Valouev et al Supplemental Information

A high-resolution, nucleosome position map of *C. elegans* reveals a lack of sequence-dictated positioning

SUPPLEMENTAL DATA AND SUPPLEMENTAL FIGURES AND LEGENDS

Oligonucleotide enrichment extrapolation.

We have investigated whether oligonucleotide frequencies inferred from genomic positions of nucleosomal cores can be predicted using frequencies of lower length oligonucleotides from the same dataset. Consider a k-mer $X_{i-k}...X_i$, that starts at position $i-k$ relative to nucleosomal start where X_j is a symbol from a nucleotide alphabet (A,C,G,T). We construct our measure of prediction quality as

$$r_{k,i} = \frac{P(X_{i-k}...X_i)}{E(X_{i-k}...X_i)},$$

where $P(X_{i-k}...X_i)$ is the frequency of oligonucleotide $X_{i-k}...X_i$ and $E(X_{i-k}...X_i)$ is the extrapolated frequency of $X_{i-k}...X_i$ inferred from lower length oligonucleotide frequencies. Under the assumptions of markovian transitions, our measure can be rewritten as

$$r_{k,i} = \frac{P(X_{i-k}...X_{i-1} \to X_i)}{P(X_{i-k+1}...X_{i-1} \to X_i)},$$

where $P(X_{i-k}...X_{i-1} \to T_i) = P(X_i \mid X_{i-k}...X_{i-1})$ is the markovian transition probability, which can be estimated using the following formula:

$$P(X_i \mid X_{i-k}...X_{i-1}) = \frac{N_{X_{i-k}...X_i}}{N_{X_{i-k}...x_{i-1}}}.$$

Here, $N_{X_{i-k}...X_i}$ is the number of times the oligonucleotide $X_{i-k}...X_i$ occurs in a dataset. Using this calculation, $r_{k,i}$ also captures nontrivial dependence present in *C. elegans* genome that is not specific to nucleosome positioning. To address this, it is important to use a corrected measure of the following form:

$$h_{k,i} = \log(r_{k,i} / c_{k,i}),$$

where $c_{k,i}$ is calculated in exactly the same way as $r_{k,i}$ except from a dataset obtained by sequencing randomly sheared genomic DNA from *C. elegans*. By plotting $h_{k,i}$, we have obtained heat maps for $k = \{2,3,4,5\}$, and can conclude strongly that 2-mers contain additional information compared to 1-mers (Satchwell et al. 1986), and that additional information is contained in 3-mers and 4-mers.

2mer

	Green (<= -0.05)	Red (>=0.05)
core		
linker	TT, CG	TA

3mer

	Green (<=-0.05)	Red (>= 0.05)
core	CAC, CTC, GAG, GTG, CCC, GGG, GCG, CGC	
linker	CAG	AAG, CCT, CTT

4mer

	Green (<=0.05)	Red (>=0.05)
core		

linker	GGGG, GTAC, GTAG	GTAA, GGGC

Supplemental Figure Legends

Supplemental Figure S1. Histogram of mononucleosome core lengths as assayed by Sanger sequencing.

Supplemental Figure S2. Example of positioning stringency calculation. Blue (from forward nucleosome reads) and green (from reverse nucleosome reads) arrows depict putative dyads for nucleosomes, overlaid on the custom browser tracks (unlike Figure 3 in the main text, each nucleosome read is given its own feature regardless of the number of reads starting at the same genomic position) . Dyads falling inside the 23bp window (marked by a red-dotted rectangle) are designated as positioned nucleosome instances. Dyads falling outside the 23bp window but inside the 301bp window (marked by a black- dotted rectangle) represent nucleosomes that are discordant with the putative positioned nucleosome. The red line represents a putative positioned nucleosome footprint with the dyad indicated by the black asterisk. In this example the stringency value is 51.2% = 100% x 35/68, resulting from having 35 dyads within the red-dotted rectangle and 68 dyads within the black-dotted rectangle.

Supplemental Figure S3. Comparison of nucleosome positioning stringency graphs for two dyad window sizes. The 23bp-window data is identical to Figure 5 in the main text with blue representing nucleosome, red representing control and green representing the difference between the two. The 11bp-window graph was generated by reducing the dyad window from 23 to 11bps, demonstrating concordance with the 23bp results.

For the 11bp-window data, light blue represents nucleosome-derived data, magenta represents control data and light green depicting the difference. The inset graph displays the same data from the region of 30% to 60% stringency.

Supplemental Figure S4. 5-mer enrichment for 1-pile data. Displayed as described in Figure 7C (main text).

Supplemental Figure S5. 5-mer enrichment for 5-pile data. Displayed as described in Figure 7D (main text).

Supplemental Figure S6. Alternate representation of over and under representation of nucleotide words using a wider range of colors. Positions are represented relative to the nucleosome start along the horizontal axis and k-mers are represented vertically. Each point of the heatmap represents the degree of over or under representation of the specified k-mer that starts at that position along the 998 base pair sequence. Degree of over or under representation is depicted, from most under represented to most over represented, as white, light blue, dark blue, light green dark green, yellow, orange, and red. Over and under representation is determined by calculating the relative frequency of each k-mer at each position, normalized by the total occurrences of each k-mer across the entire individual data set. Displayed graphically, each color represents a standard deviation from the mean frequency of all k-mers in the data set (e.g., light green represents a range of 0 to (-1) standard deviations below the mean and dark green correspond to 0 to (+1) standard deviations above the mean). Each possible k-mer is denoted in the key at the left of the heat map, A as green, C as blue, G as yellow, and T as red.

Supplemental Figure S7. Oligonucleotide enrichment extrapolation. The heatmaps show oligonucleotide enrichment extrapolation from lower length oligonucleotides, more precisely plotting $h_{k,i}$ as a heatmap for k=2,3,4. Panel 2-to-1 shows comparison of dinucleotide frequencies to extrapolated dinucleotide frequencies from mononucleotides in the 3-pile mononucleosome dataset, panel 3-to-2 shows comparison of trinucleotide frequencies to extrapolated trinucleotide frequencies from dinucleotides, and panel 4-to-3 shows comparison of tetranucleotide frequencies to extrapolated tetranucleotide frequencies from trinuclotides as observed in 3-pile mononucleosome dataset. Each panel shows mononucleosomal core fragments with some of the adjacent sequence. Horizontal position 0 represents the start of a nucleosomal fragment as inferred from the start of the sequenced read such that +20, +40, … correspond to 20, 40, … nucleotides into the mononucleosomal core. Each of the 3 panels has nucleotide positions along the horizontal axis and oligonucleotides ordered alphabetically along the vertical axis. The panels contain 16, 64, and 256 rows respectively that correspond to all possible oligonucleotides of length 2, 3, and 4. Color intensities represent how over or underpredicted a particular oligonucleotide frequency is relative to its predicated frequency at a given position in or around the nucleosome core. Red color corresponds to underpredicted oligonucleotide frequencies, while green color corresponds to overpredicted oligonucleotide frequency. The brightest shade of red corresponds to $h_{k,i} = \log(1.15)$ while the brightest shade of green corresponds to $h_{k,i} = \log(1/1.15)$, black color corresponds to a situation when predicted enrichment is the same as the actual enrichment observed in the data, i.e. $h_{k,i} = \log(1.0)$

References

Satchwell, S.C., H.R. Drew, and A.A. Travers. 1986. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology* **191:** 659-675.