m)}$ and $\lambda_{\neg e}^{(m)}$.

2. Update of the allocations $Z_1^{(m)}, Z_2^{(m)}$ (conditional on $\Omega_1^{(m)}, \Omega_2^{(m)}$):

   Sample $z_{1,i}^{(m)}, z_{2,i}^{(m)}$ independently for each i from the conditional posterior distributions $\mathrm{p}(z_{1,i}^{(m)} \mid \Omega_1^{(m)}, w_i^+), \mathrm{p}(z_{2,i}^{(m)} \mid \Omega_2^{(m)}, w_i^-)$:

   $\mathrm{p}(z_{1,i}^{(m)} = j \mid \Omega_1^{(m)}, w_i^+) \propto Pois(w_i^+ \mid \lambda_j^{(m)}) p_j^{(m)}$,

   $\mathrm{p}(z_{2,i}^{(m)} = k \mid \Omega_2^{(m)}, w_i^-) \propto Pois(w_i^- \mid \lambda_k^{(m)}) p_k^{(m)}$.

Supplementary references

1. Richmond, T.J. and Davey, C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145-150.
2. Frühwirth-Schnatter, S. (2006), *Springer Series in Statistics,*. Springer Science + Business Media, LLC, New York, NY.