Genome Biology

OPINION

Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology

Noam Kaplan[1], Timothy R Hughes[2,3], Jason D Lieb[4], Jonathan Widom[5*] and Eran Segal[1,6*]

Abstract

We propose definitions and procedures for comparing nucleosome maps and discuss current agreement and disagreement on the effect of histone sequence preferences on nucleosome organization *in vivo*.

The organization of nucleosomes in living cells is non-random and conserved across similar cells [1], and it affects several processes, most notably transcription [2-5]. It is therefore important to understand what factors govern the organization of nucleosomes on DNA. Given that transcription changes dynamically across different cellular states, one would also expect nucleosome organization to be dynamic and governed, at least in part, by dynamic factors. However, because histones have different affinities for different DNA sequences [6-8], one might also expect the static DNA sequence to have a role in determining the organization of nucleosomes. Clearly, given the static nature of both nucleosome sequence preferences and the DNA sequence, these two factors cannot be the only determinants of *in vivo* nucleosome organization. However, the magnitude of the effect of histone DNA sequence preferences on nucleosome organization *in vivo* could, in principle, range from negligible to highly significant.

In recent years, the DNA sequence preferences of nucleosomes and their contribution to *in vivo* nucleosome organization have received much attention. A major difficulty in addressing this question is that current experimental methods cannot directly measure nucleosome organization but rather only certain aspects of it, averaged over a cell population. Another issue is that, despite intensive research, the terminology and analysis methods used in the field vary, leading to ambiguity and confusion. We believe that this has created an incorrect appearance of a major controversy in the field, with seemingly contradictory paper titles such as 'A genomic code for nucleosome positioning' [9], 'The DNA-encoded nucleosome organization of a eukaryotic genome' [10], 'A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning' [11], and 'Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*' [12]. However, these works largely agree with each other both on the various experimental measurements and on most of the conceptual conclusions. Although many scientific debates and interesting questions are still open, we believe it is generally agreed that histone sequence preferences have a central role in nucleosome organization *in vivo*, and our view is that much of the remaining debate revolves around semantic and quantitative issues rather than conceptual differences.

Here, we attempt to organize clearly the various terms, measures, experimental issues, and results in the field and to state which results are relatively established and which questions remain open. Specifically, we propose definitions for nucleosome position, nucleosome configuration, nucleosome organization, nucleosome occupancy and nucleosome positioning; we discuss how the various quantities are measured experimentally and estimated; we discuss aspects of how nucleosome maps can be compared; and finally, we discuss the effect of histone sequence preferences on nucleosome organization *in vivo*, summarizing current evidence, what is generally agreed on and what is not.

Definitions

Nucleosome position

A nucleosome position consists of 147 consecutive base pairs that are wrapped around a nucleosome, assuming no partial wrapping of DNA. A nucleosome position can be specified by the nucleosome start (the first of the

*Correspondence: j-widom@northwestern.edu; eran.segal@weizmann.ac.il
[5]Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, 2153 Sheridan Road, Evanston, IL 60208, USA
[6]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel
Full list of author information is available at the end of the article

147 base pairs, that is, the base pair with the lowest coordinate), center (the 74th base pair) or end (the 147th base pair).

Nucleosome configuration

A nucleosome configuration is a set of non-overlapping nucleosome positions on a single DNA molecule of defined length. The requirement for non-overlapping positions is motivated by steric exclusion, which does not allow a DNA base pair to be simultaneously wrapped around more than one nucleosome. Thus, a nucleosome configuration can be represented by a binary vector that, for each base pair, specifies whether a nucleosome starts at that base pair (assigned '1'), with the steric hindrance constraint such that if a base pair is in state 1, then both the preceding and following 146 base pairs (bp) must be '0'. Formally:

$$c \in \{0,1\}^N \text{ s.t. } \forall i{:}c_i = 1 \Rightarrow c_{i-146}, \ldots, c_{i-1}, c_{i+1}, \ldots, c_{i+146} = 0$$

where c is the nucleosome configuration, N is the length of the DNA molecule, c_i is the ith coordinate of c and $c_i = 1$ represents a nucleosome starting at base pair i. Example nucleosome configurations on single DNA molecules are shown in Figure 1a.

Nucleosome organization

We define a nucleosome organization as a probability distribution over nucleosome configurations, that is, as a set of nucleosome configurations in which each configuration is assigned a probability and the sum over the set of configurations is 1. Formally:

$$P : C \rightarrow R \text{ s.t. } \forall c \in C : P(c) > 0 \text{ and } \sum_{c \in C} P(c) = 1$$

where P is the nucleosome organization, C is a set of nucleosome configurations, and $P(c)$ represents the probability of configuration c. Thus, a nucleosome organization can specify a complete description of all nucleosome configurations on a DNA sequence across an isogenic cell population. Figure 1a illustrates this concept.

Nucleosome occupancy

We define the nucleosome occupancy of a base pair as the sum of the probabilities of the configurations in which the base pair is covered by a nucleosome. Formally:

$$Occ(x) = \sum_{c \in C} \sum_{i=x-146}^{x} P(c)c_i$$

where $Occ(x)$ is the occupancy of base pair x, P is the nucleosome organization and C is the set of nucleosome configurations in P. Thus, a base pair covered by a nucleosome in all configurations will have 100% occupancy and a base pair that is not covered in any of the configurations will have 0% occupancy. The nucleosome occupancy of a base pair has important functional implications because it reflects how accessible the base pair is. Figure 1b shows the nucleosome occupancy that would result from the example nucleosome organization in Figure 1a. Note, however, that different nucleosome organizations can result in the same nucleosome occupancy.

Nucleosome positioning

Nucleosome positioning is a commonly used term but its exact meaning is often left vague or undefined. Typically, it attempts to quantify the degree to which the positions of individual nucleosomes vary across the different configurations of a nucleosome organization. It is generally agreed that a perfectly positioned nucleosome is one that adopts the same position across all measured configurations. However, unlike nucleosome occupancy, which describes a physical quantity and is thus intuitive, the meaning of '30% positioning', for instance, is typically unclear. We thus define two kinds of nucleosome positioning that have a physical interpretation and that relate to quantities that were previously suggested [11,12]: absolute and conditional.

Absolute nucleosome positioning

We define the absolute nucleosome positioning at base pair x as the probability of a nucleosome starting at base pair x, equal to the sum of probabilities of the configurations in which a nucleosome starts at base pair x (Figure 1c). Formally:

$$Pa(x) = \sum_{c \in C} P(c)c_x$$

where $Pa(x)$ is the absolute nucleosome positioning at base pair x, P is the nucleosome organization and C is the set of nucleosome configurations in P. Notably, absolute positioning uniquely determines nucleosome occupancy, and different landscapes of absolute positioning can yield identical occupancy at a given position. However, absolute positioning does not uniquely determine the nucleosome organization, because information regarding the individual nucleosome configurations is not retained.

Conditional nucleosome positioning

Some investigators focus on a different positioning metric, which asks about the absolute positioning at base pair x divided by the probability that a nucleosome starts anywhere within a larger (for example, nucleosome-length) region centered on x [11,12]. For definiteness we therefore also define the conditional nucleosome positioning at base pair x as the probability of a nucleosome starting at x (absolute nucleosome positioning) divided by the sum of probabilities of the configurations

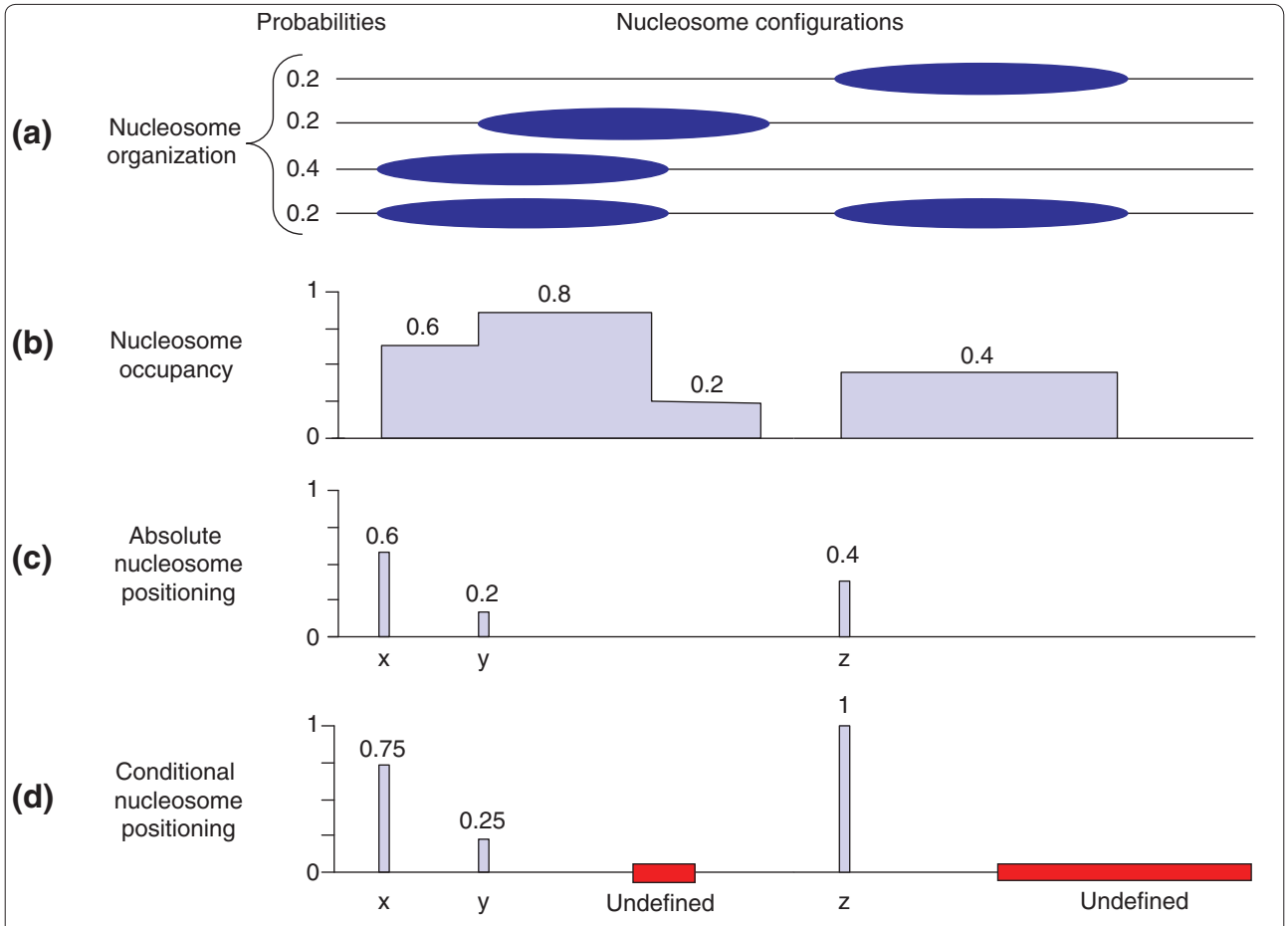

Figure 1. Illustration of the proposed definitions. (a) Four nucleosome configurations, which, together with their respective probabilities, constitute a nucleosome organization. **(b-d)** The derived (b) nucleosome occupancy, (c) absolute positioning and (d) conditional positioning measures. The configuration with a probability of 0.4 is weighed twice as heavily as the other configurations in the derived occupancy (b) and positioning (c,d) measures. Note how the rightmost nucleosome that appears in the same position in two of the four configurations has a relatively low absolute positioning value (0.4 in (c)) but a high conditional positioning value (1 in (d)), whereas the leftmost nucleosome is relatively well positioned by both the absolute and conditional positioning measures (0.6 in (c) and 0.75 in (d)). Also note that owing to the existence of another nucleosome close to the leftmost nucleosome, the rightmost nucleosome has a higher conditional positioning value than the leftmost nucleosome (d), even though the rightmost nucleosome has an overall lower probability across all four nucleosome configurations. Red boxes in (d) represent regions in which the conditional positioning is undefined.

in which a nucleosome starts at the 147 base pairs centered on base pair x (Figure 1d). Formally:

$$Pc(x) = \frac{\sum_{c \in C} P(c) c_x}{\sum_{i=x-73}^{x+73} \sum_{c \in C} P(c) c_i} = \frac{Pa(x)}{\sum_{i=x-73}^{x+73} Pa(i)}$$

where $Pc(x)$ is the conditional nucleosome positioning at base pair x, $Pa(x)$ is the absolute positioning at x, P is the nucleosome organization and C is the set of nucleosome configurations in P. Thus, the conditional nucleosome positioning at base pair x is the probability that a nucleosome starts at x given that a nucleosome starts somewhere between x - 73 and x + 73. Conditional positioning at x is undefined if the absolute positioning of all 147 base pairs around x is zero.

Absolute positioning versus conditional positioning

To illustrate the difference between absolute and conditional positioning, consider the example shown in Figure 1a. This shows a base pair x with absolute positioning 0.6, and another base pair y with absolute positioning 0.2, which is proximal to x (that is, y is within the 147-bp window around x). Because a nucleosome starts at x in a large fraction (60%) of the configurations, we might consider it to be well positioned. Base pairs proximal to x are not explicitly considered in computing the absolute positioning at x, although the values of x and y are not independent (because the occupancy is limited to 1). In the case of base pair x, although its conditional nucleosome positioning is also high, computed as $Pc(x)$ = 0.6/(0.6 + 0.2) = 0.75, the conditional nucleosome positioning value offers a different interpretation: the nucleosome

at x is well positioned because, across all configurations in which a nucleosome appears in the window around x, it starts at x in 0.75 of the cases.

This difference in interpretation may also result in very different values for absolute and conditional positioning, as is the case of base pair z shown in the example, which has absolute positioning 0.4 but conditional positioning of 1 (because no other nucleosome is positioned in its vicinity). Base pair z demonstrates the dependence of conditional positioning on absolute positioning in its vicinity, because a base pair with absolute positioning of 0.4 can have conditional positioning anywhere within the range from 1 (if no nucleosome is in its vicinity) down to 0.4 (if high-probability nucleosomes are in its vicinity). Absolute positioning provides a lower bound on conditional positioning because, at most, one nucleosome can start in a 147-bp window:

$$Pc(x) = \frac{Pa(x)}{\sum_{i=x-73}^{x+73} Pa(i)} \geq \frac{Pa(x)}{1}$$

Thus, high absolute positioning necessarily means high conditional positioning, but the converse is not implied. Similarity in absolute positioning of two nucleosome maps implies similarity in nucleosome organizations. Alternatively, two nucleosome maps may have low similarity in absolute positioning but high similarity in conditional positioning, suggesting that whereas the fraction of configurations in which a nucleosome appears differs greatly between the maps, the position of the nucleosome when it does appear is similar between the maps. These important differences between absolute and conditional positioning suggest that they could each be useful for addressing different biological questions.

Experimental measurement and estimation of nucleosome organization

Because nucleosome organization is a probability distribution over nucleosome configurations, ideally one would like to estimate it by measuring the nucleosome configuration of a single cell and then repeat this measurement for many cells. Unfortunately, such measurements are not currently possible. Instead, existing methods sample nucleosome positions from the entire nucleosome organization, in which each nucleosome position measured can come from a different cell in the population. Although such methods do not directly measure the nucleosome organization, they do allow us to estimate occupancy and positioning.

Experimental technology

A popular method for nucleosome mapping is digestion of chromatin by micrococcal nuclease (MNase), an endonuclease that preferentially cuts linker DNA rather than DNA wrapped around a nucleosome. Thus, DNA that is highly digested is relatively depleted of nucleosomes, and loci that are under-digested are relatively protected by nucleosomes. The resulting digestion pattern can then be measured by methods such as primer extension and real-time PCR with gel electrophoresis or low-throughput sequencing of nucleosome-protected DNA segments. More recently, high-throughput technologies were used to measure nucleosome positions on a genome-wide scale, first using DNA microarrays [13-15,16] and then using deep sequencing [10,11,15,17-24]. Importantly, deep sequencing can potentially provide measurements of many individual nucleosome positions, whereas microarrays have lower resolution and can only provide measurements of nucleosome occupancy.

Experimental biases

Genome-wide nucleosome mapping experiments have two main steps, DNA isolation and DNA measurement, and both can introduce noise and biases. The DNA isolation step typically includes MNase digestion followed by extraction of the approximately 147-bp mononucleosome DNA band that results. One bias of this step arises from the sequence specificity of MNase, because MNase has a preference to having a TA/AT dinucleotide as its cleavage site [18,25]. The appearance of a discrete mononucleosome band after MNase digestion shows that nucleosome protection, not MNase specificity, is the dominant factor in the digestion. In addition, because the specificity of MNase is low, a preferred cleavage site is found frequently. Thus, biases arising from MNase specificity will mostly result in imprecise mapping of nucleosome ends, but the extracted mononucleosomes still correspond to nucleosome-bound DNA. Nevertheless, the use of MNase limits the accuracy of the resulting nucleosome maps, which are certainly not at single-base-pair resolution.

Another bias introduced by MNase digestion arises from the length variability of the extracted mononucleosomes. Studies that fully sequenced the extracted mononucleosomes showed that their lengths vary by tens of base pairs even within the same experiment [11,17,18, 22]. This length variability limits the mapping accuracy, especially in more recent maps that used short-read sequencing with only one nucleosome end sequenced. Indeed, for this reason, these recent maps, which currently constitute most of the nucleosome data available, are actually less accurate than earlier maps in which mononucleosomes were sequenced in their entirety.

The DNA measurement stage includes primer ligation and DNA amplification, followed by application of microarrays or deep sequencing. All of these steps have sequence-specific biases [26,27] that will manifest as

inaccuracies in the resulting intensity (microarrays) or number of reads (deep sequencing) obtained. However, such biases can sometimes be sporadic and not reproducible between experiments, as seen, for instance, in a small number of genomic positions that have extremely high read coverage in only a subset of the replicates [10].

Experimental control of biases

Control experiments are the most direct way to account for the above experimental biases, but unfortunately they are not straightforward in the case of nucleosome mapping. In one type of control experiment, naked DNA is digested by MNase, followed by size selecting specific DNA lengths (such as approximately nucleosome length, about 150 bp), and DNA measurement using microarray [15] or deep sequencing [11,12]. In principle, such an experiment could account for biases that arise from the sequence specificity of MNase, or from microarrays and deep sequencing. However, we believe that such an experiment is not valid as a control, for the following reasons.

First, because there are no nucleosomes to protect against MNase digestion in the naked DNA experiment, if one uses the same concentration of nuclease for the same time as used with the chromatin, the naked DNA will be digested completely, down to tiny oligonucleotides. Therefore, the extent of MNase digestion must be far lower on the naked DNA than on the chromatin. Sequence specificities are generally more pronounced in lower enzyme concentrations, so this means that the MNase sequence specificities that appear in the naked DNA experiment will exaggerate the true effect of MNase sequence specificities on chromatin. The appearance of the sharp band of about 147 bp after MNase digestion in the real experiment but not in the naked DNA experiment indicates that the band in the real experiment truly reflects nucleosomes, and not just favored MNase sites, and further evidence of this includes the following observations: that the products of MNase digestion on chromatin yield nucleosomes that crystallize and whose structure, determined at 7 Å resolution by X-ray crystallography [28], is commensurate with that of later analyses of reconstituted nucleosomes imaged at atomic resolution [29]; and that the ladder-like distribution of oligonucleosome DNAs created during the nuclease digestion, and its evolution during the course of the digestion, mirror exactly the distribution of actual oligonucleosomes as imaged by electron microscopy [30].

Second, another problem with such a 'control' experiment is that linker DNAs *in vivo* are generally AT-rich (a result that was also shown by methods that did not use MNase; see [31] for details), and MNase preferentially cleaves TA/AT dinucleotides. Thus, we expect that on naked DNA, MNase will have more cleavage sites in regions that are true linkers in the real nucleosome samples, leading to similarities between the real and control experiment. Indeed, it is well known that MNase cleavage sites on naked DNA are related to true linkers [32-34], and these observations are reinforced by the new genome-wide analyses. Normalizing by such a 'control' dataset would artifactually reduce the real nucleosome positioning signals.

Considering both these caveats, we propose that results of nucleosome mapping experiments are best validated by independent methods that do not use MNase. Many such approaches are available, including: (i) using selection-based methods to define the nucleosome sequence preferences [10]; (ii) using high-resolution imaging to map nucleosome locations without nuclease [35,36]; (iii) using chemical probes, such as methidium propyl EDTA [37,38] or 1,10-phenanthroline-cuprous complex [39], to define linker regions; (iv) using chemical probes to define specific locations within the nucleosome, such as the nucleosome center [40] or the nucleosome ends [41]; and (v) using other enzymes that may have reduced, or at least different, sequence specificities compared with MNase [42].

Estimating nucleosome organization from experimental measurements

The final result of a sequence-based nucleosome mapping experiment is a set of uniquely mapped nucleosome-bound DNA segments, which we can represent as a vector r whose entries are the number of nucleosome reads that start at each base pair. From these reads, we then estimate quantities of interest, such as nucleosome occupancy and nucleosome positioning. The given estimations are not probabilities, as converting these quantities into probabilities is not trivial.

Estimating nucleosome occupancy

Estimating the nucleosome occupancy of a base pair is straightforward because it is simply the sum of reads covering that base pair, that is, the sum of reads in the 147 bp preceding the base pair. Formally:

$$Occ_e(x) = \sum_{i=x-146}^{x} r_i$$

where $Occ_e(x)$ is the estimated (unnormalized) occupancy at base pair x and r is the reads vector. Although this quantity is commonly referred to as nucleosome occupancy, it is not a probability and thus is not exactly occupancy as defined above. In addition, nucleosome occupancy is relatively robust to errors in the precise mapping of true nucleosome ends, which typically arise from the use of MNase.

Estimating absolute nucleosome positioning

In principle, nucleosome positioning could be computed from the number of nucleosome reads that start at each

base pair. However, because nucleosome maps have considerable noise at the single-base-pair level, such estimation would be noisy, as seen by the poor reproducibility of nucleosome positioning obtained from two randomly chosen subsets of the same nucleosome map (Figure 2). A reasonable solution, which considerably increases this reproducibility, is to convolute the raw reads with some smoothing function (Figure 2). Formally, we propose that absolute nucleosome positioning should be estimated as:

$$Pa_e(x) = \sum_{i=-d}^{d} w_i r_{x+1}$$

where $Pa_e(x)$ is the estimated (unnormalized) absolute positioning at base pair x, w is a weights vector representing the smoothing function, d determines the dimension of w and r is the reads vector. Typical smoothing functions are uniform smoothing, which is a simple moving average of reads in some window, and Gaussian smoothing, in which the moving average assigns higher weight to closer base pairs. Both functions have a single parameter representing the width of the smoothing window, and we suggest that its value should be tens of base pairs to accommodate for inaccuracies in mapping precise nucleosome positions. Unlike nucleosome occupancy, absolute positioning estimation depends on an arbitrary parameter choice (width) and its estimation may thus be less robust. In addition, estimating absolute positioning with uniform smoothing over 147-bp windows is exactly equal to nucleosome occupancy (shifted by 73 bp), showing that the two terms are strongly related.

Estimating conditional nucleosome positioning

Having estimated the absolute positioning of a base pair, it is straightforward to compute its conditional positioning, as it is simply the ratio between its absolute positioning and the sum of the absolute positioning in the 147-bp window surrounding the base pair. However, conditional positioning is more sensitive to experimental noise. For example, consider a base pair x whose surrounding 147 bp contain no other read starts, where in one case one nucleosome read starts at x and in another case 1,000 reads start at x. Although in both cases the conditional positioning of x is 1, a single additional nucleosome read in the vicinity of x (not within the range of the smoothing function) will change the conditional positioning at x to 0.5 in the first case but to 0.999 in the second case. Thus, estimation of conditional positioning is less robust to noise than estimation of absolute positioning and nucleosome occupancy, especially in regions with relatively few reads. Notably, this problem stems from the estimation process rather than from the definition of conditional positioning.

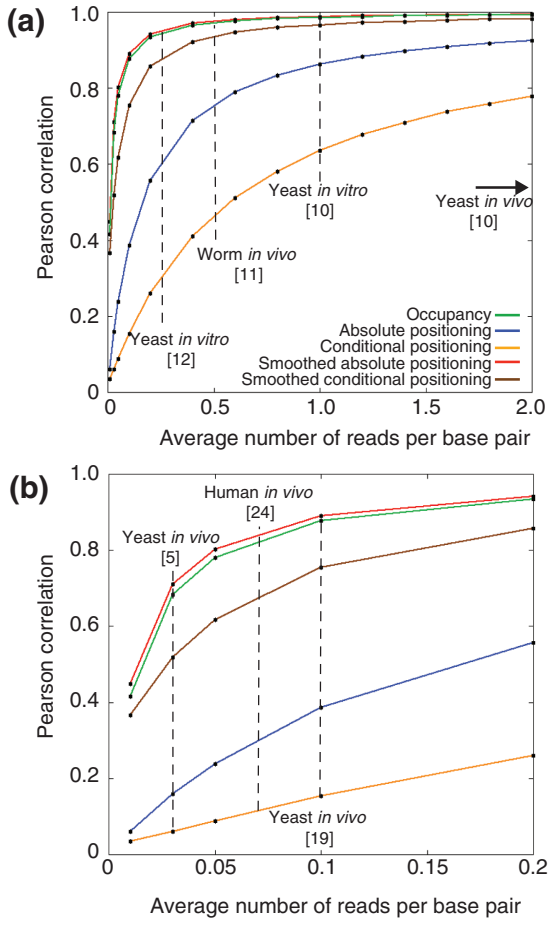

Figure 2. The effect of the number of sequence reads on the comparison of nucleosome maps using different measures. **(a,b)** We randomly sampled *in vivo* nucleosome data from yeast at different levels of genomic coverage [10]. At each level, five pairs of nucleosome maps were generated and, for each map, we estimated five different quantities: nucleosome occupancy, absolute nucleosome positioning (without smoothing), conditional positioning (without smoothing), smoothed absolute positioning and smoothed conditional positioning (both using a Gaussian filter with 20-bp standard deviation). For each pair of maps and every estimated measure, the Pearson correlation between each pair of maps was computed; this simulates the comparison of two replicates with the same level of coverage and thus shows the difference between two random samples from the same experiment with the same number of reads. The black arrow indicates an average read number beyond the scale of the y-axis. (b) An expansion of (a) for low numbers of reads. Standard deviation at all plotted points is smaller than 0.001. The coverage of the full *in vivo* map was about 2.2 nucleosome read starts per base pair. This simulation addresses only the error introduced by sampling and does not simulate the effect of other sources of errors in the experiments. These include the effect of variability in the extracted lengths of nucleosome-protected sequences, to which measures such as positioning are especially sensitive. Vertical dashed lines indicate the approximate amount of uniquely mapped reads in various studies [5,10-12,18,24], suggesting that sequencing coverage in several of these studies might lead to underestimation of similarities among maps, depending on the estimated quantity.

One practical solution is to ignore regions with a low read coverage when calculating conditional positioning.

Comparing nucleosome maps

Effects of experimental issues on nucleosome map comparisons

Several experimental issues can affect the similarity between nucleosome maps. First, the experimental sequence-specific biases and length variability of the measured nucleosome-bound sequences can lead to overestimation or underestimation of similarity between maps, respectively. Second, comparisons of maps are sensitive to the number of reads measured because each map is sampled from the distribution of nucleosome positions, and thus even two random samples of reads from the same experiment will differ, with the difference being inversely proportional to the number of reads (Figure 2). Even with deep sequencing, the current coverage of existing maps is relatively low, totaling about 2 nucleosome read starts per base pair in a yeast *in vivo* map [10], about 0.1 to 1 in yeast *in vitro* maps [10,12], and only about 0.07 in a human *in vivo* map [24]. Thus, reported similarities are likely to increase as maps with more reads are measured (Figure 2).

Finally, there are often many experimental differences in how the maps are measured, which can lead to underestimation of the maps. For example, *in vivo* maps differ from *in vitro* maps in temperature, salt concentrations, histone concentrations and even in the histones themselves, because *in vitro* maps used histones from chicken erythrocytes [10] or fly embryos [12].

Comparing nucleosome occupancy

We propose several methods for comparing nucleosome occupancies of maps. One direct method for comparing the occupancies is by computing the Pearson correlation between their respective occupancy vectors. The correlation of unrelated maps is expected to be close to zero, although the background model is non-trivial because of dependencies between adjacent positions. Alternatively, the Spearman correlation can be computed, which is the same as Pearson except that each value is converted to its rank in the data. Although the conversion to ranks loses information, it is less sensitive to non-linear scaling errors and to outlier regions that have an abnormally large number of reads, which are occasionally observed [10].

A third method used to compare nucleosome occupancy is receiver operating characteristic (ROC) analysis, where the aim is to quantify the degree to which one map (the predictor map) can discriminate high occupancy regions from low occupancy regions in the other map (the target map) [18,43,44]. First, a set of high occupancy regions and a set of low occupancy regions are derived from the target map by choosing two thresholds, such that high occupancy regions are defined as those consecutive genomic regions in which all base pairs are above one threshold, and low occupancy regions as consecutive regions where all base pairs are below the second threshold. When the two thresholds are close (or equal) to each other, then the defined regions will cover most (or all) of the genome. Next, each high and low occupancy region defined in the target map is assigned a single occupancy value using the other (predictor) occupancy map - for instance, by taking the mean occupancy that each region has in the predictor map. Finally, the degree to which the predicted occupancy values of the target regions discriminate between high and low occupancy regions can be computed using the area under curve (AUC) metric, or the Mann-Whitney-Wilcoxon statistic [45]. A perfect discrimination, in which all high occupancy regions have higher predicted occupancy values than all low occupancy regions, has an AUC value of 1. Random discrimination has an AUC value of 0.5, allowing statistical significance to be computed. An advantage of the AUC analysis is that it is less sensitive to coverage- and resolution-related experimental noise because its computations are done on regions that are typically much larger than single base pairs. Thus, the AUC is especially suited to comparisons of maps with relatively low coverage.

Finally, we note that all of the above methods are insensitive to linear scaling and may thus overestimate the similarity of maps that are not on the same scale. For example, one could compare a map whose occupancy values are between 95 and 105 reads per base pair with a map whose occupancy values range from 0 to 1,000 reads per base pair. Although the maps are very different, they could still show a perfect correlation of 1. Thus, we emphasize the importance of examining the occupancy distributions, and in cases of scaling differences we suggest using a non-scaling metric such as the fraction of variance unexplained (FVU) statistical measure. Simply put, this measure quantifies the mean squared error between two vectors relative to the mean squared error between one vector and its mean.

Comparing nucleosome positioning

As with occupancy, absolute and conditional positioning can both be represented by per-base-pair positioning vectors and thus can be compared using the Pearson or Spearman correlation. However, there are two important differences. First, because a well positioned nucleosome produces a narrow peak whose width is equal to the smoothing window, positioning is more localized than occupancy and can thus be more sensitive to experimental noise that causes small shifts in the location of read starts (such as the amount of MNase used to digest the chromatin). In addition, the amount of data used to

estimate the nucleosome positioning of a base pair is significantly smaller than the amount of data used to estimate its occupancy, because when estimating positioning each read start provides data for a single base pair, whereas in occupancy estimation each read start provides data for the 147 bp covered by its corresponding nucleosome. Thus, positioning comparisons are less robust than occupancy comparisons, especially given the relatively low coverage of existing maps (Figure 2). For these reasons, even at a higher read coverage, the regions of low occupancy would produce less reliable positioning comparisons. The problem is especially pronounced in comparisons of conditional positioning because this quantity may be undefined in low occupancy regions.

Comparing locations of well positioned nucleosomes

Even if every nucleosome has a different position *in vivo* and *in vitro*, histone sequence preferences may still be important determinants of nucleosome positions *in vivo*. For example, consider a case in which every well positioned nucleosome *in vivo* is shifted *in vitro* by a couple of dozens of base pairs in an arbitrary direction. This type of correspondence between the maps might not be captured by correlation analysis. Instead, we can ask whether the locations of nucleosomes that are well positioned in the two maps are closer to each other than expected by chance.

The first step in such an analysis is to identify well positioned nucleosomes in each map; the resulting set of well positioned nucleosomes will depend on the details and parameters of the method. After selecting well positioned nucleosomes in each map, we designate nucleosomes of one map as the target set and those of the other as the predicted set. For each target nucleosome start, we then find the distance to the closest predicted nucleosome start. We then plot the fraction of target nucleosomes that have a predicted nucleosome start at most d base pairs away, for all possible distances d. To assess significance, we repeat the computations after shuffling nucleosome locations in the predicted set; in this shuffling we maintain the original distribution of pairwise distances between neighboring nucleosomes and shuffle only in the uniquely mappable regions over which the original map is defined. Finally, for each distance d we can estimate the fraction, above random expectation, of target nucleosomes that have a predicted nucleosome less than d base pairs away by subtracting the shuffled value from the actual value and dividing the result by one minus the shuffled value. The division is needed to scale the resulting value such that the highest attainable value is 1.

One important limitation of this nucleosome distance analysis is that although the concept of well positioned nucleosomes is intuitive and simple for analysis, it is an oversimplification because there is no requirement for such nucleosomes to be especially prevalent in nucleosome maps. In principle, a nucleosome occupancy map could have regions of high occupancy and regions of low occupancy without having a single well positioned nucleosome. Whether or not nucleosomes are well positioned in a given genomic region *in vivo* is not indicative of whether or not the positions of nucleosomes in the region are governed by histone sequence preferences. This is because regions with well positioned nucleosomes and regions with weakly positioned nucleosomes can both be encoded by histone sequence preferences - for example, through a peaked nucleosome affinity landscape in the former case and a relatively flat affinity landscape in the latter.

Current agreements and open questions

Nucleosome maps are similar across different technologies

In vivo nucleosome maps have been measured by different laboratories and using tiling microarrays and multiple different deep-sequencing technologies [10,13-15,17-20]. In general, these maps show significant correspondence. Although formally, in the absence of a gold standard, we cannot say that deep sequencing is better than tiling microarrays, this is generally believed to be the case because deep sequencing maps data at single-base-pair resolution and it allows direct measurement of positioning rather than just occupancy. *In vitro* nucleosome maps, produced by reconstitution of histones on naked DNA, have been measured using deep sequencing and by a wholly independent single-molecule microscopy assay that does not use MNase [10,12,35]. Here, too, the maps are significantly similar. However, some technologies may share sequence biases that can contribute to similarities between maps. In conclusion, although technological differences clearly lead to differences in the results obtained, in general, both *in vitro* and *in vivo* maps are reliable.

DNA sequence is significantly predictive of nucleosome organization *in vitro* and *in vivo*

Numerous studies [9,13,18,43,44,46,47] have used genome-wide *in vivo* maps of nucleosomes, generated by either microarrays or deep sequencing, to construct sequence-based computational models that predict nucleosome occupancy or nucleosome positions using only DNA sequence information. Importantly, these models were evaluated in a cross-validation manner, that is, they were trained on part of the *in vivo* data and their predictions were tested on the other (held-out) parts of the data. These methods were evaluated in various ways, including correlation, AUC and nucleosome distance analyses; all methods performed significantly better than random expectation, thereby demonstrating that DNA sequence

is significantly predictive of various aspects of the *in vivo* nucleosome organization. More recently, these sequence-based nucleosome models were constructed using *in vitro* nucleosome maps, so their sequence signals are more likely to reflect histone sequence preferences, and, again, these models had strong predictive power [10,48]. It is thus clear that DNA sequence is significantly predictive of aspects of nucleosome organization both *in vitro* and *in vivo*.

Histone sequence preferences are major determinants of nucleosome organization *in vivo*

A direct way of assessing the effect of histone sequence preferences on the nucleosome organization *in vivo* is through comparisons of *in vivo* nucleosome maps with *in vitro* maps produced by reconstitution of histones on naked DNA. Given that the *in vitro* map is governed only by the histone sequence preferences, similarity of this map to the *in vivo* map would suggest that histone sequence preferences contribute to *in vivo* organization. Because nucleosome organizations cannot be measured directly, these comparisons must be done using the occupancy and positioning measures. Although actual numbers may depend on technical details, comparison of nucleosome occupancy by correlation and AUC analyses clearly demonstrate that this aspect of nucleosome organization is significantly similar *in vitro* and *in vivo* [10,12].

Estimation of nucleosome positioning is less reliable because it depends on arbitrary thresholds and requires higher coverage to be robustly estimated, and we thus expect that its correspondence between the two maps will be lower. Although direct comparisons of positioning were not done between the *in vivo* and *in vitro* maps, different versions of nucleosome distance analyses were performed, yielding estimates that, after accounting for random expectation, about 20 to 50% of the well positioned nucleosomes *in vivo* are in close proximity to matching nucleosomes *in vitro* [12,49].

Even though these comparisons could overestimate or underestimate the actual similarity, the *in vitro* and *in vivo* maps clearly show significant similarity even when they are measured with different technologies, confirming that their similarity is robust. Thus, the results of all studies so far indicate that histone sequence preferences have a considerable effect on nucleosome organization *in vivo*, both in terms of occupancy and positioning.

Which sequence signals are important for nucleosome organization?

The exact sequence preferences of nucleosomes are not known. Because it is not feasible to measure directly the nucleosome affinities of all of the possible 147-bp sequences, one approach to comprehensively characterize nucleosome affinities is to construct a mathematical model that attempts to generalize from a smaller set of nucleosome affinity measurements. Ideally, such models should be learned from affinity measurements of sequences that represent a random sample of the sequence space. However, large collections of nucleosome affinity measurements are currently available only for the yeast genome. This may bias our current understanding of nucleosome sequence preferences and limit the ability of current models to correctly predict the nucleosome affinities of many types of sequences that do not exist in the yeast genome.

Periodic patterns of dinucleotides, initially observed in alignments of nucleosome-bound sequences *in vivo* [50], were the basis of the first models of nucleosome sequence preferences that were used for genome-wide prediction [9,47,51,52]. Because relatively few nucleosome affinity measurements and nucleosome-bound sequences were available at the time, these initial models were trained and evaluated on a relatively small amount of data (on the order of hundreds of sequences). The use of deep sequencing for measuring nucleosome positions increased the available data by about 100,000-fold, prompting the development of a new generation of models whose performance was drastically better than that of the initial models [18,43,44]. Prominent features of these newer models include poly(dA:dT) sequences, which are strongly predictive of low nucleosome occupancy, and high GC content, which is strongly predictive of high nucleosome occupancy. These features are unlikely to be a consequence of sequence-specific experimental biases as they were observed by many different measurement technologies [10,35]. Notably, although a model based only on poly(dA:dT) frequency and GC content is highly predictive of nucleosome occupancy, models with more features are significantly more predictive [53]. Moreover, it is not clear whether these simplified models can accurately predict detailed nucleosome positions, mainly because such evaluations are difficult to perform given the limited accuracy of nucleosome position measurement. Periodic dinucleotide patterns are also evident in large nucleosome collections derived from deep sequencing, so they are also likely to have an important effect.

Finally, we again note that because our current understanding is mainly based on measurements of yeast sequences, it may be biased. Specifically, it is not clear whether the correlation between GC content and nucleosome affinity also holds for sequences with very high GC content because such sequences are rare in yeast and their affinity has thus not been measured on a large scale. Thus, although we clearly understand some of the sequence signals that are important for determining nucleosome organizations, we expect that better mapping technology and measurements on more diverse

sequences will improve our understanding and allow better models to be developed.

What causes the long-range ordering of nucleosomes over genes?

Studies have observed a long-range ordering of nucleosomes downstream of gene start sites, which decays with the distance from the start of the gene [15,18,19]. Although the functional significance of this ordering is not known, it is a prominent feature of nucleosome organization *in vivo*, and it is thus important to understand its cause. Kornberg and Stryer [54,55] suggested that this phenomenon can be explained by boundary elements that restrict nucleosomes from binding to specific regions (for example, DNA-bound transcription factors or polymerase, whose presence sterically occludes a nucleosome from adjacent base pairs). They showed [54,55] that given a high concentration of nucleosomes along DNA and steric hindrance between nucleosomes, a simple model based on statistical mechanics predicts that a single boundary constraint is sufficient to generate a long-range ordering of nucleosomes, where the ordering, or 'statistical positioning', is greatest immediately adjacent to the boundary and decays with the distance away from it.

One possibility is that *in vivo* factors cause the long-range nucleosome ordering. For example, the boundary constraint for nucleosome binding could be caused by the binding of transcription factors and of the transcriptional initiation machinery upstream of gene starts. This possibility is supported by our previous observation [10] that alignment of *in vitro* nucleosome data relative to gene starts does not show long-range nucleosome ordering over genes.

A second possibility, which is not mutually exclusive, is that nucleosome-disfavoring sequences cause long-range ordering. Given that many sequences, most notably poly(dA:dT)-like elements, strongly disfavor nucleosome formation, one possibility is that these sequences themselves, rather than just bound proteins, constitute the boundary constraint required to generate the observed long-range ordering of nucleosomes. Indeed, many yeast genes have such nucleosome-disfavoring sequences just upstream of their start site [15,18,19]. Even if such a mechanism were partly responsible for generating the long-range ordering, some might argue that it should be regarded as a contribution from histone sequence preferences. However, this argument is purely semantic because it would still be a case in which nucleosome positions are determined in part by DNA sequence, even though here the influencing positioning sequences act negatively and reside in linker regions and not where the nucleosomes are bound. Notably, by incorporating nucleosome concentration and steric hindrance between nucleosomes, some of the current sequence-based models for predicting nucleosome organization from DNA sequence can capture the long-range effect of nucleosome-disfavoring sequences [10,18,36]. Thus, the long-range ordering of nucleosomes could be caused in part by histone sequence preferences, through the indirect effect of nucleosome-disfavoring sequences. Indeed, nucleosome-disfavoring sequences have been shown in principle to be able to induce statistical ordering effects [36]. However, current genome-wide *in vitro* maps do not show statistical positioning.

Thus, although the exact reason for the long-range ordering of nucleosomes over genes *in vivo* is still unclear, most of the current evidence suggests that it is mostly due to statistical positioning emanating from boundary constraints for nucleosome binding, where the boundary constraints may be caused by the binding of various factors *in vivo* and by nucleosome-disfavoring sequences.

Can the effect of histone sequence preferences on nucleosome organization be called a major determinant or a code?

Although studies mostly agree with each other on the results of many of the analyses, some studies have argued about the semantics of whether or not the effect of histone sequence preferences can be called a 'code' or a 'major' determinant of *in vivo* nucleosome organization. For example, one paper [56] stated that if histone sequence preferences account for about 25% of *in vivo* nucleosome positions, then this is 'considered to be too low for the existence of a nucleosome positioning code'. Although this is clearly a subjective statement about which we make no judgment, even if the estimate of about 25% were true, then we are not aware of any other single factor that affects the genome-wide nucleosome organization to a greater or even similar extent. For example, poly(dA:dT) sequences are significantly more predictive of nucleosome depletion *in vivo* than is any known transcription factor [18]. In this sense, histone sequence preferences are a major determinant of *in vivo* nucleosome organization, even if the aforementioned conservative estimates of 25%, which we believe to be underestimates, are used.

Regarding the use of the term 'code', it has been argued that, by analogy to the genetic code, a biological code must be deterministic. Clearly, some aspects of the *in vivo* nucleosome organization are encoded in the DNA sequence through histone sequence preferences. However, it is also clear that histone sequence preferences do not, and in fact cannot, completely determine the *in vivo* nucleosome organization. However, the term 'code' has also been used in several biological contexts to describe non-deterministic information flow; prominent examples include the transcriptional code, the histone code and the

splicing code. Whether the DNA sequence preferences of nucleosomes should or should not be called a code is, in our view, an inconsequential semantic issue.

In parallel to the active research and fast progress that the nucleosome positioning field is experiencing, it is important to develop clear definitions and standards that researchers agree on, and with which the key issues can be addressed. We believe that the lack of such standards has contributed to a perceived disagreement regarding the role of histone sequence preferences in determining nucleosome organization *in vivo*. Here, we propose definitions and procedures for measuring and comparing nucleosome maps and try to summarize clearly which points are currently in agreement, and which questions are still open. Our conclusion is that most studies agree on the key points. Specifically, although there are indeed some quantitative and semantic disagreements, it is generally agreed that histone sequence preferences do indeed have a central effect on nucleosome organization *in vivo*.

Author details
[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Terrence Donnelly Centre for Cellular and Biomolecular Research, 160 College Street, Toronto, Ontario M5S 3E1, Canada. [3]Banting and Best Department of Medical Research, 112 College St, Toronto, Ontario M5S 1L6, Canada. [4]Department of Biology and the Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [5]Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, 2153 Sheridan Road, Evanston, IL 60208, USA. [6]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel.

Published: 30 November 2010

References

1. Wittig B, Wittig S: **A phase relationship associates tRNA structural gene sequences with nucleosome cores.** *Cell* 1979, **18:**1173-1183.
2. Lam FH, Steger DJ, O'Shea EK: **Chromatin decouples promoter threshold from dynamic range.** *Nature* 2008, **453:**246-250.
3. Miller JA, Widom J: **Collaborative competition mechanism for gene activation *in vivo*.** *Mol Cell Biol* 2003, **23:**1623-1632.
4. Han M, Grunstein M: **Nucleosome loss activates yeast downstream promoters *in vivo*.** *Cell* 1988, **55:**1137-1145.
5. Field Y, Fondufe-Mittendorf Y, Moore IK, Mieczkowski P, Kaplan N, Lubling Y, Lieb JD, Widom J, Segal E: **Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization.** *Nat Genet* 2009, **41:**438-445.
6. Gencheva M, Boa S, Fraser R, Simmen MW, A Whitelaw CB, Allan J: ***In vivo* and *in vitro* nucleosome positioning on the ovine beta-lactoglobulin gene are related.** *J Mol Biol* 2006, **361:**216-230.
7. Thåström A, Bingham LM, Widom J: **Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning.** *J Mol Biol* 2004, **338:**695-709.
8. Lowary PT, Widom J: **New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning.** *J Mol Biol* 1998, **276:**19-42.
9. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang JZ, Widom J: **A genomic code for nucleosome positioning.** *Nature* 2006, **442:**772-778.
10. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, Leproust EM, Hughes TR, Lieb JD, Widom J, Segal E: **The DNA-encoded nucleosome organization of a eukaryotic genome.** *Nature* 2009, **458:**362-366.
11. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM: **A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning.** *Genome Res* 2008, **18:**1051-1063.
12. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K: **Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*.** *Nat Struct Mol Biol* 2009, **16:**847-852.
13. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C: **A high-resolution atlas of nucleosome occupancy in yeast.** *Nat Genet* 2007, **39:**1235-1244.
14. Whitehouse I, Rando OJ, Delrow J, Tsukiyama T: **Chromatin remodelling at promoters suppresses antisense transcription.** *Nature* 2007, **450:**1031-1035.
15. Yuan G, Liu Y, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: **Genome-scale identification of nucleosome positions in *S. cerevisiae*.** *Science* 2005, **309:**30.
16. Ozsolak F, Song JS, Liu XS, Fisher DE: **High-throughput mapping of the chromatin structure of human promoters.** *Nat Biotechnol* 2007, **25:**244-248.
17. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF: **Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome.** *Nature* 2007, **446:**572-576.
18. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E: **Distinct modes of regulation by chromatin encoded through nucleosome positioning signals.** *PLoS Comput Biol* 2008, **4:**e1000216.
19. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF: **A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome.** *Genome Res* 2008, **18:**1073-1083.
20. Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N: **High-resolution nucleosome mapping reveals transcription-dependent promoter packaging.** *Genome Res* 2010, **20:**90-100.
21. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, Gilmour DS, Albert I, Pugh BF: **Nucleosome organization in the *Drosophila* genome.** *Nature* 2008, **453:**358-362.
22. Nishida H, Motoyama T, Yamamoto S, Aburatani H: **Genome-wide maps of mono- and di-nucleosomes of *Aspergillus fumigatus*.** *Bioinformatics* 2009, **25:**2295-2297.
23. Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ: **Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin.** *Genome Res* 2006, **16:**1505-1516.
24. Schones DE, Cui K, Cuddapah S, Roh T, Barski A, Wang Z, Wei G, Zhao K: **Dynamic regulation of nucleosome positioning in the human genome.** *Cell* 2008, **132:**887-898.
25. Hörz W, Altenburger W: **Sequence specific cleavage of DNA by micrococcal nuclease.** *Nucleic Acids Res* 1981, **9:**2643-2658.
26. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA: **Evaluation of next generation sequencing platforms for population targeted sequencing studies.** *Genome Biol* 2009, **10:**R32.
27. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS: **Model-based analysis of tiling-arrays for ChIP-chip.** *Proc Natl Acad Sci USA* 2006, **103:**12457-12462.
28. Richmond TJ, Finch JT, Rushton B, Rhodes D, Klug A: **Structure of the nucleosome core particle at 7 Å resolution.** *Nature* 1984, **311:**532-537.
29. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 A resolution.** *Nature* 1997, **389:**251-260.
30. Finch JT, Noll M, Kornberg RD: **Electron microscopy of defined lengths of chromatin.** *Proc Natl Acad Sci USA* 1975, **72:**3320-3322.
31. Segal E, Widom J: **Poly(dA:dT) tracts: major determinants of nucleosome organization.** *Curr Opin Struct Biol* 2009, **19:**65-71.
32. Keene MA, Elgin SC: **Micrococcal nuclease as a probe of DNA sequence organization and chromatin structure.** *Cell* 1981, **27:**57-64.
33. Flick JT, Eissenberg JC, Elgin SC: **Micrococcal nuclease as a DNA structural probe: its recognition sequences, their genomic distribution and correlation with DNA structure determinants.** *J Mol Biol* 1986, **190:**619-633.
34. Keene MA, Elgin SC: **Patterns of DNA structural polymorphism and their evolutionary implications.** *Cell* 1984, **36:**121-129.
35. Visnapuu M, Greene EC: **Single-molecule imaging of DNA curtains reveals intrinsic energy landscapes for nucleosome deposition.** *Nat Struct Mol Biol* 2009, **16:**1056-1062.
36. Milani P, Chevereau G, Vaillant C, Audit B, Haftek-Terreau Z, Marilley M, Bouvet P, Argoul F, Arneodo A: **Nucleosome positioning by genomic excluding-**

energy barriers. *Proc Natl Acad Sci USA* 2009, **106:**22257-22262.
37. Cartwright IL, Elgin SC: **Chemical footprinting of 5S RNA chromatin in embryos of *Drosophila melanogaster*.** *EMBO J* 1984, **3:**3101-3108.
38. Cartwright IL, Hertzberg RP, Dervan PB, Elgin SC: **Cleavage of chromatin with methidiumpropyl-EDTA. iron(II).** *Proc Natl Acad Sci USA* 1983, **80:**3213-3217.
39. Cartwright IL, Elgin SC: **Analysis of chromatin structure and DNA sequence organization: use of the 1,10-phenanthroline-cuprous complex.** *Nucleic Acids Res* 1982, **10:**5835-5852.
40. Flaus A, Richmond TJ: **Base-pair resolution mapping of nucleosomes *in vitro*.** *Methods Mol Biol* 1999, **119:**45-60.
41. Kassabov SR, Bartholomew B: **Site-directed histone-DNA contact mapping for analysis of nucleosome dynamics.** *Methods Enzymol* 2004, **375:**193-192.
42. Widlak P, Garrard WT: **Unique features of the apoptotic endonuclease DFF40/CAD relative to micrococcal nuclease as a structural probe for chromatin.** *Biochem Cell Biol* 2006, **84:**405-410.
43. Yuan G, Liu JS: **Genomic sequence is highly predictive of local nucleosome depletion.** *PLoS Comput Biol* 2008, **4:**e13.
44. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z: **Nucleosome positioning signals in genomic DNA.** *Genome Res* 2007, **17:**1170-1177.
45. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143:**29-36.
46. Gupta S, Dennis J, Thurman RE, Kingston R, Stamatoyannopoulos JA, Noble WS: **Predicting human nucleosome occupancy from primary sequence.** *PLoS Comput Biol* 2008, **4:**e1000134.
47. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF: **Nucleosome positions predicted through comparative genomics.** *Nat Genet* 2006, **38:**1210-1215.
48. Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Field Y, Lieb JD, Widom J, Segal E, Hughes TR: **High nucleosome occupancy is encoded at human regulatory sequences.** *PloS ONE* 2010, **5:**e9129.
49. Kaplan N, Moore I, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, Hughes TR, Lieb JD, Widom J, Segal E: **Nucleosome sequence preferences influence *in vivo* nucleosome organization.** *Nat Struct Mol Biol* 2010, **17:**918-920.
50. Satchwell SC, Drew HR, Travers AA: **Sequence periodicities in chicken nucleosome core DNA.** *J Mol Biol* 1986, **191:**659-675.
51. Levitsky VG: **RECON: a program for prediction of nucleosome formation potential.** *Nucleic Acids Res* 2004, **32:**W346-W349.
52. Levitsky VG, Podkolodnaya OA, Kolchanov NA: **DNA: calculation and promoters analysis.** *Bioinformatics* 2001, **17:**998-1010.
53. Tillo D, Hughes TR: **G+C content dominates intrinsic nucleosome occupancy.** *BMC Bioinformatics* 2009, **10:**442.
54. Kornberg RD, Stryer L: **Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism.** *Nucleic Acids Res* 1988, **16:**6677-6690.
55. Kornberg R: **The location of nucleosomes in chromatin: specific or statistical?** *Nature* 1981, **292:**579-580.
56. Stein A, Takasuka TE, Collings CK: **Are nucleosome positions *in vivo* primarily determined by histone-DNA sequence preferences?** *Nucleic Acids Res* 2010, **38:**709-719.

doi:10.1186/gb-2010-11-11-140
Cite this article as: Kaplan N, *et al.*: Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biology* 2010, **11:**140.