Supplementary Material for

Absolute nucleosome occupancy map for the *Saccharomyces cerevisiae* genome

Elisa Oberbeckmann[1]*, Michael Wolff[2]*, Nils Krietenstein[1,3], Mark Heron[4,5], Jessica L. Ellins[6], Andrea Schmid[1], Stefan Krebs[7], Helmut Blum[7], Ulrich Gerland[2] and Philipp Korber[1#]

[1]Molecular Biology Division, Biomedical Center, Faculty of Medicine, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany; [2]Physik Department, Technische Universität München, Garching, Germany; [3]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts, Medical School, Worcester, USA; [4]Quantitative and Computational Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany; Gene Center, [5]Faculty of Chemistry and Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany; [6]Department of Biochemistry, University of Oxford, UK; [7]Laboratory of Functional Genome Analysis (LAFUGA), Gene Center, Faculty of Chemistry and Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany

* These authors contributed equally to the work.
[#] Corresponding author: pkorber@lmu.de

Contents

Supplementary Methods

ORE-seq method development

We observed an RE (restriction enzyme)-dependent distribution of fragment ends close to the respective RE sites (Supplemental Fig. S1D) which likely stemmed from resection of RE cut ends by copurified endogenous yeast exonucleases. We developed a window-based counting algorithm to score resected RE cut ends as RE cuts, too (see below). The resecting exonucleases are likely 5'-3' exonucleases as resection was more pronounced for REs generating 5'-overhangs than blunt ends (Supplemental Fig. S1E). Since resection increased with incubation time (Supplemental Fig. S1E) and digestion by endogenous yeast endonucleases became a problem at longer incubation times (see below), we used data from the 30-minute incubation time points as plateau values (30-minute incubation was sufficient for saturation of digestion (Supplemental Fig. S1B)).

For each RE site, we calculated the ratio of fragments ending at or near this site over the sum of these fragments plus twice the fragments spanning over this site. Note that each cut leads to two ends that are detected independently, while each non-cut leaves one fragment. Therefore, the uncut fragments must be counted twice, if every cut end is counted once. This ratio should directly give the absolute accessibility at each RE site, and one minus this absolute accessibility gives the absolute occupancy of whatever blocked access to this site. To test this so-called "cut-uncut method" (Fig. 1A, left), we purified yeast genomic DNA (gDNA) and fully digested it with each RE at a time. Such predigested gDNA was mixed with undigested gDNA at 10%, 30%, 50%, 70% and 90% uncut fraction within pipetting accuracy. Quantification of these mixtures and of undigested (0%) and fully digested (100%) gDNA yielded a calibration curve that confirmed no and full digestion, but fell below the expected diagonal for in-between cut fractions (Supplemental Fig. S1G, left). This deviation (7.6% average error) corresponds to too low measured occupancy (= too high measured accessibility) values, which may stem from some technical bias towards DNA fragments with RE cut ends.

To avoid this bias, we devised a second method ("cut-all cut method") that scores only the fragments with RE cuts (Supplemental Fig. S1F). Instead of measuring the number of all molecules as sum of cut and uncut fragments, we now used the number of cuts in a parallel "all cut" sample. After RE digestion of chromatin and DNA purification, the samples were split into two halves. One half was digested again with the respective RE thereby generating an "all cut" sample. For both halves, the number of DNA fragments with RE cut ends was determined by sequencing as in the cut-uncut method. As now only the same fragment type with RE cut ends was scored, all biases should be the same and cancel out upon taking a ratio. The ratio of RE cut ends in the half without over the half with second RE digest should directly give the absolute accessibility at each RE site in chromatin. Potential loss of material during parallel treatment of both halves was corrected for by normalization to RE digested *Schizosaccharomyces pombe* gDNA spiked-in before splitting into the two halves. The calibration curve using the "cut-all cut method" agreed very well with the expected diagonal (2.5% average deviation, Supplemental Fig. S1G, center). However, standard deviations between sites (error bars in Supplemental Fig. S1G) were larger for this than for the cut-uncut method at lower prepared fractions of uncut DNA molecules, measuring a larger spread for individual sites. This was due to the accessibility being estimated by the ratio of two on average similarly large (for low occupancy) cut counts from independent samples,

whereas the cut-uncut method only uses one sample and has the mathematical advantage that the standard deviation decreases towards 0% as well as towards 100% occupancy.

We returned to the cut-uncut method and introduced a correction factor for each RE (Supplemental Fig. S1H, see below) such that the calibration curves fit very well the diagonal with an average error of 1% (Supplemental Fig. S1G, right). The mean absolute occupancy values measured by the corrected cut-uncut versus the cut-all cut method for yeast chromatin centered around the same plateau (Supplemental Fig. S1I) arguing that the correction factor derived from the gDNA calibration samples was appropriate for ex vivo chromatin, too. Here, the spread of mean absolute occupancy values was again lower for the corrected cut-uncut than for the cut-all cut method (Supplemental Fig. S1I) justifying our preference for the former method.

The mean absolute occupancy values of all three ORE-seq methods and the mean resection lengths for each sample are presented in Supplemental Table S3. As an orthogonal control for the cut-uncut method, we analyzed DNA purified from ex vivo samples after the same RE digestion as for ORE-seq also by classical secondary cleavage and Southern blotting for occupancy at a BamHI site in the *PHO5* promoter and a HindIII site in the *PHO8* promoter (Supplemental Fig. S1J). The occupancies determined by ORE-seq versus Southern blotting agreed well within 7% for the BamHI site, but differed by 20% for the HindIII site, which we consider an unusually large discrepancy that may occur at individual sites even though the mean standard deviation between replicates for ORE-seq was 5-6% (Supplemental Table S1). In the end, the corrected cut-uncut method was used for the ORE-seq analyses in our study.

Bioinformatic analysis for ORE-seq

Cut and uncut fragment count and resection length

1) Filter reads based on basecalling quality with FastqFilter (quality threshold 10)
2) Map fragments with BWA using the combined *Saccharomyces cerevisiae* and *S. pombe* reference genome
3) Remove read pairs with unreasonable bam flags (rules of readGAlignmentPairs)
4) Extract paired-end read information: chromosome, start, end and strand information, with end positions shifted by +1 bp
5) Remove fragments that are longer than 500bp
6) Remove fragments on the loci of rDNA genes:

   S. cer. chr. 12: 451500 – 495000

   S. pom. chr. 3: 0 – 30000

   S. pom. chr. 3: 2430000 – 2452883
7) For each genomic inter-base position x, i.e. between base $x-1$ and base x, count the starting/ending fragments on plus and minus strand $c_\tau(x)$, with $\tau = 1, 2, 3, 4$ denoting starts on plus, starts on minus, ends on plus and ends on minus strands. Doing this for the sample without 2nd RE digest and the sample with 2nd RE digest gives $c_\tau^1(x)$ and $c_\tau^2(x)$. We assume that one single given fragment with cut and/or sheared fragment ends supplied to PCR and Illumina sequencing will on average result in p_τ^x counts with x being the position of the fragment's start or end.

8) Similarly, count the uncut fragments in the sample without 2nd RE digest, i.e. fragments that start before x and end after x, $u_\tau^1(x)$, also using the index τ, to distinguish plus, $\tau = 1,3$ and minus strand, $\tau = 2,4$. We assume that one single given fragment starting before x and ending after x supplied to PCR and Illumina sequencing will on average result in q_τ^x counts.

9) Determine the cut site positions, i. e. the recognition motif of the RE, on both genomes including generation of the actual DNA ends by end polishing in the following way. We define x^i as the position of the first base pair recognition motif of cut site i plus half the length of the recognition motif (they have always even length).

   HindIII as an example with 'I' denoting the cut in both strands:

   Position: x^i

   ```
   + strand:  5'-...A|A G C T T...-3'
   - strand:  3'-...T T C G A|A...-5'
   ```

   If the 5' end of a fragment is longer than the 3' end after cutting, the 3' end is elongated to match the 5' end (by polymerase). If the 3' end of a fragment is longer than the 5' end after cutting, the 3' end is digested to match the 5' end. After the cut and elongation (as in the case of HindIII) or digestion we get the following double stranded fragments:

   Position: x^i (+1 shift for ends) x^i

   ```
   + strand:  ending: 5'-...A A G C T-3'  and starting:  5'-A G C T T...-3'
   - strand:  ending: 3'-...T T C G A-5'  and starting:  3'-T C G A A...-5'
   ```

   Δs is the shift from the pattern centre to the cut position of the + strand in upstream direction (here $\Delta s = +2$).

   Then, for now assuming we get read counts only from perfectly cut fragments, we have for site i:

   Counts of starting reads on + strand: $c_1^1(x^i - \Delta s)$ and $c_1^2(x^i - \Delta s)$

   Counts of starting reads on – strand: $c_2^1(x^i - \Delta s)$ and $c_2^2(x^i - \Delta s)$

   Counts of ending reads on + strand: $c_3^1(x^i + \Delta s)$ and $c_3^2(x^i + \Delta s)$

   Counts of ending reads on – strand: $c_4^1(x^i + \Delta s)$ and $c_4^2(x^i + \Delta s)$

   For easier notation, we define $x_1^i = x_2^i = x^i - \Delta s$ and $x_3^i = x_4^i = x^i + \Delta s$ giving the cut and uncut fragment counts at site i as $c_\tau^1(x_\tau^i)$, $c_\tau^2(x_\tau^i)$ and $u_\tau^1(x_\tau^i)$ for $\tau = 1,2,3,4$.

10) For counting cut and uncut fragments, ignore cut sites with one neighbour less than 200bp away or an upstream and downstream neighbour less than 300bp away. We denote the set of left over sites with I and J, for the *S. cerevisiae* and the *S. pombe* genome, respectively. We ignore sites with any neighbour within 200bp, because fragments with length below 100bp are very unlikely to be amplified and then sequenced and because uncut fragments counts are increased at sites with any neighbour within approx. 150bp.

11) Also ignore start/end cut counts of a cut site and near the cut site when the next cut site downstream/upstream is closer than 300 bp, respectively. We ignore these fragments at these sites because they showed a strong increase in cut counts (as calculated below) in the direction of the close neighbour at up to 250 bp distance, with the peak at a distance of approx. 150 bp.

12) Due to the presence of endogenous exonucleases in the chromatin preparations that may trim DNA ends after restriction enzyme cleavage, some fragments ends do not match the cut site positions any more, even though they were generated by restriction enzyme. Therefore, we need to count the starting and ending fragments not only at the exact cut positions, but also at some distance from it. The amount of strand resection varies from sample to sample, so its treatment needs to be tailored to each pair of samples without and with 2nd RE digest.

Given w (the algorithm how to find w is described below), we define $W_1 = W_2 = \{0, 1, 2, \dots, w\}$ to apply a window in downstream direction for read starts and $W_3 = W_4 = \{0, -1, -2, \dots, -w\}$ in upstream direction for read ends for every cut site i (in both genomes). We define for the samples without and with 2nd RE digest ("cut" and "all-cut"):

$C_\tau^i = \sum_{a \in W_\tau} c_\tau^1(x_\tau^i + a)$ and $A_\tau^i = \sum_{a \in W_\tau} c_\tau^2(x_\tau^i + a)$

w is determined using the sample without 2nd RE digest and for the data of the *S. cerevisiae* genome and then the same value is used for the sample with 2nd RE digest and the *S. pombe* genome as well. In principle *S. pombe* should not show any resection at all, and in deed if this algorithm is applied to the *S. pombe* cut sites we do not see any resection. For completely uncut samples, we set $w = 5$ to average over fluctuations in the very low cut counts at a single position.

In the case of ignored start counts of step 11), we set $C_\tau^i = NA$ and $A_\tau^i = NA$ for $\tau = 1,2$ and the same for $\tau = 3,4$ in the case of ignored end counts.

For normal samples, we use the following algorithm, which makes sure that increasing w by $1, 2, 3, 4$ or 5 bp doesn't increase the summed counts until w by more than 1%, correcting for cut counts from shearing.

Calculate the mean counts at each position -200 bp to 200 bp away from the average cut site for starts and ends counts and both strands. Ideally it shows a single peak at 0, but normally there also is a decreasing shoulder downstream/upstream for starts/ends. Calculate the average of start and end counts and both strands (end counts need to be mirrored at 0 first). This is plotted for exemplary samples in Supplemental Fig. S1D. The corrected cut counts $\bar{c}(d)$ are the average counts at 0 bp to 99 bp minus the mean of the counts at 100 bp to 200 bp. Calculate the cumulative sum of cut counts $\bar{c}(d)$ from 0 bp onwards. Then w is the smallest distance from 0 bp where for all $n \in \{1, 2, 3, 4, 5\}$, the sum of the counts of the next n positions is smaller than 1% of the sum of counts from 0 bp to w bp. In our samples, typical values for w were from zero to 20, going up to 40 for samples with very strong resection.

The mean resection length is defined as $\sum_{d=0}^{w} d\, \bar{c}(d)$.

13) The uncut fragment counts at any RE cut site are not affected by endogenous exonucleases since they are still occupied by a nucleosome (or other protein) that blocked the restriction enzyme. Thus, we can define the uncut counts by $U_\tau^i = u_\tau^1(x_\tau^i)$.

Occupancy estimation by cut-all cut method

We treat C_τ^i and A_τ^i as random variables with the following expectation values

$$E\left[C_\tau^i\right] = N_C \mu^i \bar{p}_\tau^i \quad \text{with} \quad \mu^i = \alpha^i + \left(1-\alpha^i\right)s$$
$$E\left[A_\tau^i\right] = N_A \bar{p}_\tau^i$$

N_C and N_A are the number of cell cores in the samples without and with 2nd RE digest, respectively, and $\bar{p}_\tau^i$ is a factor that combines the sequencing probabilities and the PCR multiplication of fragments of type τ in the window W_τ at cut site i and is an effective average of the p_τ^x with $x \in x_\tau^i + W_\tau$ described earlier.

s is the probability that a given (longer) fragment will be cut by shearing within a fixed region of length $w+1$ within the fragment and α^i is the real accessibility at cut site i, that we want to estimate.

$\mu^i = \alpha^i + \left(1-\alpha^i\right)s$ considers that the restriction enzymes act before the shearing step, thus only the fraction that has not been cut by the enzyme, can be cut in the shearing step in the chromatin sample.

We assume that in the sample with 2nd RE digest, all counts near a cut site came from a cut of the enzyme and all counts far away from cut sites occurred due to shearing.

We use these four estimators for $\mu^i$and α^i: $\hat{\mu}_\tau^i := \frac{C_\tau^i}{A_\tau^i}\frac{N_A}{N_C}$ and $\hat{\alpha}_\tau^i := \frac{\hat{\mu}_\tau^i - s}{1-s}$.

We have $E\left[\hat{\alpha}_\tau^i\right] = \frac{E[\hat{\mu}_\tau^i]-s}{1-s} = \frac{1}{1-s}\left(E\left[C_\tau^i\right]E\left[\frac{1}{A_\tau^i}\right]\frac{N_A}{N_C} - s\right) \approx \alpha^i$, because the X_τ^i and H_τ^i are statistically independent as they originate from different samples and $E\left[\frac{1}{A_\tau^i}\right] \approx \frac{1}{E\left[A_\tau^i\right]}$. Note that the two sets $\{C_\tau^i\}$ as well as $\{A_\tau^i\}$ within themselves, however, are statistically dependent.

We set $\hat{\alpha}_\tau^i = NA$, if $A_\tau^i = 0$ or $A_\tau^i = NA$ (due to a close neighbour in direction of τ).

To obtain N_A/N_C, we use the *S. pombe* spike-in cut sites, which are completely cut in both samples:

$$\frac{N_A}{N_C} = \frac{\langle A_\tau^i \rangle_{i\in J,\tau}}{\langle C_\tau^i \rangle_{i\in J,\tau}}$$

with $\langle X_{i,\tau} \rangle_{i,\tau}$ denoting the mean of $X_{i,\tau}$ over i and τ, ignoring NA values.

Alternatively, the ratio of the number of sequenced *S. pombe* reads could be used, but gave slightly worse results in our calibration runs.

To estimate the probability s, we look at the set Z of all genomic positions in *S. cerevisiae* that are further away than 300 bp from any cut site (including the ones with a close neighbor). At these positions, all counted starts and ends originate from shearing. Thus, $\langle c_\tau^1(x) \rangle_{i\in Z,\tau}$ is an estimator for $N_C s_1 \langle p_\tau^x(x) \rangle_{i\in Z,\tau}$, with $s_1$being the probability of shearing a long fragment at one fixed position.

Since $\bar{p}_\tau^i$ are effectively averages of p_τ^x in the cut site windows, so their averages over large regions of the genome are (with good approximation) the same:

$$\langle p_\tau^x(x) \rangle_{x\in Z,\tau} \cong \langle \bar{p}_\tau^i(x) \rangle_{i\in I,\tau} = \frac{1}{N_A}\langle E[A_\tau^i] \rangle_{i\in I,\tau}$$

The average of $E\left[A_\tau^i\right]$ over i and τ can be well approximated by the average of A_τ^i, leading to our value for s_1:

$$s_1 = \frac{N_A}{N_C}\frac{\langle c_\tau^1(z) \rangle_{z\in Z,\tau}}{\langle A_\tau^i \rangle_{i\in I,\tau}}$$

Thus, s_1 is the (correctly normalized) ratio of the average fragment number at genomic positions where cuts can happen only by shearing (counts in the chromatin sample away from cut sites) and the average fragment number at genomic positions where cuts have to happen by the restriction enzyme (counts in the 100% cut sample at the cut sites).

Using s_1, we can calculate the probability that a fragment is sheared at least once within a fixed window of length $w+1$: $s = (w+1)s_1$. If a fragment is sheared more than once within a window of length $w+1$, the new fragments within the window will be too small and filtered out before the PCR and sequencing steps.

Due to the stochasticity in the values for C_τ^i and A_τ^i for fixed i and τ, the estimators $\hat{\alpha}_\tau^i$ can be smaller than 0 or larger than 1, even though the values they estimate, i.e. α^i are between 0 and 1.

The lowest possible value of $\hat{\alpha}_\tau^i$ is $-\frac{s}{1-s}$, with $s \leq 0.15$ in most samples.

It is not useful to restrict the estimators $\hat{\alpha}_\tau^i$ to $[0;1]$ before averaging, because a 100% accessibility test sample has measured accessibilities distributed around 1 in both directions. Capping the estimators $\hat{\alpha}_\tau^i$ at 1 would then give a mean value lower than 1.

However, very large outliers influence the mean very strongly, even though the real value cannot be greater than 1, thus we cap the values for $\hat{\alpha}_\tau^i$ at 1.5 when averaging over τ, $\hat{\alpha}^i = \langle \min(\hat{\alpha}_\tau^i, 1.5) \rangle_\tau$, to obtain one accessibility estimate for each cut site i. If $\hat{\alpha}_\tau^i = NA$, it is ignored during the averaging step.

To obtain the global accessibility, we average over all sites:

$$\hat{\alpha} = \langle \hat{\alpha}^i \rangle_{i \in I}$$

When comparing the accessibility values of individual sites with the measured values from other assays, for example bisulfite, it does make sense to restrict the values of $\hat{\alpha}^i$ to $[0;1]$, since this gives the best estimate for each individual site.

Occupancy estimation by cut-uncut method

In the following we only use data from the sample without 2nd RE digest to estimate the accessibility and use the ratio of the cut counts C_τ^i and the counts of uncut fragments U_τ^i. We choose to only consider different PCR biases and sequencing biases between cut and uncut fragments, giving all cut fragments sequencing probability p and all uncut fragments sequencing probability.

Summing up cut counts and uncut counts, we set

$$C^i := C_1^i + C_2^i + C_3^i + C_4^i \text{ and } U^i := 2\left(U_1^i + U_2^i\right)$$

for sites without any neighbor within 300 bp and

$$C^i := C_1^i + C_2^i \text{ or } C^i := C_3^i + C_4^i \text{ and } U^i := U_1^i + U_2^i$$

for sites with one upstream/downstream neighbor within 300bp, respectively. Then define

$$\hat{\kappa}^i := \frac{C^i}{U^i}$$

If the denominator is zero, we set $\hat{\kappa}^i = \infty$, which will lead to an accessibility of 1. If nominator and denominator are „independent enough“ (provided by the low sequencing probabilities) then,

$$E[\hat{\kappa}^i] \approx \frac{E[C^i]}{E[U^i]} = \frac{4N_C p}{4N_C q} \frac{(\alpha^i + (1-\alpha^i)s_1(w+1))}{(1-(\alpha^i + (1-\alpha^i)s_1))}$$

using the same expectation value as for C_τ^i as in the cut all-cut method and assuming that the uncut fragment counts are given by fragments that have not been cut by the restriction enzyme at x_τ^i and after that also not been cut by shearing at x_τ^i. The ratio of sequencing probabilities of cut and uncut fragments, the "uncut correction factor" $\gamma := \frac{p}{q}$, is fitted to the calibration samples as described in the section below.

We then obtain the following estimator for α^i: $\hat{\alpha}^i := 1 - \frac{1+\sigma}{\frac{\hat{\kappa}^i}{\gamma} + 1 - \sigma w}$, thus

$$\hat{\alpha}^i = \frac{C^i - \sigma(w+1)U^i\gamma}{C^i - \sigma(w+1)U^i\gamma + (1+\sigma)U^i\gamma} = \frac{C_{eff}^i}{C_{eff}^i + U_{eff}^i}$$

with $\sigma := \frac{s_1}{1-s_1} = \frac{1}{\gamma}\frac{\langle c_\tau^1(z)\rangle_{z\in Z,\tau}}{\langle u_\tau^1(z)\rangle_{z\in Z,\tau}}$ being the corrected ratio of all cut counts away from all cut sites and all uncut fragment counts away from all cut sites. s_1 is the shearing probability per base pair s_1, i. e. the ratio of all cut counts away from cut sites and the sum of all cut and uncut fragment counts away from cut sites.

$C_{eff}^i = C^i - \sigma(w+1)U^i\gamma$ and $U_{eff}^i = (1+\sigma)U^i\gamma$ are the effective counts of cut and uncut fragments, respectively, both corrected for cuts in the shearing step.

$C_{eff}^i + U_{eff}^i$ gives an "effective coverage" of cut and uncut fragments at the site i and we ignore sites with an effective coverage below 40.

Finally, the genome-wide average accessibility is given by $\hat{\alpha} = \langle \hat{\alpha}^i \rangle_{i\in I}$.

Fit of γ using prepared calibration samples for restriction enzyme digests:

For each restriction enzyme (AluI, BamHI and HindIII) and each calibration sample s with 0%, 10%, 30%, 50%, 70%, 90%, and 100% prepared fraction of uncut DNA molecules, i.e. prepared occupancy $\omega_s = 1 - \alpha_s$, we calculate the measured genome-wide average occupancy $\hat{\omega}_s(\gamma) = 1 - \hat{\alpha}_s(\gamma)$ for varying γ. We then choose γ for each restriction enzyme such that $\langle(\omega_s - \hat{\omega}_s(\gamma))^2\rangle_s$ is minimized. Supplemental Fig. S1H shows the relative error $\left(\langle(\omega_s - \hat{\omega}_s(\gamma))^2\rangle_s / \langle(\omega_s - \hat{\omega}_s(1))^2\rangle_s\right)^{1/2}$ for AluI, BamHI and HindIII with varying γ. Additionally, we did a combined fit over all calibration samples of the three enzymes to use for enzymes without specific calibration samples. The following table shows the best values for γ:

Enzyme	AluI	BamHI	HindIII	combined
γ_{min}	1.555	1.699	1.680	1.642

ODM-seq method development

First, we checked if DNA methylation reaches a plateau or if it eventually invades nucleosomes, and if the plateau was like the ORE-seq plateau. For this comparison, we used chromatin reconstituted in vitro by salt gradient dialysis (SGD) for a yeast whole genome plasmid library (Krietenstein et al. 2012). SGD chromatin offers several advantages over ex vivo chromatin in this case. First, it consists only of canonical nucleosomes so that REs and DNMTs (DNA

methyltransferases) could be compared specifically regarding their action towards nucleosomes without the presence of potentially confounding other factors in ex vivo chromatin. Second, SGD chromatin comes without nucleases. Therefore, there was no end-resection at the RE sites (Supplemental Fig. S1D, lower right graph). Third, absolute nucleosome occupancy could easily be varied via the DNA to histones ratio during reconstitution. This way we could compare if REs and DNMTs reached the same saturation plateau for two predictably different absolute nucleosome occupancies. Note that also for genome-wide in vitro reconstituted chromatin it was so far not possible to measure nucleosome occupancy in absolute terms (Krietenstein et al. 2012; Lieleg et al. 2015). Therefore, this SGD chromatin experiment only assessed if REs and DNMTs reached the same occupancy plateau for two different nucleosome densities that should differ roughly by a factor of two as twofold less histones were used per DNA, but we could not know beforehand to which absolute occupancy values these densities corresponded. We used the same buffer pioneered for DNMTs in the Kladde group (Darst et al. 2012) but added 1.5 mM $MgCl_2$ to allow Mg^{2+}-dependent RE cleavage. Data were selected according to quality criteria, like bisulfite conversion rate and coverage (see METHODS). For both, high and low nucleosome density SGD chromatin, mean absolute nucleosome occupancy measurements by the RE BamHI, the CpG DNMT M.SssI and the GpC DNMT M.CviPI showed saturation and reached very similar plateaus with mean difference between lowest and highest value of 5% and maximal difference of 7%, no matter if all cut/modified sites were compared (Supplemental Fig. S2A, left) or only BamHI sites within 8 bp of respective CpG (Supplemental Fig. S2A, center) or GpC (Supplemental Fig. S2A, right) methylation sites. For unknown reasons, the GpC DNMT had a systematic tendency of giving mean absolute occupancy values that were on average 5% higher at GCG sites (substrate for both DNMTs), than those obtained via the CpG DNMT (Supplemental Fig. S2B). Nonetheless, this comparison showed that REs and DNMTs could be equally used to measure genome-wide absolute nucleosome occupancy within this error range, i.e., both enzymatic approaches complement and crossvalidate each other.

After establishing ODM-seq with SGD chromatin, we incubated two of the same ex vivo chromatin replicates used for ORE-seq also with DNMTs for different times. We returned to the original Kladde group buffer without Mg^{2+} as the methylation by M.SssI is processive without magnesium (Matsuo et al. 1994) and as Mg^{2+}-dependent endogenous yeast exo- and endonucleases, which copurify with ex vivo chromatin, may interfere with the analysis, especially during long incubations. As positive control, we spiked-in SGD assembled plasmids where a 25mer array of the Widom 601 nucleosome positioning sequence (Lowary and Widom 1998) generates predictable nucleosome positions during SGD (Lieleg et al. 2015).

Such methylation time courses for ex vivo chromatin did not plateau for the WT4 replicate and yielded much lower mean absolute occupancies than measured by ORE-seq for both replicates (Supplemental Fig. S2C, left). In contrast, occupancies of the spike-in chromatin mostly did reach plateau values of the expected relative magnitude, i.e., high for sites in the 601 sequence, where nucleosomes preferentially assemble, intermediate for backbone sites, which monitor a mixture of nucleosomal and non-nucleosomal regions, and low for linker sites, where nucleosomes are scarce (Supplemental Fig. S2D). Note that respective spike-in plateaus differed between experiments as different SGD chromatin preparations with likely different nucleosome densities were spiked into each replicate. Note also that the GpC DNMT was sometimes less efficient in reaching saturation, e.g., in linker methylation. Conversely, complete methylation of 25mer 601

array linker DNA, i.e., less than 20% occupancy, could be used as an internal control for saturating methylation. The results for the spike-in chromatin (Supplemental Fig. S2D) argued against compromised nucleosome integrity in the ex vivo chromatin due to diffusible nucleases or proteases, and they argued, together with those obtained for genomic plasmid library SGD (Supplemental Fig. S2A), against methylation by DNMTs invading nucleosomes. Nonetheless, the ODM-seq mean absolute occupancy values for ex vivo chromatin (Supplemental Fig. S2C, left) did not agree with the ORE-seq results for the corresponding replicates (Fig. 1D). As overmethylation and enzymatic degradation of nucleosomes was ruled out by the spike-in chromatin control, we surmised that yeast endogenous nucleosomes may be intrinsically unstable due to their known structural properties (White et al. 2001) and in contrast to nucleosomes generated with fly histone octamers for the SGD assemblies.

We wondered why the yeast nucleosomes appeared stable during incubation with REs but not during incubation with DNMTs. Maybe this was due to the presence of Mg^{2+} in the RE digest buffers. Mg^{2+} concentrations above 1 mM compact chromatin into higher order structures (Schwarz and Hansen 1994) that may stabilize the yeast nucleosomes. Indeed, the presence of Mg^{2+} during methylation by different concentrations of DNMTs and for different times led to saturation and mean absolute occupancies of similar magnitude as via ORE-seq, at least for the CpG DNMT (Supplemental Fig. S2C, center, versus Fig. 1D), while the GpC DNMT yielded higher occupancy values. For unknown reasons, maybe due to overcompaction of the circular plasmid SGD chromatin, Mg^{2+} seemed to inhibit methylation of the spike-in SGD chromatin as even the linker DNA was hardly methylated in this experiment.

Even though Mg^{2+} seemed to stabilize the nucleosomes, the amplitude of corresponding +1 nucleosome-aligned composite plots (Supplemental Fig. S2E, center) was less pronounced than for the respective ORE-seq pattern (Fig. 1B), and some absolute occupancy values, especially in promoter NFRs, were much higher. This pointed to preferential loss of accessible DNA fragments, i.e., corresponding to NFRs and linkers, for example, by nuclease digestion in the presence of Mg^{2+}. We saw already endogenous exonucleases resecting RE cuts (Supplemental Fig. S1D,E) and suspected now that the combination of the presence of Mg^{2+} with long incubation times needed for saturating DNA methylation exacerbated digestion by endogenous exo- and also endonucleases and thereby loss of accessible DNA regions. Indeed, +1 nucleosome-aligned composite plots of the mere sequencing coverage showed striking patterns increasingly reminiscent of MNase-seq nucleosome positioning patterns with increasing incubation time and only in the presence of Mg^{2+} (Supplemental Fig. S3A, top left). High resolution fragment length distribution plots showed a superimposed 10 bp periodicity pattern (Fig. S3A, bottom left) that is diagnostic of exonuclease cleavage into nucleosome flanks (Riley and Weintraub 1978). As no nucleases were added to these samples and as no such patterns were seen for pure SGD samples despite the presence of Mg^{2+} (Fig. S3A, top and bottom second from left), these patterns confirmed the presumed presence of both endo- and exonucleases endogenous to the ex vivo chromatin. The recurring overall shape of the sequencing coverage plots with a minimum around promoter NFRs was also seen for the SGD samples and therefore unlikely reflected a yeast chromatin property, but rather a sequence-dependent bias in amplification/sequencing efficiency, maybe due to the high AT-content in these regions (Yuan et al. 2005). This bias did not compromise our methods as they rely on ratios of fragments of the same region so that the biases cancel out. Bisulfite conversion was also no concern in this regard as this overall shape was very

similar for DNMT vs. RE treated samples (Supplemental Fig. S3A, top and bottom right) and as analyzing the same DNMT-treated DNA by Nanopore sequencing, which does not involve bisulfite conversion and by EM-seq (Enzymatic Methyl-sequencing), which did not show this coverage bias (Supplemental Fig. S3A, second from right) gave equivalent results (Fig. 1C). The endogenous nuclease activity explained the apparently increased occupancies in NFRs and linker regions for Mg^{2+}-containing samples, i.e., the lower amplitude of composite plots (Supplemental Fig. S2E, center), but did not affect much the mean absolute occupancy plateau values (compare Supplemental Fig. S2C, center, with Fig. 1D) as these are dominated by the nucleosomes, which cover a far larger fraction of the genome than NFRs and linkers. As RE digestion also included Mg^{2+}, respective coverage and fragment length distribution plots also showed a trend towards MNase-seq-like and 10 bp periodicity patterns, respectively, for long incubation times of 120 min (Supplemental Fig. S3A, right). Nonetheless, the nuclease effects were virtually absent for the 30-minute RE incubation samples (Supplemental Fig. S3A, right) and this incubation time was sufficient for RE digest saturation (Supplemental Fig. S1B). Therefore, we included only ORE-seq data from 30-minute incubations for our absolute occupancy map (see also above).

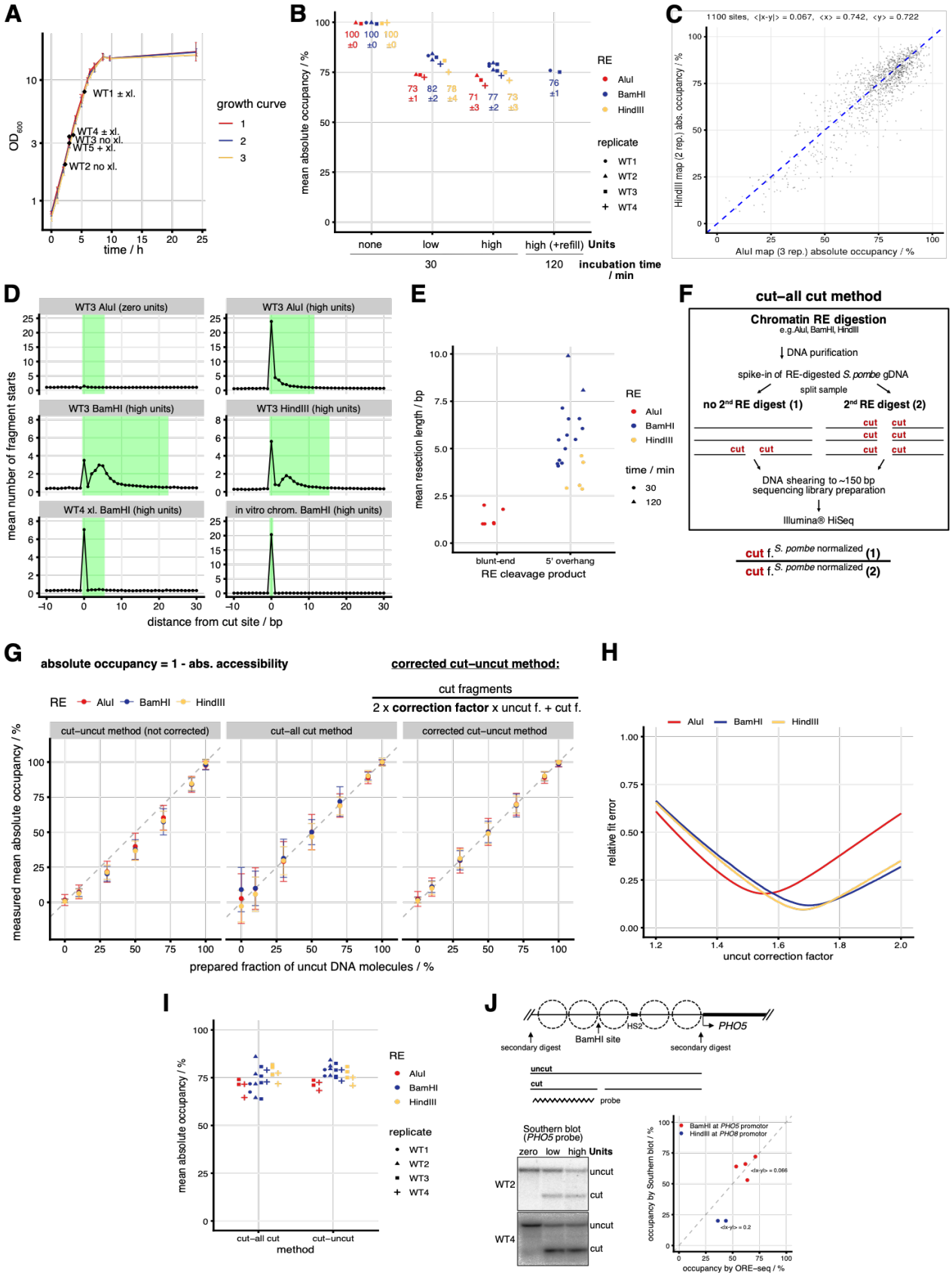
Collectively, incubation in the presence of Mg^{2+} activated nucleases endogenous to ex vivo chromatin to an extent that probably depended on the chromatin preparation. During the prolonged time required to saturate methylation these nucleases preferentially removed NFRs and linkers from the analysis such that the measured absolute occupancy amplitude was notably diminished. In contrast, up to 30-minute incubation times sufficient for ORE-seq did not matter much.

We sought a way that both stabilizes yeast nucleosomes as well as inhibits endogenous nucleases and found that in vivo formaldehyde crosslinking provided both. Ex vivo chromatin prepared from formaldehyde crosslinked cells incubated with DNMTs for different times (Supplemental Fig. S2C, right) mostly showed saturation of methylation and mean absolute occupancy plateaus very similar to the ORE-seq plateaus. Just the GpC methylation of the crosslinked WT4 replicate was not saturating according to the occupancy values for the linker region of the 25mer 601 SGD spike-in chromatin (Supplemental Fig. S2D) and was therefore excluded from further analysis. Sequencing coverage plots of DNMT-treated crosslinked ex vivo chromatin did not show any sign of nuclease activity (Supplemental Fig. S3A, second from right). We concluded that formaldehyde crosslinking did stabilize the nucleosomes and also inhibited the endogenous nucleases such that ODM-seq generated similarly pronounced amplitudes as ORE-seq (compare Supplemental Fig. S2E, right, with Fig. 1B). Variation of crosslinking time between one and 20 minutes did not make a difference (Supplemental Fig. S3B).

Note that measurements of absolute occupancies by both ORE-seq and ODM-seq for the combination of Mg^{2+} and formaldehyde crosslinking led to higher values (Supplemental Fig. S3C), probably due to some sort of overcompaction, which may decrease enzyme accessibility not just via nucleosome occupancy but also via higher order chromatin compaction. For ORE-seq applied to crosslinked ex vivo chromatin, we noted that resection by exonucleases was completely absent (Fig. S1D, bottom left) again confirming that crosslinking inactivated the nucleases.

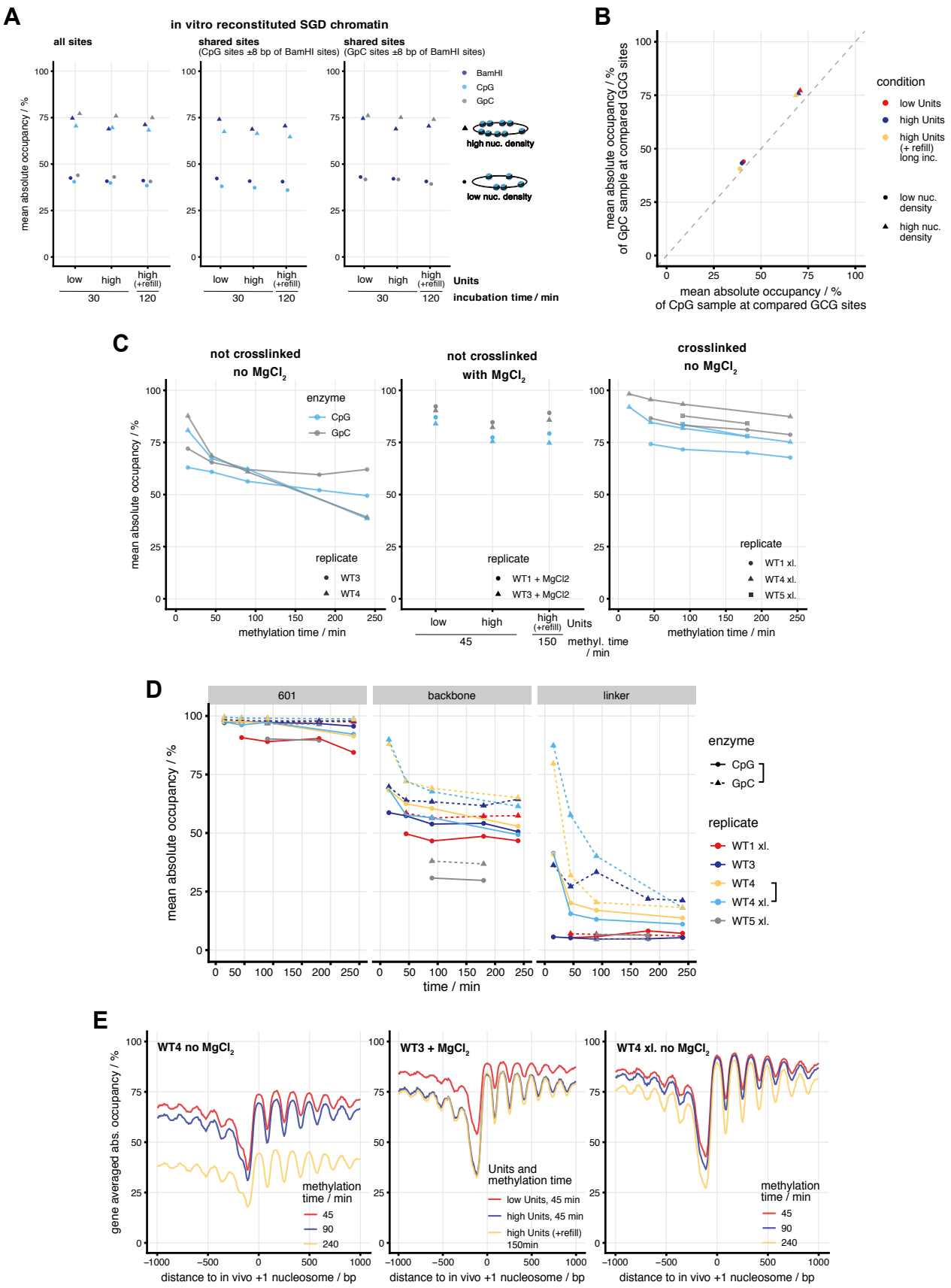
The mean absolute occupancy values of all ODM-seq samples including the spike-in occupancies are presented in Supplemental Table S4.

Supplementary Figures



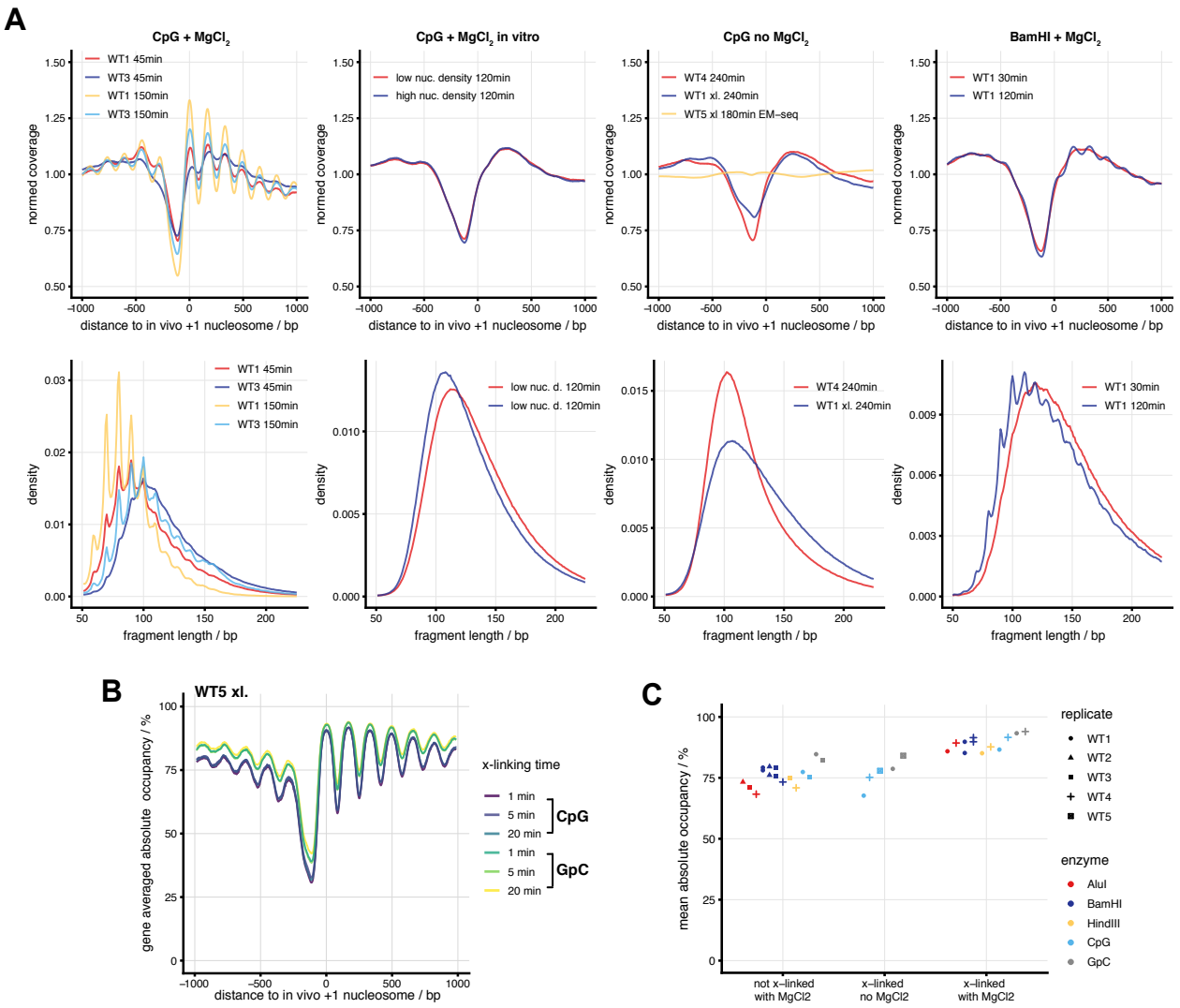
Supplementary Figure S1. Figure caption on next page.

Supplementary Figure S1. (A) Three growth curves for BY4741 in YPDA medium monitored with Zeiss PMQII photometer. OD_{600} of cultures harvested for biological replicates WT1 to WT5 ± in vivo formaldehyde crosslinking (xl.) are indicated. **(B)** Absolute occupancy averaged over all sites (mean absolute occupancy) obtained by ORE-seq for the indicated REs and biological replicates (WT1 to WT4) under the indicated conditions of RE concentration (Units) and incubation time. For longer incubation, fresh enzyme was refilled after 60 minutes (+refill). Numbers give average ± standard deviation of replicates. (**C)** Correlation plot of ORE-seq data at HindIII sites obtained by AluI versus HindIII digestion. <|x-y|> denotes the mean absolute difference and <x> and <y> denote the mean absolute occupancy measured by AluI or HindIII, respectively, over 1100 compared sites. **(D)** Histograms of DNA fragments with 3' ends at the indicated distance to the indicated RE cut sites for the indicated ORE-seq samples (all 30-minute incubations). Read ends within green areas are counted as "cut by the RE" in the analysis (Supplemental Methods). **(E)** Mean RE cut site resection lengths in dependency of RE type and incubation time. **(F)** Scheme analogous to Fig. 1A, left, but for cut-all cut method. **(G)** Calibration curves for different ways of genome-wide detection of RE accessibility and absolute occupancy derived thereof. Yeast gDNA preparations pre-cut or not with the stated RE were mixed to yield the indicated fraction of uncut DNA (x-axis), which was measured as absolute occupancy (y-axis) by the indicated method variants. Error bars correspond to the standard deviation between sites. For details see Supplemental Methods. **(H)** Dependency of the relative fit error for the calibration curve in panel G on a chosen uncut correction factor for the indicated REs. **(I)** Comparison of mean absolute occupancy values for the indicated REs and ex vivo chromatin replicates (WT1 to WT4) always after 30-minute incubation time and for low and high RE concentrations but analyzed either by the cut-all cut or the cut-uncut method. **(J)** Top: Scheme of the nucleosome organization at the yeast *PHO5* promoter. Vertical line denotes DNA sequence, stippled circles positioned nucleosomes, broken arrow *PHO5* TSS, HS2 labeled box the position of hypersensitive site 2. The position of the BamHI cleavage site as well as cleavage sites of secondary HaeIII digest are indicated by vertical arrows. The length of a DNA fragment uncut by BamHI but excised by HaeIII and detected by the indicated Southern blot hybridization *PHO5* probe is shown relative to the lengths of the two corresponding DNA fragments generated by BamHI cleavage. Bottom left: Phosphorimager scan of Southern blot of the indicated ex vivo chromatin replicates (WT2 and WT4) treated with no, low or high concentration (Units) of BamHI. Bottom right: Correlation of absolute occupancy at indicated sites determined by ORE-seq versus by Southern blot. The mean absolute difference for each site (<|x-y|>) is indicated.

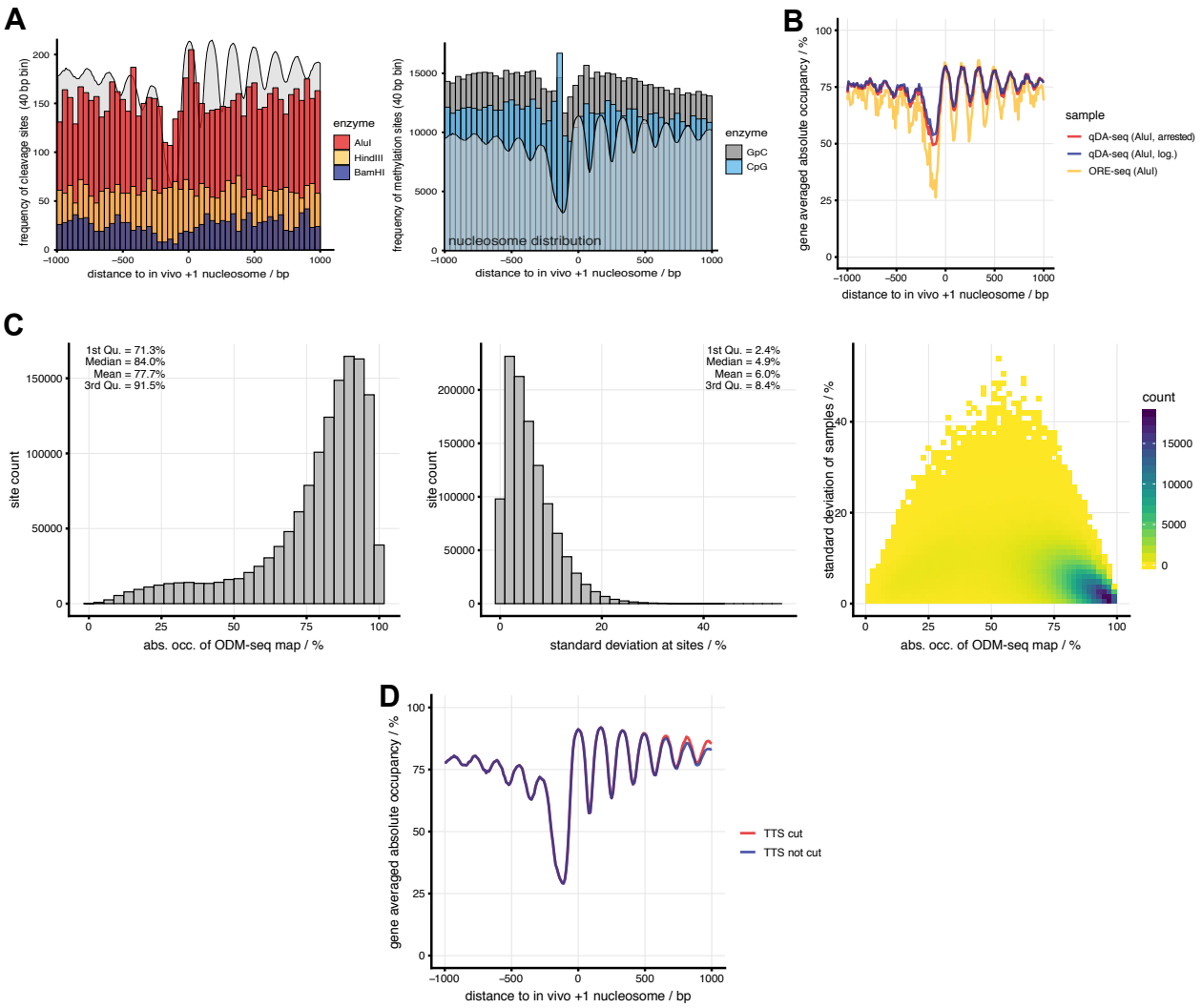


Supplementary Figure S2. Figure caption on next page.

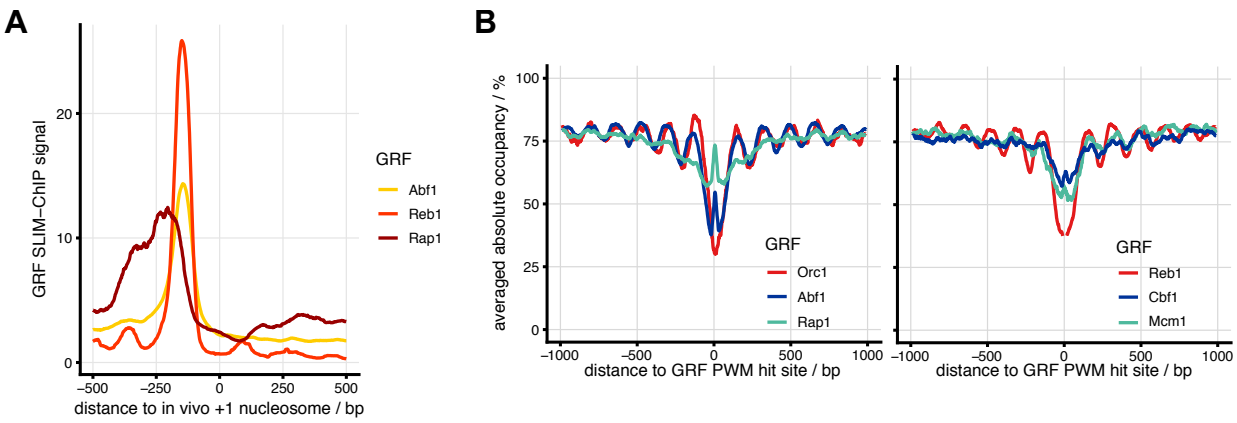
Supplementary Figure S2. (A) Comparison of data averaged over the indicated sites between ODM-seq (bisulfite-seq) using M.SssI (CpG) or M.CviPI (GpC) and ORE-seq using BamHI for genome-wide in vitro reconstituted salt gradient dialysis (SGD) chromatin of high or low nucleosome density and for indicated enzyme concentrations (Units) and incubation times. For longer incubation, fresh enzyme and for DNMTs fresh SAM was refilled after 60 minutes (+refill). **(B)** Correlation of mean absolute occupancy data for CpG and GpC methylated SGD chromatin as in panel A but only at GCG sites. **(C)** ODM-seq for ex vivo yeast chromatin replicates (WT1, WT3-WT5) under the indicated conditions and for the indicated DNMTs. “xl.” abbreviates “in vivo formaldehyde crosslinked”. **(D)** As panel C but for indicated regions on the 25mer 601 array plasmid SGD chromatin spiked into the indicated ex vivo chromatin replicates. Vertical brackets denote where the same spike-in chromatin was used, i.e., the same for CpG and GpC methylation of each replicate and the same for the comparison of ± crosslinking for WT4. **(E)** +1 nucleosome-aligned composite plots of gene averaged absolute occupancy measured for the indicated ex vivo chromatin replicates, conditions and time points corresponding to panel C.



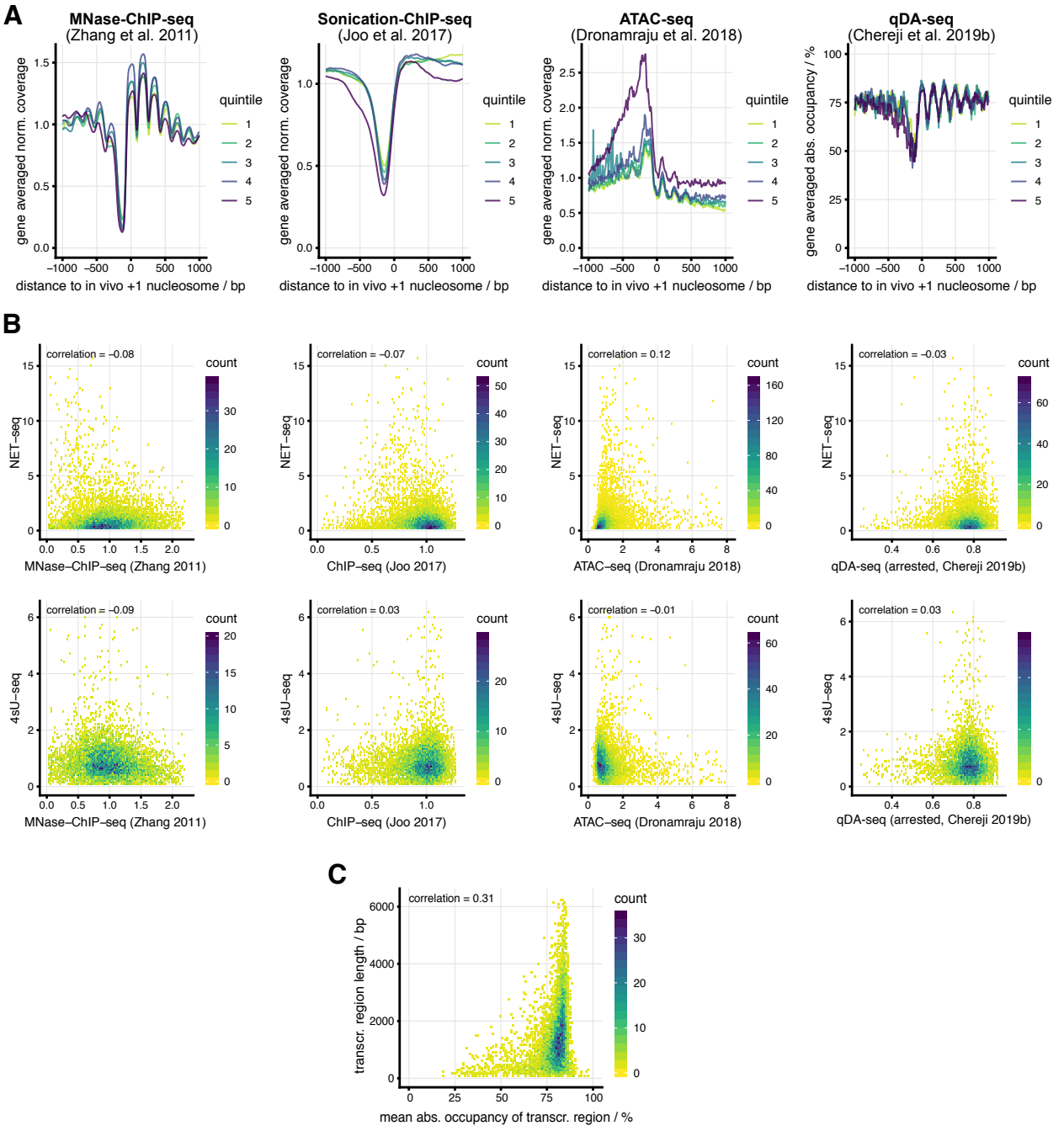
Supplementary Figure S3. (A) Top: +1 nucleosome-aligned composite plots of normalized sequencing coverage for the indicated chromatin replicates, enzymes and conditions. EM-seq stands for Enzymatic Methyl-sequencing. Bottom: Histogram of fragment length distributions for the same samples as in top graphs. **(B)** As Fig. 1C but for WT5 replicate and the indicated DNMTs and crosslinking times. **(C)** As Supplemental Fig. S1B but including ODM-seq samples, crosslinked (x-linked) chromatin preparations and only for 30-minute incubation time and high RE concentration.



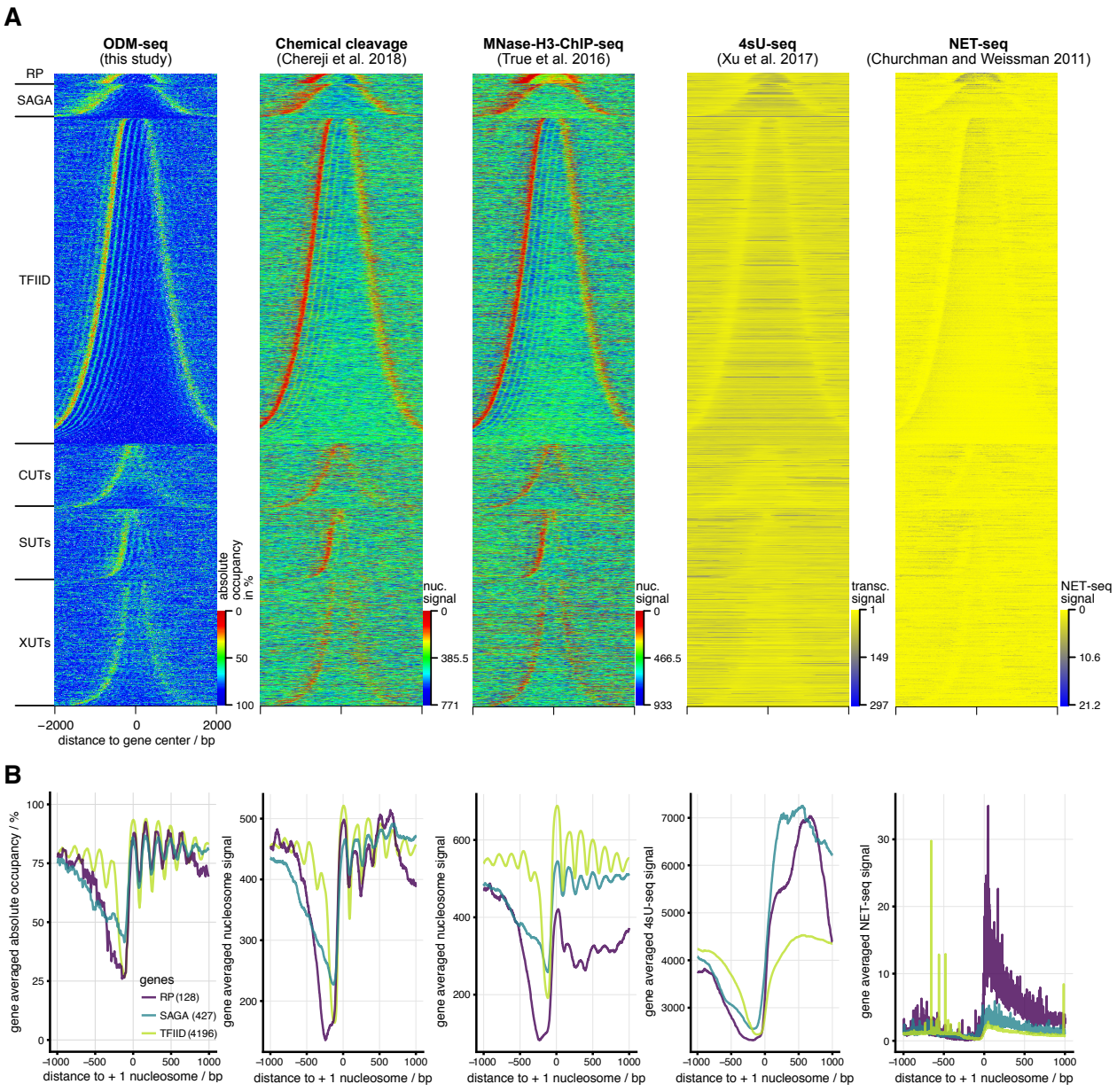
Supplementary Figure S4. (A) +1 nucleosome-aligned composite plot of site distributions of the indicated REs (left) and DNA methylation sites (right) within 40 bp bins and of ODM-seq absolute occupancy map (grey background). **(B)** As Fig. 1C but comparing our ORE-seq AluI data with qDA-seq AluI data for G1-arrested and logarithmically growing (log.) cells from Chereji et al. 2019b. **(C)** Left: histogram of absolute occupancies at DNA methylation sites. Center: histogram of standard deviation between replicates at DNA methylation sites. Right: correlation plot (color indicates number of occurrences) for absolute occupancy versus standard deviation of samples at DNA methylation sites. **(D)** As Fig. 1C but for ODM-seq data averaged using all sites on the complete downstream region (TTS not cut) or only until the TTS of each gene (TTS cut).



Supplementary Figure S5. (A) +1 nucleosome-aligned composite plots of SLIM-ChIP data for the indicated GRFs. **(B)** ODM-seq map aligned at GRF sites according to PWM hit sites called with the recommended dataset from ScerTF (Spivak and Stormo 2012).



Supplementary Figure S6. (A) As Fig. 4A but for the indicated data sets (Supplementary Table S2). **(B)** As Fig. 4B but for the indicated data sets, corresponding to panel A. **(C)** Correlation of average ODM-seq absolute occupancy on transcribed regions with region lengths.



Supplementary Figure S7. (A) Heat maps of the indicated data sets (Supplemental Table S2) aligned at gene centers. Rows were grouped (Vinayachandran et al. 2018) according to ribosomal protein (RP), SAGA-, TFIID-dependent genes as well as cryptic unstable transcripts (CUTs), stable unannotated transcripts (SUTs) and Xrn1-dependent unstable transcripts (XUTs) and sorted within each group from top to bottom by increasing gene length. **(B)** +1 nucleosome-aligned composite plots of data as in panel A for the RP, SAGA- and TFIID-dependent genes.

Supplementary References

Darst RP, Nabilsi NH, Pardo CE, Riva A, Kladde MP. 2012. DNA methyltransferase accessibility protocol for individual templates by deep sequencing. *Methods in enzymology* **513**: 185-204.

Krietenstein N, Wippo CJ, Lieleg C, Korber P. 2012. Genome-wide in vitro reconstitution of yeast chromatin with in vivo-like nucleosome positioning. *Methods in enzymology* **513**: 205-232.

Lieleg C, Ketterer P, Nuebler J, Ludwigsen J, Gerland U, Dietz H, Mueller-Planitz F, Korber P. 2015. Nucleosome spacing generated by ISWI and CHD1 remodelers is constant regardless of nucleosome density. *Molecular and cellular biology* **35**: 1588-1605.

Lowary PT, Widom J. 1998. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology* **276**: 19-42.

Matsuo K, Silke J, Gramatikoff K, Schaffner W. 1994. The CpG-specific methylase SssI has topoisomerase activity in the presence of Mg2+. *Nucleic acids research* **22**: 5354-5359.

Riley D, Weintraub H. 1978. Nucleosomal DNA is digested to repeats of 10 bases by exonuclease III. *Cell* **13**: 281-293.

Schwarz PM, Hansen JC. 1994. Formation and stability of higher order chromatin structures. Contributions of the histone octamer. *The Journal of biological chemistry* **269**: 16284-16289.

Spivak AT, Stormo GD. 2012. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic acids research* **40**: D162-168.

White CL, Suto RK, Luger K. 2001. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. *The EMBO journal* **20**: 5207-5218.

Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. 2005. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, NY)* **309**: 626-630.