ORE-seq: genome-wide absolute occupancy measurement by restriction enzyme accessibilities

Elisa Oberbeckmann[1#], Michael Roland Wolff[2#], Nils Krietenstein[3], Mark Heron[4], Andrea Schmid[5], Tobias Straub[6], Ulrich Gerland[2], Philipp Korber[5*]

1 Department of Molecular Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

2 Department of Physics, Technical University of Munich, Garching, Germany

3 The Novo Nordisk Center for Protein Research (CPR), Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

4 Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany; Gene Center, Faculty of Chemistry and Pharmacy, Ludwig-Maximilians-Universität München, Munich, Germany

5 Biomedical Center (BMC), Division of Molecular Biology, Faculty of Medicine, LMU Munich, Martinsried, Germany

6 Core Facility Bioinformatics, Biomedical Center (BMC), Faculty of Medicine, LMU Munich, Martinsried, Germany

shared first authors

* corresponding author: pkorber@lmu.de

Summary/Abstract

Digestion with restriction enzymes is a classical approach for probing DNA accessibility in chromatin. It allows to monitor both the cut and the uncut fraction and thereby the determination of accessibility or occupancy (= 1- accessibility) in absolute terms as the percentage of cut or uncut molecules, respectively, out of all molecules. The here presented protocol takes this classical approach to the genome-wide level. After exhaustive restriction enzyme digestion of chromatin, DNA is purified, sheared and converted into libraries for high throughput sequencing. Bioinformatic analysis counts uncut DNA fragments as well as DNA ends generated by restriction enzyme digest and derives thereof the fraction of accessible DNA. This straight forward principle is technically challenged as preparation and sequencing of the libraries leads to biased scoring of DNA fragments. Our protocol includes two orthogonal approaches to correct for this bias, the “corrected cut-uncut” and the “cut-all cut” method, so that accurate measurements of absolute accessibility or occupancy at restriction sites throughout a genome are possible. The protocol is presented for the example of *S. cerevisiae* chromatin but may be adapted for any other species.

Key Words

chromatin, DNA accessibility, absolute occupancy, restriction enzyme, high throughput sequencing

1. Introduction

Nucleases are often used to measure DNA accessibility in chromatin. DNA accessibility and the corresponding cleavability by nucleases is mainly modulated by DNA binding factors, of which histones are the most common in chromatin. DNA binding of a factor is characterized by the position where along the DNA sequence the factor binds (“peak position”) and the occupancy, i.e., which fraction of DNA molecules is occupied by the factor at this position (“peak height” or, more exactly, “area under the peak”). Occupancy is related to accessibility as the sum of both amounts to 100%. Many techniques that map DNA binding of a factor are good at determining the position but are limited regarding occupancy measurements. This is because they are often yield methods, i.e., they score either the bound (e.g., MNase-seq [1]) or unbound (e.g., ATAC-seq [2,3]) sub-population but not both. This allows, at best, to compare occupancies at different conditions relative to each other (relative occupancy) but not to measure occupancy in absolute terms. Absolute occupancy is defined as the percentage of DNA molecules bound by a factor. The measurement of absolute occupancy requires monitoring a) simultaneously the bound and unbound state, b) under saturating conditions and at c) sufficiently frozen binding-unbinding-dynamics to avoid a shift of the bound:unbound ratio during mapping. A prominent counter-example where mapping of absolute occupancy is not possible is the genome-wide mapping of nucleosomes, i.e., binding of histone octamers along DNA, by digestion of chromatin with Micrococcal nuclease (MNase). This technique relies on a carefully titrated and limited MNase digestion degree that removes most non-nucleosomal DNA while not yet cleaving the DNA wrapped around the histone octamers. This protected DNA is detected by high-throughput sequencing (MNase-seq). While MNase-seq readily determines nucleosome positions, it does not track the unbound state and does not work at saturation and therefore cannot measure absolute nucleosome occupancy.

Our here presented ORE-seq protocol offers a genome-wide and complementary technique for the determination of absolute occupancies. It employs type II restriction endonucleases (restriction enzymes, REs) that cleave DNA as dimers by embracing [4] defined short DNA

sequences (RE sites) with high specificity. This has the advantages that their cleavage is prevented by most DNA binders, leads to double strand breaks (DSBs) at predictable sites and with predictable DNA end properties and can be carried out to saturation without losing any DNA. Saturated digestion can be ensured by following a digestion time course or by comparing digestion with two sufficiently different RE concentrations over the same digestion time. The combination of both can be used to ensure that chromatin dynamics are sufficiently frozen, i.e., if two different RE concentrations yield similar accessibilities at two different time points each. After cleavage of all accessible RE sites on the level of chromatin, the DNA is purified so that all occluding DNA binders are removed, and DSBs flanking the RE sites are quantitatively introduced in a secondary cleavage step. This generates DNA fragments with all combinations of either one or two ends generated by the RE (RE end) and/or by the secondary cleavage. From the viewpoint of one particular RE site, there can be DNA fragments that span the RE site (uncut fragments) or that start/end with this RE site (cut fragments). As the RE cleavage generates two cut fragments that is compared to only one uncut fragment, the absolute accessibility at a certain RE site is given by the formula "number of cut fragments/(number of cut fragments + 2x number of uncut fragments)" and the absolute occupancy is "1-absolute accessibility".

This RE accessibility approach that relies on the ratio between cut and uncut fragments was established early on for single loci via Southern blot detection [5]. Here we adopt this approach for genome-wide measurements and call it the "cut-uncut" method. Secondary cleavage is achieved by DNA shearing to a degree that the DNA fragments are short enough for scoring their ends by Illumina high throughput sequencing. To allow a fine-grained determination of absolute occupancy, the sequencing coverage has to be at least 40fold. In theory, this should measure absolute accessibility or occupancy by counting the respective DNA ends right away.

However, in practice, we identified confounding factors and established ways to correct for them in our first application of this protocol [6]. Scoring too many RE ends due to fortuitous shearing at RE sites is controlled for by analysis of genome regions far away of RE sites and

is usually not a major problem. Conversely, scoring too few RE ends due to DNA end resection by contaminating exonucleases in chromatin is quite common but taken into account by scoring RE ends around RE sites within a window size determined from the fragment end distributions around RE sites. The Illumina sequencing platform shows a bias against long fragments (>500 bp). Therefore, only fragments of <500 bp size are included in the bioinformatics analysis. Finally, by using calibration samples consisting of mixtures between undigested and totally digested genomic DNA (gDNA) at known percentages we observed a bias that may reflect, at least in part, a preference for RE ends over DNA ends generated by shearing (shearing ends). Such a preference may be due to different ligation efficiencies at RE versus shearing ends as they differ in DNA sequence and as shearing ends generated by sonication may have an unusual chemistry [7] and be no or poor substrates for the enzymes used in DNA end polishing and adapter ligation during the preparation of sequencing libraries. In contrast, the REs provide 5'-phosphate DNA ends right away, which are, as blunt ends, required for the ligase. Our speculation that the bias observed with the calibration samples is linked to end polishing and adapter ligation is supported as this bias became less pronounced when we used a commercial and highly optimized sequencing library preparation kit (NEB) in the protocol presented here in contrast to our own enzyme mix used in our prior study [6]. But this could not completely eliminate this bias, which was also true in a similar approach that used in addtion a commercial DNA repair enzyme kit [8]. Nonetheless, this bias can be compensated by a correction factor derived from the calibration curve.

Alternatively, this bias can be circumvented by what we call the "cut-all cut" method. For this, a bioinformatically tractable spike-in gDNA, e.g., from a sufficiently different species, that was completely digested with an $RE^{spike\text{-}in}$ different from the RE used on chromatin is added to the purified DNA after RE digestion of chromatin and prior to DNA shearing. This mixture is divided into two aliquots and one aliquot is digested again with the same RE ("second RE cleavage", not to be confused with the secondary cleavage by sonication) as was used at the level of chromatin. As all DNA binders that could block cleavage are gone at this stage, this

second RE cleavage will cleave all RE sites (“all cut” or “100% accessibility” or just “1”). The other aliquot gets only a mock second RE digest, therefore retains only the RE cuts that were generated in the presence of chromatin and represent the unknown and to be measured “X%” of accessibility. Both the “100%” and the “X%” samples are processed in parallel by Illumina sequencing and the accessibility at a given RE site corresponds to the number of cut RE ends normalized to the number of cut $RE^{spike\text{-}in}$ ends in the “X%” sample divided by the number of cut RE ends per number of cut $RE^{spike\text{-}in}$ ends in the “100%” sample. This approach compares only the same kind of RE ends so that DNA end biases in the downstream processing pipeline cancel out.

Nonetheless, there still remains a source of bias as RE sites may have closely (<500 bp) neighboring RE sites, so that the “100%” sample contains respective DNA fragments that are delimited by the RE at both ends and may be on average shorter compared to the “X%” sample where 100%-X% of the corresponding DNA fragments have only one RE end and the other end is generated by shearing at variable and often longer distances than that of the neighboring RE site. This bias is due to differences in fragment lengths and maybe also due to a preference for RE over shearing ends and is corrected for by excluding such “next neighbor” RE sites.

An overview of the “cut-uncut” versus “cut-all cut” protocols is given in Figure 1, the performance of both methods is shown with their respective calibration curves in Figure 2 and an example of ORE-seq measurements with three different REs for an *S. cerevisiae* genome region is shown in Fig. 3.

Our ORE-seq protocol is presented here at a scale for use of one RE, including the mock digest control and using two different RE concentrations. This can be scaled up for using several REs as the resolution of the absolute occupancy map increases with the number of REs, and especially with the use of REs with 4 bp RE sites.

The protocol is described for chromatin prepared from *S. cerevisiae* and for an *S. pombe* gDNA spike-in. It can be readily adapted for chromatin preparations and spike-in gDNA from other species. Especially for metazoan genomes, the required high sequencing coverage

becomes rather costly. Nonetheless, we suggest that sequencing costs can be reduced if biotinylated adapters are ligated only to the RE ends prior to shearing and if the biotinylated fragments are enriched by immunoprecipitation after shearing and prior to sequencing library preparation. In this case the “cut-all cut” method must be used. This enrichment of RE ends over shearing ends via ligation with tagged adapters is not part of our protocol but an interesting modification and akin to methods that were developed to measure DSBs in the context of DNA repair studies [9].

2. Materials

2.1 Cells and buffers for preparation of *S. cerevisiae* chromatin

1. *S. cerevisiae* strain of interest and media and growth conditions according to your requirements

2. cold (0-4°C) distilled or deionized water (dH_2O)

3. preincubation solution: 0.7 M β-mercaptoethanol, 2.8 mM EDTA pH 8 (Note 1). Add 2.5 ml 14.3 M β-mercaptoethanol and 278 µl 0.5 M EDTA pH 8 to a 50 ml tube. Add dH_2O to 50 ml. Wipe off any spills from the tube and wrap it with parafilm as this solution smells. Store at -20°C.

4. 1 M sorbitol: Top up 182.2 g sorbitol with 1 l dH_2O. Store at -20°C.

5. Sorbitol + β-mercaptoethanol solution: 1 M sorbitol, 5 mM β-mercaptoethanol (Note 1). Top up 17.5 µl 14.3 M β-mercaptoethanol with 1 M sorbitol to 50 ml. Prepare freshly.

6. 0.1 M EGTA pH 8.0: Add 3.8 g EGTA (Titriplex VI) to a 100 ml beaker. Add 50 ml dH_2O. Adjust to pH 8 with 5 M KOH, otherwise EGTA will not dissolve completely. Adjust volume in measuring cylinder to 100 ml with dH_2O. Store at room temperature (RT).

7. Zymolyase: Weigh in Zymolyase (ICN Biochemicals, 100.000 u/mg) in a 1.5 ml tube. Top up with dH_2O (RT) to obtain a concentration of 20 mg/ml. Prepare freshly just before use. Zymolyase does not dissolve, therefore mix this suspension gently before pipetting.

8. Ficoll solution: 18% Ficoll (Sigma Ficoll Type 400 F4375), 20 mM KH_2PO_4, 1 mM $MgCl_2$, 0.25 mM EGTA, 0.25 mM EDTA. Add 90 g Ficoll, 1.36 g KH_2PO_4, 0.5 ml 1 M $MgCl_2$, 1.25 ml 0.1 M EGTA pH 8 and 250 µl 0.5 M EDTA pH 8 to a 0.5 l beaker. Add 400 ml dH_2O and stir the solution over night at RT as it dissolves very slowly. Cover the beaker with parafilm or foil. After dissolution, adjust pH to 6.8 with KOH. Use a 5 M KOH stock in the beginning and switch to a 1 M stock later to prevent overtitration. Adjust volume to 0.5 L with dH_2O in a 0.5 l measuring cylinder. Mix and aliquotize to 50 ml tubes. Store at -20°C.

2.2 Cells and buffers for preparation of *S. cerevisiae* and *S. pombe* genomic DNA (gDNA)

The *S. pombe* spike-in gDNA is only necessary if the cut-all cut method is applied.

1. *S. cerevisiae* and *S. pombe* wild type strains (in our case BY4741 and h- 972)

2. *S. cerevisiae* YPD medium: 20 g/l Bacto Peptone (Becton, Dickinson and Company), 10 g/l yeast extract (Biolife), 20 g/l D-glucose, autoclave, store at RT.

3. *S. pombe* YES medium: 5 g/l yeast extract (Difco), 30 g/l D-glucose, 0.7 g/l amino acid mix (0.1 g/l each of adenine, leucine, histidine, uracil, lysine, arginine, glutamate), use Millipore treated or equivalent water quality, filter sterilize, store at RT.

4. Blood & Cell Culture DNA Midi Kit (QIAGEN), including Buffer Y1 (including 14 mM β-mercaptoethanol, Note 1), Buffer G2 (including RNaseA), Buffer QBT, Buffer QC, Buffer QF and QIAGEN Proteinase K

2.3 Buffers and enzymes for digestion of chromatin and *S. cerevisiae* / *S. pombe* gDNA with restriction enzymes and DNA purification

1. Suitable restriction enzymes (REs; e.g. NEB) and corresponding RE-buffers, e.g., BamHI-HF and 10x CutSmart-Buffer. Store at -20°C. Dilute 10x CutSmart-Buffer with double distilled water (ddH_2O) to 1xCutSmart- Buffer (50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 100 µg/ml BSA, pH 7.9). Store at -20°C.

2. 10x STOP-Buffer: 4% SDS, 100 mM EDTA, 50 mM Tris-HCl pH 7.5. Add 2.5 ml 1 M Tris-HCl pH 7.5, 10 ml 0.5 M EDTA pH 8 and 27.5 ml ddH_2O to a 50 ml tube. Mix and add 10 ml 20% SDS. Store at RT.

3. 20 mg/ml proteinase K (Genaxxon) solution in ddH_2O. Store at -20°C. Aliquots may be refrozen.

4. 10 mg/ml RNase A (Roche) in ddH_2O. Remove DNases by incubation at 95°C for 15 min. Store at -20°C. Aliquots may be refrozen.

5. 5 M $NaClO_4$ in ddH_2O. Store at RT.

6. 100% and 70% ethanol. Store at RT.

7. Isopropanol. Store at RT.

8. Phenol for DNA extraction, equilibrated at pH ~8 (Sigma). Store 50 ml aliquots at -20°C.

9. Chloroform/isoamylalcohol (24:1): Under the fume hood, add 20 ml isoamylalcohol to 480 ml chloroform. Store at RT under the hood.

10. TE-buffer: 5 mM Tris-HCl pH 8, 1 mM EDTA. Store at RT.

11. 1 M KOAc in ddH_2O. Store at RT.

12. 0.2 M EDTA pH 8. Store at RT.

2.4 DNA shearing and purification after RE digestion and preparation of DNA sequencing libraries

1. MicroTUBE AFA Fiber Pre slit 6x16 mm

2. Covaris S220 sonicator

3. AmPure XP beads (Beckman Coulter)

4. 5 mM Tris-HCl pH 8.0

5. magnetic rack (Invitrogen)

6. agarose gel electrophoresis or Bioanalyzer or TapeStation or Fragmentanalyzer

7. Qubit™ Fluorometer and Qubit™ dsDNA HS Assay Kit

8. NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®, including NEBNext Ultra II End Prep Enzyme Mix, NEBNext Ultra II End Prep Reaction Buffer, NEBNext Ultra II Ligation Master Mix, NEBNext Ultra II Ligation Enhancer, NEBNext Ultra II Q5 Master Mix

9. Thermocycler

10. NEBNext Multiplex Oligos, including NEBNext Adaptor for Illumina and USER Enzyme

11. Illumina sequencer, including all sequencing reagents

3. Methods

3.1 Preparation of *S. cerevisiae* chromatin ("nuclei", Note 2)

1. Grow 1.5 * 10^9 cells (corresponding to ca. 30 µg gDNA) of your chosen *S. cerevisiae* strain at your chosen biological condition (Note 3).

2. Harvest cells by centrifugation for 10 min at 8959 g (e.g., 6000 rpm, Beckman JLA 8.1 rotor) at 4°C (Note 4). Discard supernatant.

3. Resuspend cell pellet(s) in 45 ml cold dH_2O per liter of cell culture. Twirling with an inoculation loop and vortexing helps to resuspend the often rather compact pellets.

4. Centrifuge for 10 min at 3220 g (e.g., 4000 rpm Eppendorf 5810R) and 4°C, discard supernatant and determine mass of this wet cell pellet. This is the "g wet cell pellet" that is referred to in the following steps (Note 5).

5. Resuspend cell pellet in 2 ml preincubation solution (Note 1) per g wet cell pellet and incubate for 30 min at 30°C while shaking.

6. Centrifuge for 5 min as in step 4, discard supernatant, resuspend in 40 ml cold 1 M sorbitol per g wet cell pellet.

7. Centrifuge for 5 min as in step 4, discard supernatant, resuspend in 5 ml 1 M sorbitol + β-mercaptoethanol per g wet cell pellet.

8. Dilute an aliquot (e.g. 50 µl) of this suspension with dH_2O such that OD_{600} can be measured in a reasonable range in your photometer, e.g., around 0.5.

9. Add 100 µl 20 mg/ml zymolyase suspension per g wet cell pellet and incubate for 30 min at 30°C while shaking.

10. Control spheroplasting efficiency by repeating step 8, which should yield at most 40% of the OD_{600} value measured in step 8. (Note 6).

11. Centrifuge for 8 min as in step 4, discard supernatant, resuspend in 40 ml cold 1 M sorbitol per g wet cell pellet (Note 7).

12. Centrifuge for 8 min as in step 4, discard supernatant, resuspend in 7 ml cold Ficoll solution per g wet cell pellet.

13. Make aliquot(s) corresponding to 0.3 g wet cell pellet and centrifuge each aliquot for 30 min at 21,546 g (15,000 rpm Beckmann, JA20.1) at 4°C.

14. Decant supernatant (Note 8), close tube with parafilm or lid and shock freeze chromatin pellets for 10 min in dry ice/ethanol bath (Note 9).

3.2 Preparation of *S. cerevisiae* and *S. pombe* gDNA (Note 10)

1. Grow 7 * 10^9 cells of *S. cerevisiae* or *S. pombe* wild type strain in YPD or YES medium, respectively, to a density of approx. 3 * 10^8 cells/ml.

2. Purify gDNA with Blood & Cell Culture DNA Midi Kit (QIAGEN) using 100/G QIAGEN Genomic-tip (Note 11). Start by harvesting yeast by centrifuging at 3000-5000 g for 5-10 min at 4°C.

3. Remove supernatant and resuspend cell pellet in 4 ml of TE-buffer by vortexing.

4. Centrifuge as in step 2, discard supernatant and thoroughly resuspend pellet in 4 ml Buffer Y1 supplemented with β-mercaptoethanol by vigorous vortexing.

5. Add 250 µl zymolyase suspension and incubate for 30 min at 30°C.

6. Centrifuge for 10 min at 5000g and 4°C.

7. Remove supernatant and thoroughly resuspend cell pellet in 5 ml Buffer G2 (including RNaseA) by inverting tube or vortexing.

8. Add 100 µl of QIAGEN Proteinase K and incubate for 30 min at 50°C.

9. Centrifuge as in step 6.

10. Keep supernatant and discard the pellet.

11. Equilibrate QIAGEN Genomic-tip 100/G with 4 ml Buffer QBT by gravity flow.

12. Vortex supernatant of step 10. for 10 s at top speed and load onto the QIAGEN Genomic-tip equilibrated in step 11.

13. Wash QIAGEN Genomic-tip with 2x 7.5 ml Buffer QC by gravity flow.

14. Elute with 5 ml Buffer QF by gravity flow into a clean tube.

15. Add 3.5 ml (0.7 volumes) isopropanol to the eluate of step 14 and mix.

16. Centrifuge at >5000g for at least 15 min at 4°C.

17. Remove supernatant without disturbing or losing the glassy DNA pellet.

18. Wash DNA pellet with 2 ml cold 70% ethanol by brief vortexing.

19. Centrifuge at >5000 g for 10 min at 4°C.

20. Repeat step 17.

21. Resuspend air-dried (5-10 min) DNA in 0.5 ml of TE-buffer and incubate over night at RT or for 1 to 2 h at 55°C for complete dissolution.

22. Determine gDNA concentration with Qubit.

3.3 Optional: Restriction enzyme digest of *S. pombe* gDNA

Digestion of *S. pombe* gDNA is necessary only for the cut-all cut method to obtain the *S. pombe* gDNA spike-in.

1. Digest 20 µg gDNA with 100 Units of your chosen RE in 200 µl of respective 1x RE buffer, e.g., 1x CutSmart Buffer (NEB), for 1.5 h at the temperature according to used RE, e.g., 37°C. Do NOT use the same RE as for chromatin digestion.

2. Stop the digest by addition of 50 µl 10x STOP-Buffer and proteinase K to a final concentration of 0.5 µg/µl.

3. Incubate for 45 min at 37°C. Store gDNA at 4°C. (Note 12)

3.4 Chromatin digest with restriction enzymes and DNA purification

1. Per RE, thaw chromatin pellet corresponding to 0.3 g wet cells and keep on ice. Chromatin pellet corresponding to 0.1 g wet cells is used per individual RE-digest or mock sample, e.g., per zero/low/high RE concentration (Note 13).

2. Resuspend chromatin pellet corresponding to 0.3 g wet cells in 2 ml ice-cold 1x RE-Buffer (e.g., 1x CutSmart or specific 1x RE-buffer) by vortexing. Make sure that the sample does not get too warm and that no clumps remain.

3. Centrifuge for 5-8 min at ~750 g (2000 rpm Eppendorf 5810R) and 4°C.

4. Decant supernatant, resuspend pellet in 0.6 ml 1x RE-Buffer by vortexing and aliquotize into three 200 µl aliquots in 1.5 ml tubes.

5. Add RE to desired concentration, e.g., 0, 5 and 20 µl 20 U/µl RE (NEB) (Note 14). To each sample, add the same total volume of either RE or RE storage buffer.

6. Incubate for 0.5 h (Note 15) at the temperature according to used RE, e.g., 37°C.

7. Stop reaction with 1/10 volume of 10x STOP-Buffer, vortex, and add 1/20 volume proteinase K (Note 16).

8. Incubate for 0.5-1 h or up to over night at 37°C.

9. Optional: For cut-all cut method, add 12 µl (amounts to ca. 1 µg, i.e., ca. 10% of RE digested chromatin sample's DNA mass) of RE-digested *S. pombe* gDNA spike-in prepared in section 3.3.

10. Add 1/5 volume 5 M $NaClO_4$, vortex.

11. Add one volume phenol (Note 17), vortex for 5 sec.

12. Add one volume chloroform/isoamyl alcohol, vortex for 5 sec.

13. Centrifuge for 5 min at 21230 g (15.000 rpm Eppendorf 5424R) at RT.

14. Transfer the upper aqueous phase to a fresh 1.5 ml tube and repeat steps 11-14 once.

15. Add 2.5 volumes 100% ethanol, mix by inverting and incubate on ice for 5 min.

16. Centrifuge for 20 min at ≥ 20,000 g at 4°C.

17. Discard supernatant and add 700 µl 70% ethanol to pellet.

18. Centrifuge for 5 min at ≥ 20,000 g at 4°C.

19. Discard supernatant, centrifuge again briefly and remove remaining ethanol with a pipette tip. Avoid disturbing the pellet.

20. Air-dry for 2 min or until ethanol is gone. Avoid overdrying as this makes resuspension more difficult.

21. Resuspend the pellet in 200 µl TE buffer.

22. Add 5 µl RNase A, mix and incubate sample for 5 min at 65°C to increase digestion of some folded RNA.

23. Incubate on ice for 1 min and proceed with RNase-digest for 1 h at 37°C.

24. Add 1/10 volume 1 M KOAc, vortex.

25. Add 0.8 volumes isopropanol, mix by inverting, incubate for 2 min at RT.

26. Centrifuge for 10 min at 21230 g (15.000 rpm Eppendorf 5424R) at RT.

27. Repeat once steps 17-20.

28. Resuspend the pellet in 85 µl ddH_2O. DNA solution can be stored at -20°C.

3.5 Optional: Second RE digest for cut-all cut method

1. Take two 40 µl aliquots from the purified DNA sample after step 28 in section 3.4 and add 5 µl 10x RE-Buffer (e.g., 10x CutSmart Buffer (NEB)) to each aliquot. Vortex.

2. Label one aliquot as “100% digested” and add the same RE as used for chromatin digestion of this sample, e.g., 4 µl 20 U/µl RE. Mix gently.

3. Lable the other aliquot as “X% digested” and add same volume of RE storage buffer as the volume of RE added in step 2.

4. Incubate both samples for 1.5 h at the temperature appropriate for the RE, e.g., 37°C.

5. Stop reaction by adding 6 µl 0.2 M EDTA pH 8. Optional: heat-inactive the RE (Note 18). Samples may be stored at -20°C.

3.6 DNA shearing and purification

1. Add 120 µl TE buffer to 10 µl sample (corresponding to about 0.8 to 1 µg DNA) of either step 28 in section 3.4 (without second RE digest) or from step 5 in section 3.5 (with second RE digest).

2. Transfer to a MicroTUBE AFA Fiber Pre slit 6x16 mm and shear sample in a Covaris S220 sonicator. Settings: Dutyfactor: 10%, 175 Watt, 200 cycles per burst, 180 sec; 4°C.

3. Transfer the sheared sample to a 1.5 ml tube.

4. Add 300 µL AMPure XP Beads (2.5x), vortex, spin down briefly to collect sample and incubate for 10 min at RT.

5. Collect beads in a magnetic rack and remove the supernatant with pipet.

6. Add 500 µl freshly prepared 80% ethanol (RT).

7. Incubate for 30 sec at RT. Keep tubes in magnetic rack.

8. Decant supernatant and repeat steps 4-7 once.

9. Decant supernatant, spin down briefly at RT and place tubes back in the magnetic rack.

10. Remove remaining ethanol with a 10 µl pipette.

11. Remove tubes from rack, add 100 µl 0.1x TE, vortex and spin down briefly.

12. Incubate for 2 min at RT, put tubes back to the magnetic rack and transfer 98 µl of the cleared solution to a fresh 1.5 ml tube.

13. Determine DNA concentration with Qubit. DNA can be stored at -20°C.

3.7 Sequencing library preparation

1. Use 100-200 ng (Note 19) DNA of step 12 in section 3.6 for library preparation with NEBNext Ultra II DNA Library Prep Kit. Adjust volume to 50 µl with 1x TE-buffer.

2. Add 7 µl NEBNext Ultra II End Prep Reaction Buffer and 3 µl NEBNext Ultra II End Prep Enzyme Mix and mix thoroughly by pipetting up and down.

3. Incubate in thermocycler with lid to at least 75°C for 30 minutes at 20°C, 30 minutes at 65°C and hold at 4°C.

4. Add 2.5 µl NEBNext Adaptor for Illumina, 30 µl NEBNext Ultra II Ligation Master Mix (mixed by pipetting prior to addition, very viscous, ensure proper mixing) and 1 µl NEBNext Ligation Enhancer. Mix well by pipetting up and down, spin briefly to collect sample. Ligation Master Mix and Ligation Enhancer may be combined as a master mix, but not the adaptor.

5. Incubate for 15 min at 20°C in thermocycler without lid heating.

6. Add 3 µl USER Enzyme, mix thoroughly and incubate for 15 min at 37°C in thermocycler with lid heated to at least 47°C.

7. Prewarm AMPure XP Beads for 30 min at RT, add 87 µl of these beads to the sample, mix thoroughly by pipetting or vortexing and incubate for 5 min at RT.

8. Place in magnetic rack to collect beads for ca. 5 min or until solution is clear and discard supernatant

9. Add 200 µl of freshly prepared 80% ethanol to tube that is still in the magnetic rack, incubate for 30 sec at RT and discard supernatant.

10. Repeat once step 9.

11. Remove traces of ethanol with pipet tip, air-dry for up to 5 min at RT.

12. Outside of the magnetic rack, elute DNA from beads by adding 17 µl with 5 mM Tris-HCl pH 8. Mix by pipetting, incubate for >2 min at RT, briefly spin down and put back into magnetic stand.

13. Incubate for 5 min or until solution is clear and remove 15 µl of supernatant into clean PCR tube. May be stored at -20°C.

14. Add 25 µl NEBNext Ultra II Q5 Master Mix, 5 µl chosen Index Primer, 5 µl Universal PCR Primer (or chosen i5 Primer). Mix thoroughly by pipetting. Briefly spin down.

15. Incubate in thermocycler with the following program: 30 sec at 98°C, 3-8 cycles (Note 20) of [10 sec 98°C, 75 sec 65°C], 5 min at 65°C, hold at 4°C.

16. Prewarm AMPure XP Beads for 30 min at RT, add 45 µl of these beads to the sample, mix thoroughly by pipetting or vortexing and incubate for 5 min at RT.

17. Repeat once steps 8-13, but elute with 33 µl 0.1xTE in step 12 and remove 30 µl in step 13. Determine DNA fragment size distribution and concentration on an Agilent Bioanalyzer High Sensitivity DNA chip.

18. Sequence the library by Illumina sequencing in 42 or 50 bp paired-end mode such that >40fold genome coverage is ensured, e.g., 5-10 * 10^6 reads for on average 100-200 bp long yeast genome fragments.

3.8 Calibration curve

1. Treat 25 µl of step 3 in section 3.3 (corresponding to 2 µg RE-digested *S. pombe* gDNA spike-in) according to steps 15-20 in section 3.4 and resuspend in 20 µl TE-buffer.

2. Mix 8 µg *S. cerevisiae* gDNA prepared in step 21 of section 3.2 with 0.8 µ*g S. pombe* gDNA spike-in that was ethanol precipitated according to step 1 and that is already digested with an RE (= $RE^{spike\text{-}in}$) that is NOT any of the REs, for which the calibration curve shall be generated (see section 3.3).

3. Digest 4 µg of this gDNA mixture with the RE, for which the calibration curve shall be generated, following steps 1-3 in section 3.3, but scaled down fivefold.

4. Mock digest as in step 2, but with RE-storage buffer instead of RE.

5. Stop both the real and the mock digestion by addition of 1/20 volume 0.2 M EDTA.

(Note 21).

6. Mix defined (Note 22) percentages of cut gDNA according to the scheme given in Table 1.

Table 1. Mixing scheme for calibration curve samples.

Percent cut	**0%**	**10%**	**30%**	**50%**	**70%**	**90%**	**100%**	**total amount**
uncut gDNA (= mock digest)	1 µg	0.9 µg	0.7 µg	0.5 µg	0.3 µg	0.1 µg	0 µg	3.5 µg
RE-digested gDNA	0 µg	0.1 µg	0.3 µg	0.5 µg	0.7 µg	0.9 µg	1 µg	3.5 µg

7. Top up volume to 130 µl with TE-buffer and follow section 3.6 starting at step 2.

8. Prepare sequencing libraries from these purified DNA samples as in section 3.7.

3.9 Bioinformatics analysis

Overviews of the bioinformatics steps for the cut-all cut and the cut-uncut method are shown in Figs. 4 and 5.

1. Map fragments with bowtie2 using the combined *S. cerevisiae* and *S. pombe* reference genome (see `reference_genome/ScerAndSpomWithMT.fsa`: with chromosomes named as follows: chr01 – chr16 for the sixteen *S. cerevisiae* and chrI, chrII, chrIII for the three *S. pombe* chromosomes). Use alignment parameters: `-X 500 --no-discordant --no-mixed --no-unal`.

2. Remove multiply mapped reads using `samtools view -hf 0x2`.

3. Index BAM file using `samtools index`.

4. Download this repository (https://github.com/gerland-group/ORE-seq_analysis). (Note 23)

5. Install R & packages detailed in `restriction_enzyme/RE_Rprofile.R`.

6. Make new folder `<Example>` in folder 'restriction_enzyme' for your analysis.

7. Make new folders `data/bam` within `<Example>`.

8. Put bam and bai files into `<Example>/data/bam`. Name files according to the following rules.

Sample naming rules

Bam files within `data/bam` need to follow these naming conventions:

* The script needs bam files for both samples (X% cut and 100% cut) with identical file name except the ending: Samples with one RE digest end with `_X.bam`, while samples with second RE digest end with `_1.bam`.

* If there was no second digest and only cut-uncut analysis is wanted, the `_1.bam` file can be a copy / hard link of the `_X.bam` file and the cut-all cut results should then be ignored.

* File names must contain the RE name of the enzyme present in the sample after a "_" sign e.g. `<Strain>_BamHI-HF_<RE units>_X.bam`, where the information in `<Strain>` and `<RE units>` is not used by the script and `<RE units>` could be ommitted.

* If the spike-in (if present) used a different RE, then add `<spike-in RE>-norm`, e.g. `<Strain>_AluI_EcoRI-norm_<RE units>_X.bam`.

* Usable RE names can be checked and added in the `RE_info.txt` file.

* Multiple REs can be used on the main genome (not the spike-in): `<Strain>_BamHI-HF_<RE units>_KpnI_<RE units>_X.bam`, which will be analysed accordingly.

 * To only get results of one RE if others where present in the same sample, set parentheses to ignore REs: `<Strain>_BamHI-HF_<RE units>_(KpnI_400)_X.bam`. This will be analysed similarly to `<Strain>_BamHI-HF_<RE units>_X`, with the following difference:

 * The sites of the ignored RE (and their neighbourhoods) will still be excluded when calculating the background.

* The sites of the ignored RE might exclude sites of the "main" enzyme when they are close to each other.

* For calibration samples to be used for fitting the uncut correction factor, add the cut percentage with `X_pct_cut` as in this example: `<Strain>_AluI_10_pct_cut.bam`.

9. Start R and set working directory to `<Example>`.

10. `source("../RE_analysis.R")` or run `RE_analysis.R` step by step in RStudio from within `<Example>`.

11. Find desired output files and plots:

* The script creates a folder structure beginning with the folder `analysis_results` which contains different subfolders for each type of plot and result files.

* Depending on the parameters chosen in the script the main results path is `analysis_results/window_limit_times_1_max_length_500/close_distances_200_300/background_Michael` (in the following called `MAIN`) with plot folders for certain intermediate results along the way.

* Genomic mean accessibilities are saved in `MAIN/acc_site_means_simple.txt` with all_mean = cut-all cut results, cut_uncut_all_3 = uncorrected cut-uncut result, cut_uncut_4 = corrected cut-uncut result.

* Histograms of site accessibilities are saved in `MAIN/accessibility_histograms/` for plus/minus strand and starting/ending fragments as well as combined results (last column) with all_mean = cut-all cut results, cut_uncut_all_3 = uncorrected cut-uncut result, cut_uncut_4 = corrected cut-uncut result.

* Individual site results are saved in an R dataframe in `MAIN/occs_df_list.RData` with columns chr, enzyme, pos, eff_coverage, eff_cuts, occ_X_1 (= cut-all cut), occ_cut_uncut_2 (= uncorrected cut-uncut) and occ_cut_uncut_4 (=corrected cut-uncut).

How to use other REs

Add the required information to `RE_info.txt`, see `RE_info_README.txt`.

How to use other genomes

Other genomes than *S. cerevisiae* (and *S. pombe* for spike-in) are not supported by default, as there are unfortunately several references to the chromosome names within the code. If these are treated properly, the script should run for other genomes as well.

How to fit uncut correction factors

Run section `3.1.1 Calc and plot deviation from calibration samples` in the script that is skipped by default.

4. Notes

Note 1

β-mercaptoethanol is toxic. Use appropriate personal safety equipment (lab coat, safety glasses, gloves) and handle in fume hood. Probably β-mercaptoethanol can be replaced by DTT.

Note 2

The preparation of yeast chromatin by such methods is sometimes called “nuclei”, although this protocol will not yield nuclei in the clean sense, for example if checked by electron microscopy, as obtained, for example, for HeLa cell nuclei. For comparison with other protocols for the preparation of yeast chromatin see for example [10-15].

Note 3

Do not crosslink cells with formaldehyde. This will impair chromatin digestion by REs and lead to too high occupancy values [6]. This amount of cells is sufficient for preparing chromatin for digestion with one type of RE. Linear upscaling is readily possible if enough chromatin shall be prepared for analysis with several REs.

Note 4

Cells are still alive until the cell lysis in step 12. So if you need to monitor some special conditions, e.g., high temperature for a temperature-sensitive mutant, you may have to keep up these conditions during chromatin preparation [16]. The temperatures given in the protocol are for our routine chromatin preparations from wild type or mutant cells, like an *isw1* or *gcn5* mutant, grown to log phase at 30°C.

Note 5

This is a somewhat shaky measure as it depends on the residual amount of water in the pellet. Nonetheless, for most purposes it works just fine. If a more exact measure is needed, you could go by cell number via OD_{600} or cell counting.

Note 6

If lysis is not efficient, as happens for example with stationary cells or very sick strains, either use more zymolyase (more effective, but may introduce more protease contamination alongside with zymolyase) or incubate longer (may not help much, but longer incubation usually does not harm as cells are still closed at this point) or go to 35°C (optimal temperature for zymolyase).

Note 7

From here on, pellets are not so tight and stable anymore. Take care not to lose the pellet! Maybe reduce vortex speed from now on to about half. Resuspending is now more difficult due to clumps. Use less volume first and stirring with inoculation loop for better resuspension efficiency.

Note 8

Usually, there remains some sticky stuff at the tube walls that can not be poured off and that is also not part of the pellet. Remove, e.g., with cotton swab or tissue paper wrapped around spatula, this sticky stuff without disturbing the pellet. This also allows to remove residual Ficoll solution.

Note 9

Tube labeling may wash off in dry ice/ethanol bath.

Note 10

S. pombe gDNA is only required for the cut-all cut method.

Note 11

Refer to the manufacturer's detailed instructions for using this kit. This protocol states the basic steps only.

Note 12

EDTA from the STOP-buffer chelates all Mg^{2+} and other divalent cations so that DNases that usually depend on divalent cations are inhibited and the gDNA is stable at 4°C. As the proteinase K will inactivate the RE and as the *S. pombe* gDNA spike-in is usually added to crude chromatin, purification of the *S. pombe* gDNA is not necessary at this point. Omitting the purification and storage at 4°C instead of -20°C avoids unnecessary DNA breaks in the very long gDNA. Nonetheless, the *S. pombe* gDNA is purified in the context of the calibration curve (see step 1 in section 3.8) as this is done also with purified *S. cerevisiae* gDNA and not crude chromatin.

Note 13

It is of principal importance to ensure saturated digestion. We routinely use two sufficiently different, e.g., fourfold different, RE concentrations. Only if both of them yield approximately the same accessibility value within the same digestion time, e.g., within 5 percent points or some other criterion depending on the application and desired accuracy, then RE digestion was not limiting. This logic only applies if the dynamics of the chromatin substrate were sufficiently frozen, i.e., DNA accessibility did not change over time. This can only be tested by following digestion time courses. Sufficient chromatin stability is usually the case in the absence of ATP and presence of >2 mM Mg^{2+}, but may be an issue for *S. cerevisiae* chromatin at low Mg^{2+}concentrations [6]. If several different REs are used, the mock digestion sample does have to be prepared for all of them.

Note 14

RE preparations and RE storage buffer usually contain a high glycerol concentration. Upon addition of RE or RE storage buffer, avoid a final concentration of >5% glycerol if the RE is prone to star activity.

Note 15

The incubation time has to be long enough to reach saturation for the chosen RE concentrations. However, this RE digestion occurs in crude chromatin so that overlong incubation increases side effects. For example, chromatin preparations may contain endogenous exo- and endonucleases, which may resect DNA ends introduced by the RE or lead to not RE-caused DNA breaks, respectively. The former is detected and controlled for in the bioinformatic analysis (section 3.9) and the latter by comparison with the mock digestion. The influence of endogenous nucleases can be dampened by lowering RE digestion time and/or temperature and/or Mg^{2+} concentration. While all these measures will also compromise the RE digestion, this can be compensated by increasing RE concentration. The RE digestion conditions given in the protocol here were found to work well for wild type *S. cerevisiae* chromatin [6], but may have to be modified for other applications.

Note 16

If the 1x RE-buffer contains enough potassium to cause precipitation at low temperature in the presence of SDS from the STOP-buffer, as the case for 1x CutSmart buffer (NEB), keep the sample always at 37°C until phenol/chloroform extraction.

Note 17

Phenol is toxic. β-mercaptoethanol is toxic. Use personal safety equipment (lab coat, safety glasses, gloves) and handle in fume hood.

Note 18

Heat inactivation is not always possible depending on the RE. But just stopping by EDTA addition is also sufficient as this will reliably inactivate the RE digest and the RE is later on removed after shearing by AMPure bead purification.

Note 19

We noted (see Fig. 2) a scoring bias that may stem from a bias towards RE-generated versus shearing-generated DNA ends and attribute this to incomplete DNA end repair and impaired sequencing adapter ligation of the latter in comparison to the former DNA ends. This may be due to incomplete removal of 3' phosphate groups or other chemical modifications generated by shearing [7] but not by REs and that are not efficiently processed by end repair enzymes in the library preparation kit. Additional pre-treatment with the DNA repair PreCR kit (NEB, M0309) may help, but others still observed a similar bias in their calibration curve despite using this kit [8]. To incrase the likelihood of sufficient DNA end repair, we recommend to increase the ratio of repair enzymes relative to DNA and therefore to use only 100-200 ng although the kit's manufacturer recommend using up to 1 µg of DNA.

Note 20

Avoid using too many PCR cycles as this will increase the occurrence of clonal PCR duplicates. A maximum of 8 cycles is recommended, usually 6 cycles are sufficient and preferred.

Note 21

Some REs, especially the HF (high fidelity) version sold by NEB may stick to the DNA ends after cutting their site. This may inhibit the adapter ligation reaction during sequencing library preparation. Therefore, it is recommended that the RE is removed by proteinase K digestion and further DNA purification.

Treatment of mock digested gDNA mix in parallel ensures that the DNA concentration for RE-digested and undigested gDNA stays the same, i.e., generating the defined percentages in the following step 6 can be done via mixing corresponding volumes without measuring DNA concentrations.

Note 22

Your pipetting precision and accuracy will decide how well defined these percentages are. Follow the recommendations of the manufacturer of your pipets. Use the same pipet and setting for each percentage if you prepare calibration curves for several REs. The DNA masses given in Table 1 can be estimated from steps 2 and 3. The exact DNA amount is less important than the exact volume ratio of cut versus uncut gDNA (see Note 20).

Note 23

The bioinformatics analysis is a modified version based on the analysis in our first application [6] and in MRW's PhD thesis [17]. Our script deposited at GitHub (https://github.com/gerland-group/ORE-seq_analysis) is written for use with several REs (thoroughly tested with AluI, BamHI and HindIII) and the *S. cerevisiae* genome and with an *S. pombe* spike-in for the cut-all cut method. For custom applications of other REs and especially other genomes, the script has to be modified or newly written by a bioinformatics expert. As we cannot foresee future applications by other users, we just state in the following in detail the underlying rationale and mathematical background.

In the following, steps marked with * are only needed for the cut-all cut method, which for example needs a normalization between the cut and all cut sample using an *S. pombe* gDNA spike-in (Fig. 4). Likewise, steps marked with ° are only needed for the cut-uncut method, like the counts of uncut fragments at a given cut site (Fig. 5). Note that this description and our script is an all-in-one solution that calculates the outcome according to both methods at the same time.

First follow the steps for mapping / indexing and download of the script files as described in 3.9 and have a look at the readme file of the repository.

The script then performs the following actions:

Map / filter reads

1) Extract paired-end read information: chromosome, start, end and strand information, with end positions shifted by +1 bp.
2) Remove fragments that are longer than 500 bp.
3) Remove rDNA fragments by excluding the following loci:

   S. cerevisiae chr. 12: 451500 – 495000

   S. pombe chr. 3: 0 – 30000

   S. pombe chr. 3: 2430000 – 2452883

Count cut and uncut° fragments

4) Count the starting/ending fragments on plus and minus strand $c_\tau(x)$ for each genomic position x with $\tau = 1,2,3,4$ denoting starts on plus, starts on minus, ends on plus and ends on minus strands, respectively. For starting reads, we count the position of the first base pair, for ending reads we count the position after the last base pair (i.e. end positions are shifted by +1 bp). We use the notation $c_\tau^1(x)$ and $c_\tau^2(x)$ for the sample without and with second RE digest, respectively. For later modeling, we assume that one single given fragment with RE-cut or sheared fragment start or end at x will on average yield p_τ^x counts after PCR and Illumina sequencing.
5) For the cut-uncut method, we need the uncut fragments for fixed genomic positions x, i.e. fragments that start before $x - d$ and end after $x + d$ (end positions are shifted by +1 bp) in the sample without second RE digest. The extension by d is needed due to the fact that not all RE cut both strands at the same position, as explained later. We denote this number of uncut fragments with $u_\tau^1(x, d)$, also using the index τ as in step 4 to differentiate between plus ($\tau = 1$ or 3) and minus strand ($\tau = 2$ or 4). We assume

that one such uncut fragment at x will on average results in q_τ^x counts after PCR and Illumina sequencing.

Determine cut site positions with RE motif

6) Determine the cut site positions, i. e. the positions of the RE recognition motif, 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 of the recognition motif of cut site i plus half the length of the recognition motif, which usually has an even length.

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

   Position x^i given by the underlined base:

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

   In case of a 5' overhang, the 3' end is elongated to match the 5' end during DNA end polishing by a DNA polymerase. Conversely, a 3' overhang is digested to match the recessed 5' end during DNA end polishing by an 5'-3' exonuclease. For such end-polished HindIII ends we get the following double stranded fragment ends:

   Position x^i given by the underlined base:

   ```
   + 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'
   ```

   Let Δs be the shift length from the pattern center to the cut position of the + strand in upstream direction, which corresponds to half the length of the 5' overhang of the cleavage product in bp. For HindIII, $\Delta s = +2$, $\Delta s = 0$ for blunt end cutting RE whereas in case of an RE with 3' overhangs, Δs is negative.

RE	**Recognition motif (vertical line indicates cut position)**	**Shift length Δs**
AluI	AG|CT	0
BamHI	G|GATCC	2

HindIII	A|AGCTT	2
EcoRI	G|AATTC	2
HhaI	GCG|C	-1
KpnI	GGTAC|C	-2

Assuming proper end polishing of cut fragments as described above, we have the following counts 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 $c_4^2(x^i + \Delta s)$

Uncut read on + strand°: $u_1^1(x^i, \Delta s)$

Uncut read on – strand°: $u_2^1(x^i, \Delta s)$

To obtain the number of fragments not cut° by the RE at a given site, we count all fragments that start before $x^i - \Delta s$ and end after $x^i + \Delta s$, yielding $u_\tau^1(x^i, \Delta s)$. For easier notation, we set $x_1^i = x_2^i = x^i - \Delta s$ and $x_3^i = x_4^i = x^i + \Delta s$, yielding the cuts at site i as $c_\tau^1(x_\tau^i)$, $c_\tau^2(x_\tau^i)$, $\tau = 1, 2, 3, 4$.

If the cut-all cut method is used, this step needs to be done on both the *S. cerevisiae* and the *S. pombe* genomes. In the *S. pombe* gDNA spike-in, a different RE can be used.

Remove RE sites with close neighbor RE sites

7) Especially for X% samples derived from RE digestion of chromatin, uncut fragment counts are increased at cut sites with any neighboring cut site within approx. 150 bp. Thus, we ignore RE sites completely if they have a neighbor within 200 bp in either direction. This cut-off may be adjusted depending on given samples. We denote the

set of left-over sites with I and J, for the *S. cerevisiae* and the *S. pombe** genome, respectively.

8) As shown in Fig. 6, we often saw dependencies between the fragment counts C_τ^i and A_τ^i (defined below) and the distance to the next neighboring RE site, ranging up to 300-500 bp, e.g. for starting reads and the downstream distance to the next neighbor RE site (Fig. 6B). Thus we ignore start or end cut counts of an RE site and near the RE site (see RE site window approach below), if the next RE site downstream or upstream, respectively, is closer than 300 bp, respectively. Note that this value can be further tuned to the experimental conditions, although for our calibration samples shown in Fig. 2, there was hardly any difference between this limit set to 300 bp or a more conservative 500 bp. In general, the higher the degree of shearing, i.e., the shorter the average fragment length, the lower this limit can be. See also legend to Fig. 6.
9) In our protocol here, we modified the original protocol [6] such that different REs are used for digesting *S. cerevisiae* chromatin or *S. pombe* gDNA spike-in, for example BamHI and EcoRI, respectively. In this case, the EcoRI sites in the *S. cerevisiae* genome are not considered when determining close RE sites. However, since the second RE digest is applied after including the *S. pombe* gDNA spike-in, the BamHI sites need to be considered when removing the close EcoRI site on the *S. pombe* genome, i.e. remove all EcoRI sites with a close EcoRI neighbor or a close BamHI neighbor as described in this section above.

Collect cut and uncut° counts within window near cut sites to correct for resection

10) Due to endogenous exonucleases that may be present in the chromatin preparations and trim DNA ends after RE cleavage, some fragment ends do not match the RE cut site positions any more, even though they were generated by the RE. Thus 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 between samples, so its correction needs to be tailored to each pair of samples without and with* second RE digest.

We define count windows for each fragment type: For read starts, $W_1 = W_2 = \{0, 1, 2, \dots, w\}$ to apply a window in the downstream direction and for read ends, $W_3 = W_4 = \{0, -1, -2, \dots, -w\}$ to apply a window in the upstream direction. The algorithm to find the optimal value for w is described at the end of this step. C_τ^i denotes the number of cut fragments in the sample without second RE digest (``cut sample'') and $A_\tau^i$ denotes the number of cut fragments in the sample with second RE digest* (``all cut sample''):

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

w is determined using the sample without second RE digest and for the data of the *S. cerevisiae* genome and then* the same value is used for the sample with second RE digest and the *S. pombe* genome as well. For mock digested 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 8), 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 does not increase the summed counts within w by more than 1%, correcting for cut counts from shearing.

Calculate the mean counts (averaged over all cut sites) at each position -200 bp to 200 bp away from the average cut site for starts and ends counts and both strands. These cut counts near the average cut site usually show a single peak at 0, but depending on the conditions there is also a decreasing shoulder downstream/upstream for starts/ends, respectively. Averaging the different types and strands (end counts need

to be mirrored at 0 first) yields $m(d)$, d being the distance to the average cut site. The cut counts need to be corrected by the average shearing cut counts, which we obtain 100 bp to 200 bp away from the cut site: $m^c(d) = m(d) - \langle m(d) \rangle_{d=100,\dots 200}$ ($\langle \dots \rangle$ indicating the average). We define the cumulative sum of counts by $S(w) = \sum_{d=0}^{w} m^c(d)$. Finally, we set w equal to the first integer starting from 0 such that for all $n \in \{1, 2, 3, 4, 5\}$, the sum of the counts of the next n positions, $S(w+n) - S(w)$, is less than 1% of $S(w)$. In our samples typical values for w ranged from 0 to 20, going up to 40 for samples with very strong resection.

Uncut fragment counts at any RE cut site are not influenced by endogenous exonucleases as they are still occupied by a nucleosome or other protein that blocked the RE. For easier notation we define the uncut counts at site i by

$$U_\tau^i = u_\tau^1\left(x^i, \Delta s\right)$$

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

Occupancy estimation by cut-all cut method with background correction and normalization

We seek to estimate the real accessibility α^i at cut site i using the cut counts of the cut and all cut samples taking into account a bias towards RE versus sheared fragment ends and effective sequencing probabilities. We begin with viewing C_τ^i and A_τ^i as random variables with the 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 second 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. The probability that a given (longer) fragment will

be cut by shearing within a fixed region of length $w+1$ within the fragment is denoted by s (not to be confused with the site shift variable Δs).

Since the REs act before the shearing step, only the fraction that has not been cut by the RE can be cut in the shearing step, leading to

$\mu^i = \alpha^i + (1-\alpha^i)s.$

We assume that in the sample with second RE digest, all counts near a cut site came from a cut of the RE 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 $\hat{\alpha}_\tau^i := \frac{\hat{\mu}_\tau^i - s}{1-s}$.

The estimators for α^i are approximately unbiased, as

$$E\left[\hat{\alpha}_\tau^i\right] = \frac{E[\hat{\mu}_\tau^i]-s}{1-s} = \frac{1}{1-s}\left(E\left[C_\tau^i\right]E\left[\frac{1}{A_\tau^i}\right]\frac{N_A}{N_C} - s\right) \approx \alpha^i,$$

because the C_τ^i and A_τ^i are statistically independent as they originate from different samples and $E\left[\frac{1}{A_\tau^i}\right] \approx \frac{1}{E[A_\tau^i]}$. 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* gDNA 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 window of site x^i, so their averages over large regions of the genome are (with good approximation) the same:

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

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 sample without second RE cleavage ("X%") away from cut sites) and the average fragment number at genomic positions where cuts have to happen by the RE (counts in the sample with second RE cleave ("all cut" or "100%") 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 PCR and sequencing.

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,

$$\hat{\alpha}^i = \langle \min\left(\hat{\alpha}_\tau^i, 1.5\right)\rangle_\tau,$$

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, 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 with background correction

In the following we only use data from the sample without second 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 the sequencing probability p and all uncut fragments the sequencing probability q.

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 300 bp, respectively. Then define the ratio of cut and uncut fragments,

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

If the denominator is zero, we set $\hat{\kappa}^i = \infty$, which will lead to an accessibility of 1.

Similar to the previous section we have $E\left[C^i\right] = 4N_C p\left(\alpha^i + \left(1 - \alpha^i\right)s_1(w+1)\right)$ with s_1 being the shearing probability per base pair, but now calculated only using the cut sample, i. e. the ratio of all cut counts away from sites and the sum of cut and uncut fragment counts away from cut sites. For U^i we assume that the uncut fragment counts are given by fragments that have

not been cut by the RE at x_τ^i and after that also not been cut by shearing at x_τ^i. The generally very low sequencing probabilities justify the assumption that C_τ^i and U_τ^i are ``independent enough'' to make the following approximation:

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

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.

$$C_{eff}^i = C^i - \sigma(w+1)U^i\gamma \text{ and}$$

$$U_{eff}^i = (1+\sigma)U^i\gamma$$

are the effective counts of cut and uncut fragments, respectively, both corrected for cuts in the shearing step and different sequencing probabilities of cut and uncut fragments. $C_{eff}^i + U_{eff}^i$ gives an "effective coverage" of cut and uncut fragments at the site i and we ignore sites with an effective coverage below 40. This limit may be adapted for different applications. 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 RE digests:

For each RE (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 RE such that $\langle(\omega_s - \hat{\omega}_s(\gamma))^2\rangle_s$ is minimized. Additionally, we did a combined fit over all calibration samples of the three REs to use for REs,

for which no specific calibration samples were measured. The following table shows the best values for γ:

RE	AluI	BamHI	HindIII	combined
γ_{min}	1.282	1.279	1.333	1.300

4. Figure captions

Figure 1. Flow chart for cut-uncut and cu-all cut method. For details see text. “f.” abbreviates “fragments”.

Figure 2. Calibration curves for the indicated REs. Mixtures of given percentages of *S. cerevisiae* gDNA cut with the indicated REs were prepared as in 3.8 and analyzed both via the cut-uncut method (uncorrected or corrected by individual factor derived for each RE or by a combined correction factor derived from the combined RE calibration samples, Note 23) and via the cut-all cut method. Circles denote the average and error bars the standard deviation over all RE sites in the *S. cerevisiae* genome included in the analysis. Data are deposited at GEO under the accession number GSE189142.

Figure 3. Absolute occupancies as measured for the indicated REs by ORE-seq for an exemplary region of the *S. cerevisiae* genome. Data for ORE-seq with BamHI, HindIII, AluI and their combination are taken from [6]. Nucleosome mapping data by chemical cleavage [18] is shown for reference.

Figure 4. Flow chart for bioinformatic analysis steps for cut-all cut method.

Figure 5. Flow chart for bioinformatic analysis steps for cut-uncut method.

Figure 6. Scoring bias due to close next neighbor RE sites. Exemplary selection from the “cut_counts_vs_nn_distance” plots for AluI 50% cut calibration sample (as in Fig. 2) are shown. Our script automatically generates such plots for the “X%” (A, B) and “100%” (C, D) samples (X or 1 in y-axis label) that show the number of reads (position of the marker along the y-axis) that map to the indicated combinations (y-axis label) of the plus strand of the chromosome and start (A, B) or end (C, D) (y-axis label) at a given RE site (0 on x-axis) that have a next neighbor (nn) site for the same RE at a given upstream (A, C) or downstream (B,

D) distance (x-axis label) in bp (position of the marker along the x-axis). Analogous plots are also generated for minus end reads. We found that the strand identity does not matter, but rather the orientation (upstream versus downstream) relative to whether a read starts or ends at the RE site. (E, F) Plots for sequencing reads analogous to those in panels A and B, but for uncut fragments where the given RE site was not cut. The green lines correspond to the average at a given x-axis position. Our interpretation of the observed curve shapes is as follows. (A) If the sequencing reads stem from fragments starting with the RE site, then next neighbor RE sites upstream are irrelevant for scoring efficiency measured via the obtained read number as they are not contiguous with the sequenced fragment anymore due to the RE cut and therefore do not affect scoring by adapter ligation and sequencing. (D) The same is true for next neighbor RE sites downstream of an RE site where a read ends. In contrast, next neighbor RE sites downstream (B) or upstream (C) of an RE site where a read starts or ends, respectively, may be cut (are indeed cut to 100% in our calibration samples shown here) and therefore generate DNA fragments with two RE cut ends and a consistent length that may be shorter than the average length generated by the combiation of one RE and one shearing end. Such fragments with two RE ends are scored more efficiently (= above averages shown in panels A and D) than fragments with one RE and one sheared end. The paucity in fragments <100 bp reflects the DNA fragment length cut-off of the AMPure bead purification during this particular sequencing library preparation. Note that fragments of >500 bp length are excluded from the analysis as Illumina sequencing becomes biased against longer fragments, which explains that the curves level off to the average level (similar to green line in panels A and D) beyond 500 bp next neighbor distance. If shearing is more extensive, i.e., if the average fragment length is much shorter than 500 bp, then the curve will approach the average level at a next neighbor distance close to the average fragment length as next neighbor sites beyond the average fragment length will not be contiguous anymore. Note that especially for the “X%” samples generated from chromatin digestion the x-axis need not reflect the actual fragment length that gave rise to a certain sequencing read, but denotes a property of the genome sequence (distance to the next neighbor RE site).

Nonetheless, for calibration samples shown here the x-axis does mostly reflect actual fragment length as virtually all RE ends stem from the 100% cut *S. cerevisiae* gDNA that was mixed with uncut *S. cerevisiae* gDNA. Fortuitous ends at RE sites due to shearing are negligible (e.g., less than 10 counts on y-axis here).

Finally, the bias due to next neighbor RE sites can also be apparent for uncut fragments where there is a potentially cut RE site within approx. 150 bp of the view point RE site (0 on x-axis) in either the upstream (E) or downstream (F) direction. While this is not much pronounced for the samples shown here as the uncut fragments stem from mock digests in these calibration samples, it may be considerable in chromatin samples and also calls for excluding such next neighbor sites.

The bioinformatics procedure that corrects for the next neighbor site bias by excluding these RE sites is detailed in Note 23.

5. References

1. Wal M, Pugh BF (2012) Genome-wide mapping of nucleosome positions in yeast using high-resolution MNase ChIP-Seq. Methods in enzymology 513:233-250. doi:10.1016/b978-0-12-391938-0.00010-0
2. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ (2015) ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current protocols in molecular biology 109:21.29.21-21.29.29. doi:10.1002/0471142727.mb2129s109
3. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523 (7561):486-490. doi:10.1038/nature14590
4. Kim YC, Grable JC, Love R, Greene PJ, Rosenberg JM (1990) Refinement of Eco RI endonuclease crystal structure: a revised protein chain tracing. Science (New York, NY) 249 (4974):1307-1309. doi:10.1126/science.2399465
5. Gregory PD, Barbaric S, Horz W (1999) Restriction nucleases as probes for chromatin structure. Methods in molecular biology (Clifton, NJ) 119:417-425. doi:10.1385/1-59259-681-9:417
6. Oberbeckmann E, Wolff M, Krietenstein N, Heron M, Ellins JL, Schmid A, Krebs S, Blum H, Gerland U, Korber P (2019) Absolute nucleosome occupancy map for the Saccharomyces cerevisiae genome. Genome research 29 (12):1996-2009. doi:10.1101/gr.253419.119
7. Ohtsubo Y, Sakai K, Nagata Y, Tsuda M (2019) Properties and efficient scrap-and-build repairing of mechanically sheared 3' DNA ends. Communications biology 2:409. doi:10.1038/s42003-019-0660-7
8. Chereji RV, Eriksson PR, Ocampo J, Prajapati HK, Clark DJ (2019) Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation. Genome research 29 (12):1985-1995. doi:10.1101/gr.249326.119
9. Biernacka A, Skrzypczak M, Zhu Y, Pasero P, Rowicka M, Ginalski K (2021) High-resolution, ultrasensitive and quantitative DNA double-strand break labeling in eukaryotic cells using i-BLESS. Nature protocols 16 (2):1034-1061. doi:10.1038/s41596-020-00448-3
10. Martinez-Campa C, Kent NA, Mellor J (1997) Rapid isolation of yeast plasmids as native chromatin. Nucleic acids research 25 (9):1872-1873
11. Aris JP, Blobel G (1991) Isolation of yeast nuclei. Methods in enzymology 194:735-749. doi:10.1016/0076-6879(91)94056-i
12. Kizer KO, Xiao T, Strahl BD (2006) Accelerated nuclei preparation and methods for analysis of histone modifications in yeast. Methods (San Diego, Calif) 40 (4):296-302. doi:10.1016/j.ymeth.2006.06.022
13. Reese JC, Zhang H, Zhang Z (2008) Isolation of highly purified yeast nuclei for nuclease mapping of chromatin structure. Methods in molecular biology (Clifton, NJ) 463:43-53. doi:10.1007/978-1-59745-406-3_3
14. Zhang Z, Reese JC (2006) Isolation of yeast nuclei and micrococcal nuclease mapping of nucleosome positioning. Methods in molecular biology (Clifton, NJ) 313:245-255. doi:10.1385/1-59259-958-3:245
15. Kiseleva E, Allen TD, Rutherford SA, Murray S, Morozova K, Gardiner F, Goldberg MW, Drummond SP (2007) A protocol for isolation and visualization of yeast nuclei by scanning electron microscopy (SEM). Nature protocols 2 (8):1943-1953. doi:10.1038/nprot.2007.251
16. Schmid A, Fascher KD, Horz W (1992) Nucleosome disruption at the yeast PHO5 promoter upon PHO5 induction occurs in the absence of DNA replication. Cell 71 (5):853-864
17. Wolff MR (2020) Nucleosome occupancy and dynamics in yeast: genome-wide and promoter-level analyses and modeling. PhD, LMU München, München
18. Chereji RV, Ramachandran S, Bryson TD, Henikoff S (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. Genome biology 19 (1):19. doi:10.1186/s13059-018-1398-0