Supporting Information

Moyle-Heyrman et al. 10.1073/pnas.1315809110

SI Materials and Methods

1. H4S48C *Saccharomyces pombe* Strain. The H4S48C *Saccharomyces pombe* strain was engineered in the KGY425 (American Type Culture Collection 96155) (*h- his3-D1 leu1-32 ura4-D18 ade6-M210*) background. The *HHF1* and *HHF2* genes encoding two of the three copies of H4 were mutated from serine to cysteine at position 48 by a standard pop-in/pop-out two-step method using linear PCR-generated cassettes targeted to endogenous loci through homologous recombination and selected for a gain/loss of the ura4 marker. The third copy of the H4 gene (*HHF3*) was not altered. The H4S48C *S. pombe* strain demonstrated normal growth at 30 °C on rich medium.

2. In Vivo Chemical Mapping and Preparation of Mapped DNA Fragment for Applied Biosystems Sequencing by Oligonucleotide Ligation and Detection. The H4S48C *S. pombe* strain was grown at 30 °C in 500 mL of yeast extract with supplements (YES) medium (1) to a mid-log phase (OD_{600} = 0.7) and chemically mapped as previously described (2, 3) with the following modification in the spheroplasting and lysis steps. Cells were collected by centrifugation and washed twice with water; after removing as much supernatant as possible, the weight of the pellet was determined. After 10 min of incubation at 30 °C in 20 mL of preincubation buffer [20 mM citric acid, 20 mM Na_2HPO_4, 30 mM β-mercaptoethanol (BME), 40 mM EDTA], cells were resuspended in permeabilizing buffer [1 M sorbitol, 50 mM Tris·HCl (pH 7.4), 5 mM BME with 5 mg of lyticase (L5263vc-200KU; Sigma)] using 10 mL per gram of original cell pellet and spheroplasted at 30 °C with gentle shaking. Spheroplast formation was terminated at an OD_{600} of 30–40% of the original value by centrifugation at 4 °C. Spheroplasts were washed twice in 5 mL of 1 M sorbitol and 50 mM Tris·HCl, and were resuspended in 2 mL of labeling buffer [1 M sorbitol, 50 mM NaCl, 10 mM Tris·HCl (pH 7.5), 5 mM $MgCl_2$, 0.5 mM spermidine, 0.15 mM spermine, 0.075% Nonidet P-40]. After the chemical mapping reaction, the lowest molecular weight DNA band corresponding to the center-to-center distance of two nucleosome neighbors was purified and DNA was prepared for sequencing as previously described (2, 3). The libraries prepared from two biological replicates were paired-end sequenced on Sequencing by Oligonucleotide Ligation and Detection (SOLiD) 4 and SOLiD 5500 systems (Applied Biosystems).

3. Alignment. The paired-end sequencing reads from the two experiments were mapped to the *S. pombe* genome (PomBase version 16.28, released on March 10, 2012) using Bowtie (http://bowtie-bio.sourceforge.net/index.shtml). In the first stage of alignment, we mapped the paired-end reads by allowing for zero to three mismatches progressively. If any reads were aligned for the given mismatch tolerance, they were not realigned in the following steps. For paired-end reads that were not aligned in the first stage, we aligned each end separately by allowing zero to three mismatches progressively. If one sequencing read (paired or single end) was aligned to n multiple locations, each location was assigned with a reads frequency of $1/n$. By allowing nonunique reads, we are able to provide a nucleosome map in the heterochromatin regions, which are rich in repetitive elements. The two stages of alignment yielded a total of 99 million cleavages on each strand from the two experiments together.

To characterize the cleavage patterns around the dyad, we followed the model training approach of Brogaard et al. (2) and identified four distinct cleavage patterns, each of which is specified by the presence or absence of a base A or T at the −3 or +3 position, respectively. These four templates, denoted as A(−3)T(+3), A(−3)−T(+3), −A(−3)T(+3), and −A(−3)−T(+3) [where A(−3)−T(+3) stands for the presence of "A" at the −3 position and the absence of "T" at the +3 position, etc.], were built into a Poisson cluster model (2). We applied the deconvolution method developed by Brogaard et al. (2) and computed the nucleosome center positioning (NCP) score at every genomic location on the Watson and Crick strands. Each NCP score on each strand essentially reflects the amount of cleavage that complies with the characteristic cleavage pattern defined in the corresponding template. We thus defined the NCP score at a given genomic location, denoted $\overline{k_i}$, as the average of the Watson and Crick NCP scores. For every genomic position i, we used the average of the lower half of the observed cut frequency in the region of $i \pm 73$ bp of each strand to represent the noise level. We denoted the average noise level from the two strands as λ_i^0. The average NCP score-to-noise ratio, $\overline{k_i}/\lambda_i^0$, was used as the index for the nucleosome positioning signal to define unique nucleosomes. We used a greedy algorithm to make nucleosome calls sequentially based on the magnitude of $\overline{k_i}/\lambda_i^0$ as follows. On each chromosome, the position that had the largest $\overline{k_i}/\lambda_i^0$ was first called as the center of a nucleosome. Another position with the largest $\overline{k_i}/\lambda_i^0$ among all positions that were at least ±107 bp away from the first selected nucleosome was then called as a nucleosome center. This step was repeated such that every selected nucleosome in the current step was at least ±107 bp away from any previously selected nucleosomes. The algorithm stopped when no further nucleosomes could be called. By this approach, we allowed a maximum overlap of 40 bp between two neighboring nucleosomes. Based on all defined nucleosomes, we selected the top 90% of nucleosomes with the highest NCP score/noise ratio as the final unique set (a total of 73,945 nucleosomes) to represent the most premium nucleosomes (Dataset S2). The smallest $\overline{k_i}/\lambda_i^0$ value from the unique set was used as a threshold value to define further a redundant map of 411,590 nucleosomes (Dataset S3).

4. Nucleosome Occupancy Scores. The NCP score defined from the deconvolution algorithm provides a measure of the relative amount of nucleosomes centered at every genomic location. In some regions, particularly regions with long linkers [e.g., upstream of transcription start site (TSS)], the cleavage cut frequency on one strand can be substantially lower than that on the other due to the gel selection in the experiments (2). To correct for this bias, we calculated the ratio of cleavage frequency of the two strands for every 147-bp region. If the logarithm of the ratio exceeds 2 SDs of the genome-wide average, the larger NCP score from either strand will be regarded as the average NCP score at location i (i.e., $\overline{k_i}$; see above).

Throughout the paper, we defined the nucleosome occupancy score as the total NCP score of the redundant nucleosomes in the ±60-bp region of every genomic location for both *S. pombe* and *Saccharomyces cerevisiae* to make them comparable. A similar window size has been used in the literature for *S. pombe* data (e.g., ±60 bp and ±50 bp were used, respectively, in refs. 4 and 5). The defined nucleosome occupancy is an evidence-based relative measure of the amount of nucleosome coverage at each position. This is analogous to the reads occupancy commonly used in the micrococcal nuclease (MNase) mapping, in which the nucleosome occupancy is measured by the number of reads that cover each genomic position. For the single-end MNase data used in this paper (4, 5), we need to project the nucleosome center based on each read. Because the linker length and the MNase digestion extent may vary, the distance between the nucleosome center and the read start position also varies at different nucleosome locations. Instead of projecting the nucleosome center by shifting the read start position downstream by a constant distance (as done in refs. 4 and 5), we took an adaptive approach (as done in ref. 6) to project the center of nucleosomes based on every sequencing read. The nucleosome occupancy at any genomic location was defined as the total number of nucleosome centers in the ±60-bp regions. For paired-end MNase data for *S. cerevisiae* (7), the nucleosome occupancy at any genomic location is calculated as the total number of reads of length between 142 and 152 bp that were centered in the ±60-bp region.

When calculating the nucleosome occupancy over poly (dA-dT) or poly (dG-dC) tracts, we made one exception (Fig. 2*D* and Fig. S2 *H–J*). It is possible that these polymer tracts may prefer to be positioned at the nucleosome edges; thus, we defined the occupancy based on the NCP scores in ±73-bp region for the chemical map. Likewise, for MNase maps, the nucleosome occupancy at any given location was defined as the total projected nucleosome centers in the ±73-bp region.

5. Nucleosome Fuzziness. For any given unique nucleosome at position i, define the normalized NCP score as follows:

$$k_i' = k_i \Bigg/ \sum_{j=i-73}^{i+73} k_j.$$

The fuzziness score at position i is defined as the Shannon's entropy based on the normalized NCP score:

$$s_i = -\sum_{j=i-73}^{i+73} k_j' \log\left(k_j'\right).$$

The fuzziness score achieves the maximum if the NCP scores in the ±73-bp region are all equal, which implies the fuzziest positioning.

6. Additional Details for Different Figures. ***Fig. 1.*** In Fig. 1*B*, based on the redundant map, we first calculated the distance between every pair of nucleosomes. Because the NCP score roughly reflects the abundance of nucleosomes centered at each genomic location, we also calculated the NCP score-weighted internucleosomal distance frequency as follows. Let I be the indices of genomic positions on one chromosome, and let $R \subset I$ be the set indexing the redundant nucleosome positions on the same chromosome. Let k_i be the NCP score defined at position $i \in I$. Define $k_i' = k_i$ if $i \in R$ and $k_i' = 0$ otherwise. The NCP score-weighted internucleosome distance frequency is calculated as follows:

$$f(d) = \sum_{i=1}^{I} k_i' k_{i+d}'.$$

The calculated $f(d)$ from all chromosomes is summed and then normalized in the range of $d = 1, \ldots, 100$.

In Fig. S1*C*, for the MNase map (4), we followed the method of Yigit et al. (6) to project the center of the nucleosome based on each sequencing read. Based on the projected nucleosome center of each read, we assigned a Gaussian weight of $\exp[-0.5^*(d/20)\hat{}2]$ to a position d bp away from the center of the sequence for $d \leq 73$. The aggregated weight at each position gives a center-weighted reads occupancy score genome-wide. We identified 74,126 well-defined peaks on the reads occupancy curve as putative nucleosome centers by controlling the peak height and steepness simultaneously. The distance was measured between the 74,126 putative nucleosomes and the nearest nucleosomes from the unique map of the chemical experiment.

Fig. 2. In Fig. 2*B*, the A/T fraction as a function of position within nucleosomes was calculated based on the unique nucleosomes. The curves were normalized by the genome-averaged A/T fraction and further smoothed using the ±10-bp moving average.

In Fig. 2*C*, we divided the entire *S. cerevisiae* and *S. pombe* genomes into 20-bp consecutive bins. The occupancy and A/T fraction at each bin were calculated. The bins were then divided into 18 categories according to the A/T fraction within each bin with a 5% stepwise increment. The average nucleosome occupancy within each category was plotted for both species.

In Fig. S2*G*, for each 20-bp bin defined in Fig. 2*D*, we calculated the distance between the bin center and the nucleosome centers in the ±80-bp region. Considering that multiple nucleosome positions typically exist in most genomic regions, we calculated the distance distribution based on the redundant nucleosome map weighted by the NCP score. For example, for bins from a given A/T fraction category, let $X = \{X_j : j = 1, \ldots, n\}$ be the centers of bins, let $R_j = \{r_{jm} : m = 1, \ldots, M\}$ be the center positions of M nucleosomes defined within ±80 bp of bin i, and let $K_j = \{k_{jm} : m = 1, \ldots, M\}$ be the corresponding NCP scores. For the jth bin centered at X_j, the distance between the bin center and nucleosome center defined by $|r_{jm} - X_j|$ is associated with the NCP score, k_{jm}, for $m = 1, \ldots, M$. The box plot for each A/T fraction category was generated using the NCP score k_{jm} as a frequency weight for the corresponding distance $|r_{jm} - X_j|$.

In Fig. 2*D* and Fig. S2 *H–J*, we denoted the poly (dA) tract without mismatches as A0, the poly (dA) tract with one internal mismatch as A1, and so forth. The immediate upstream or downstream base pair of a poly (dA) tract must be a non-A nucleotide. For example, a sequence TAAAAAG contains an A0 of length 5, but a sequence TAAAAAAG cannot be counted as two A0 tracts of length 5; instead, it is counted as an A0 tract of length 6. Likewise, an A1 tract of length 5 is referred to a sequence of five nucleotides, among which four are A's, but with one internal non-A [position 2 or 3 or 4, not in the first and last positions of the 5-mer (e.g., AAGAA or AGAAA or AAAGA)]. Again, an A1 tract of length 5 must be preceded or followed immediately by a non-A nucleotide. The same rule applies to poly (dG) tracts.

We first identified every poly (dA) or poly (dG) tract of a given length in the uniquely mappable regions genome-wide. The nucleosome occupancy score at every base pair of a given tract was first averaged. The average occupancy was further averaged across all tracts of the same kind identified genome-wide. The final average occupancy score was normalized by genome average occupancy (from the uniquely mapped regions) and presented in the $\log_2$ scale. Each curve was truncated at the length above which the polymers are observed in less than 15 cases genome-wide.

In Fig. 2 *E* and *F* and Fig. S2*K*, based on the unique nucleosome map, we calculated the distance from the center of every poly (dA) or poly (dT) tract on both strands to the nearest nucleosome center. Because of the reverse complementarity of the two strands, the curves are strictly center-symmetrical. All curves were smoothed using a Gaussian kernel with an SD equal to 2 bp.

Fig. 3. In Fig. 3*A,* the occupancy from chemical experiments was calculated as described in *SI Materials and Methods*, section 4. The occupancy from the ±1,000-bp region of 4,013 TSSs was first averaged and then normalized by the genome average occupancy. The plotted occupancy was the normalized averaged occupancy in the $\log_2$ scale.

Fig. 3*B* is the same as Fig. 3*A*, but the NCP score was averaged instead of the nucleosome occupancy score. The NCP score occupancy is analogous to the dyad occupancy score used in the

MNase data, which indicates the likelihood of each position being occupied by a nucleosome center.

In Fig. 3*C*, the RNA-sequencing data were obtained from Wilhelm et al. (8). We first downloaded the *S. pombe* transcript annotation file (.gtf file) from ftp://ftp.ensemblgenomes.org/pub/fungi/release-15/gtf/schizosaccharomyces_pombe/. We used TopHat (http://tophat.cbcb.umd.edu/) software for alignment of reads by setting the minimum and maximum intron lengths as 50 bp and 50,000 bp, respectively. The Cufflinks software was used to define the gene expression value in fragments per kilobase of transcript per million mapped reads (FPKM) using the above-referenced transcript file (http://cufflinks.cbcb.umd.edu; Dataset S4). Log_{10} transformation of the FPKM value was used in the analyses. Some of the expression values in FPKM were essentially zero. Log transformation would result in extremely negative values. To avoid such outliers, if the gene expression in FPKM was less than 0.1 (in a total of 35 genes), the $\log_{10}$ of FPKM was set as -1.

In Fig. 3*D*, based on the unique nucleosomes, we trained a fourth-order duration hidden Markov model (dHMM) proposed by Xi et al. (9). The dHMM for the nucleosome position prediction requires the input of a nucleosome profile model, a linker profile model, and a linker length distribution. We trained the nucleosome profile as a fourth-order position-specific Markov chain based on the mapped unique nucleosomes and the linker profile as a homogeneous fourth-order Markov chain. The linker length distribution used in the prediction was a uniform distribution defined in the range of 1, 2. . ., 500 bp. The dHMM outputs the probability for each genomic position to be the center of a nucleosome. To be consistent with the definition of occupancy score based on the chemical map or MNase map, we defined the occupancy score from the dHMM at a given location as the aggregated probabilities in the ±60-bp region. The occupancy scores from the dHMM and from the chemical map were both normalized by the genome average and plotted in the $\log_2$ scale.

Fig. 3*E* is the same as Fig. 3*D*, except that it is for *S. cerevisiae*.

In Fig. 3*F*, we calculated the MNase occupancy score from the published studies (4, 5) as described in *SI Materials and Methods*, section 4. The occupancy scores from 217 DNA replication origins [also known as autonomously replicating sequences (ARS); coordinates from ref. 5] were averaged and normalized by the genome-wide average occupancy score.

Fig. 4. The heterochromatic regions are typically rich in repetitive elements. In the alignment, if a read starting at a given position is mapped to k multiple locations, the mappability of this given position is $1/k$. The average weight in every genomic location shown in Fig. 4 was the averaged mappability score in the ±10-bp region.

In Fig. 4*B*, the occupancy curves from the two MNase maps were constructed exactly in the same way as described in *SI Materials and Methods*, section 4. The NCP score was normalized by the average NCP score of the redundant nucleosomes, and the MNase occupancy was normalized by the genome average occupancy from respective MNase maps.

In Fig. S4*I*, from the unique nucleosome map, we calculated the linker DNA length genome-wide. The cumulative distribution of the linker DNA length was based on the frequency of linker length observed in the range from 0–100 bp.

1. Forsburg SL, Rhind N (2006) Basic methods for fission yeast. *Yeast* 23(3):173–183.
2. Brogaard K, Xi L, Wang JP, Widom J (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature* 486(7404):496–501.
3. Brogaard KR, Xi L, Wang JP, Widom J (2012) A chemical approach to mapping nucleosomes at base pair resolution in yeast. *Methods Enzymol* 513:315–334.
4. Tsankov A, Yanagisawa Y, Rhind N, Regev A, Rando OJ (2011) Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome Res* 21(11):1851–1862.
5. Givens RM, et al. (2012) Chromatin architectures at fission yeast transcriptional promoters and replication origins. *Nucleic Acids Res* 40(15):7176–7189.
6. Yigit E, et al. (2013) High-resolution nucleosome mapping of targeted regions using BAC-based enrichment. *Nucleic Acids Res* 41(7):e87.
7. Floer M, et al. (2010) A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. *Cell* 141(3):407–418.
8. Wilhelm BT, et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453(7199):1239–1243.
9. Xi L, et al. (2010) Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics* 11:346.

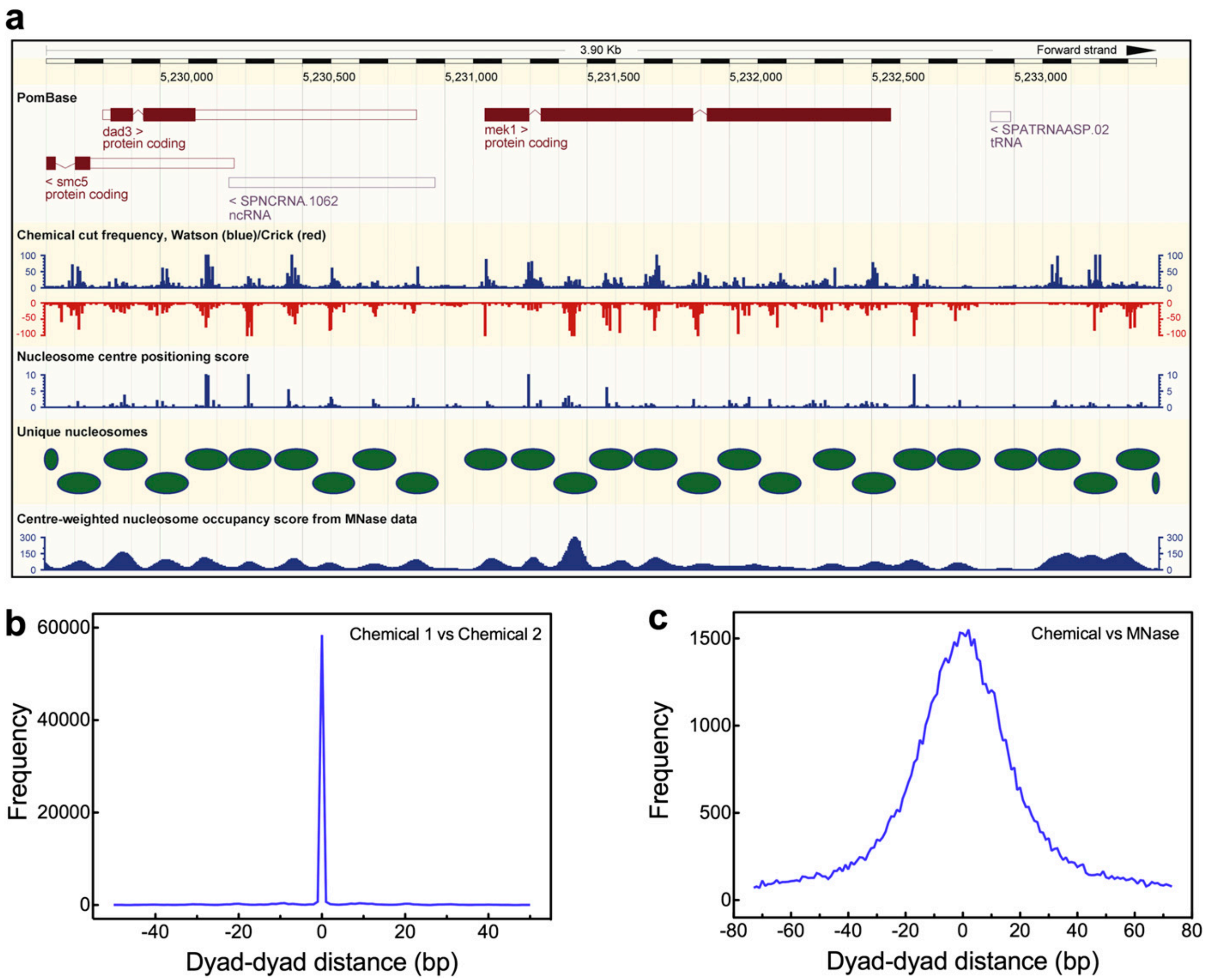

Fig. S1. Overview of the chemical map for *S. pombe*. (*A*) Screen shot of raw data, NCP scores from deconvolution algorithm, unique nucleosomes called, and center-weighted nucleosome occupancy score from published MNase data (4). (*B*) Plot of the distance between the unique nucleosomes from two independent mapping experiments, suggesting the high consistency and reproducibility of the chemical mapping technique. (*C*) Plot of the distance between unique nucleosomes from the chemical map and from the MNase map (4), suggesting a general consistency of the two maps in defining the nucleosome centers.

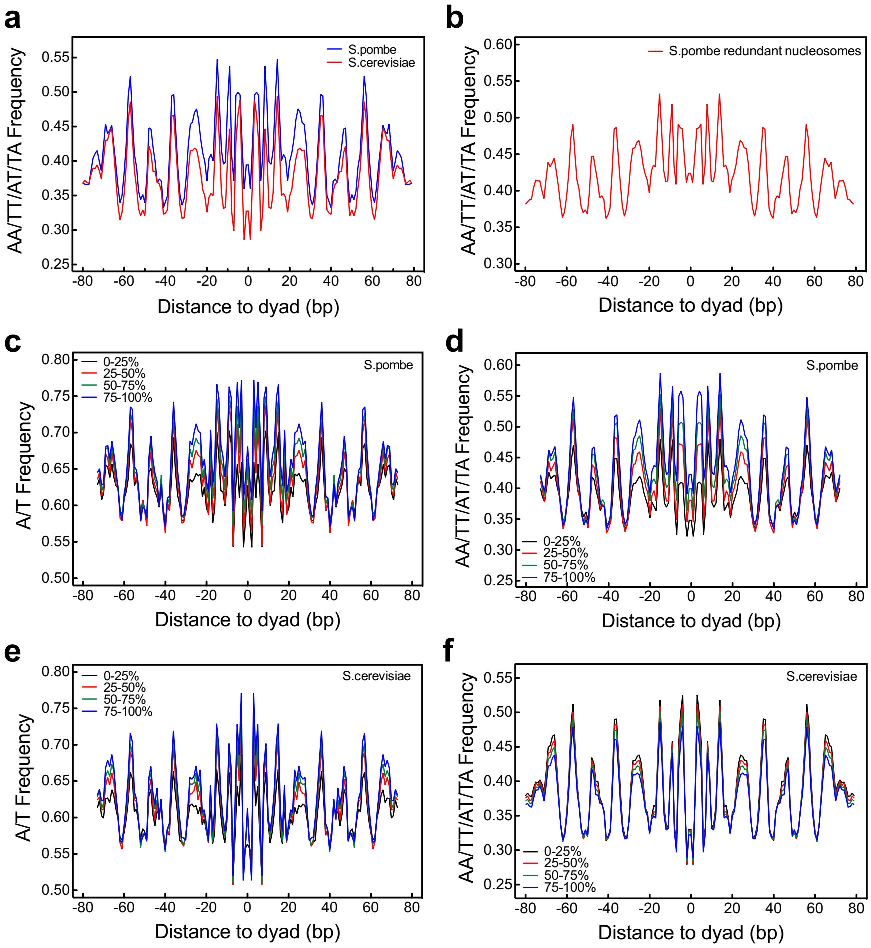

Fig. S2. Continued

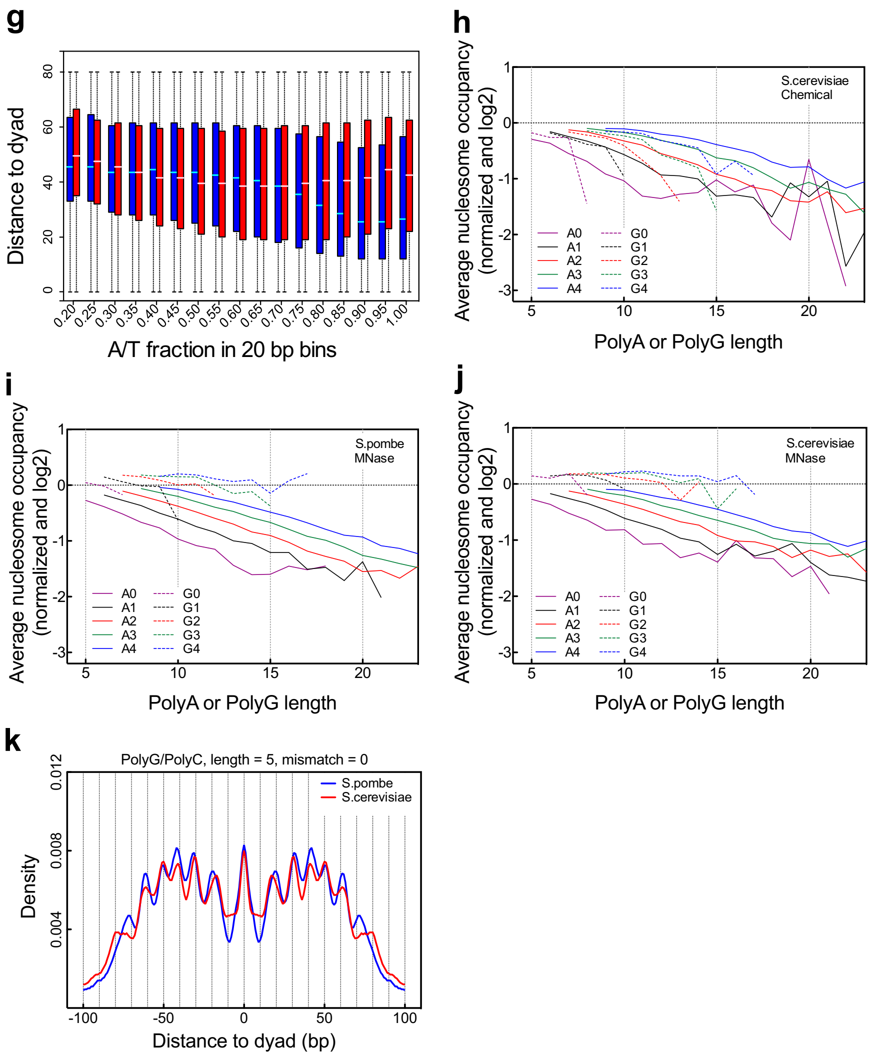

Fig. S2. Distinct sequence features of *S. pombe* nucleosomes. Plot of AA/TT/AT/TA dinucleotide frequency in unique nucleosomes is defined in the chemical maps of *S. pombe* vs. *S. cerevisiae* (*A*) and in the redundant nucleosomes of *S. pombe* (*B*). (*C*) Mononucleotide fraction in the nucleosome region by quartiles of the NCP score/noise ratio in the unique nucleosome map in *S. pombe* (100% is the highest NCP score/noise ratio). (*D*) Same as in *C*, but for the AA/TT/AT/TA dinucleotide. (*E* and *F*) Mono- and dinucleotide plots are the same as in *B* and C, respectively, but are for *S. cerevisiae* (2). (*G*) Box plot of the average distance of 20-bp bins to the dyad as a function of A/T fraction shows that the A/T richer sequences are preferentially positioned closer to the dyad (*SI Materials and Methods*) in *S. pombe* (blue), whereas they are moderately disfavored by the dyad region in S. cerevisiae (red). (*H*) Normalized nucleosome occupancy of poly (dA-dT) in the $\log_2$ scale as a function of polymer length from the chemical map of *S. cerevisiae* shows that poly (dA-dT) tracts only slightly degrade occupancy as polymer length increases, whereas poly (dG-dC) tracts deplete nucleosome occupancy at a much more rapid rate. The zero line is the genome average occupancy. The A0/G0 curve stands for poly (dA-dT)/poly (dG-dC) with no mismatch, the A1/G1 curve stands for tracts with one internal mismatch, etc. (*I* and *J*) Average nucleosome occupancy normalized by genome average as a function of the length of poly(dA-dT) and poly(dG-dC) tracts plotted on the $\log_2$ scale. *S. pombe* single-end MNase data (4) (*I*) and *S. cerevisiae* paired-end data (7) (*J*) are shown. The zero line is the genome average occupancy. A0 stands for the poly (dA-dT) tract with straight A's or straight T's, and A1 stands for the poly (dA-dT) tract with one internal mismatch (e.g., AAAAGAAA or AAAATAAA for A1 of length 8, etc.). (*K*) Distance between the nucleosome dyad and centers of poly (dG-dC) tracts of length 5 in *S. pombe* vs. *S. cerevisiae*.

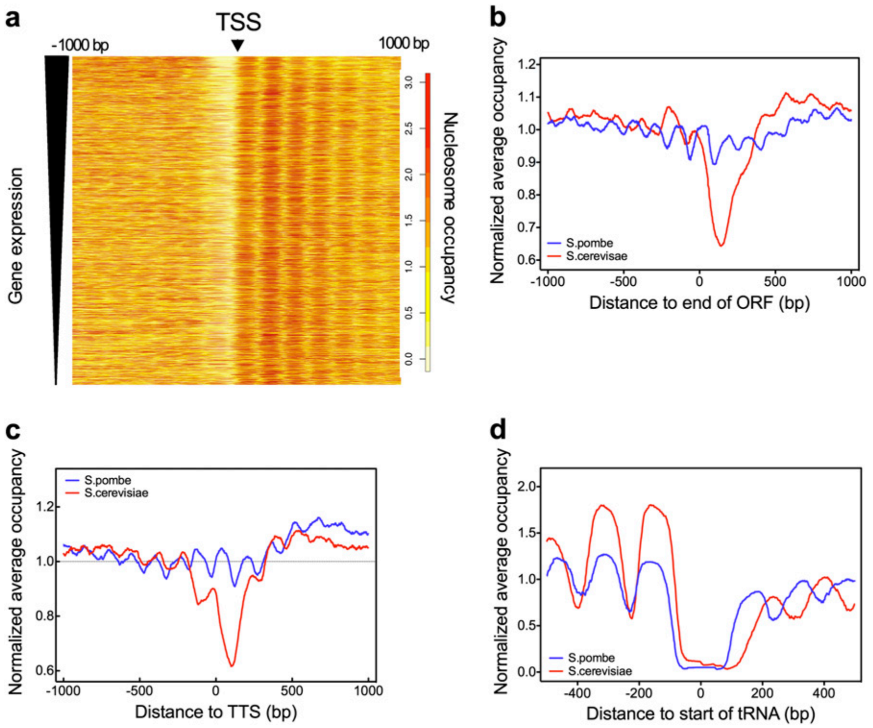

Fig. S3. Nucleosome occupancy at genetic landmarks. (*A*) Heat map of nucleosome occupancy aligned at the TSS, shown in decreasing order of gene expression (8) from top to bottom. A weak negative correlation exists between gene expression and nucleosome occupancy in both the promoter and gene body regions. (*B* and *C*) Nucleosome occupancy at the transcription termination site (TTS) and ORF end, respectively, for *S. pombe* vs. *S. cerevisiae* based on the chemical map. (*D*) Nucleosome occupancy at tRNA start for *S. pombe* vs. *S. cerevisiae* based on chemical maps.

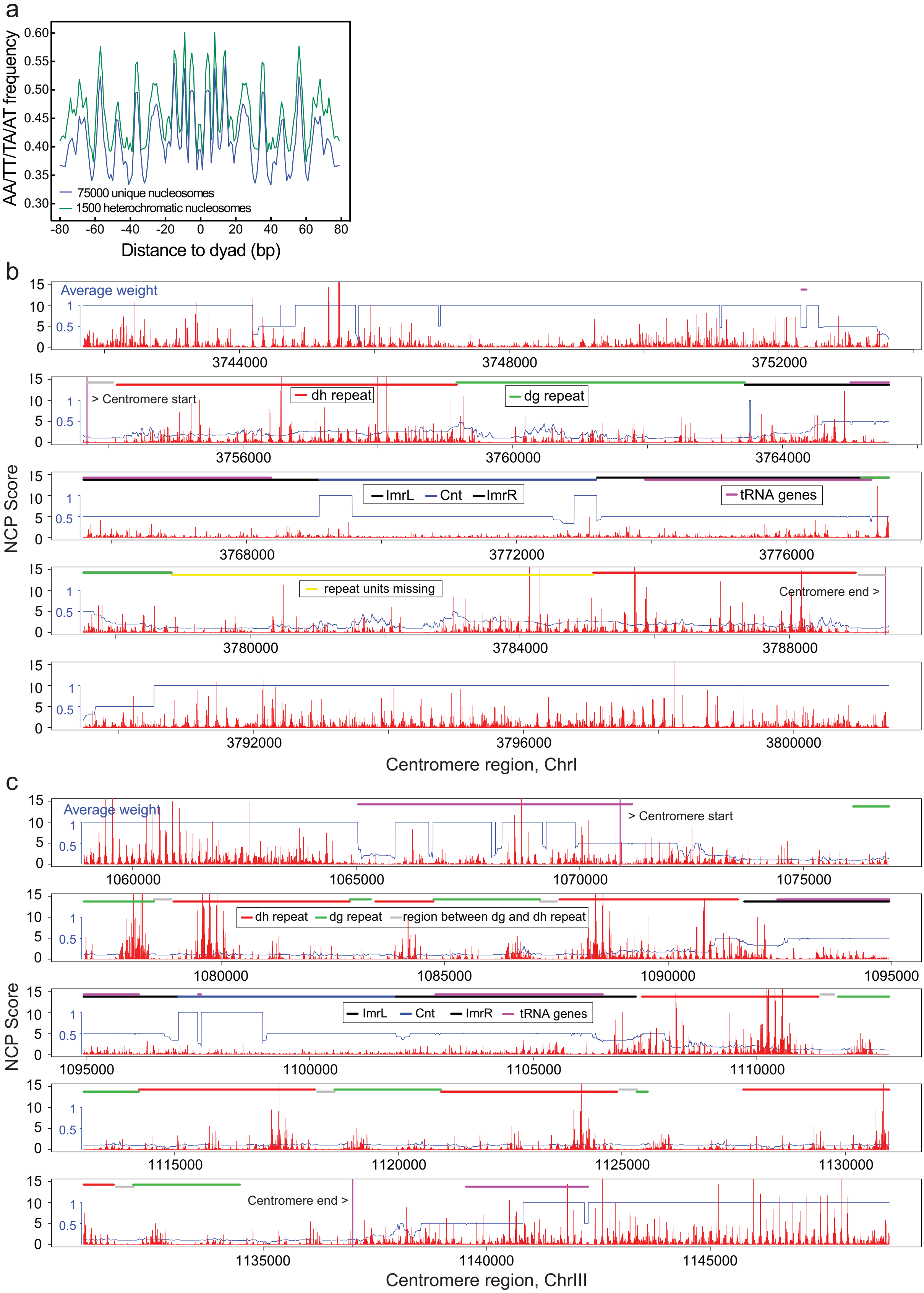

Fig. S4. Continued

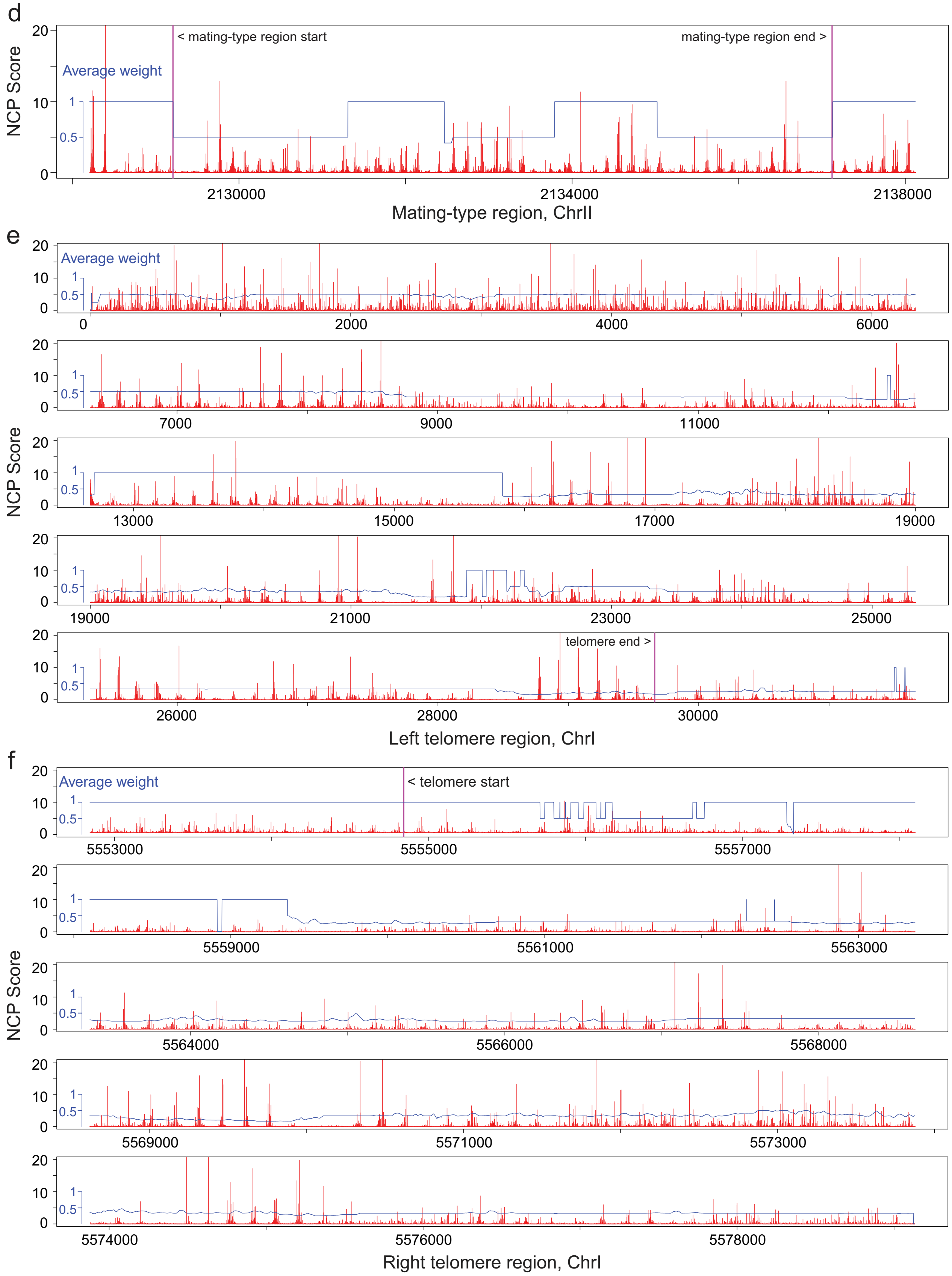

Fig. S4. Continued

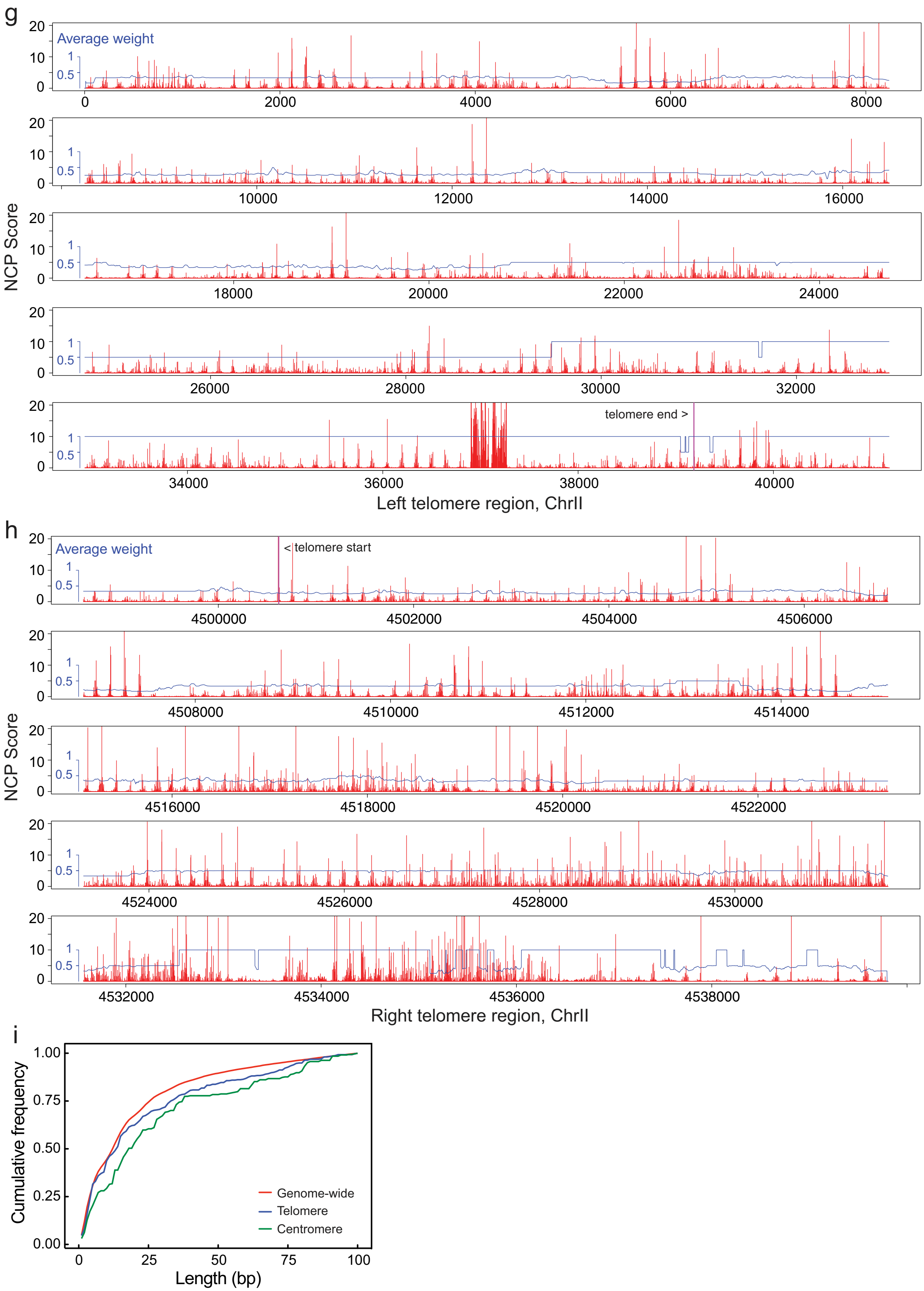

Fig. S4. Nucleosome occupancy in heterochromatin regions. (*A*) Plot of AA/TT/AT/TA dinucleotide frequency in nucleosomes from heterochromatin regions compared with that from the entire set of unique nucleosomes. (*B–H*) NCP score plot in different heterochromatin regions indicated as in the labels of the *x* axis. (*I*) Cumulative linker length distribution in centromere and telomere regions shows more sparse positioning in the heterochromatic regions. Chr, chromosome; Cnt, central core; dg and dh, two outer repeat elements; ImrL/R, inner repeat left/right.

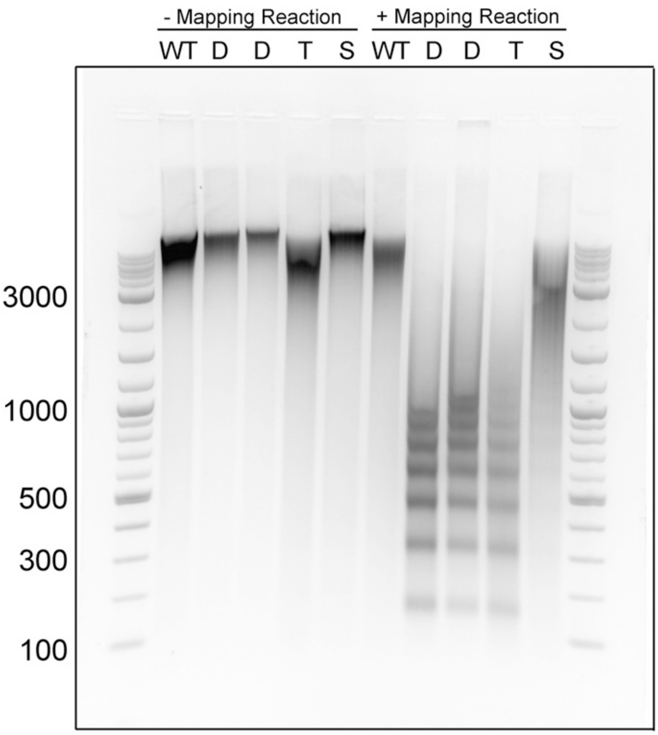

Fig. S5. Chemical mapping of H4S48C *S. pombe* strains. Ethidium bromide-stained agarose gel of chemical mapping reaction products in WT, a single mutant (S; *HHF2*-mutated), a double mutant (D; *HHF1*- and *HHF2*-mutated), and a triple mutant (T, all three H4 genes mutated) in the *S. pombe* strain. Efficient chemical cleavage was observed for the double- and triple-mutant *S. pombe* strains, but mapping was less efficient for the single-mutant *S. pombe* strain.

Dataset S1. NCP score and NCP score/noise ratio for unique nucleosomes defined genome-wide

Dataset S1

Dataset S2. NCP score and NCP score/noise ratio for redundant nucleosomes genome-wide

Dataset S2

Dataset S3. Nucleosome occupancy for poly (dA-dT) and poly (dG-dC) tracts in *S. pombe* and *S. cerevisiae* by chemical maps

Dataset S3

Dataset S4. Gene expression values in FPKM by Cufflinks based on RNA-seq data of Wilhelm et al. (1)

Dataset S4

1. Wilhelm BT, et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453(7199):1239–1243.

Dataset S5. Tag count mapped onto its own region and other regions

Dataset S5