Supplemental Methods

An Assessment of Prediction Algorithms for Nucleosome Positioning

YOSHIKI TANAKA\textsuperscript{1,2}  
ytanaka@hgc.jp

KENTA NAKAI\textsuperscript{1,2,3}  
knakai@ims.u-tokyo.ac.jp

\textsuperscript{1} Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
\textsuperscript{2} Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
\textsuperscript{3} Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency, 5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan

Detection of nucleosome locations

\textbf{Supplemental figure 1. A detection method of nucleosome locations.} (A) The gradients of signals. (B) The diagram of HMM. (C) The emission probability of each hidden state. A linker state is represented by pink color. The 5' end and the 3' end nucleosome states are represented by green and waterblue color, respectively.
To define the positions of nucleosomes from the distribution of nucleosome DNA tags in the genome, we focused on the gradient of signals and developed Hidden Markov model (HMM) like Yuan et al. [1]. At first, 73bp was added to each position of the 5’ end of mapped tags for representing the position of corresponding nucleosome centers. For each nucleosome center $i$, the counts of corresponding tags $c(i)$ were recorded. Then, at each discrete position $I_j (I_j = 5 + 10j, j = 0, 1, 2, \ldots)$ in the genome, the nucleosome signal $s(I_j)$ is calculated with the following formula:

$$
s(I_j) = \log_2 \left( \frac{\sum_{i=I_j-73}^{I_j+73} w(i,I_j)c(i)}{\sum_{i=I_j-73}^{I_j+73} w(i,I_j)} \right) + 1
$$

where $w(i, I_j)$ is the Gaussian distribution with mean $I_j$ and standard deviation 20 as introduced in Albert et al. [2]. The nucleosome signals were further normalized to the Z-scores. In the tiling array data, we used the $\log_2$ (intensity) as the nucleosome signal.

Next, for each discrete position $L$ ($L = 10, 20, 30, \ldots$), the minimum range that encompasses it is defined as follows:

$$
I_m < L - 5 < I_{m+1} \text{ and } I_{n-1} < L + 5 < I_n
$$

Note that the above formulae are applicable to both of the data. It is possible that $m = n - 1$ where the tiling array is sparse.

Then, we define the gradient (more accurately, the sine) $g(L)$ as:

$$
g(L) = \frac{10(s(I_n) - s(I_m))}{\sqrt{10^2 (s(I_n) - s(I_m))^2 + (I_n - I_m)^2}}
$$

g(L) takes the values from -1 to +1. The value “10” is a scaling coefficient. In the linker region, the gradient signals are almost zero, whereas the signals are positive/negative on the 5’/3’ side of nucleosome region, respectively (Supplemental fig. 1A).

Our HMM contains three types of hidden states, each of which outputs the gradient signal value: one linker node (L), seven nucleosome nodes on the 5’ side (5N1-5N7), and seven nucleosome nodes on the 3’ side (3N1-3N7) (Supplemental fig. 1B). In addition to one self-connecting loop on the L node for allowing various lengths of linker regions, two self-loops were added to the 5N7 and 3N7 nodes for the detection of wider peaks. The emission probability function in each node is represented by a Gaussian distribution $N(\mu, \sigma)$, where $\mu$ and $\sigma$ are parameters that take the same value in each type of all states (Supplemental fig. 1C). All model parameters were estimated from a sliding window of 100 consecutive positions by Baum-Welch algorithm using the gradient signal. The averaged parameters from all windows were used for the estimation of hidden states by Viterbi algorithm.

**Supplemental References**
