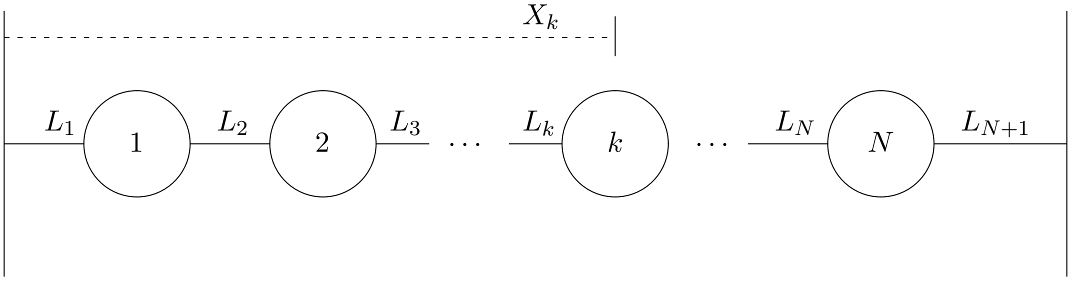
SUPPLEMENTARY DATA

Quantifying the Role of Steric Constraints in Nucleosome Positioning

H. Tomas Rube and Jun S. Song

S.1 Supplementary Methods: Statistical Positioning (Nucleosome Location and Its Variance)

Statistical positioning describes how the combination of (1) barriers confining nucleosome movement and (2) inter-nucleosomal steric hindrance affects the distribution of densely packed nucleosomes. In this framework, the nucleosomes are modeled as one-dimensional gas, with N non-overlapping finite-sized particles distributed at random in a finite region of length w, where all valid configurations are given equal probability. The effective size d of the particles, determining how tightly they can be packed, parametrizes the effect of steric constraints. Because of the discrete nature of the genomic sequence, all quantities take integer values. Let X_k denote the center location of the k-th nucleosome (numbered sequentially from left to right; see figure below). The primary goal of this section is to calculate $E[X_k]$ and $\mathrm{Var}[X_k]$.



An equivalent problem that can be easily solved is to examine the distribution of the gaps between nucleosomes. Let L_k denote the length of the gap to the left of the k-th nucleosome, and let L_{N+1} be the length of the gap between the N-th nucleosome and the right barrier. Any configuration can be uniquely specified in terms of these gaps, and the position of the k-th nucleosome is

$$X_k = (k - 1/2)d + Y_k, \tag{S1}$$

where $Y_k = \sum_{i=1}^{k} L_i$ is the partial sum of gap lengths to the left of the k-th nucleosome. Because all configurations have the same total gap length, i.e.

$$L_1 + L_2 + ... + L_{N+1} = w - Nd \equiv L,$$

there is a one-to-one map between the set of valid configurations and the set of weak compositions of L into $N + 1$ ordered non-negative integers. Thus, there are $\binom{L+N}{N}$ distinct configurations in total. Noting that the number of configuration with $Y_k = y$ is equal to the number weak compositions of y into k parts times the number of weak compositions of $(L - y)$ into $N - k + 1$ parts, the probability of Y_k is

$$P(Y_k = y) = \frac{\binom{y+k-1}{k-1}\binom{L-y+N-k}{N-k}}{\binom{L+N}{N}}. \tag{S2}$$

S.1.1 Statistical Positioning with a Single Barrier

Before calculating $\mathrm{Var}[X_k]$ for double barriers, we give a simplified derivation of the most important result: $\mathrm{Var}[X_k]$ increases linearly in k at the rate $\ell(\ell+1)$ away from a single barrier, where ℓ is the mean gap length between nucleosomes. This relation is true even if the barrier is not firm, but merely imposes a partial restriction on a nucleosome. The linear scaling further generalizes to a broader class of models where nucleosomes interact through some interaction potential V, although how fast variance changes between nucleosomes may be different. (In the case of statistical positioning, this potential vanishes unless the nucleosomes overlap, in which case it is infinite.)

To prove the linear scaling, note that

$$X_k = X_{k-1} + L_k + d,$$

where L_k is the length of the gap between the two nucleosomes. We now assume that a restriction is imposed on the 0'th nucleosome, causing $\mathrm{Var}[X_0]$ to be finite. We further assume that all subsequent nucleosomes are allowed to move freely, except for steric interactions with immediate neighbors. If the array of nucleosomes continues indefinitely, the gap length L_k becomes independent of the position X_{k-1} of the nucleosome to the left; note that this independence does not hold for a finite nucleosome array confined to a finite region delimited by two barriers. Using that X_k is the sum of two independent random variables then gives

$$\mathrm{Var}[X_k] = \mathrm{Var}[X_{k-1}] + \mathrm{Var}[L_k]. \tag{S3}$$

Furthermore, for $k \geq 1$, $E[L_k]$ is independent of k for a semi-infinite array, giving

$$E[X_k] = E[X_0] + kE[L_k] + kd, \qquad \mathrm{Var}[X_k] = \mathrm{Var}[X_0] + k\mathrm{Var}[L_k].$$

Thus, the variance increases linearly in this class of models, and the variance gradient, defined as $\mathrm{Var}[X_{k+1}] - \mathrm{Var}[X_k]$, corresponds to the "wiggle room" that the interaction allows.

In the case of statistical positioning, $\mathrm{Var}[L_k]$ can be evaluated by noting that $P(L_k = y) = P(Y_1 = y)$. In the limit $L \to \infty$ with the ratio $L/N = \ell$ kept fixed, $P(Y_1 = y)$ in Eq. S2 becomes

$$P(L_k = y) = P(Y_1 = y) \to \frac{\ell^y}{(\ell+1)^{y+1}}. \tag{S4}$$

It is then straightforward to show that $E[L_k] = \ell$ and $\mathrm{Var}[L_k] = \ell(\ell+1)$, yielding

$$\boxed{\begin{aligned} \mathrm{Var}[X_k] &= \mathrm{Var}[X_0] + k\ell(\ell+1) \\ E[X_k] &= E[X_0] + k(\ell+d) \end{aligned}}\,. \tag{S5}$$

S.1.2 Statistical Positioning with a Pair of Fixed Barriers

Next, we evaluate $E[X_k]$ and $\mathrm{Var}[X_k]$ in the case of a pair of fixed barriers. This is is simplified by expressing $E[Y_k]$ and $\mathrm{Var}[Y_k]$ in terms of un-normalized moments S_a defined as

$$S_a = \sum_{y=0}^{L} y^a \binom{y+k-1}{k-1}\binom{L-y+N-k}{N-k}.$$

Using Eq. S2, we can express $E[Y_k]$ and $\mathrm{Var}[Y_k]$ as

$$E[Y_k] = \frac{S_1}{S_0}, \text{ and}$$

$$\mathrm{Var}[Y_k] = E[Y_k^2] - E[Y_k]^2 = \frac{S_2}{S_0} - \frac{S_1^2}{S_0^2}.$$

One practical method for evaluating these sums is to use the generalized binomial expansion

$$\frac{1}{(1-z)^k} = \sum_{y=0}^{\infty} \binom{y+k-1}{k-1} z^y. \tag{S6}$$

By acting on this equation with $(z\frac{d}{dz})^a$ and multiplying by $\frac{1}{(1-z)^{N-k+1}}$, it is straightforward to show that

$$\sum_{y=0}^{\infty} y^a \binom{y+k-1}{k-1} z^y \times \sum_{j=0}^{\infty} \binom{j+N-k}{N-k} z^j$$
$$= \left(\left(z\frac{d}{dz} \right)^a \frac{1}{(1-z)^k} \right) \times \frac{1}{(1-z)^{N-k+1}}.$$

Because the moment S_a is the L'th order term in the z-expansion of left-hand side, it can be evaluated by algebraically simplifying the right-hand side and extracting the L'th order term using Eq. S6. For example, to evaluate S_0 we set $a = 0$ and expand the right hand side using Eq. S6:

$$\frac{1}{(1-z)^k} \frac{1}{(1-z)^{N-k+1}} = \sum_{j=0}^{\infty} \binom{j+N}{N} z^j.$$

Extracting the L'th order term gives $S_0 = \binom{L+N}{N}$, proving that the probability in Eq. S2 is correctly normalized.

Next, to calculate S_1, we extract the L'th order term from

$$\left(z\frac{d}{dz} \frac{1}{(1-z)^k} \right) \times \frac{1}{(1-z)^{N-k+1}} = \sum_{j=0}^{\infty} k \binom{j+N}{N+1} z^j,$$

yielding $S_1 = k\binom{N+L}{N+1}$. We thus get

$$E[Y_k] = \frac{S_1}{S_0} = \frac{k\binom{N+L}{N+1}}{\binom{N+L}{N}} = \frac{kL}{N+1} = k\ell, \tag{S7}$$

where $\ell = \frac{L}{N+1}$ is the average gap length. Using Eq. S1 and Eq. S7, the expectation value of X_k is

$$\boxed{E[X_k] = k(\ell + d) - d/2}.$$

Similarly, S_2 is evaluated by extracting the L'th order term from

$$\left(z\frac{d}{dz} z\frac{d}{dz} \frac{1}{(1-z)^k} \right) \times \frac{1}{(1-z)^{N-k+1}} = \frac{kz(1+kz)}{(1-z)^{N+3}}$$
$$= \sum_{j=0}^{\infty} \left[k\binom{j+N+1}{N+2} + k^2 \binom{j+N}{N+2} \right] z^j,$$

giving

$$E[Y_k^2] = \frac{S_2}{S_0} = \frac{k\binom{L+N+1}{N+2} + k^2\binom{L+N}{N+2}}{\binom{N+L}{N}} = \frac{kL(1 + k(L-1) + L + N)}{(N+1)(N+2)}.$$

The variance of Y_k is then

$$\mathrm{Var}[Y_k] = E[Y_k^2] - E[Y_k]^2 = \frac{kL(N+1-k)(L+N+1)}{(N+1)^2(N+2)}.$$

Since X_k and Y_k differ only by a constant, we have

$$\boxed{\mathrm{Var}[X_k] = \mathrm{Var}[Y_k] = \ell(\ell+1)\frac{k(N+1-k)}{N+2}}\,. \tag{S8}$$

In the limit of large N with ℓ kept fixed, this expression reduces to

$$\boxed{\mathrm{Var}[X_k] \to k\ell(\ell+1)}\,, \tag{S9}$$

which is Eq. S5 with $\mathrm{Var}[X_0] = 0$. Note that this linear scaling also holds near the boundaries of a long confining interval.

S.1.3 Single Barrier Limit of the Grand Canonical Ensemble

It is well known in statistical physics that the Grand Canonical Ensemble (GCE) and the Canonical Ensemble (CE) are equivalent in the thermodynamic limit of large N with fixed density w/N. The N-nucleosome term in the GCE partition function is $Z_N = \binom{w-Nd+N}{N} e^{N\mu}$, where w is the length of the interval bounding the nucleosomes and μ the chemical potential. To see the equivalence of the two distributions in the thermodynamic limit, initially assume that w is fixed and find N that maximizes Z_N, or equivalently, $\log Z_N$. For large N,

$$\log Z_N \approx N\mu - N\log\left(\frac{N}{w-Nd+N}\right) - (w-Nd)\log\left(\frac{w-Nd}{w-Nd+N}\right).$$

Thus,

$$\frac{d\log Z_N}{dN} \approx \mu - \log\left(\frac{N}{w-Nd+N}\right) + d\log\left(\frac{w-Nd}{w-Nd+N}\right),$$

and N^* that maximizes Z_N thus satisfies

$$\mu \approx \log\left(\frac{N^*}{w-N^*d+N^*}\right) - d\log\left(\frac{w-N^*d}{w-N^*d+N^*}\right).$$

Then, in the thermodynamic limit of $N^* \gg 1$ and $w = (N^*+1)\ell + N^*d$, we get

$$\mu \to \log\left(\frac{(\ell+1)^{d-1}}{\ell^d}\right). \tag{S10}$$

The size of the fluctuation $\delta N = N - N^*$ can be estimated using the saddle-point method and the expression

$$\frac{d^2\log Z_N}{dN^2} \approx -\frac{w^2}{N(w-N(d-1))(w-Nd)} \to -\frac{(d+\ell)^2}{\ell(\ell+1)}\frac{1}{N},$$

giving $\mathrm{Var}[N] \propto N^*$. The fractional fluctuation in particle number thus behaves like $\delta N/N^* \propto 1/\sqrt{N^*}$ and is negligible for $N^* \gg 1$. A GCE with the chemical potential μ in Eq. S10 is thus equivalent to a CE of N^* nucleosomes with mean gap length ℓ.

S.1.4 Statistical Positioning with Dynamic Barriers

The previous section described statistical positioning in its original form, corresponding to non-overlapping nucleosomes distributed between two fixed barriers. The assumption of fixed barriers, however, may not be realized in nature, and we are led to consider a more general model with dynamic barriers: we here consider the statistical positioning of N free nucleosomes (indexed sequentially as $1, ..., N$) flanked by two partially restricted nucleosomes (indexed as 0 and $N+1$) that function as moving barriers. For example, the restricted nucleosomes could be subjected to chromatin remodeling or a free energy barrier caused by a Poly(dA:dT) stretch. Let X_k denote the center location of the k-th nucleosome. The effect of restrictions on the flanking nucleosomes is encoded in the marginal distribution $P(X_0, X_{N+1})$. The in-between nucleosomes are then positioned according to statistical positioning conditioned on X_0 and X_{N+1}. The goal of this section is to express $\mathrm{Var}[X_k]$ in terms of $P(X_0, X_{N+1})$ and compare the resulting expression to the fixed barrier case in Eq. S8.

The evaluation of $\mathrm{Var}[X_k]$ is facilitated by first conditioning on X_0 and X_{N+1}. For example, consider any quantity $Q(X_0, X_{N+1}, Y_1, \ldots, Y_{N+1})$, where as defined in the previous section, Y_k denotes the partial sum of gap lengths to the left of the k-th nucleosomes. Then, we can compute the expectation of Q as

$$E[Q] = E_{X_0,X_{N+1}} E_{Y|X_0,X_{N+1}}[Q].$$

Because the flanking nucleosomes can be thought of as barriers delimiting the N in-between nucleosomes, the first step of computing the conditional expectation uses the fixed barrier results from the previous section with

$$L = X_{N+1} - X_0 - (N+1)d.$$

To average over the free nucleosomes, we use $X_k = X_0 + kd + Y_k$ to expand

$$\mathrm{Var}[X_k] = \mathrm{Var}[X_0] + \mathrm{Var}[Y_k] + 2\,\mathrm{Cov}[X_0, Y_k]. \tag{S11}$$

Here only the last two terms depend on the in-between nucleosomes through Y_k. To evaluate $\mathrm{Var}[Y_k]$, we rewrite

$$E[(Y_k - E[Y_k])^2] = E_{X_0,X_{N+1}} E_{Y|X_0,X_{N+1}}[Y_k^2] - \left(E_{X_0,X_{N+1}} E_{Y|X_0,X_{N+1}}[Y_k]\right)^2. \tag{S12}$$

Using the fixed barrier results in Eq. S7 and Eq. S8, we get

$$\begin{aligned}
E_{X_0,X_{N+1}} E_{Y|X_0,X_{N+1}}[Y_k] &= E_{X_0,X_{N+1}}\left[\frac{kL}{N+1}\right] = \frac{kE[L]}{N+1} \\
E_{X_0,X_{N+1}} E_{Y|X_0,X_{N+1}}[Y_k^2] &= E_{X_0,X_{N+1}}\left[\frac{k(N+1-k)L(L+N+1)}{(N+1)^2(N+2)} + \left(\frac{kL}{N+1}\right)^2\right] \\
&= \frac{k(N+1-k)E[L(L+N+1)]}{(N+1)^2(N+2)} + \frac{k^2E[L^2]}{(N+1)^2}.
\end{aligned}$$

Substituting these expressions into Eq. S12, we get

$$\mathrm{Var}[Y_k] = \frac{k(N+1-k)E[L](E[L]+N+1)}{(N+1)^2(N+2)} + \frac{k(k+1)\mathrm{Var}[L]}{(N+1)(N+2)}.$$

The last term in Eq. S11 is similarly simplified using the results from the previous section:

$$\begin{aligned}
2\,\mathrm{Cov}[X_0, Y_k] &= 2E_{X_0,X_{N+1}}[(X_0 - E[X_0])(E_{Y_k|X_0,X_{N+1}}[Y_k] - E[Y_k])] \\
&= \frac{2k}{N+1}\mathrm{Cov}[X_0, X_{N+1} - X_0] \\
&= \frac{k(\mathrm{Var}[X_{N+1}] - \mathrm{Var}[X_0] - \mathrm{Var}[L])}{N+1}.
\end{aligned} \tag{S13}$$

Putting everything together finally gives

$$\boxed{\mathrm{Var}[X_k] = \frac{k(N+1-k)}{N+2}\left(E[\ell](E[\ell]+1) - (N+1)\mathrm{Var}[\ell]\ \right) + \frac{(N+1-k)\mathrm{Var}[X_0] + k\mathrm{Var}[X_{N+1}]}{N+1}}$$

where we defined $\ell = L/(N+1)$, so that $E[\ell]$ is the mean gap length. In the limit of large N with $E[\ell]$ and $\text{Var}[L]$ kept fixed, this expression reduces to

$$\boxed{\text{Var}[X_k] \to \text{Var}[X_0] + kE[\ell](E[\ell]+1)}, \tag{S14}$$

which is just Eq. S5, with ℓ replaced by its expectation. In this limit, the only effect of replacing the fixed barriers with partially constrained flanking nucleosomes is a constant offset to the fixed barrier formula in Eq. S9. As a result, the variance gradient is again constant in the semi-infinite limit or near the boundaries of a long confining interval.

S.1.5 Directionally Packed Nucleosomes

To model packing, we here reformulate statistical positioning in terms of an isothermal-isobaric ensemble and then generalize the case by adding packing forces. Statistical positioning models nucleosomes as dense bidirectional fluid, where movement is restricted only by steric constraints and barriers. Because the nucleosomes are assumed to be free, they are on average evenly spaced. Alternatively, ATP-dependent directional packing could increase nucleosome density near one of the barriers. We will thus model such packing as position-independent, but nucleosome-specific, forces acting on the nucleosomes and derive the variance profile in Eq. (4-6).

The probability distribution of fixed-barrier statistical positioning is, in the language of statistical physics, the canonical ensemble (CE) with zero and infinite energy for non-overlapping and overlapping nucleosomes, respectively. The single barrier case can be equivalently expressed in terms of the isothermal-isobaric ensemble (IIE). In this distribution, the region length w is allowed to fluctuate and the weight of an allowed configuration is e^{-wP}, where P is the pressure (we here set the temperature $T = 1$ for convenience). The single barrier case is retrieved in the thermodynamic limit $w, N, L \gg 1$ with $\ell = L/N$ kept fixed. The IIE and the CE are equivalent near the barrier in this limit, and their respective parameters are related through $P = \ln \frac{\ell+1}{\ell}$.

We now model packing as constant (i.e. position-independent) and nucleosome-specific forces f_k pulling the nucleosomes towards a single barrier. In terms of the IIE, each allowed configuration has weight $e^{-Pw-\sum_k f_k X_k}$, where the second term in the exponent is the work done by the forces. To calculate $P(L_k|X_{k-1})$, note that incrementing L_k by a suppresses the weight of a state by

$$\frac{P(L_1, ...L_{k-1}, L_k + a, L_{k+1}, .., w + a)}{P(L_1, ...L_{k-1}, L_k, L_{k+1}, .., w)} = e^{-p_k a}, \qquad p_k = P + \sum_{k'=k}^{\infty} f_{k'},$$

where p_k is a nucleosome-specific pressure. Normalizing gives the geometric distribution

$$P(L_k|X_{k-1}) = (1 - e^{-p_k})e^{-p_k L_k}.$$

Using this distribution, and noting that X_{k-1} and L_k are independent, gives

$$\text{Var}[L_k] = E[L_k](E[L_k]+1), \qquad e^{-p_k} = \frac{E[L_k]}{E[L_k]+1}. \tag{S15}$$

The independence of X_{k-1} and L_k also implies that Eq. S3 is true for this type of packing forces, finally giving

$$\text{Var}[X_k] = \text{Var}[X_{k-1}] + E[L_k](E[L_k]+1). \tag{S16}$$

Thus, the effect of constant packing forces is to modulate the nucleosome pressure p_k. While this modulation in turn changes the nucleosome spacing and variance gradient, the relationship between the two is the same as in statistical positioning.

S.2 Supplementary Methods: Enrichment Analysis

The enrichment analysis considers an ordered set of N genes and tests whether a fixed subset preferentially contains either high-ranking or low-ranking genes. To assess statistical significance, the N genes are first ranked between 0 and 1 as $i/(N-1)$, $i = 0, \ldots, N-1$, according to their associated data values. Then, the median rank x of the n genes in the fixed subset is calculated.

More precisely, consider a data set $\Omega = \{y_0, y_1, \ldots, y_{N-1}\}$ of N distinct values, sorted such that $y_0 < y_1 < \cdots < y_{N-1}$. The rank of y_i is thus i. To get a measure of rank that is independent of the size of the data set, we define the rescaled rank of y_i to be $r_i = i/(N-1)$, so that $0 \leq r_i \leq 1$. Let $S = \{y_{i_1}, \ldots, y_{i_n}\} \subset \Omega$ be a subset of n elements, where $y_{i_1} < y_{i_2} < \ldots < y_{i_n}$, and define the median rank of S to be median$\{r_{i_1}, \ldots, r_{i_n}\}$. We will develop an enrichment analysis that tests whether S preferentially contains either high-ranking or low-ranking values by comparing its observed median rank with the null distribution of the median rank of all possible distinct subsets of size n, where each distinct subset has uniform probability $\binom{N}{n}^{-1}$ of being sampled.

If $n = |S|$ is odd, then the median rank of S is equal to $x = r_{i_{(n+1)/2}}$. Note that $x(N-1) = i_{(n+1)/2}$ is an integer. The probability of x is thus

$$P(x) = \binom{(N-1)x}{(n-1)/2}\binom{(N-1)(1-x)}{(n-1)/2} \bigg/ \binom{N}{n},$$

where the numerator gives the total number of ways of picking $(n-1)/2$ values less than $y_{i_{(n+1)/2}}$ and $(n-1)/2$ values greater than $y_{i_{(n+1)/2}}$.

If $n = |S|$ is even, then the median rank x of S is

$$x = \frac{r_{i_{n/2}} + r_{i_{1+n/2}}}{2} = \frac{i_{n/2} + i_{1+n/2}}{2(N-1)}.$$

Thus, x is an integer multiple of $\frac{1}{2(N-1)}$, and the possible values of $x(N-1)$ are $\frac{n-1}{2}, \frac{n}{2}, \frac{n+1}{2}, \ldots, N - \frac{n+1}{2}$. Defining $k = x(N-1)$, it can be shown that the probability of x is given by

$$P(x) = \binom{N}{n}^{-1} \sum_{m=0}^{M} \binom{\lceil k-1 \rceil - m}{n/2 - 1}\binom{N - 1 - (\lfloor k \rfloor + m + 1)}{n/2 - 1},$$

where $M = \min(\lceil k-1 \rceil + 1 - n/2, N - n/2 - \lfloor k \rfloor - 1)$, $\lfloor z \rfloor$ is the greatest integer less than or equal to z, and $\lceil z \rceil$ is the smallest integer greater than or equal to z.

Two-sided p-values were calculated from these probability functions.

S.3 Supplementary Figures

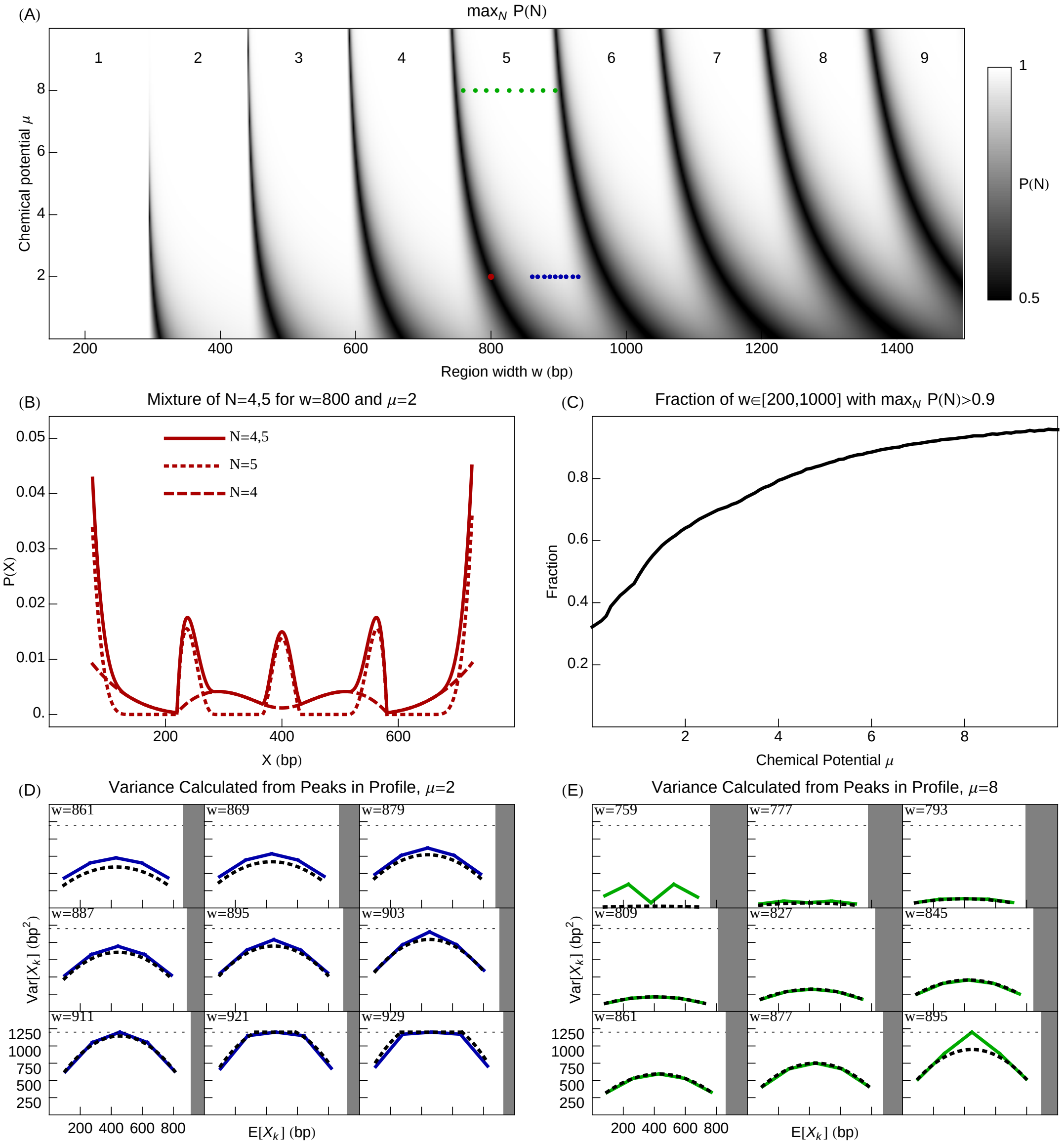


Figure S1: **Properties of statistical positioning and the grand canonical ensemble (GCE).** (A) Probability of most likely nucleosome count (i.e. max$_N P(N)$ where $P(N) \propto e^{\mu N}\binom{w-Nd+N}{N}$) as a function of region length w and chemical potential μ, using $d = 147$. Red, blue and green dots correspond to values of μ and w used in (B), (D) and (E) respectively. (B) Example of a nucleosome distribution (solid) that is a mixture of two values of N ($N = 4$ dashed and $N = 5$ dotted). The length w was chosen such that $P(N = 4) = P(N = 5) = 0.5$. (C) Fraction of w-values (in the range $[200, 1000]$) for which one value of N dominates the GCE (max$_N P(N) > 0.9$) as a function of μ. (D) Values of Var$[X_k]$ calculated from theoretical nucleosome distribution using peak calling (solid blue) or Eq. (7) (dashed), using $\mu = 2$ and varying w. The nine w-values where chosen such that $P(N = 5) > 0.9$. Var$[X_k]$ is truncated at 1200bp^2, as described in Methods. (E) Same as in (D) but with $\mu = 8$.

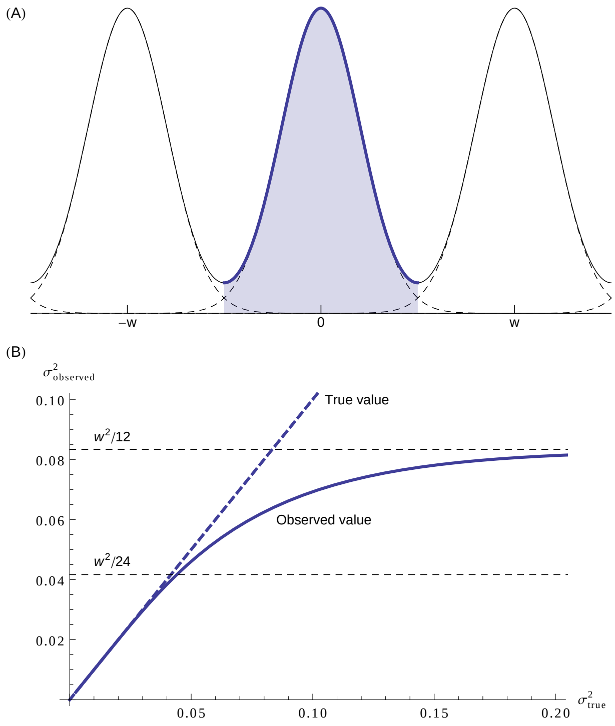


Figure S2: **Modeled effect of nucleosome overlap on variance estimate.** (A) The sampling density of reads associated with nucleosomes (dashed lines) are represented by an array of Gaussians, offset by w and with variance σ^2 (here set to $w^2/24$, which is half the variance of a uniform random variable defined on $[0, w]$). In practice, only the marginal read distribution (solid line) is observed. Reads are assigned to the closest nucleosome peak (e.g. blue filled region for the middle nucleosome). (B) Solid line shows the estimated variance $\sigma^2_{\text{observed}}$ of the reads in the shaded region in (A), for varying values of the actual variance σ^2 of the middle peak in (A). For $\sigma^2 \lesssim w^2/24$, the observed variance agrees well with the true value of σ^2 (dashed line), but it deviates from the true value for larger σ^2. This analysis justifies our choice of $w^2/24$ as a maximal cutoff σ^2_{max} for calling positioned nucleosomes whose variance we can confidently estimate.

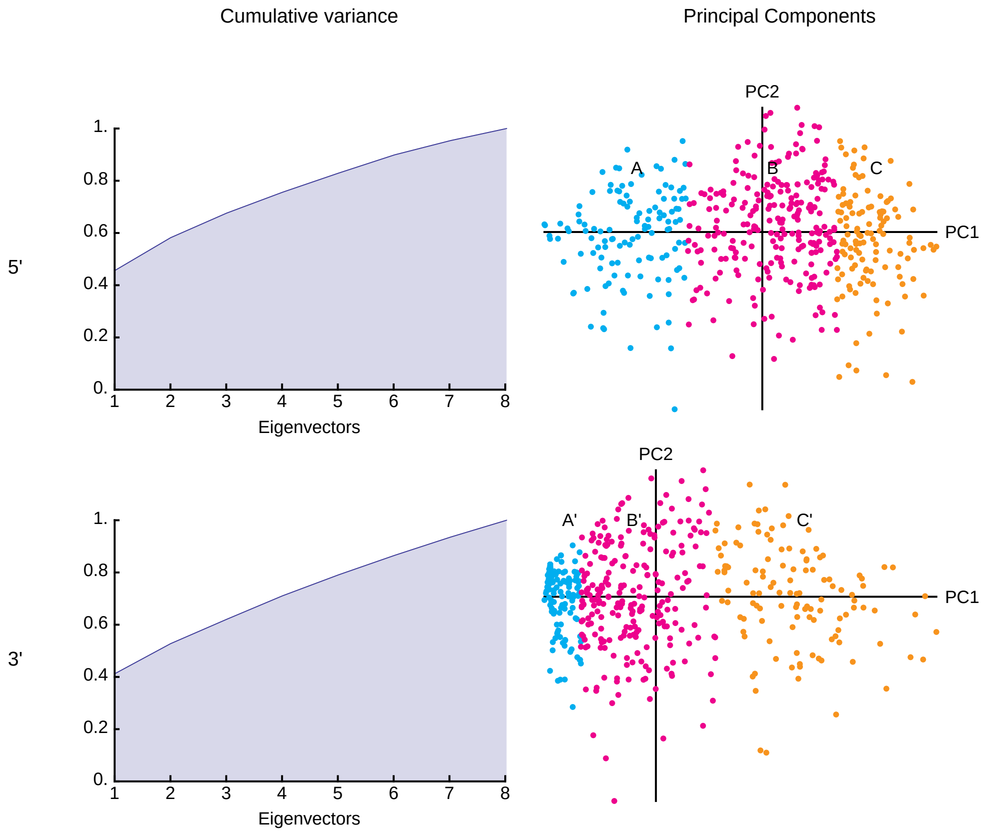


Figure S3: **Principal component analysis of the fuzziness profiles of 8 nucleosomes in the 5'- and 3'-ends of long genes.** Left: Cumulative variance explained as a function of principal components. Right: Distribution of fuzziness in the first two principal components. Genes are grouped based on the quartiles of the first principal component. The groups A and A' correspond to the first quartile, B and B' to the two central quartiles, and C and C' to the last quartile.

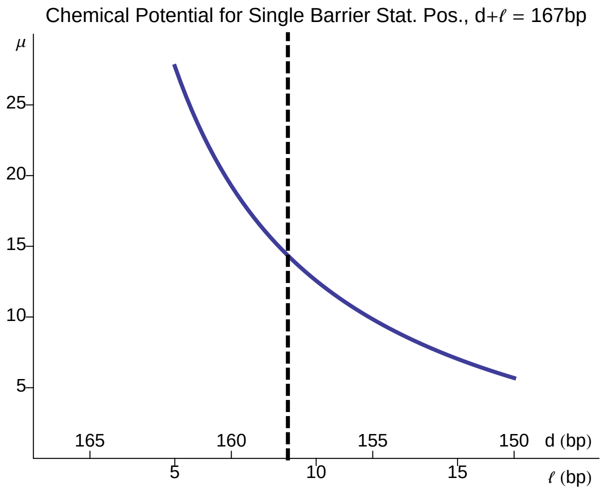


Figure S4: **Relation between chemical potential μ and nucleosome spacing for single barrier statistical positioning.** The chemical potential was calculated from the formula $e^{\mu} = (\ell+1)^{d-1}/\ell^{d}$, using that $\ell + d = 167$bp is the mean nucleosomes spacing. This relation is derived in Supplementary Methods by taking the thermodynamic limit of the grand canonical ensemble.

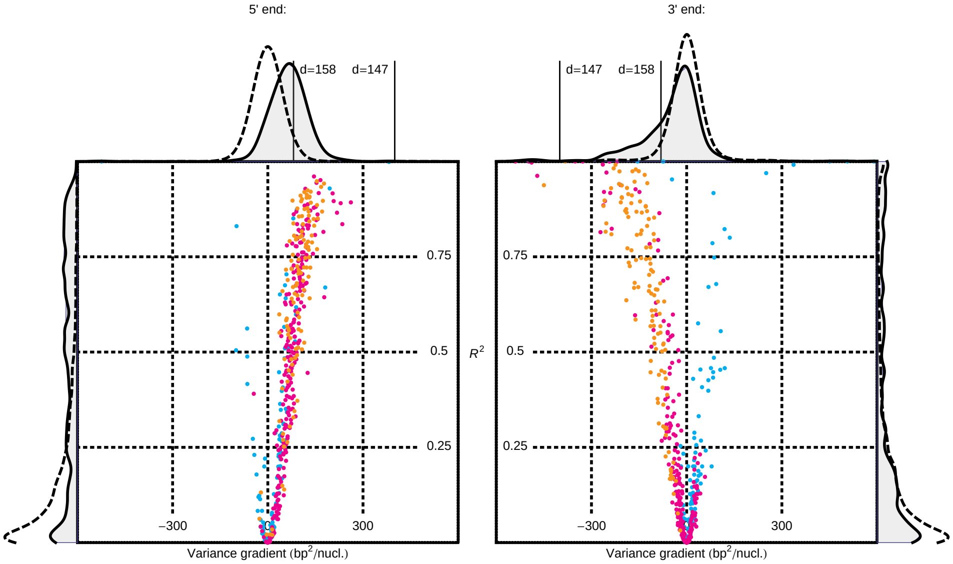


Figure S5: **Plot of variance gradient vs. goodness-of-fit R^2 of statistical positioning for long genes.** Scatter plots show the distribution of variance gradient (x-axis) and R^2-fit (y-axis) calculated from 8 nucleosomes in the 5'-end and 3'-end of long genes (16 or more nucleosomes). Solid lines indicate the marginal distributions. Dashed lines show the marginal distributions obtained by randomly permuting nucleosomes 10 times and refitting. Vertical lines in the top marginal plots indicate the theoretical variance gradient in a single-barrier statistical positioning model for two different values of d, using an inter-nucleosome distance of 168bp.

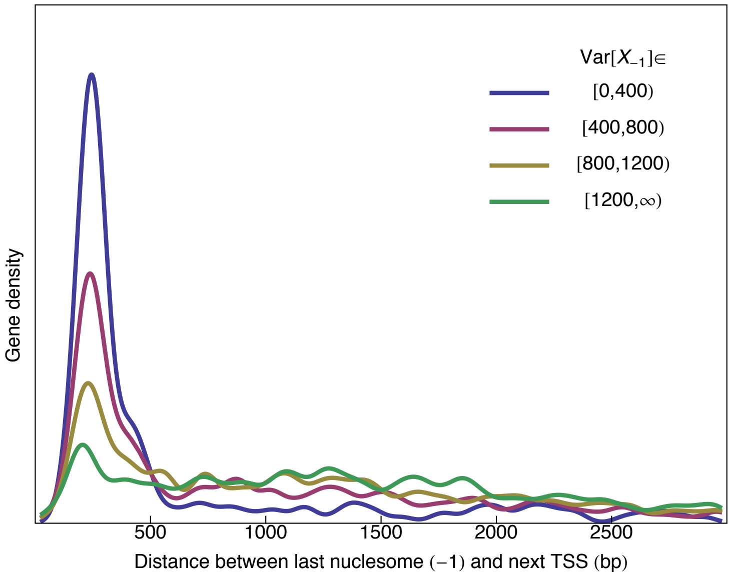


Figure S6: **Distribution of the distance between the last (−1) nucleosome in a gene and the TSS of the nearest downstream gene.** The colored lines correspond to genes with fuzziness of the last nucleosome Var$[X_{-1}]$ in the intervals $[0, 400)$ (blue), $[400, 800)$ (purple), $[800, 1200)$ (yellow), and $[1200, \infty)$ (green). This figure shows that well-positioned last nucleosomes tend to be near the TSS of an adjacent gene.

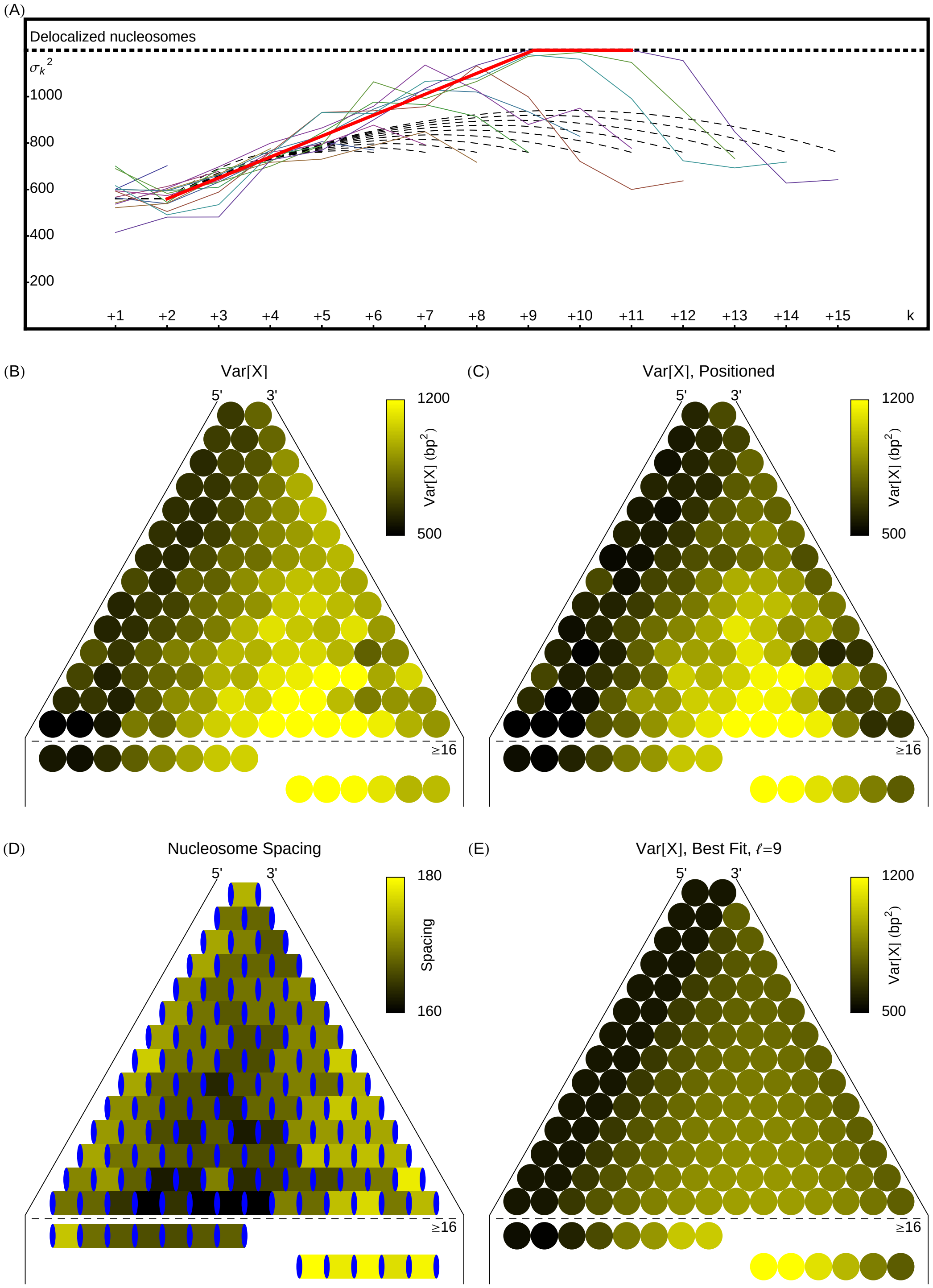


Figure S7: **Nucleosome fuzziness and spacing for short- and medium-long genes.** (A) Thin lines show median nucleosome fuzziness σ_k^2 of the k-th nucleosome for genes with $2 \leq \hat{N} \leq 15$ nucleosomes, $|\hat{w} - \hat{w}_{\hat{N}}| \leq 40$bp and σ_{+1}^2, $\sigma_{-1}^2 < 1200\text{bp}^2$. The extrapolation $\hat{w}_{\hat{N}} = -164 + 169\hat{N}$ was used for $\hat{N} > 9$. Dashed lines correspond to the double-barrier model in Figure 5D. Solid red line corresponds single barrier prediction with $\ell = 9$bp . (B) Variance calculated as in (A), but without the constraint $\sigma_{+1}, \sigma_{-1} < \sigma_{\max}$. (C) Median variance σ_k^2 for genes in (A). (D) Spacing beween nucleosomes in genes in (B). (E) Double-barrier model from Figure 5D.