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

Anton Valouev, Jeffrey Ichikawa, Thaisan Tonthat, et al.

Genome Res. published online May 13, 2008
Access the most recent version at doi:10.1101/gr.076463.108

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

Anton Valouev[1], Jeffrey Ichikawa[2], Thaisan Tonthat[1], Jeremy Stuart[2], Swati Ranade[2], Heather Peckham[2], Kathy Zeng[1], Joel A. Malek[2], Gina Costa[2], Kevin McKernan[2], Arend Sidow[1], Andrew Fire[1,3] and Steven M. Johnson[1]

[1]Departments of Pathology and Genetics, Stanford University School of Medicine, Stanford, CA 94305
[2]Applied Biosystems, 500 Cummings Ctr., Beverly, MA 01915
[3]Corresponding author

contact information:
Dr. Andrew Fire
Departments of Pathology and Genetics
Stanford University School of Medicine
300 Pasteur Drive, Room L235
Stanford, CA 94305-5324
Tel: (650)723-2885
Fax: (650)724-9070
email: afire@stanford.edu

running title: *C. elegans* high-resolution nucleosome map

key words: nucleosome code, nematode

ABSTRACT

Using the massively parallel technique of Sequencing by Oligonucleotide Ligation and Detection (SOLiD) we have assessed the *in vivo* positions of more than 44 million putative nucleosome cores in the multicellular genetic model organism *Caenorhabditis elegans*. These analyses provide a global view of the chromatin architecture of a multicellular animal at extremely high density and resolution. While we observe some degree of reproducible positioning throughout the genome in our mixed stage population of animals, we note that the major chromatin feature in the worm is a diversity of allowed nucleosome positions at the vast majority of individual loci. While absolute positioning of nucleosomes can vary substantially, relative positioning of nucleosomes (in a repeated array structure likely to be maintained at least in part by steric constraints) appears to be a significant property of chromatin structure. The high density of nucleosomal reads enabled a substantial extension of previous analysis describing the usage of individual oligonucleotide sequences along the span of the nucleosome core and linker. We release this dataset, via the UCSC Genome Browser, as a resource for the high-resolution analysis of chromatin conformation and DNA accessibility at individual loci within the *C. elegans* genome.

INTRODUCTION

The regulation of genetic information within eukaryotic cells involves a high degree of specificity both in the availability of individual DNA-binding factors in individual cells, and in availability in the genome of DNA sequences that are their potential targets. Modulating the accessibility of individual DNA sequences are many complex interactions, the most prevalent of which are the interactions between histone octamers and DNA in compacted chromosomes. Each histone core interacts with 147 base pairs of DNA, which coil 1.7 times around the histone octamer (Davey et al. 2002; Luger et al. 1997) to form the basic unit of chromatin structure, the nucleosome. Since the first description over three decades ago (Kornberg 1974), the nucleosome and its role in gene regulation has been the subject of intensive study and speculation.

As we have an increasingly detailed functional view of the genome, the tools of high throughput molecular characterization have been used to begin obtaining a genome-wide description of nucleosome positions (Albert et al. 2007; Johnson et al. 2006; Lee et al. 2007; Peckham et al. 2007; Satchwell et al. 1986; Schones et al. 2008; Shivaswamy et al. 2008; Yuan et al. 2005). These data have in turn been used in attempts to reveal nucleosome positioning signals in DNA sequence (Ioshikhes et al. 1996; Satchwell et al. 1986; Segal et al. 2006; Yuan and Liu 2008). Although of great interest and value, sequence-based predictions of nucleosome position have been limited to date in their accuracy and resolution.

In the literature, nucleosome positioning has been defined as the ability of sequences in the genome to encode a nucleosome organization, with such features regulating the access of non-histone proteins to DNA *in vivo* (e.g. Segal et al. 2006). A variety of global estimates for sequence-directed nucleosome positioning have been described for the single-cell eukaryote *Saccharomyces cerevisiae* (Segal et al. 2006; Lee et al. 2007; Peckham et

al. 2007; Yuan and Liu 2008; Shivashwamy et al. 2008). More complex multicellular systems with multiple tissue and cell types provide a number of interesting challenges and opportunities for the global study of chromatin structure. In this study, we use the new ultra-high throughput SOLiD-sequencing technology to characterize nucleosome positions in a mixed-stage mixed-tissue population of *C. elegans* cells. This work provides both a substantial *in vivo* dataset of nucleosome positions for a metazoan organism and an experimentally-derived high resolution map of nucleosome constraints for the nematode *C. elegans*.

Modes of possible nucleosome positioning

From the outset one can conceive of four different combinatorial modes or patterns for the individual and relative positioning of nucleosomes (Fig. 1).

Pattern 1: Positioned and Uniformly Spaced

The first pattern consists of nucleosomes that are reproducibly spaced and positioned in the genome so they occupy the same uniformly spaced set of positions regardless of cell type or developmental stage (Fig. 1A).

Pattern 2: Not Positioned but Uniformly Spaced

The second pattern consists of nucleosomes that are spaced at uniform intervals relative to each other within individual cells, but which lack any reproducibility in comparisons of absolute positioning between different cells (Fig. 1B). Differences in absolute positions could be specific to individual cell types or completely stochastic.

Pattern 3: Positioned but not Uniformly Spaced

The third pattern consists of nucleosomes where the precise positioning is maintained as a function of the underlying DNA sequence, but for which there is no consistent pattern of spacing along the chromosome (Fig. 1C).

Pattern 4: Not Positioned and not Uniformly Spaced

The fourth pattern invokes a scenario without any consistent positioning and spacing, so that nucleosomes are variably positioned on the genome with respect to both the underlying DNA sequence and each other (Fig. 1D).

It is important to note that the first and second patterns of nucleosome arrangement involve uniformly spaced nucleosomes and that either would be sufficient to explain the tight laddering bands of mono-, di-, tri-, and higher order nucleosome DNA fragments that are seen as a result of Micrococcal nuclease (MNase) digestion (Noll 1974); whereas they would yield very different results in sequence analysis looking at reproducibility of individual positions. Additionally, across an individual genome all four types of patterns could co-exist at different loci, and similarly across cell types and states, all four patterns could exist and overlap at an individual locus. In order to dissect these possibilities we took advantage of a new ultra-high-throughput sequencing technology, SOLiD.

RESULTS

Sequencing and coverage

SOLiD sequencing (solid.appliedbiosystems.com) is a massively parallel, ligation-mediated sequencing method (Figure 2) that extends a number of recent developments in nucleic acid chemistry (Dressman et al. 2003; Embleton et al. 1992; Ghadessy et al. 2001; Kwoh et al. 1989; Kwoh and Kwoh 1990; Macevicz 1998; McKernan et al. 2006; Shendure et al. 2005; Tawfik and Griffiths 1998) . Tens of millions of 25-50 nucleotide reads are generated by a single run of this sequencing technology.

To sample the global spectra of allowed positions for nucleosome occupancy within the *C. elegans* genome *in vivo*, we isolated DNA fragments associated with nucleosome cores from a mixed stage population of *C. elegans* (Johnson et al. 2006). Sanger sequencing of a limited number of these nucleosome core

fragments revealed the average length to be 149.5 base pairs (Supplemental Fig. S1 available in the Supplemental Data). *C. elegans* nucleosomal fragment libraries for SOLiD sequencing were constructed from this DNA population. A single experimental run on the SOLiD platform produced a total of 107 million 50 base pair reads from the nucleosome core fragment experiment (Table 1). Allowing up to five mismatches (and no insertions or deletions), we were able to place 56.4 million reads using the standard SOLiD mapping pipeline onto the *C. elegans* (WS170 freeze) or the *E. coli* (K12) genome (*E. coli* is the food source for *C. elegans* and hence a DNA contaminant). The 50 nucleotide reads that contained a unique best match with least number of mismatches (using these criteria) were assigned to corresponding locations in the genome with 44.0 million reads being uniquely placed on the worm genome (core read data set). Of these 44.0 million, ~10.7 million reads matched perfectly, ~9.0 million reads had one mismatch, ~7.5 million reads had two mismatches, ~6.4 million reads had three mismatches, ~5.5 million reads had four mismatches, and ~4.8 million reads had five mismatches. As an experimental control and reference data set, we used the SOLiD technology to sequence *C. elegans* genomic DNA digested with MNase. The control library consisted of whole genome DNA lightly digested with MNase and size selected in a range of 400-900 base pairs. Fragments were circularized to produce a paired-end library and subjected to paired-end SOLiD sequencing resulting in 25 base pair reads from each end . Mate pairs were mapped onto the reference genome using standard SOLiD mapping pipeline which produced 51.91 million mapped pairs in the correct size range matching either the worm or bacterial genomes. A total of 51.89 million reads were uniquely placed on the *C. elegans* genome: ~27.9 million reads matching perfectly, ~9.5 million with one mismatch, ~5.7 million reads had two mismatches, ~3.7 million reads had three mismatches, ~2.5 million reads had four mismatches, ~1.6 million reads had five mismatches, and ~0.9 million containing six mismatches between the two 25 base pair ends. By placing 50 nucleotide reads uniquely on the genome, we exclude about 9 million base pairs of the *C. elegans* genome (i.e., 50 base pair regions of the genome that are not unique and have an exact match elsewhere or have a second best match within one

mismatch). Using this approach, our 44 million placed putative nucleosome core reads correspond to, on average, one read every two base pairs of the genome. Since each placed read represents the position of a putative nucleosome covering 147 base pairs of DNA, our nucleosome coverage density of the genome is 71x or ~484 nucleosomes per kbp of unique genomic sequence.

General caveats

With such a large data set sampling of nucleosome and control sequences across and throughout the genome, it is important to be aware of potential limitations of the tools used to generate the sampling. As part of the SOLiD sequencing sample prep, a library population or sample undergoes linker ligation, preliminary template amplification, and subsequent emulsion PCR to attach and extend clonal template species onto a single bead. In addition to the standard concerns of ligation and PCR bias, the emulsion PCR adds several possible caveats including duplication of templates (multiple beads per reactor) and polyclonal beads (multiple templates per reactor). The latter could result in the obfuscation of individual sequence reads when multi-templated beads are removed from analysis due to multiple channels of fluorescence coincident on polyclonal beads; and the former could result when more than one bead extends the same molecule to result in overrepresentation of individual sequences. Additionally, in this data set, there is an underrepresentation of coverage (both control and experimental) coincident with AT-rich regions of the genome (mainly in introns and intergenic regions). This underrepresentation might reflect inefficient sequencing in these genomic loci due to the lower melting point of A/T rich oligonucleotides in the ligation sequencing. In any case, for our analysis sequencing biases such as that against A/T rich DNA are greatly reduced by normalization with control experiment data.

The other major tool used in this study with major caveats is micrococcal nuclease (MNase). While this enzyme is universally used to isolate nucleosome core DNA by preferentially digesting linker DNA to liberate the mononucleosome cores, its activity is not without sequence biases (Horz and Altenburger 1981; McGhee

and Felsenfeld 1983; Wingert and Von Hippel 1968). These biases can result in the imprecise trimming of nucleosome core DNAs, cutting into the core DNA by a few nucleotides, leaving behind a few linker DNA nucleotides, and the ability of MNase to cleave at the preferred sequences. Such preferences could result in "blurring" of the resolution of the ends of nucleosome core DNAs by a few nucleotides. One manifestation of this is the strong bias for A or T at the first and G or C at the second nucleotide position in the beginning of nucleosome core DNAs (Johnson et al. 2006). MNase also has the potential to cleave nucleotides within the core of the nucleosome itself when highly accessible sites are on the outer surface of the core thus de-enriching any mononucleosome core DNA library for cores which contain these preferred sites internally (McGhee and Felsenfeld 1983).

Genome browser and individual loci

Our current core read data set is of sufficient density to allow analysis of the nucleosome footprint at any unique locus in the genome. To facilitate the global browsing of unique loci we have converted the data to a genome browser format which can be viewed using the UCSC genome browser with the WS170 version of the *C. elegans* genome (http://genome.ucsc.edu/cgi-bin/hgGateway). There are nine custom tracks used to display our data and subsequent analysis. The first three tracks display the raw data from control and nucleosome core isolation experiments. Tracks two and three show the start of each placed nucleosome core read, with forward reads in blue and reverse reads in green. Each feature represents putative nucleosome cores extending 147 nucleotides with a thick bar of 50 nucleotides in length demarking one end of and the portion of the nucleosome core actual sequenced, followed by a thin bar ending in a line which extends an additional 97 nucleotides in the forward or reverse direction representing the portion of the genome protected by the entire idealized putative nucleosome core and its respective extrapolated opposite end. Multiple reads starting at the same nucleotide were collapsed into the same feature and colored according to number of contributing reads (1 read, 2 reads, 3-5

reads, 6-10 reads and greater than 10 reads). The control data (orange) was displayed in a similar fashion. The analysis tracks include coverage and nucleosome positioning stringency plots. At every nucleotide, we provide coverage values from the control (orange) and nucleosome experiments (blue and green). They can be used to evaluate the frequency of nucleosome instances at every position. The nucleosome positioning stringency track (purple) evaluates the degree of nucleosome positioning at every base pair (see below for formulaic description). Finally, the adjusted nucleosome coverage track (pink) provides the measure of abundance of nucleosome instances relative to control data to account for any local enzymatic and sequencing biases. When evaluating the nucleosome coverage of any particular locus, it is important to assess both the nucleosome and control read coverage. To aid in identifying regions of the genome with repetitive structure, we provided an additional browser track which shows masking of 25 base segments of genomic DNA which are not unique in *C. elegans* (Fire et al. 2006). A region devoid of nucleosome starts, but containing control starts would indicate a bona fide nucleosome depleted region, or be devoid of both nucleosome and control starts indicating a non-repetitive region recalcitrant to sequencing by the SOLiD technology (discussed below).

Using the nucleosome browser we see multiple loci with reproducibly positioned and regularly spaced nucleosome and areas of nucleosome eviction (Fig. 3A) and regions with no strong positional preference (Fig. 3B). The latter scenario is the predominant feature of *C. elegans* chromatin as assessed by our global analysis of neighboring nucleosomes (Start-to-Start and Start-to-End analysis see below).

Global analysis of nucleosome positioning and spacing in *C. elegans*

The dense nucleosome read coverage across the genome allowed us to perform a high resolution global analysis of *C. elegans* chromatin, searching in particular for features that could not be seen by lower density analysis. We derived a nucleosome read start set (1-pile data set) of 22,546,201 unique nucleosome core start site positions in the *C. elegans* genome by taking all the individual read start sites in the genome where one or more

nucleosome core reads are detected. We observed some degree of nucleosome positioning constraint detected by using the 1-pile data set to compare distances between read starts on different strands (Start-to-End distances) (Fig. 4A top). This positioning was evidenced by a major peak 147 nucleotides downstream from the start site representing the opposite end of another nucleosome being sequenced from the opposite direction, indicating that there is at least a subset of genomic sites where nucleosomes exhibit a preferential localization. Additionally, more subtle peaks are detected downstream of this major peak at a periodicity of ~175 nucleotides providing evidence for constrained positioning and spacing of neighboring nucleosomes in certain regions of the genome. Comparison of distances between starts of core reads on the same strand (Start-to-Start distances) revealed a consistent set of characteristics. (i) The major feature is the relative uniformity of the distribution indicating that nucleosome positions for most of the *C. elegans* are flexible, (ii) a series of peaks at ~175 nucleotides down-stream and at further 175-nucleotide intervals indicates some situations where positions are constrained and spacing is uniform (Fig. 4B top).

We can focus on regions of the genome where positioning of nucleosomes is partially or fully constrained by analyzing positions where at least five core reads define the same genome position. This "5-pile data set" contains 1,539,014 different start sites. On this data set, the "Start-to-End" analysis described above reveals an even stronger peak at ~150 base pairs, independently reinforcing the preferred positioning of these nucleosomes. This plot also provides evidence of preferred spacing of neighboring nucleosomes, since we observe strong additional peaks at intervals of ~175 base pairs (Fig. 4A bottom). Start-to-Start analysis on the 5-pile data set reveals a strong peak at ~175 base pairs (with echoes at 175 base pair intervals) again supportive of preferred spacing of neighboring nucleosomes (Fig. 4B bottom). Comparison of the analyses of the 5-pile versus the 1-pile data sets demonstrates an enrichment for positioned nucleosome signal (the amplitudes of the Start-to-End ~150 base pair peaks are 156% versus 26.4% of the baseline amplitude for the 5-pile versus the 1-pile data, respectively) and demonstrates that reproducible positioning tends to occur in neighboring clusters of

nucleosomes (the amplitudes of the Start-to-End neighboring nucleosome peaks are 34.6% versus 6.58% and the amplitudes of the Start-to-Start neighboring nucleosome peaks are 45.4% for the first and 19.1% for the second versus 9.18% and 3.66% for the respective peaks in the 5-pile versus the 1-pile analyses, respectively).

In addition to the strong evidence for some degree of preferential positioning and uniform spacing of neighboring nucleosomes in the *C. elegans* genome, the Start-to-End and Start-to-Start analyses in both the 1-pile data and the 5-pile data sets reveal features that are consistent with rotational setting of nucleosomes. There is a striking 10-base periodicity of subpeaks that surround the major 147 base peak in the Start-to-End analysis (Fig. 4A). These peaks are very pronounced moving upstream and are much more subtle moving downstream of the major peak in the Start-to-End graph. The same evidence, consistent with rotational setting, exists in the Start-to-Start data with subpeaks extending out from position zero and continuing at a period of 10 bases through position 130, thereafter being lost (Fig. 4B). These Start-to-Start subpeaks are somewhat blurred by a subtle but persistent three-base periodicity that is superimposed on the underling data (presumably reflecting codon biases in protein-coding regions and their effects on true and/or measured nucleosome positions). While these data do argue in favor of rotational positioning/setting of nucleosomes, there are other mechanical explanations of these 10 base periodic subpeaks which will be discussed below.

We used a separate analysis to quantify the degree of positioning as a fraction of total genomic sequence. This analysis is based on a compiled list of predicted dyad (center of nucleosome) sites (obtained by adding [for forward reads] or subtracting [for reverse reads] 73bps from the coordinate representing each read). Positioning stringency for any site was calculated by counting putative dyads falling within a 23bp window (the site in question plus 11 nt on each side) and dividing by the total number of infringing dyads within the flanking 150bps (a 301bp window). Sites were only considered as potentially positioned if both a forward-based and a reverse-based dyad were present within the 23bp window. The analysis is illustrated for an exemplary locus in Supplemental Figure S2. Figure 5 shows the portion of the genome occupied by positioned nucleosomes as a

function of the stringency (Blue Line). Control reads processed using an identical analysis are shown (Red Line), with the difference between the nucleosome and control data (Green Line) indicating the portion of the genome with putative positioned nucleosomes that is not due to random sampling. For low values of the cutoff score (e.g. less than 20%), the differences between the control and experimental curves may or may not be significant. The largest difference between the experimental and control curves (29.4%) is obtained only at a very low stringency (13.5%; so >85% of nucleosomes in such regions don't adopt the position of interest). At more significant stringency cutoffs of 20%, 25%, 33%, 50%, 66.5% and 75% the net portion of the genome with positioned nucleosomes (experimental versus control) is calculated as 16.1% (20.7% - 4.6%), 7.49% (8.64% - 1.15%), 2.09% (2.30% - 0.21%), 0.25% (0.27% - 0.02%), 0.036% (0.040% - 0.004%) and 0.014% (0.015% - 0.001%) respectively with the nucleosome to control differences in parentheses. Concordant results were obtained when the dyad position window was narrowed down from +/-11 bp to +/-5 bp (Supplemental Fig. S3).

The *C. elegans* genome contains approximately 20000 genes. It was of considerable interest to address where positioned nucleosomes occur relative to transcriptional start sites. To assess the relationship between these sites and promoter sequences, we first needed at least an approximation to the list of *C. elegans* transcription start sites. Note that only a handful of transcriptional start sites are known for *C. elegans*, since trans-splicing tends to mask true initiation sites (Blumenthal 1995). As an alternative approximation, we have used defined and predicted translational start sites, which in many cases (due to the general brevity of *C. elegans* out-trons and 5' UTRs) will provide a close mark for transcriptional initiation. We used moderate positioning stringency cutoffs of 20% and 33% to map the dyads of these putative positioned nucleosomes relative to the ATG of the translational start site for all RefSeq genes (UCSC genome browser) (Fig. 6). We observe a significant enrichment of positioned nucleosomes covering the translational start site but centered slightly downstream (reaching maxima of 2.37 and 1.85 fold enrichment for 20% and 33% data relative to the

mean across the respective plots). Notably, there is a pronounced depletion of positioned nucleosomes at ~100bps upstream of the start of the coding sequence (bottoming out with 6.13 and 2.40 fold depletion for 20% and 33% data relative to the mean across the respective plots). Similar results were observed genome-wide relative to transcriptional start sites in yeast (Albert et al. 2007; Lee et al. 2007; Shivaswamy et al. 2008; Whitehouse et al. 2007; Yuan and Liu 2008; Yuan et al. 2005) as well as in human promoters (Lin et al. 2007; Schones et al. 2008). The width of the enrichment peak and the lack of complete absence of positioned nucleosomes in the trough likely reflects a combination of variation in transcriptional activity between tissues, real variation in setting relative to the transcription start within individual tissues, and our use of translational instead of transcriptional start sites (since the former is not a fixed distance from the latter there will be some 'fuzzing' of the resulting positioning data). We also note a series of interesting enrichment peaks both upstream and especially downstream of the above mentioned features.

Global sequence characteristics of nucleosome core segments

The 44.0 million sequence reads represent ~22.5 million unique 147-base pair putative nucleosome core sequences, which we designate the "22.5M core sequence data set". These provide a data-intensive framework to evaluate models for sequence-based nucleosome positioning. Many previous analyses of smaller nucleosome core data sets have reported a periodicity of k-mer usage within the DNA of the core of the nucleosome (e.g., Albert et al. 2007; Ioshikhes et al. 1996; Johnson et al. 2006; Muyldermans and Travers 1994; Satchwell et al. 1986), which is thought to be important in the positioning of nucleosomes due to the ability of certain regularly-spaced di-nucleotides to more easily facilitate the extreme curvature of the double stranded DNA around the histone octamer (Luger et al. 1997). Consistent with these previous studies, a simple di-nucleotide analysis of our core sequence data set reveals an oscillating ~10-base periodicity of AA/TT with a counter phased ~10-base periodicity for GC.

Our 22.5M core sequence data set is sufficiently large to permit analysis of the tri-, tetra-, penta-, and even hexa-nucleotide (k-mer) prevalence as a function of position in nucleosome cores from *C. elegans*. To this end we analyzed the over- or under-representation of each of the 64, 256, 1024, and 4096 tri-, tetra-, penta-, and hexa-nucleotide combinations (words) at each position in the nucleosome core relative to their prevalence in the control sequence data set inferred from the control read data set. To render the results in a visually accessible format, we generated intensity plots ('heat maps') composed of grids in which each column represents a position in the nucleosome core and each row represents an individual k-mer word. At any point on the plot, the over-representation or under-representation of a particular k-mer may be assessed by the intensity of the color with brightest yellow being a 1.3 fold or greater over-representation, black being no enrichment and brightest cyan being a 1.3 fold or greater under-representation of the k-mer relative to the control set, with the intermediate colors representing a logarithmic scale within this range. Separate plots were generated for 2-mer, 3-mer, 4-mer (Fig. 7A, B), 5-mer (Supplemental Figs. S4 and S5), and 6-mer (available on request) using both the 1-pile data (inclusive) and the 5-pile data (positioned-nucleosome-enriched set). An alternative means of representing the data that uses a wider color spectrum for enhanced visualization of subtle enrichment patterns is shown in Supplemental Figure S6. For each plot, any individual k-mer may be evaluated across the entire nucleosome core by reading the intensities across an entire row; conversely an individual position within the core may be evaluated across all k-mers by reading down an entire column. Expanding the analysis to include the 500 nucleotides flanking the start of the core sequences allowed us to look at the k-mer frequencies in the putative linker regions and also to detect neighboring nucleosomes as evidenced by the fainter enrichment/de-enrichment patterns seen as echoes of the nucleosome core pattern both upstream and downstream of the core (Fig. 7C, D).

A visual inspection of these plots reveals a variety of distinctive features of nucleosome structure, including clear indications of the core termini, a periodic "shadow" of the individual helical turns of DNA that

wrap around the core, and a uniquely structured dyad that sits in the middle of the core region. This detailed set of position-specific sequence biases should serve to refine current models for sequence-specific prediction of nucleosome seating preferences.

One striking feature of the incidence plots is a cluster of strong anomalies in virtually all k-mers at the beginning of the nucleosome core (around position 0). These anomalies can be explained by very localized sequence specificity in micrococcal nuclease trimming adjacent to the nucleosome core (e.g., an ability of MNase to trim terminal individual bases at nucleosomal termini much more efficiently if the following bases are not G/C base pairs). This bias is minimized due to use of the control data to normalize the heat maps. The asymmetric character of the anomalies at the cleavage site (i.e., lack of a similar set of anomaly at positions 146+/-2 in the plot) argues against any definitive relationship between the cleavage-terminus sequence bias and nucleosome positioning. Conversely, the symmetric appearance of other features (such as the internal periodicity and dyad sites) is indicative of a fundamental relationship with nucleosome position.

The large dataset allowed us to also address the extent to which k-mer strings of different lengths contributed to sequence preferences in nucleosomal positioning. In particular, we asked to what extent any preferences in incidence of any given (k+1)-mer could be explained by preferences of the constituent (k)-mers. The resulting analysis, using Markov formalism, indicates a substantial contribution to nucleosomal positioning from 3-mer (as compared to 2-mer), with additional contributions of higher order (e.g. 4-mer) words (see supplemental analysis and Supplemental Fig. S7).

DISCUSSION

To sample the global spectra of allowed positions for nucleosome occupancy within the *C. elegans* genome *in vivo*, we have isolated (using MNase), sequenced (using SOLiD sequencing technology) and uniquely placed more than 44.0 million individual putative nucleosome core DNAs from a pan-cellular and mixed developmental stage sample. The density of this analysis provides a high resolution map of nucleosome positioning in the entire unique portion of the genome allowing scrutiny of any and all loci within this portion of the genome. We have made this map universally accessible through the UCSC genome browser.

We would envision the browser-based map being useful based on virtually any detailed analysis of association between DNA sequence and regulatory function *in vivo*. Circumstances in which a specific transcription factor has been shown to interact with a defined DNA sequence that is present in many places in the genome naturally lead to questions of the accessibility of the different putative sites in chromatin. Circumstances in which a region of the genome is a particularly active or inactive participant in modulating gene expression or in other activities of DNA (e.g., recombination, replication, repair, etc) similarly lead to questions of protected accessibility. While not all such circumstances would necessarily result from the simple nucleosomal coverage pattern, there is certainly strong precedent suggesting that this pattern will be of great use in evaluating and understanding the activities of individual loci.

Beyond the examination of individual loci, these data allow assessment of the global state of nucleosome positions both relative to the underlying DNA sequence and in relation to other neighboring nucleosomes. The major feature we observed with *C. elegans* chromatin has been the lack of universal sequence-dictated nucleosome positioning for a substantial majority of the genome. We note a number of differences between this mapping project and the very interesting analysis that has been reported in yeast. First, *C. elegans* is a multicellular eukaryote with a diverse variety of cell identities and tissue types. A mixed stage population of *C. elegans* thus represents a wider diversity of physiological states than is observed in the relatively uniform yeast cell populations. This allows a critical test of sequence-dictated positioning as

compared to uniformity in chromosome organization as a reflection of common metabolic states. Second, the sequencing-based analysis can provide discrete quantification of degrees of positioning in terms of fractions of nucleosomes that occupy preferred positions in a given genomic region. Critically, this analysis accurately reports situations in which a degree of constraint in a given region is superimposed on a background of relatively random positioning. Finally, the sequence based analysis is of considerable value in interpreting genomic regions where several different positioning configurations co-exist; such situations may produce relatively constant overall occupancy and thus be difficult to distinguish using standard microarray procedures from more constant "random positioning". Certainly sequencing and microarray approaches each offer unique advantages to chromatin analysis and we anticipate that data from both approaches will be of considerable utility in generating comprehensive biological surveys of chromatin dynamics.

Although flexibility in positioning of nucleosomes is a major feature for large portions of the genome, we see clear examples of (and bulk evidence for) regions with strong preferences in precise positioning. On a bulk scale, the population of sites with preferred positioning is most evident from the ~146 nucleotide "Start-to-End" peak in the analysis in Figure 4. Both bulk and individual locus analysis also show some degree of uniform spacing of adjacent and preferentially positioned nucleosomes, with a spacer length of ~28 nt. This spacing is somewhat larger than the 14-21nt spacing reported from previous whole genome biochemical analyses of (Dixon et al. 1990; Johnson et al. 2006). One intriguing possibility is that the positionally constrained subset of the genome exhibits a longer spacer length than average for the larger portion of the genome in which relative spacing of nucleosomes is regular (e.g. Figure 1B) but absolute positions show little or no constraint.

An additional salient feature of the bulk analysis (Fig. 4) is a modest 10 base periodicity in putative nucleosome residency along the genome. Although intriguing in that the 10-base periodic subpeaks can be interpreted as resulting from rotational positioning (Lu et al. 1994), it is also possible that these peaks could

result from biases in the sequencing technique and/or MNase digestion (McGhee and Felsenfeld 1983). These non-physiological explanations seem less likely (1) due to a lack of such periodicity in the control runs, which were sequenced with the same technology arguing against sequencing bias, and (2) due to the extent of periodicity across the entire 147 bases covering the putative nucleosome (and not just the region adjacent to [or 10n bases removed from] the MNase cleavage site).

Conclusion:

Our data and analysis indicate that multiple scenarios of nucleosome positioning co-exist for *C. elegans*. From the browser-based analysis of individual loci and from bulk informatic analysis, we know that there are sites in the genome where nucleosomes have strongly preferred positions and near-uniform spacing, and that there are sites where nucleosomes are preferentially positioned at one site but not organized into a reproducible local array. From the strong laddering pattern observed from partial micrococcal nuclease digestion combined with the preponderance of genomic sites where multiple positions were observed in our sequence survey, we know that variable absolute positioning is a common feature even in regions where the spacing between nucleosomes is constrained. Finally, we have observed that the degree of local plasticity of nucleosome seating is a characteristic feature of individual localities in the *C. elegans* genome and we look forward to further studies in which the details of chromatin architecture and nucleosome regularity can be associated with diverse nuclear functions on a genome wide scale.

METHODS

Isolation of mononucleosome core DNA fragments

Mixed stage, wild-type (N2) *C. elegans* were cultured on DH5α *E. coli*, flash frozen with liquid nitrogen in 0.34 M sucrose/Buffer A (15 mM Tris-HCl at pH 7.4, 15 mM NaCl, 1mM DTT, 60 mM KCl, 0.5 mM spermidine,

0.15 mM spermine, 25 mM bisulfite) and ground to a fine powder in liquid nitrogen using a mortar and pestle. After thawing on ice, $CaCl_2$ and micrococcal nuclease (Roche) resuspended at 300 U/ul were added for final concentrations of 1 mM and 25 U/ul respectively followed by incubation at 16ºC for 12 minutes to liberate the mononucleosome cores. The reaction was stopped by the addition of an equal volume of worm lysis buffer (0.1 M Tris-HCl at pH 8.5, 0.1 M NaCl, 50 mM EDTA, 1% SDS) and proteins were removed by treating with one-tenth volume proteinase K (20 mg/ml in TE at pH 7.4) for 45 minutes at 65 ºC followed by phenol, phenol/chloroform and chloroform extractions and ethanol precipitation. After RNase treatment and phenol/chloroform, chloroform extraction, separation the micrococcal nuclease-digested DNA into mono-, di-, tri- and multi-nucleosome DNAs was done using a 2% UltraPure Agarose (Invitrogen) gel run at 100 V for 4 hours and DNA from the mononucleosome DNA band was extracted from the gel using the QIAquick Gel Extraction Kit (Qiagen) following the standard protocol with the exception of allowing the isolated gel sample to incubate in Buffer QG at room temperature until dissolved.

End repair and linker ligation

The ends of isolated mononucleosome core DNAs were processed by treating 0.5 ug of the DNA sample with T4 DNA polymerase (New England Biolabs) at 12ºC for 30 minutes followed by purification (MiniElute reaction cleanup kit, Qiagen). T4 kinase treatment (New England Biolabs) for 30 min at 37ºC was then used to add a 5' phosphate and remove the 3' phosphate, followed by purification as above. Linkering was accomplished by addition of 1ul each of the appropriate SOLiD sequencing linkers (100 uM each), 3.6 ul 5x T4 ligase buffer (Invitrogen), and 2.4 ul HC T4 ligase (Invitrogen, 5 U/ul) with a 16ºC overnight incubation. The ligation reaction was separated on a 1.8% agarose gel, and the relevant band isolated (MiniElute gel-extraction kit; Qiagen).

Library preparation

Sequencing runs were conducted using cycled ligation sequencing on a SOLiD Analyzer (Applied Biosystems, Beverly, MA) and were conducted using short fragment libraries. For the *C. elegans* micrococcal nuclease library, DNA fragmented into mononucleosomal units of ~147 bp and ligated to unique forward (P1) and reverse (P2) adaptors (P1, 41 bp: 5'-CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG AT-3'; P2, 23 bp: 5'-AGA GAA TGA GGA ACC CGG GGC AG-3') was used. For control libraries, genomic DNA (50 ug) from *C. elegans* (N2) in 0.34 M sucrose/Buffer A/1X BSA (New England Biolabs) /1mM $CaCl_2$ was digested with 200 units of micrococcal nuclease (Roche) (0.4 units/ul final concentration) in a total volume of 500ul for 10 minutes at 23°C. The digestion was stopped by addition of 10ul 0.5M EDTA, followed by ethanol precipitation. The digested DNA was run on a 2% agarose gel and the smear of DNA fragments from400bp-850bp was excised from the gel and purified using the QIAquick Gel Extraction Kit (Qiagen) as noted above. The size-selected fragment molecules were then subjected to SOLiD paired-end library preparation and sequencing (Applied Biosystems Inc).

Bead preparation

Double-stranded fragment libraries were diluted to 50 pg/ul into low 1x TE buffer (1 mM Tris-OAc, 1 mM EDTA) and supplemented into a 2800 ul PCR aqueous phase containing: 1x PCR Gold buffer (Applied Biosystems), 3.5 mM dNTP mix, 25 mM, 10 uM P1 primer (IDT, Inc., Coralville, IA; 5'-CCT CTC TAT GGG CAG TCG GTG AT-3'), 3 mM P2 primer (IDT, Inc., 5'-CTG CCC CGG GTT CCT CAT TCT CT-3'), 560-22400 pg library template, 150 U AmpliTaq Gold UP (Applied Biosystems) and 1.7 billion one micron covalent P1 DNA-coated paramagnetic beads. Briefly, covalent P1 DNA beads were prepared by covalent attachment of 5'-aminated P1 oligonucleotides (P1, 41 bp: /5AmMC6/ CCA CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG AT) (IDT, Inc.) to one micron carboxylic acid-coated paramagnetic particles

(Dynal/Invitrogen). The PCR aqueous phase reaction mix containing covalent P1 beads was then added to a 50-mL vial containing 10-mL of oil phase (SOLiD reagent, Applied Biosystems). Emulsions were generated by vortexing at 2100 rpm for 9 min, 53 sec in an Emulsomatic Device (SOLiD System, Applied Biosystems). Emulsions were transferred to 96-well plates and cycled for 40 cycles in a Gene Amp PCR system 9700 (SOLiD System, Applied Biosystems). Post-amplification, emulsion beads (ePCR beads) were broken with butanol (Sigma) and washed with 1x TE buffer (10 mM Tris-OAc, 1 mM EDTA) supplemented with 0.01% Triton X-100 (Sigma). Washed ePCR beads were then enriched for template-positive beads by hybridization with P2-coated capture beads (SOLiD reagent, Applied Biosystems). Template-positive beads were collected and extended in the presence of terminal transferase and Bead Linker (SOLiD reagents, Applied Biosystems). Extended beads were then deposited and covalently attached onto 25mm x 75mm SOLiD slides at >20,000 beads per panel (SOLiD reagent, Applied Biosystems). Template bead slides were then loaded onto a SOLiD Analyzer (Beverly, MA) and cycled ligation sequencing using the SOLiD Sequencing System was performed.

Mapping of the SOLiD sequencing data

Sequence reads from both nucleosome fragment core isolation and control experiments were aligned to the reference genome using standard SOLiD mapping pipeline. Nucleosome experiment reads were aligned allowing up to 6 mismatches out of 50 bps. Reads that matched to a single position at the least number of mismatches and for which the second best match contained at least two more mismatches, contributed to the final nucleosome data. Control experiment sequences were comprised of two 25 bp reads derived from the fragment ends which were matched to the reference genome using standard SOLiD paired-end rescue pipeline allowing up to 6 mismatches between both ends and falling in the correct size range (400-850 bps). All mapping was performed in the SOLiD color space corresponding to dinucleotide encoding of the sequenced DNA.

ACKNOWLEDGEMENTS

We thank Hiram Clawson and Jim Kent for their help with compiling and hosting the custom browser tracks, members of the Fire lab for their help and suggestions over the course of this work, and acknowledge National Institutes of Health (Grant NIGMS RO1-GM37706 to A.F) and American Cancer Society (postdoctoral fellowship PF-05-121-01-DDC to S.M.J.) for financial support.

REFERENCES

Albert, I., T.N. Mavrich, L.P. Tomsho, J. Qi, S.J. Zanton, S.C. Schuster, and B.F. Pugh. 2007. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446:** 572-576.

Blumenthal, T. 1995. Trans-splicing and polycistronic transcription in *Caenohabditis elegans*. *Trends in Genetics* **11:** 132-136.

Davey, C.A., D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. 2002. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *Journal of Molecular Biology* **319:** 1097-1113.

Dixon, D.K., D. Jones, and E.P. Candido. 1990. The differentially expressed 16-kD heat shock genes of *Caenorhabditis elegans* exhibit differential changes in chromatin structure during heat shock. *DNA & Cell Biology* **9:** 177-191.

Dressman, D., H. Yan, G. Traverso, K.W. Kinzler, and B. Vogelstein. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 8817-8822.

Embleton, M.J., G. Gorochov, P.T. Jones, and G. Winter. 1992. In-cell PCR from mRNA: amplifying and linking the rearranged immunoglobulin heavy and light chain V-genes within single cells. *Nucleic Acids Research* **20:** 3831-3837.

Fire, A., R. Alcazar, and F. Tan. 2006. Unusual DNA structures associated with germline genetic activity in *Caenorhabditis elegans*. *Genetics***:** genetics.106.057364.

Ghadessy, F.J., J.L. Ong, and P. Holliger. 2001. Directed evolution of polymerase function by compartmentalized self-replication. *Proceedings of the National Academy of Sciences of the United States of America* **98:** 4552-4557.

Horz, W. and W. Altenburger. 1981. Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic Acids Research* **9:** 2643-2658.

Ioshikhes, I., A. Bolshoy, K. Derenshteyn, M. Borodovsky, and E.N. Trifonov. 1996. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *Journal of Molecular Biology* **262:** 129-139.

Johnson, S.M., F.J. Tan, H.L. McCullough, D.P. Riordan, and A.Z. Fire. 2006. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Research* **16:** 1505-1516.

Kornberg, R.D. 1974. Chromatin structure: a repeating unit of histones and DNA. *Science* **184:** 868-871.

Kwoh, D.Y., G.R. Davis, K.M. Whitfield, H.L. Chappelle, L.J. DiMichele, and T.R. Gingeras. 1989. Transcription-based amplification system and detection of amplified human immunodeficiency virus type 1 with a bead-based sandwich hybridization format. *Proceedings of the National Academy of Sciences of the United States of America* **86:** 1173-1177.

Kwoh, D.Y. and T.J. Kwoh. 1990. Target amplification systems in nucleic acid-based diagnostic approaches. *American Biotechnology Laboratory* **8:** 14-25.

Lee, W., D. Tillo, N. Bray, R.H. Morse, R.W. Davis, T.R. Hughes, and C. Nislow. 2007. A high-resolution atlas of nucleosome occupancy in yeast. **39:** 1235-1244.

Lin, J.C., S. Jeong, G. Liang, D. Takai, M. Fatemi, Y.C. Tsai, G. Egger, E.N. Gal-Yam, and P.A. Jones. 2007. Role of Nucleosomal Occupancy in the Epigenetic Silencing of the MLH1 CpG Island. *Cancer Cell* **12:** 432-444.

Lu, Q., L.L. Wallrath, and S.C. Elgin. 1994. Nucleosome positioning and gene regulation. *Journal of Cellular Biochemistry* **55:** 83-92.

Luger, K., A.W. Mader, R.K. Richmond, D.F. Sargent, and T.J. Richmond. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389:** 251-260.

Macevicz, S.C. 1998. DNA sequencing by parallel oligonucleotide extensions. Lynx Therapeutics, Inc., United States of America.

McGhee, J.D. and G. Felsenfeld. 1983. Another potential artifact in the study of nucleosome phasing by chromatin digestion with micrococcal nuclease. *Cell* **32:** 1205-1215.

McKernan, K., A. Blanchard, L. Kotler, and G. Costa. 2006. (WO/2006/084132) Reagents, methods, and libraries for bead-based sequencing. Agencourt Bioscience Corp., McKernan, K., Blanchard, A., Kotler, L., Costa, G., US.

Muyldermans, S. and A.A. Travers. 1994. DNA sequence organization in chromatosomes. *Journal of Molecular Biology* **235:** 855-870.

Noll, M. 1974. Subunit structure of chromatin. *Nature* **251:** 249-251.

Peckham, H.E., R.E. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W.S. Noble, K. Struhl, and Z. Weng. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res.* **17:** 1170-1177.

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.

Schones, D.E., K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132:** 887-898.

Segal, E., Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J.P. Wang, and J. Widom. 2006. A genomic code for nucleosome positioning. *Nature* **442:** 772-778.

Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309:** 1728-1732.

Shivaswamy, S., A. Bhinge, Y. Zhao, S. Jones, M. Hirst, and V.R. Iyer. 2008. Dynamic Remodeling of Individual Nucleosomes Across a Eukaryotic Genome in Response to Transcriptional Perturbation. *PLoS Biology* **6:** 618-630.

Tawfik, D.S. and A.D. Griffiths. 1998. Man-made cell-like compartments for molecular evolution. *Nature Biotechnology* **16:** 652-656.
Whitehouse, I., O.J. Rando, J. Delrow, and T. Tsukiyama. 2007. Chromatin remodelling at promoters suppresses antisense transcription. **450:** 1031-1035.
Wingert, L. and P.H. Von Hippel. 1968. The conformation dependent hydrolysis of DNA by micrococcal nuclease. *Biochimica et Biophysica Acta* **157:** 114-126.
Yuan, G.C. and J.S. Liu. 2008. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Computational Biology* **4:** e13.
Yuan, G.C., Y.J. Liu, M.F. Dion, M.D. Slack, L.F. Wu, S.J. Altschuler, and O.J. Rando. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309:** 626-630.

FIGURE LEGENDS

Figure 1. Possible patterns of nucleosome positioning. Diagrams show four possible combinations for the local and relative position of nucleosomes. (a) Nucleosomes occupy reproducible positions in all cells and are regularly spaced relative to one another. (b) Nucleosomes are regularly spaced relative to one another, but have no strong tendency to occupy the same position in different cells. (c) Nucleosomes occupy the same position in all cells but are not regularly spaced relative to one another. (d) Nucleosomes neither occupy the same position in all cells nor have regular spacing relative to one another.

Figure 2. An outline of SOLiD sequencing technology.

Figure 3. UCSC Genome Browser displaying custom tracks of the *C. elegans* nucleosome position map. (A) Browser shot of the *mcm-5* gene locus with several preferentially positioned nucleosomes in an organized array (indicated by red arrows). Nucleosome depleted regions upstream of the ATG are highlighted by brown boxes, a common feature of many actively transcribed genes as noted by several groups (Albert et al. 2007; Lee et al. 2007; Lin et al. 2007; Schones et al. 2008; Shivaswamy et al. 2008; Whitehouse et al. 2007; Yuan and Liu

2008; Yuan et al. 2005). (B) A representative locus (displaying gene C16C10.3) in the genome demonstrating lack of nucleosome positioning particularly across a 3.5 kb stretch (highlighted by the blue box) possibly due to random nucleosome occupancy. In both panels (A&B) the top three tracks display raw data from control (orange) and nucleosome (forward blue and reverse green) experiments. Tracks two and three represent 147 nt stretches of putative nucleosome cores from forward (blue) and reverse (green) reads with the first 50 sequenced nucleotides indicated by a thick portion of the track feature. Track one displays both forward and reverse control data (orange) in a similar fashion as the nucleosome data. The nucleosome core and genomic DNA control reads have been collapsed and colored such that multiple reads starting at the same nucleotide are represented by a single feature that varies in hue from lightest to darkest depending on the number of instances. Lightest to darkest, hues correspond to the following categories: 1 read instance, 2 read instances, 3-5 instances, 6-10 instances and greater than 10 instances, respectively. Tracks four and five (coverage of nucleosome control, sense/antisense strand reads) display the coverage by 147nt stretches from the control data. Tracks six and seven (coverage of mononucleosomal fragments, sense/antisense strand reads) show coverage by putative nucleosome cores inferred from reads that map to sense (blue) or antisense (green) strand of the reference genome. Tack eight (purple) evaluates nucleosome positioning stringency at every bp varying between 0 and 1, such that 1.0 corresponds to 100% positioning and 0.0 corresponds to no positioning or insufficient data. Track nine (adjusted nucleosome coverage, pink) displays nucleosome coverage (on a log of 2 scale) relative to control data to account for sequencing and enzymatic biases. Areas falling below 0.0 are on average more depleted for nucleosomes while areas above 0.0 have increased frequency of nucleosome instances. +/- 1 indicates a two-fold increase or depletion of putative nucleosome cores at that position.

Figure 4. Global analysis of positional relationships between individual pan-cellular and neighboring nucleosomes. (A) Start-to-End distances for reads mapped to opposite strands. The graph shows total pairs of

nucleosomes with a Start-to-End distance corresponding to the value on the x-axis. The dominant peak corresponding to position 146 (the 147th base from the start of the read) demonstrates reproducible positioning of nucleosomes at the same loci across cells. The insets graphs show the same data but highlight the 175 base periodicity reflecting the phasing on neighboring nucleosomes (green inset graph) and the 10 base periodicity indicative of rotational positioning of nucleosomes (blue inset graph). (B) Start-to-Start distances of reads mapped to the same strand. The graph shows total pairs of nucleosomes with Start-to-Start distance corresponding to the value on the x-axis. A subtle broad peak is located at approximately base 175 with echoes of this peak at a periodicity of 175 bases (green inset graph). A 10 base periodicity (blue inset graph) is also seen extending out from the start site, but is somewhat obscured by an underlying 3 base periodicity (red inset graph). In both panels A & B the top set of graphs are generated form the total data (1-pile data) and the bottom graphs are generated from data enriched for positioned nucleosomes (5-pile data). This enrichment results in a greater relative amplitude of the major signal for reproducibility in absolute nucleosome positioning [(A bottom graph) major peak at 146] and in relative positioning [(A & B bottom) green inset graphs], but slightly decreases the resolution of the 10 base periodicity [(A & B bottom) blue inset graphs] and makes the 3 base periodicity [(B bottom) red inset graph] less prominent due to greater noise in the signal from the smaller data set (only 1/28th the size of non-enriched data set).

Figure 5. Portion of the *C. elegans* genome with positioned nucleosomes. The percent of the genome (vertical axis) that falls above as specified positioning stringency cutoff (horizontal axis). The blue line is obtained from the nucleosome data, the red line is obtained from the negative control non-nucleosomal data (background) and the green line indicating the difference between the two (net positioning). The inset graph is the same data expanded between the 40% and 65% stringency cutoff levels.

Figure 6. Positioned nucleosomes relative to the translation start site. Using the positioned nucleosome data from 20% (pink) and 33% (blue) positioning stringency cutoffs, the number of positioned nucleosome dyads (vertical axis) is plotted relative to the ATG of the translational start sites of all annotated RefSeq genes (UCSC genome browser) (horizontal axis). The grey plot in the background depicts the nucleosome coverage from the complete data relative to ATG over the entire genome.

Figure 7. Over and under representation of oligonucleotide words in and around nucleosome cores. The over- (yellow) and under-representation (cyan) of k-mer words in the nucleosome core is displayed for every position within the core. Each column represents the position in or around the nucleosome core starting with the position 40 nucleotides upstream and ending with the position 220 nucleotides downstream of the start position of the core as indicated by the black numbers -40 and +220. The putative start and end positions of the core are indicated with the red numbers 0 and +146 respectively, and black dots demark positions 20 nucleotides apart and the black bracket indicates the position of the putative nucleosome dyad. Each row represents one of the 16 possible 2-mers, 64 possible 3-mers and 256 possible 4-mers with the k-mer key to the left representing A in green, C in blue, G in yellow and T in red. Thus the key for the first row indicates AA as two green boxes, the second row indicates AC as a green and then a blue box, the third row indicates AG as a green and then a yellow box, the fourth row indicates AT as a green and then a red box, etc. The over-, or under-representation of each dinucleotide may be assessed at any position in or around the nucleosome core by looking at the intersection of any column (position relative to the start of the nucleosome core) and any row (individual dinucleotide word) with the color indicating over- (yellow) or under-representation (cyan) relative to the control value (black indicating no enrichment over sequenced control DNA). (A & B) The same depiction of data derived from the 1-pile or 5-pile data sets respectively. (C & D) The same 4-mer analysis as in (A & B) expanded to show 500 positions both upstream and downstream of the nucleosome core start site. The

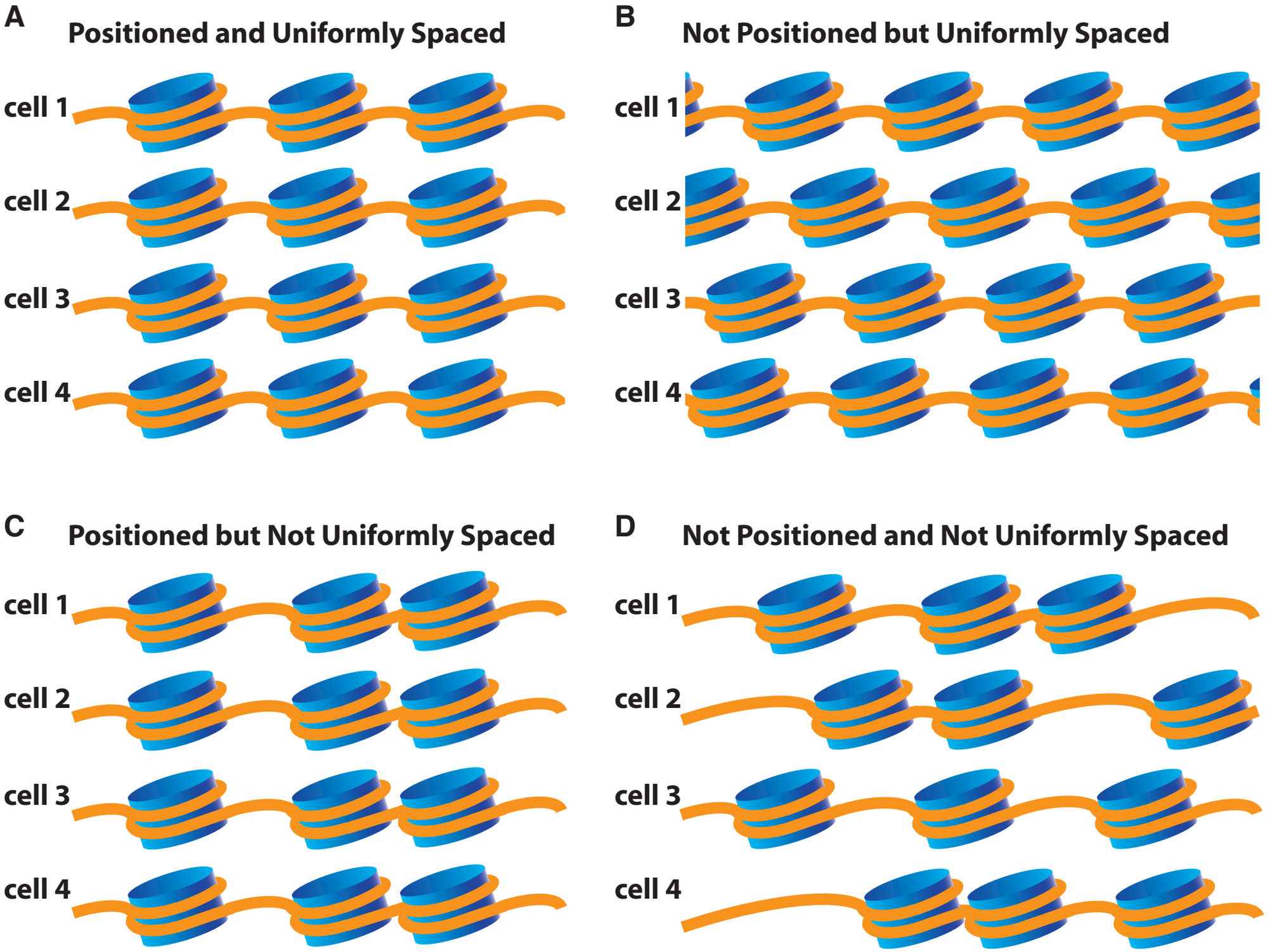
magnification in (C & D) is reduced to allow visualization of all the positions. In all panels the fold enrichment scale is the same ranging from 0.75 to 1.34 fold enrichment and the color range is shown to the far right of panels (A & B). Additionally, a cartoon depiction of the position of the nucleosome core along with the linker region relative to the graphical data is displayed at the bottom of (A & B) and at the bottom of (C & D) scaled for panels. In all panels, sequencing and enzymatic biases affect the representation of k-mer words between positions -10 and +25, while the rest of the plot is expected to be free of these biases.

Figures and Tables

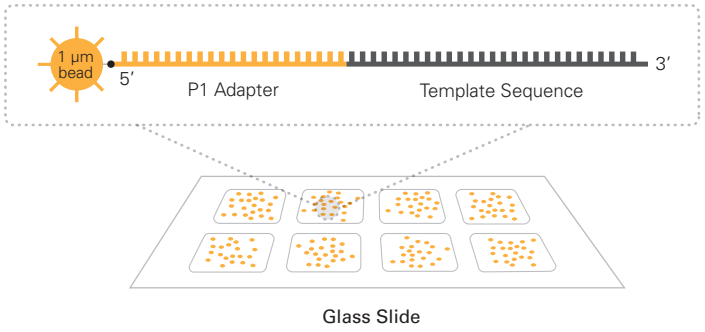
Table 1.

SOLiD read stats	**nucleosome**	**control**
total reads	107422570	NA
total matches (WS170+k12)	56384632	51916926
single best matches (WS170+k12)	45048550	51916926
single best matches 0mm (WS170+k12)	10921712	27926058
single best matches 1mm (WS170+k12)	9229265	9524044
single best matches 2mm (WS170+k12)	7708837	5696731
single best matches 3mm (WS170+k12)	6544743	3753960
single best matches 4mm (WS170+k12)	5668352	2526216
single best matches 5mm (WS170+k12)	4975641	1609135
single best matches 6mm (WS170+k12)	NA	880782
E. coli (k12) single best matches	1057009	23692
single best matches 0mm (k12)	191408	12035
single best matches 1mm (k12)	205929	4238
single best matches 2mm (k12)	192946	3286
single best matches 3mm (k12)	173107	1940
single best matches 4mm (k12)	154777	1197
single best matches 5mm (k12)	138842	652
single best matches 6mm (k12)	NA	344
C. elegans (WS170) single best matches	43991541	51893234
single best matches 0mm (WS170)	10730304	27914023
single best matches 1mm (WS170)	9023336	9519806
single best matches 2mm (WS170)	7515891	5693445
single best matches 3mm (WS170)	6371636	3752020

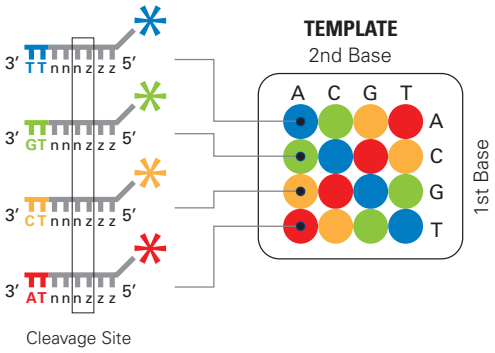
single best matches 4mm (WS170)	5513575	2525019
single best matches 5mm (WS170)	4836799	1608483
single best matches 6mm (WS170)	NA	880438


A
Positioned and Uniformly Spaced
cell 1
cell 2
cell 3
cell 4
B
Not Positioned but Uniformly Spaced
cell 1
cell 2
cell 3
cell 4
C
Positioned but Not Uniformly Spaced
cell 1
cell 2
cell 3
cell 4
D
Not Positioned and Not Uniformly Spaced
cell 1
cell 2
cell 3
cell 4

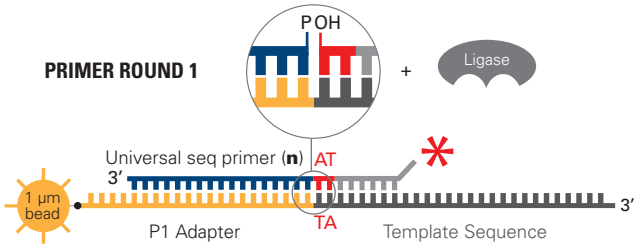
SOLiD™ Substrate



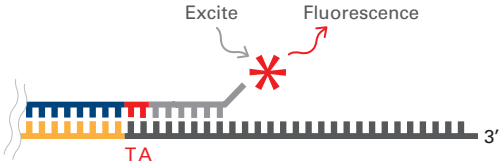
Di base Probes



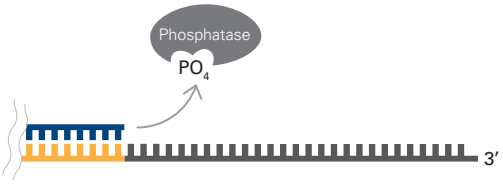
1. Prime and Ligate



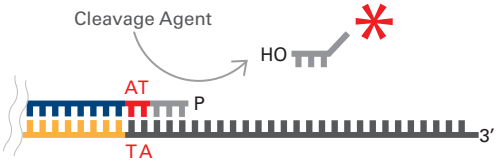
2. Image



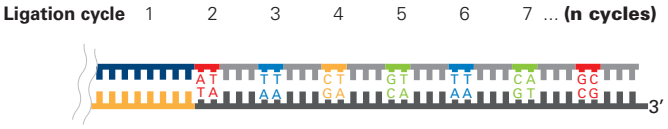
3. Cap Unextended Strands



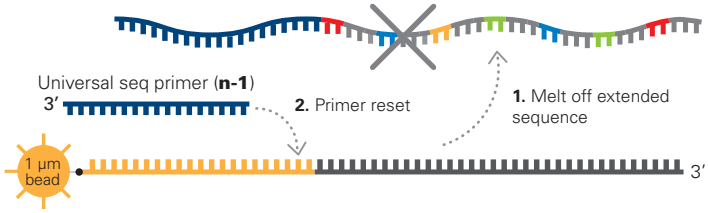
4. Cleave off Fluor



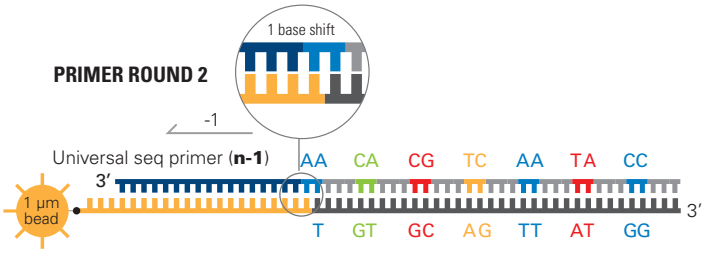
5. Repeat steps 1-4 to Extend Sequence



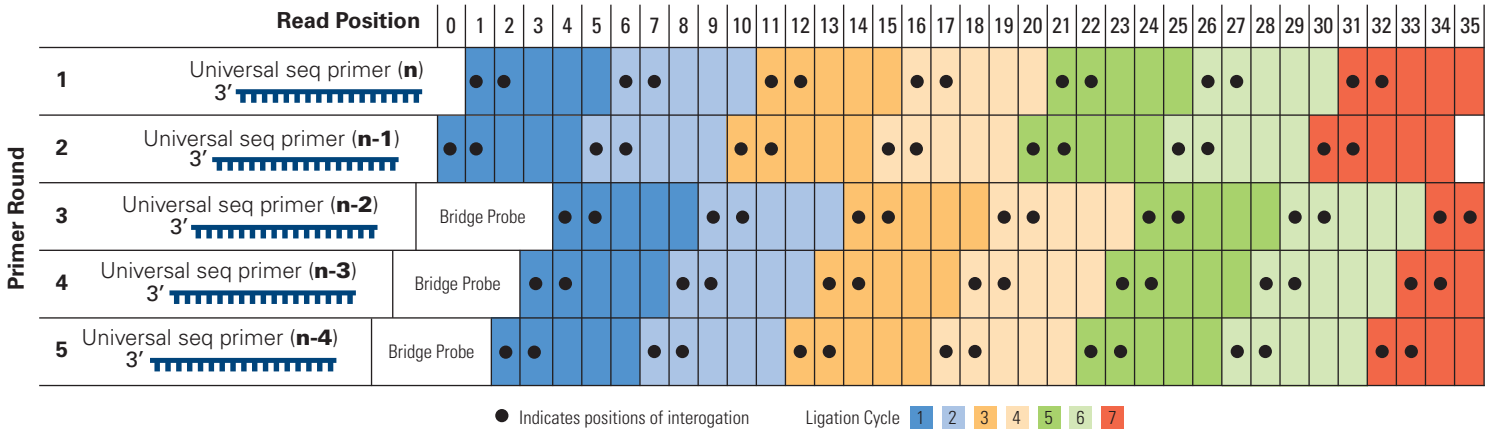
6. Primer Reset



7. Repeat steps 1-5 with new primer

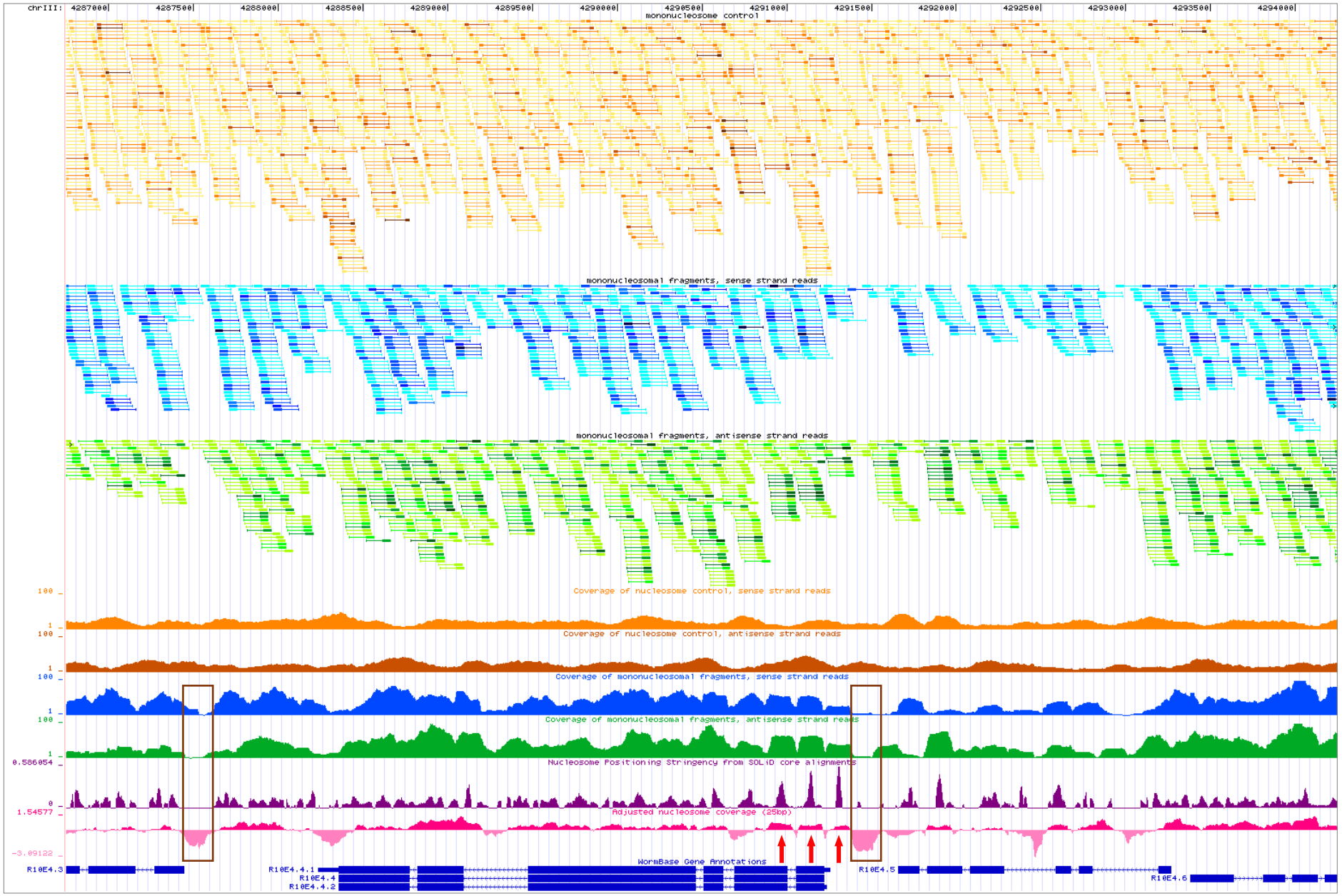


8. Repeat Reset with , n-2, n-3, n-4 primers



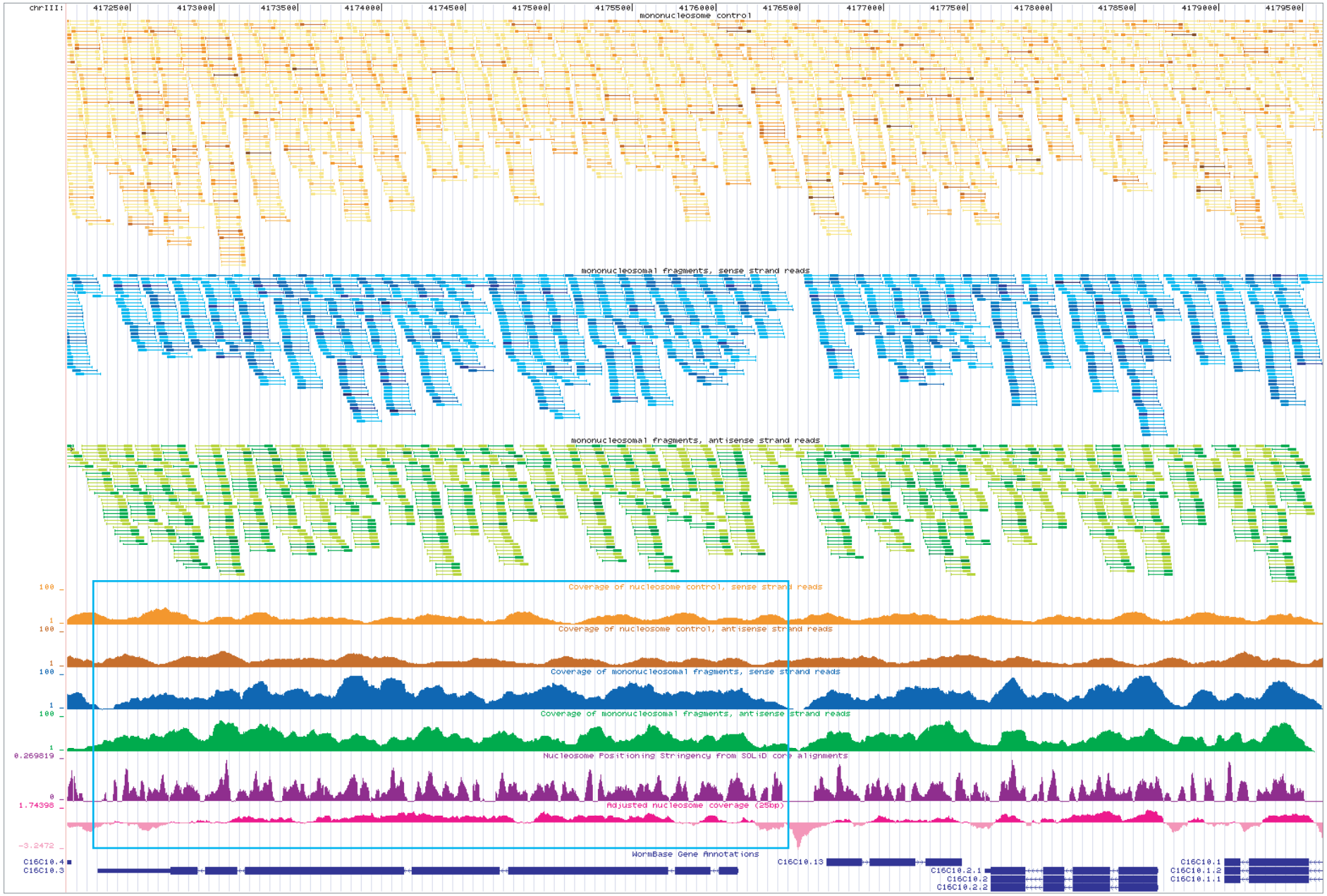
A

Positioned and Regularly Spaced

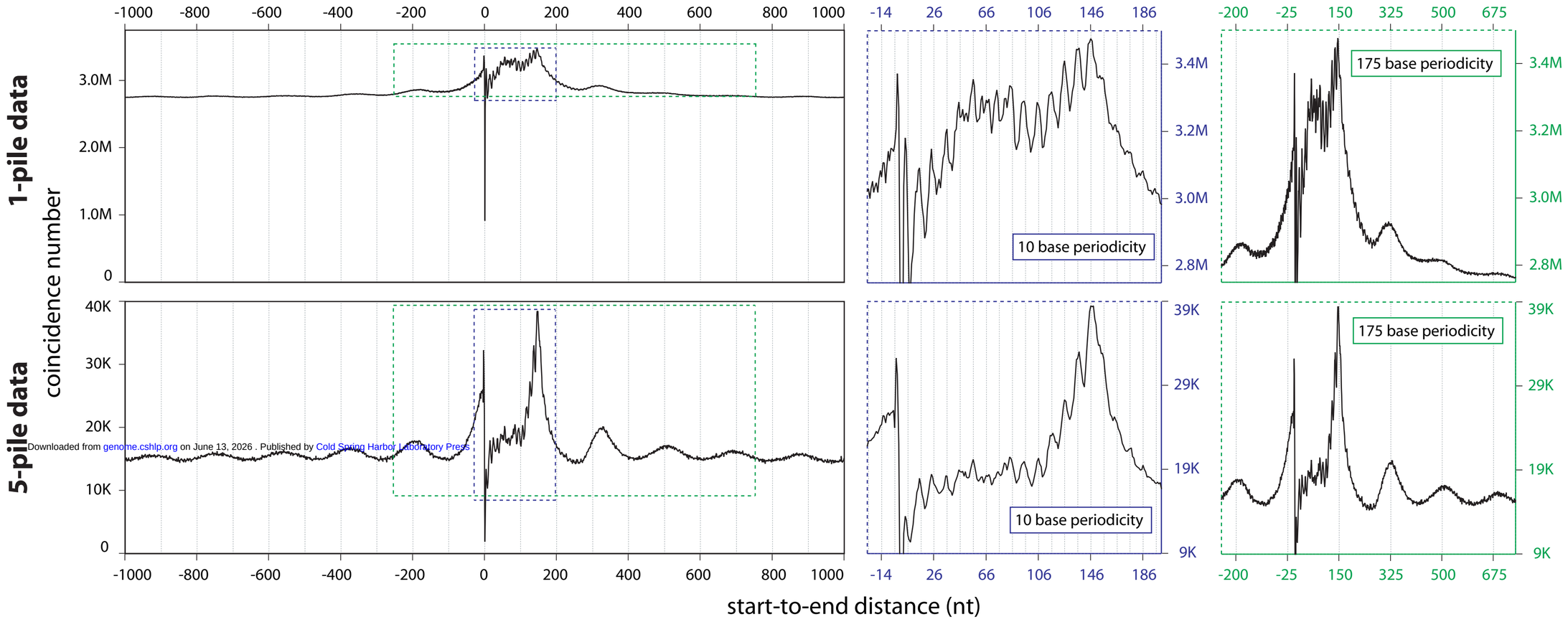


B

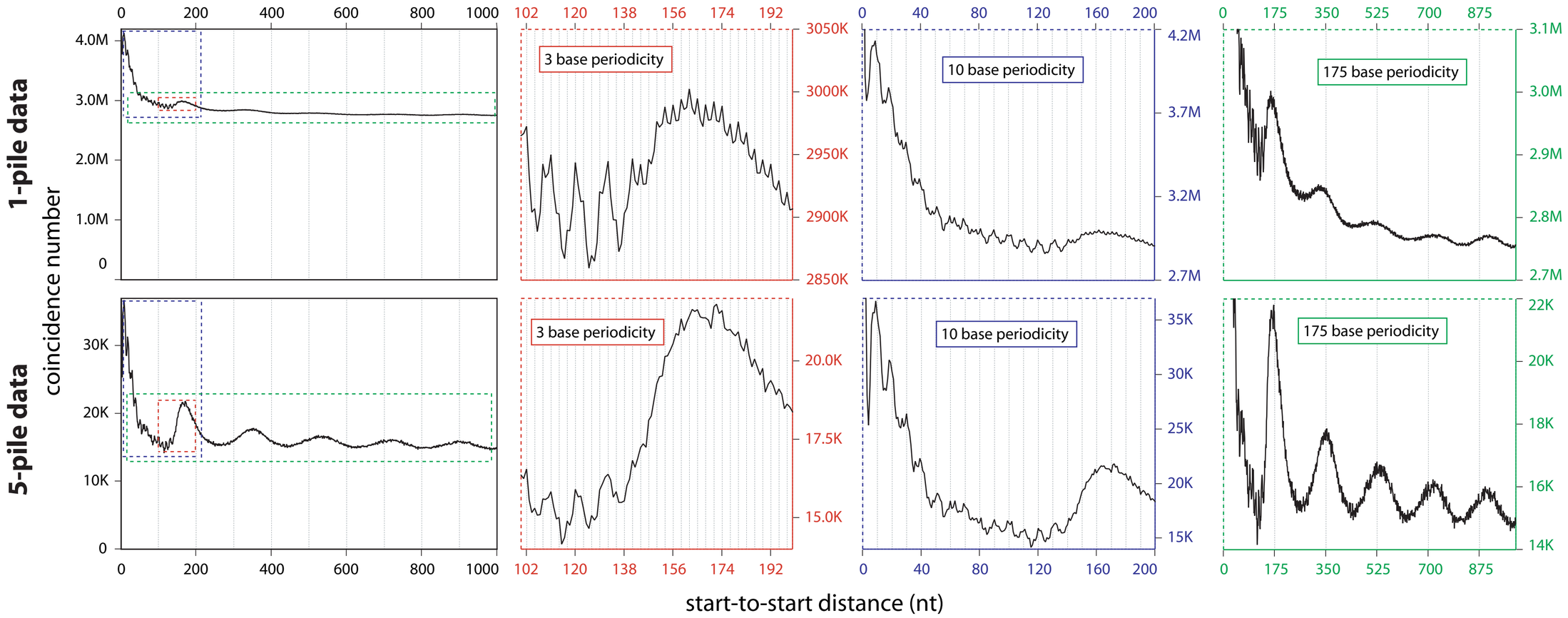
No Strong Positional Preference

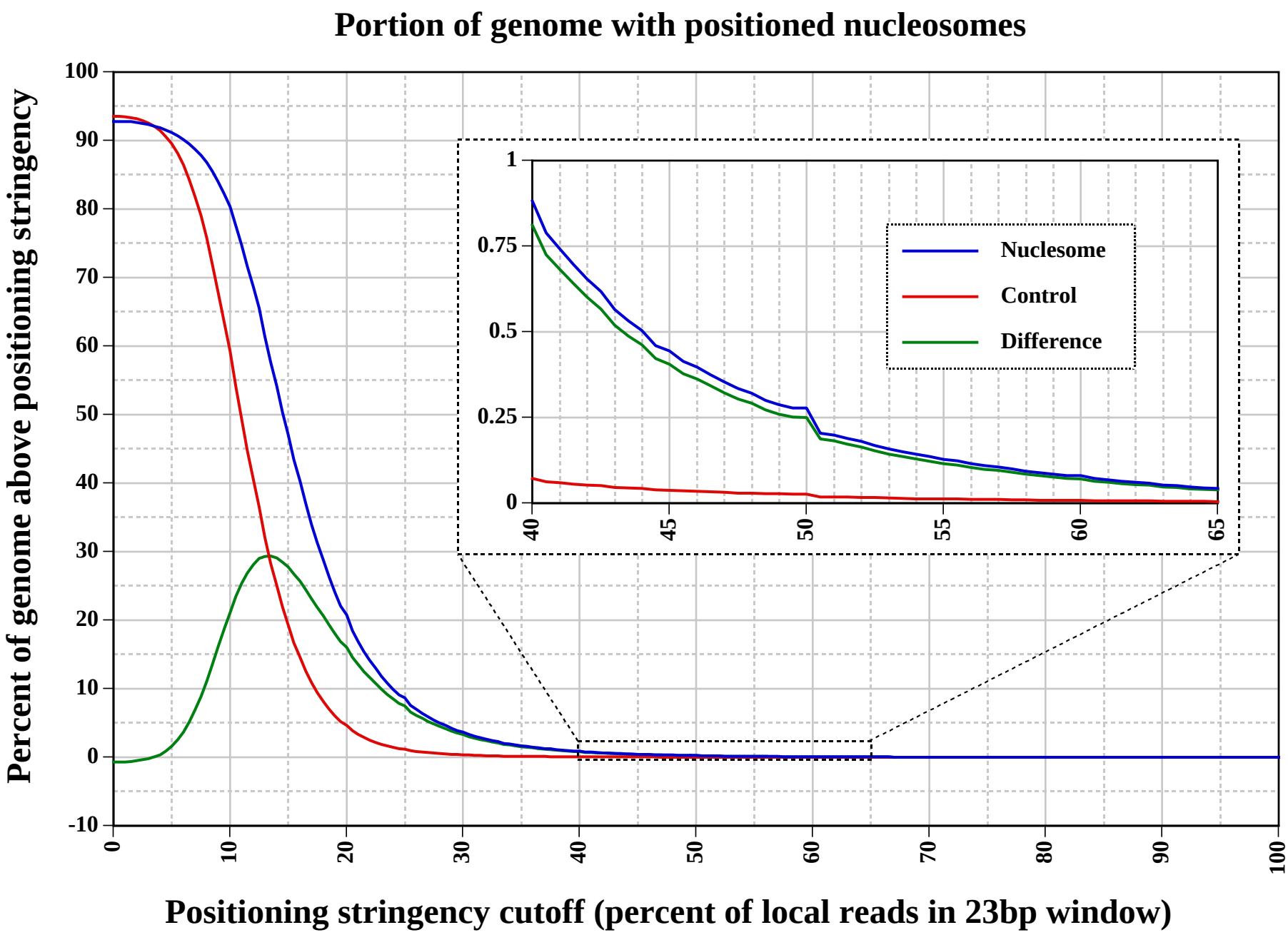


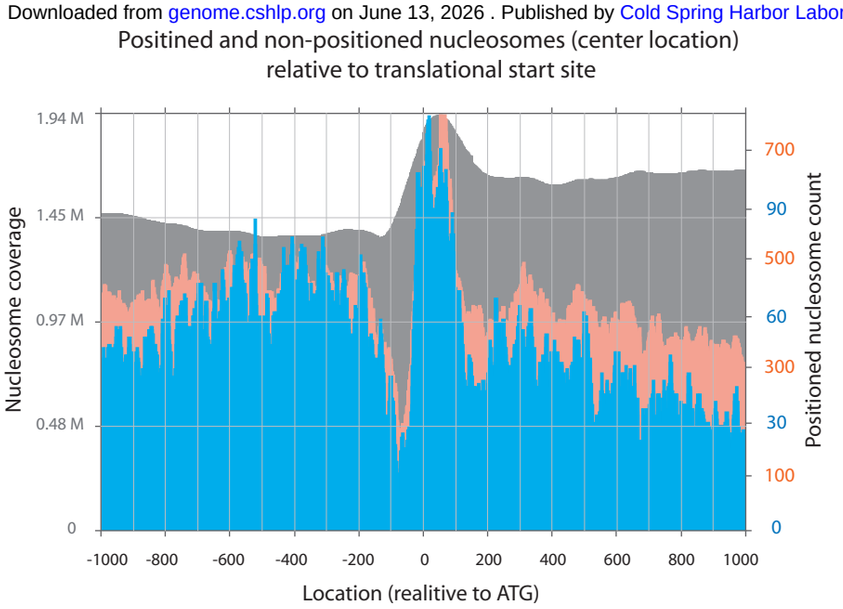
A

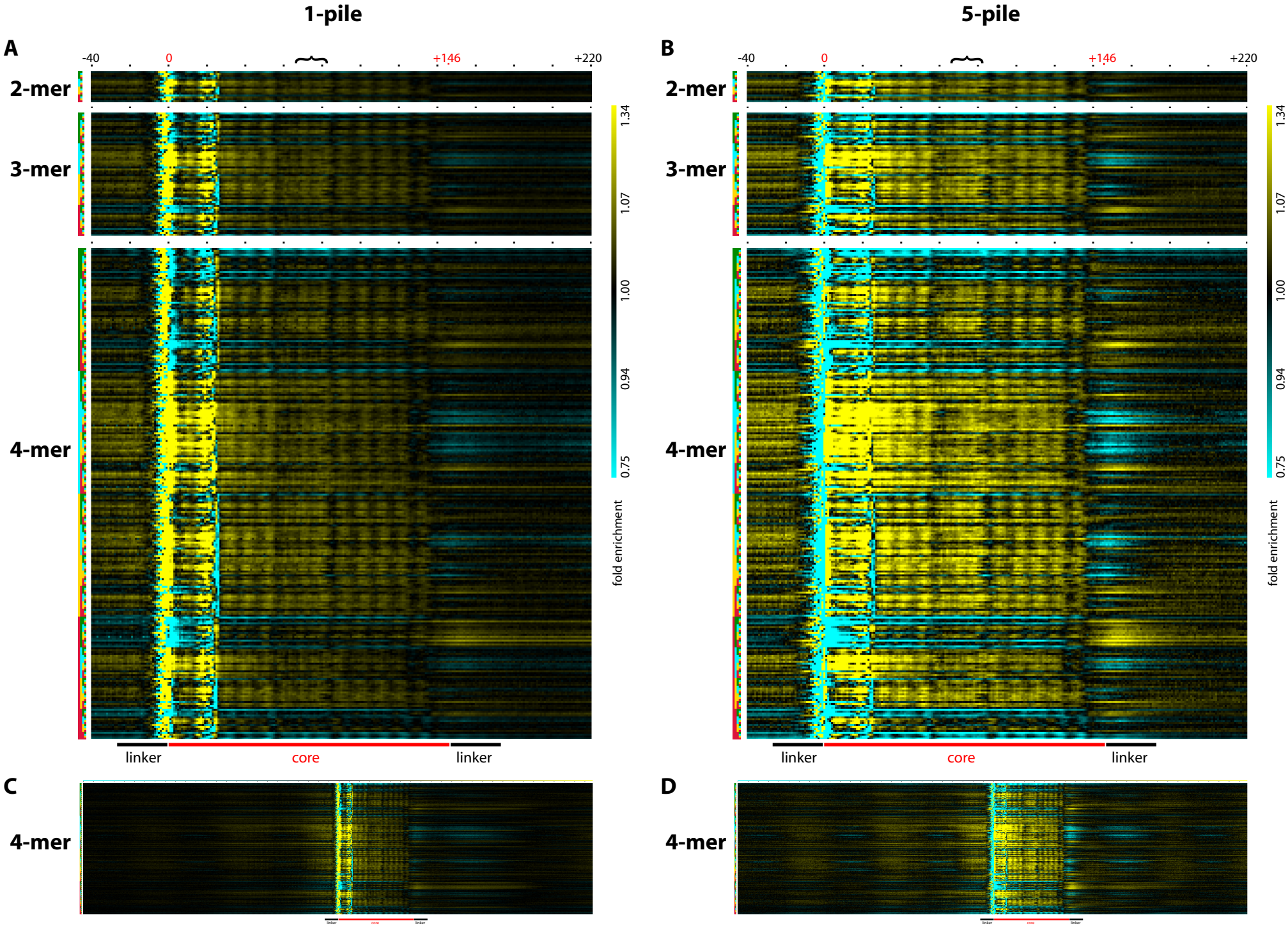


B




Portion of genome with positioned nucleosomes
Percent of genome above positioning stringency
Positioning stringency cutoff (percent of local reads in 23bp window)
Nuclesome
Control
Difference
100
90
80
70
60
50
40
30
20
10
0
-10
0
10
20
30
40
50
60
70
80
90
100
1
0.75
0.5
0.25
0
40
45
50
55
60
65




1-pile
5-pile
A
B
C
D
-40
0
+146
+220
2-mer
3-mer
4-mer
1.34
1.07
1.00
0.94
0.75
fold enrichment
linker
core
linker