Nucleosome reorganisation in breast cancer tissues

Divya R. Jacob[1,^], Wilfried M. Guiblet[2,^], Hulkar Mamayusupova[1,^], Mariya Shtumpf[1], Luminita Ruje[1], Isabella Ciuta[1], Svetlana Gretton[1,3], Clark Correa[1], Emily Dellow[1], Shivam P. Agrawal[1], Navid Shafiei[1], Anastasija Drobysevskaja [1], Chris M. Armstrong[1], Jonathan D. G. Lam[1], Yevhen Vainshtein[4], Christopher T. Clarkson[1,5], Graeme J. Thorn[1,6], Kai Sohn[4], Madapura M. Pradeepa[1,7], Sankaran Chandrasekharan[8], Greg N. Brooke[1], Elena Klenova[1], Victor B. Zhurkin[2,*] and Vladimir B. Teif[1,*]

[1] School of Life Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK
[2] Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
[3] Current address: School of Engineering, Arts, Science and Technology, University of Suffolk, James Hehir Building, University Avenue, Ipswich, Suffolk, IP3 0FS, UK
[4] Fraunhofer-Institut für Grenzflächen- und Bioverfahrenstechnik IGB, Nobelstraße 12, Stuttgart, 70569, Germany.
[5] Current address: University College London, Gower St, Bloomsbury, London WC1E 6BT, UK.
[6] Current address: Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK.
[7] Current address: Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT, UK
[8] Colchester General Hospital, East Suffolk and North Essex NHS Foundation Trust, Turner Road, Colchester, CO4 5JL, UK
^ These authors contributed equally
* Correspondence should be addressed to Vladimir B Teif (vteif@essex.ac.uk) or Victor B Zhurkin (zhurkin@nih.gov)

Abstract

Nucleosome repositioning in cancer highlights changes of many aspects of genome organisation and regulation. Understanding it is important both from the fundamental point of view and for medical diagnostics based on cell-free DNA (cfDNA), which originates from genomic DNA regions protected from digestion by nucleosomes. Here we generated ultra-high resolution nucleosome maps in paired tumour and normal tissues from the same breast cancer patients using MNase-assisted histone H3 ChIP-seq and compared them with the corresponding cfDNA from blood plasma. We detected cancer-specific single-nucleosome repositioning at key regulatory regions in a patient-specific manner, as well as common patterns across patients. The nucleosomes gained in tumour versus normal tissue were particularly informative of cancer pathways, with 20-fold enrichment at CpG islands, a large fraction of which marked promoters of genes encoding DNA-binding proteins. Tumour tissues were characterised by 5-10 bp decrease in average distances between nucleosomes (nucleosome repeat length, NRL). These effects were modulated by GC content and DNA sequence repeats and correlated with differential DNA methylation and binding of linker histone variants H1.4 and H1X. Our findings provide missing mechanistic understanding of nucleosome positioning in tumour tissues and can be valuable for nucleosomics analyses in liquid biopsies.

Keywords

Breast cancer, nucleosome positioning, cfDNA, chromatin, liquid biopsy, transcription factors binding, nucleosome repeat length, NRL, linker histones.

Introduction

Nucleosome positioning is one of the main determinants of gene regulation (1-3). Apart from its fundamental importance, nucleosome positioning-motivated analysis is promising for clinical diagnostics based on cell-free DNA (cfDNA) (4,5). cfDNA is composed of DNA fragments released to body fluids upon digestion of chromatin of their cells of origin by apoptotic nucleases present *in situ*. Such DNA fragments are protected from nuclease digestion by nucleosomes and other DNA-bound proteins, and therefore cfDNA analysis allows to reconstruct nucleosome maps. This natural process is similar to MNase-seq experiments where chromatin is digested between nucleosomes by micrococcal nuclease (MNase), and MNase-resistant DNA fragments are analysed (6-12). Nucleosome positioning is also related to chromatin accessibility profiled with experiments such as ATAC-seq (13), although the latter is focused on open chromatin regions and does not provide high-resolution nucleosome maps like MNase-seq. While the connection between nucleosome positioning in tissues and cfDNA is of particular importance, only few low- to moderate-resolution nucleosome landscapes have been previously reported in breast cell lines (14-19), and none available in paired normal/tumour primary tissues from breast cancer (BRC) patients. It is worth noting that the lack of nucleosome positioning datasets in paired normal/tumour tissues is also characteristic for other cancers.

Our recent study in chronic lymphocytic leukaemia uncovered some of the rules of cancer-specific nucleosome repositioning in B-cells from peripheral blood (20), but nucleosome repositioning in solid tissues remains enigmatic. One of the reasons for this is that chromatin in solid tissues may have very different properties from immortalised cell lines, which have been used historically as model systems for nucleosome positioning. For example, as early as 1976, Compton et al (21) compared nucleosome repeat length (NRL) values from different studies in normal and cancer cells and observed that normal cells in culture have shorter NRLs in comparison with primary cells in tissues. This means that the detection of differences in nucleosome positioning between normal and tumour cells in culture may be challenging, and it may not represent the effects happening *in situ*. To overcome these challenges, here we generated the first resource comprised of ultra-high resolution nucleosome maps in paired normal and tumour breast tissues, as well as in cfDNA from the same patients.

The dataset reported here contains about 20 billion paired-end sequenced reads, which makes it one of the largest nucleosome positioning resources of this kind. This allowed us to perform a systematic analysis of BRC-specific nucleosome repositioning *in situ*. Discrete nucleosome repositioning was particularly enriched at gene promoters encoding key BRC transcription factors (TFs), as well as at cancer-sensitive subsets of TF binding sites. In addition, we found that cancer tissues are characterised by a 5-10 bp decrease in NRL. We investigated potential molecular mechanisms of such dramatic shortening of distances between nucleosomes, which may include changes in DNA methylation and rearrangement of linker histone variants. These observations open new clinically relevant avenues based on analyses of cancer-specific nucleosome reorganisation.

Materials and Methods

Human tissues and blood. We studied matched tissue samples and blood samples from the same individual patients as described below. Primary tumour tissues together with paired peripheral tissues (also called “normal”) were collected during surgery from breast cancer patients numbered P1-P4 treated at the Colchester General Hospital, with written consent taken before surgery. Tumour tissues were taken from the center of the tumour, and normal tissues were taken from a distance of about 1 millimetre from the periphery of the tumour. For patients P1 and P3 in addition matching blood samples were drawn before the surgical procedure. Detailed clinical information for patients P1 and P2 provided by the Colchester General Hospital is available in Supplementary Table S1. Both patients P1 and P2 are ER-positive. For patient P3, we determined ERα status using Western blot (detailed below), which did not detect ERα presence in tumour. The study was approved by the local ethics committee (NRES Committee East of England, reference number MH363). The morphological and molecular characterisation of the paired normal/tumour breast tissues from the patient cohort to which the patients studied here belong has been performed in our previous publications (22-24). The paired peripheral tissues displayed morphological characteristics identical to normal breast tissues whereby the following structures were present: breast ducts and lobules with two cell layers consisting of inner luminal/epithelial cells and outer basal/myoepithelial cells, blood vessels and stromal cells. Two tissue types had distinct proliferative states. Normal breast tissues were negative for the proliferation marker Ki-67, whereas breast tumour tissues were positive (see Supplemental Figure 4 in (22)). The cell composition of the paired breast tissues from this cohort was also clarified in our previous work (22) using immunofluorescent staining with markers specific for different cell types: CK14 for myoepithelial cells and CK19 for luminal cells. Normal breast tissues were composed of myoepithelial cells and luminal cells. Tumour tissues were composed predominantly of luminal cells. Tissues were stored at -80 ^{o}C.

MNase-seq and MNase-assisted histone H3 ChIP-seq (MNase-H3)

Chromatin extraction. Frozen tissue was first ground into fine powder in liquid nitrogen. It was further homogenized by manual douncing in 5 ml ice-cold Buffer I (0.3 M Sucrose in 60 mM KCl, 15 mM NaCl, 5mM $MgCl_2$, 0.1 mM EGTA, 15 mM Tris-HCl (pH 7.5) with freshly added 0.5 mM DTT, 0.1 mM PMSF and 1XProtease Inhibitor Cocktail, Thermo Scientific Halt Protease Inhibitor Single-use Cocktail (100X). The homogenized cell suspension was centrifuged in a 15 ml Falcon conical centrifuge tube at 6000 g for 10 minutes at 4 ^{o}C. The precipitated cells were re-suspended in 500 µL of Buffer I, then 500 µL of Buffer II (same as Buffer I, but with freshly added 0.4 % (v/v) IGEPAL CA-630) was gently added on top and the lysate was incubated on ice for 3 minutes, followed by centrifugation at 3000 g for 3 minutes at 4 ^{o}C. The pellet was re-suspended in 1 ml of ice-cold NBR Buffer (85 mM NaCl, 5.5 % Sucrose, 10 mM Tris-HCl pH 7.5, 3 mM $MgCl_2$, 1.5 mM $CaCl_2$ with freshly added 0.2 mM phenylmethylsulfonyl fluoride (PMSF) and 1 mM DTT) and centrifuged at 3000 g for 3 minutes at 4 °C. The pellet then was re-suspended in 200 µl of NBR buffer. Samples were treated with 1 µl RNAseA (10 mg/ml) per 100 µl of nuclear lysate for 5 minutes at room temperature.

*MNase digestion***.** MNase restriction enzyme (Sigma, Cat No: N5386-500UN) was diluted to 1U (Sigma Boehringer Units) in MNase buffer (50 % Glycerol, 10 mM Tris p H7.6 and 50 mM NaCl). The pellets re-suspended in 200 µl of NBR buffer in the chromatin extraction step were digested with MNase enzyme as follows. Chromatin derived from paired normal and tumour tissue was diluted with NBR buffer to keep the same concentration in both preparations. A 20 µl aliquot of purified nuclei was test digested with 1 µl MNase for 7 min at room temperature to assess the optimal digestion time and enzyme concentrations. 1 µg of chromatin was digested with 1U of MNase for 7 minutes at room temperature to achieve 80:20 relative abundance of mono- and di-nucleosome fractions on the agarose gel. The MNase digestion was stopped by addition of equal volume (NBR:STOP, 1:1) of STOP Buffer (215 mM NaCl, 10 mM TrisHCl pH 8, 20 mM EDTA, 5.5 % Sucrose, 2 % TritonX 100 with freshly added 0.2 mM PMSF, 1 mM DTT and 2X Protease Inhibitors). After digestion, chromatin was incubated on ice in the fridge at 4 °C overnight to release the digested chromatin.

*Chromatin Immunoprecipitation (ChIP)***.** 10 µl of Protein A beads (Dynabeads™ Protein A; Invitrogen) equivalent to 10 µl bed volume of slurry were prepared per each ChIP reaction. Bead storage buffer was aspirated, and the beads were washed twice in 500 µl in Block solution (0.5 % BSA (w/v) in PBS supplemented with 0.1 mM PMSF). Beads sufficient for 4 ChIP samples were all combined and dissolved in 600 µl of Block solution with the addition of 5 µg of anti-H3 antibody Abcam AB1791 per ChIP reaction (20 µg for 40 µl of beads for 4 ChIP samples). The antibody-bead solution was incubated on a slowly rotating wheel at 4 °C for two hours. After incubation, the beads were again washed in 600 µl block solution to remove unbound antibodies. The chromatin was incubated overnight after MNase digestion and then centrifuged for 10 minutes at a max speed of 9800 g at 4 °C. The supernatant with released chromatin was collected and stored at -20 °C and used for the ChIP experiment. 5 % of chromatin volume used for MNase-assisted ChIP-seq was frozen at -20 °C as Input and used for MNase-seq. 300 µl of chromatin was incubated with the antibody-bound Protein A beads for 3 hours on a rotating wheel at 4 °C. Unbound chromatin was washed 4 times in 1 ml volume of ChIP-Wash buffer (150 mM NaCl, 10 mM Tris-HCl pH 8, 2 mM EDTA, 1 % NP40, 1 % Nadeoxycholate with freshly added 0.5 mM DTT, 0.1 mM PMSF and 1XProtease Inhibitor Cocktail). The initial wash of bound beads was done by gently mixing beads in the ChIP-Wash buffer for one min. For the subsequent 3 washes: the tube was rotated for 5 minutes at 4 °C on a rotating wheel, then placed in a magnetic rack for 1 min at 4 °C and then the supernatant was withdrawn. After two washes the beads with wash buffer were transferred to new tubes. And as a final step, the beads were washed with 1 ml of filtered 1x TE buffer (10 mM Tris-HCl containing 1 mM EDTA) at room temperature, after which the TE buffer was removed as above. The elution of bound chromatin was performed with 100 µl of Elution buffer (0.1 M $NaHCO_3$, 1 % SDS) supplemented with 2 µl of Proteinase K per ChIP sample tube. The beads were incubated with the elution buffer for 6-12 hours at 55 °C. Similarly, 2 µl of Proteinase K were added to the Input material and incubated at 55 °C for 6-12 hours. The incubated beads were briefly vortexed and the supernatant was collected by placing the tubes in a magnetic rack. The beads were additionally washed in 20 µl of elution buffer and the supernatant collected as above. The eluants were further purified with PCR purification columns according to the manufacturer's instructions

and eluted in 20 µl of elution buffer of the kit. These final eluants were stored at -20 °C ready for the library preparation step.

cfDNA extraction. Cell-free DNA was extracted from 500 µl of blood plasma or serum as detailed below, following the instructions of the purification kits. For patient P1, cfDNA was extracted from 500 µl of blood serum using Plasma/serum Cell-Free Circulating DNA Purification Mini Kit (Norgen Biotek Corp). For patient P3, cfDNA was extracted from 500 µl of blood plasma using Plasma/serum Cell-Free Circulating DNA purification Mini Kit (Norgen Biotek Corp) and additionally, cfDNA was extracted from 500 µl of blood plasma using DNeasy Blood and Tissue Kit (Qiagen).

Cell culture. A549 and T47D were acquired from the ATCC and cultured in Dulbecco's Modified Eagle's Medium (DMEM, Corning, NY, USA) or Roswell Park Memorial Institute (RPMI, Corning) medium 1640, respectively. Media were supplemented with 10% FCS and 2 mM L-glutamine, 100 U/ml penicillin, 100 mg/ml streptomycin (PSG, Sigma Aldrich, MI, USA) and cultured as previously described (25).

Immunoblotting. Samples were lysed in RIPA buffer (150 mM NaCl, 1% IGEPAL CA-630, 0.5% sodium deoxycholate, 0.1% SDS, 25 mM Tris pH 7.4) containing freshly added protease inhibitors (Halt Protease Cocktail, Thermo Fisher, Waltham, MA, USA) and prepared as previously described (26). The DC protein assay (BioRad, Hercules, CA, USA) was used to quantify protein concentrations. 60 µg of protein was separated using SDS-PAGE and immunoblotting performed. Primary antibodies used were rabbit anti-ERα (1:1000, sc-543, Santa Cruz, TX, USA), mouse α-tubulin (1:10,000, B-5-1-2, Sigma Aldrich). Secondary HRP-conjugated antibodies (Sigma Aldrich) were used at 1:2000. Proteins were visualised using Immobilon Forte HRP substrate (Merck Millipore, MA, USA) and a Fusion FX Chemi Imager (Vilber Lourmat, Collégien, France).

External datasets. Raw data from ChIP-seq of histone H1 variants in T47D breast cancer cell line reported in (27) was obtained from GEO entry GSE156036 and mapped using Bowtie to genome hg19, considering only uniquely mapped reads with up to 1 mismatch. The hg19 coordinates of A and B chromatin compartments in MCF-7 cells reported in (28) were kindly provided by the author. Coordinates of genomic regions with differential DNA methylation between breast cell lines MCF-10A, MCF-7 and MDA-MB-231 were obtained from (29). Whole-genome bisulfite sequencing in MCF-7 cells reported by (30) was obtained from GSM4055031 in the form of methylation calls in hg19 genome assembly. Coordinates of chromatin domains in breast cells enriched with H3K9me2/3 determined with ChIP-seq by the ENCODE consortium (31) were obtained from the following GEO entries: H3K9me3 in HMEC cells (GSM1003485), H3K9me3 in MCF-7 cells (GSE96517), H3K9me2 in MCF-7 cells (GSE96141). Binding sites of CTCF, ER and FOSL2 were defined as detailed below.

Defining TF binding sites. Consensus datasets of CTCF sites bound in normal and breast cancer states were defined as the intersection of CTCF ChIP-seq peaks present in at least 50% of the replicates of the publicly available ChIP-seq datasets per cell line (MCF10A and HMEC for normal; MCF7 and T47D for breast cancer, using 14 published datasets as detailed in Supplementary Table S5. Merged CTCF dataset, bound in a given condition (normal or breast cancer) was defined as the concatenation

of all peaks present in at least 50% of the datasets used in the analysis. Lost and gained CTCF sites were defined as those present in the consensus of one condition (normal or cancer) and not present in the merged dataset of the other condition. Common CTCF sites were defined as those present in all four cell lines. The details of the intersections of these datasets are provided in Supplementary Table S6. Lastly, CTCF motif identification within peaks was done with RSAT (32), using the JASPAR (33) matrix MA0139.1. ERα binding sites were defined based on ChIP-seq peaks in six female BRC patients reported by Severson et al (34) (GSE104399). We constructed two consensus datasets of ERα binding: more stringent dataset contained the overlapping peaks appearing in all six patients (193 peaks), and less stringent dataset of peaks confirmed in at least three patients (11,637 peaks). These datasets were further refined with RSAT to consider only ERα motifs inside these peaks. FOSL2 binding sites were defined by re-calling peaks using HOMER starting from raw ChIP-seq data in MCF-7 cells reported by the ENCODE consortium (GSM1010768). This resulted in 14,106 peaks, which were then narrowed down using the coordinates of FOSL2 motifs inside these peaks determined with RSAT's default settings.

Sequencing data processing. Breast tissue samples for patients P1-P3 were sequenced at the National Institutes of Health Frederic sequencing facility on Illumina NovaSeq S4, TruSeq ChIP Library Prep Kit (IP-202-1012), 2 x 100 bp paired-end mode. cfDNA samples for patients P1 and P3 were sequenced as above, but in 2 x 150 bp paired-end mode. Samples for patient P4 were sequenced in the Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB on Illumina HiSeq 2500 sequencer, 2 x 75 bp paired-end mode, using NEBNext Ultra II Kit from NEB for library preparation. Paired-end sequenced reads were mapped with Bowtie (35) to human genome hg19, considering only uniquely mapped reads and allowing up to one mismatch. This resulted in ~500 million mapped reads per sample (see details in Supplementary Tables S2-3). NucTools (36) pipeline was used to obtain BED files with mapped nucleosomal DNA fragments including information about DNA fragment sizes. A set of Perl scripts called cfDNAtools developed for the next steps of the analysis reported here is available at https://github.com/TeifLab/cfDNAtools. cfDNAtools script extract_nuc_sizes.pl was used to extract fractions with DNA fragments with sizes between 100-120 bp, 120-140 bp, 140-160 bp, 160-180 bp and 180 -200 bp. Intersections between regions of interest were done using BEDTools (37). Each sample was analysed individually.

Calculation of aggregate occupancy profiles. Aggregate occupancy profiles and nucleotide frequency profiles were calculated separately for each sample using the script "annotatePeaks.pl" of the HOMER package (38). Where aggregate profiles are reported averaged per condition and per experiment type, the HOMER-generated profiles, each normalised individually to the sequencing depths of the corresponding sample, were plotted together in Origin Pro (originlab.com) and the arithmetic average was determined across all samples with the same condition and the same experiment type. H1 occupancy profiles were normalised by dividing the ChIP-seq signal by the corresponding Input.

Defining stable, shifted, common, lost and gained nucleosomes. Nucleosomal DNA fragments with sizes between 120-180 bp and separately 160-180 bp were used to define the subclasses of "stable", "shifted", "common", "lost" and "gained" nucleosomes as detailed below. First, we defined stable nucleosomes based on DNA

fragments found in one patient (P2) in one condition (normal or tumour), whose coordinates were reproduced with 99% precision (allowing 1 bp mismatch) in at least one other patient. These were defined using the BEDTools command “intersect” with parameters -u -f 0.99. Stable nucleosomes which shifted less than 1% between stable tumour and normal breast tissues of patient P2 were referred to as “common nucleosomes”, whereas fragments that changed their coordinates by more than 20% in tumour were defined as “shifted nucleosomes” (BEDTools “intersect” with parameters -f 0.80 -r -v). Stable nucleosomes which were present in tumour tissue of patient P2 and not overlapped even by 1 bp with stable nucleosomes in normal tissue of patient P2 were defined as “gained” in BRC; those which were present as stable in normal tissue of patient P2 and not overlapped with stable nucleosomes in tumour tissue of patient P2 were defined as “lost” in BRC. The numbers of stable, common, shifted, lost and gained DNA fragments protected from MNase digestion identified in this analysis are indicated in Table S4.

Calculation of aggregate DNA methylation profiles. We used published bisulfite sequencing data in MCF-7 cells (30), where methylation beta-values per CpG were determined by the authors. NucTools (36) was used to split the regions into chromosomes and 5mC occurrence per CpG was calculated with a cfDNAtools script bed2occupancy.v3d.methyl.pl. Aggregate DNA methylation profiles were calculated around centres of common and shifted nucleosomes defined above using NucTools, summing up the methylation beta-values (which take values in the interval [0, 1]) for each individual CpG located at a given distance from the genomic feature of interest as in (39). When the corresponding graph refers to “DNA methylation (a.u.)”, this corresponds to the value obtained by summation of all corresponding beta-values without further normalization.

NRL calculation. The method of NRL calculation from paired-end reads was developed in our previous publication (9) and is available in NucTools software package (36). This method is conceptually similar to the method based on single-end reads (40) but offers higher precision because it’s based directly on nucleosome centers (dyads) and applies different filtering. The NucTools protocol is based on the calculation of the distribution of frequencies of distances between nucleosome dyads with single-bp resolution, separately per each chromosome and then averaged over all chromosomes. At the next step, Origin Pro software (originlabs.com) is used to detect peaks of the sinusoidal oscillations of the distribution of frequencies, followed by plotting the locations of the peak summits as a function of the peak number and then performing linear regression to determine the slope of the fit which gives the NRL. The calculations performed here were done with NucTools using up to 40 million reads per chromosome, considering dyad-dyad distances up to 2,000 bp, and filtering out locations with more than 50 mapped reads centred on the same base pair. Changing the filter threshold to 20, 30 and 40 did not affect the results. NRLs were calculated separately for each sample, both genome-wide and inside sets of genomic regions of interest. For the calculation of NRL inside genomic regions, we defined them as follows: regions surrounding promoters were defined as intervals [-5,000; 5,000] from annotated RefSeq TSSs. The genes used in NRL calculation were selected as protein-coding genes according to the UCSC Genome Browser annotation, with the length satisfying criteria 1 kb < Length < 200 kb. This resulted in 16,161 genes. Alu repeats annotated based on UCSC Genome Browser’s Repeat Masker were extended by 1,000 bp on both sides. DMR regions defined above (29) were also extended by 1,000

bp on both sides. Regions enriched with H1X were defined based on MACS2 peak calling with default parameters using ChIP-seq of H1X in the T47D breast cancer cell line reported in (27), resulting in 33,337 peaks which were extended by 1,000 bp on both sides. Coordinates of chromatin A-compartments in human genome assembly hg38 defined based on Hi-C in MCF-7 cells were kindly provided by the author of the original publication (28) and converted to hg19 using LiftOver utility of the UCSC Genome Browser. Genes enriched and depleted with protected DNA fragments of size 160-180 bp used in Figure 5E were defined by calculating the number of 160-180 bp fragments per gene normalised by the gene length, then sorting the genes based on this number and defining the top 3,000 genes as enriched with “long nucleosomes” and the bottom 3,000 genes as depleted of “long nucleosomes”. Regions with low- and high-GC content used in Figure 3D were defined by calculating average GC content with the running 10 kb window and defining equal numbers of low-GC and high-GC regions (144,758 regions of each type, threshold GC content 39.8%). Genes marked by differential nucleosome occupancy in their promoters used in Figure 3D were defined by intersecting the genes from the set 1 kb < Length < 200 kb defined above with the sets of nucleosomes gained and lost in patient P2 in BRC versus normal tissue (Table S2). This resulted in 1,423 genes that gained nucleosomes at their promoters and 401 genes lost nucleosomes at their promoters.

Enrichment analysis. BedTools “Fisher” function was used to calculate fold enrichment as well as the corresponding *P*-values of the Fisher test for the enrichment with respect to random genomic regions for gained and lost nucleosomes inside gene promoters (defined as +/-1000 bp from RefSeq-annotated transcription start sites (TSSs)); CpG islands, Alu repeats, L1 repeats, and microsatellite repeats based on the UCSC Genome Browser Repeat Masker; CTCF sites defined using ChIP-seq by the ENCODE consortium (31); DMRs (29) and H1.4 and H1X defined as detailed above (27).

Gene annotation. RefSeq-annotated gene promoters defined as +/-1000 bp from TSS were intersected with regions with “lost” and “gained” nucleosomes defined above, considering all fragment sizes between 120-180 bp. Gene Ontology analysis was performed using the R package gprofiler2 (41). Gene network was constructed with Cytoscape 3.9.1 (42) using STRING database (43), including genes with promoters marked by nucleosome gain in BRC, associated with gene ontology terms related to “DNA-binding”.

Results

Cancer-specific nucleosome repositioning in breast tissues. We determined ultra-high resolution nucleosome positioning landscapes in paired normal and tumour breast tissues from four BRC patients using MNase-seq and MNase-assisted histone H3 ChIP-seq (MNase-H3) (Figure 1A). The high sequencing coverage allowed us to investigate well-defined changes at the level of single nucleosomes (Figure 1B). We first defined ultra-sensitive nucleosomes which are consistently lost or gained in cancer as follows: regions protected from digestion in a given patient (e.g. patient P2, see Methods) were defined as stable nucleosomes if regions with the same start/end coordinates (up to 1 bp mispatch allowed) were protected from digestion in at least one other patient in the same condition (normal or tumour correspondingly). Lost/gained nucleosomes were then defined as those which are stable in one condition (normal/tumour) and do not overlap even by 1 bp with any stable nucleosomes in the other condition. As a result of this definition, we were able to arrive at a small set of ~50,000 gained/lost nucleosomes in Patient P2.

Gained nucleosomes defined above were 19-fold enriched at CpG islands and 10-fold enriched at promoters, including promoters of 316 genes encoding DNA-binding proteins (for both effects $P < 1e\text{-}100$, Fisher Test) (Figure 1C). Gained nucleosomes were also prevalent at CTCF binding sites. These punctate changes of the nucleosome landscape may define differential accessibility of DNA to regulatory proteins, underlying dysregulated cancer-specific pathways. On the other hand, nucleosomes which were stably present in healthy breast tissues but lost in tumour did not show such enrichment in cis-regulatory regions and were instead modestly enriched inside L1 repeats (Figure 1C). Gained nucleosomes were particularly predictive of breast cancer pathways. The network of genes encoding DNA-binding proteins, marked by gained nucleosomes at their promoters, was significantly enriched for protein-protein interactions ($P < 1e\text{-}16$), with major nodes represented by known BRC-associated TFs such as RXRA (44), PPARγ (45), CREBBP (46), etc. (Figure 1D).

Nucleosome changes at sub-classes of transcription factor (TF) binding sites. To complement the above analysis of promoter nucleosome occupancy, we also investigated nucleosome patterns at TF binding sites. One of the most important TFs involved in breast cancer and used for its classification/stratification is Estrogen Receptor alpha (ERα). Cancer-specific nucleosome changes at ERα binding sites were patient-specific and pronounced at narrowly defined subclasses of sites, as exemplified for patients P1 and P3 in Figure 2A-B. A class of less stringent ERα binding sites, defined as a consensus dataset of ERα ChIP-seq peaks present in at least three BRC patients out of six from the study of Severson et al (34), did not show significant change in nucleosome occupancy between normal and tumour breast tissues and cfDNA of patients P1 and P3 (Figure 2A). On the other hand, a class of more stringent ERα binding sites present in all six patients from Severson et al (34) was characterised by pronounced changes between normal/tumour breast tissues and cfDNA, and these changes were patient-specific (Figure 2B). Patient P1, which has clinically confirmed ER-positive status, showed clear change of the shape of the nucleosome profile between tumour and adjacent normal breast tissue (nucleosome depletion at the centres of ERα binding sites and oscillations of nucleosome occupancy around ERα sites in the tumour, as opposed to a peak of nucleosome occupancy at ERα binding sites in normal tissue). On the other hand, patient P3, for

which our immunoblotting did not detect the presence of ERα (Supplementary Figure S1), was characterised by similar shapes of nucleosome occupancy profiles in both paired normal and tumour tissues. Interestingly, the cfDNA profile of patient P1 resembled that of normal tissue, while the cfDNA profile of patient P3 had an inverted shape in comparison with both normal and tumour breast tissues from this patient.

Patient-specific changes were also specific to other TFs, although not as pronounced as for ERα, as exemplified by FOSL2 binding sites (Figure 2C). The 3D genome organiser CTCF presents a particular case, since a number of published CTCF ChIP-seq datasets are available in normal and cancer breast cell lines. We used nine published datasets (see Methods) to define three sub-classes of CTCF sites: common between normal and cancer, as well as lost and gained in cancer. The distinction between these three classes of CTCF sites in terms of their nucleosome signature was quite dramatic, and consistent between normal and tumour breast tissues and cfDNA (Figure 2D). To summarise the TF binding site analyses, not all experimentally confirmed TF binding sites are equally suitable as diagnostic markers, and sub-selection of such sites may play a very important role.

Nucleosome repeat length shortening in breast cancer. Next, we asked whether there is some general change of nucleosome organisation in BRC that affects all patients. We calculated the distribution of genomic distances between centres of neighbouring nucleosomes, which revealed a dramatic shift towards smaller distances in cancer in comparison with adjacent non-cancer tissues from the same patients (Figure 3A). This resulted in a significant decrease in the average genome-wide distance between neighbouring nucleosomes (the Nucleosome Repeat Length, NRL) in tumour versus normal tissues. This effect of genome-wide NRL shortening by 5-10 bp in cancer was observed in all MNase-H3 samples from this study and in MNase-seq for three out of four patients (Figure 3B). The only outlier was the “Normal” MNase-seq sample of patient P1. However, paired MNase-H3 samples from the same patient P1 were consistent with NRL shortening in other patients. Thus, the effect of NRL shortening was detected in all patients in this cohort. Overall, the NRL decrease in tumour tissues was statistically significant ($P = 0.014$, paired sample t-test). A number of previous works characterised NRL changes in different human cell types and states (21,40,47-50). However, to the best of our knowledge, this is the first report of a consistent decrease of NRL in tumour versus adjacent non-tumour breast tissue from the same patients.

Cancer-specific NRL decrease cannot be explained by shortening of DNA fragments protected from MNase or by repositioning of nucleosomes at cis-regulatory regions. What can explain the large NRL changes detected above? Firstly, we asked whether the NRL shortening in breast cancer is related to differences in the distributions of sizes of DNA fragments protected from MNase digestion. One reason for this question is that while NRL is traditionally considered a distinct biological characteristic, the sizes of DNA fragments protected from MNase digestion (hereafter referred to as “DNA fragments”) may depend on the experimental setup, and therefore it is important to uncouple these two effects. Another reason is that the effect of cfDNA fragment size shortening in cancer is already extensively used in liquid biopsies, and it is interesting to clarify whether the effect of NRL shortening that we described above is related to this effect. Figure 3C shows that genes enriched with chromatosome-size DNA fragments (160-180 bp, marked as “long nucleosomes”) have a longer NRL, while genes depleted of 160-180 bp fragments have a smaller NRL. This means that

there is indeed a connection between NRL shortening and DNA fragment size shortening. However, genes enriched with “long nucleosomes” in the tumour had a shorter NRL than genes enriched with “long nucleosomes” in normal tissue. Similarly, genes depleted of “long nucleosomes” in the tumour had a shorter NRL than genes depleted of long nucleosomes in normal tissue. This means that cancer-specific NRL shortening cannot be explained by the DNA fragment size shortening *per se*.

Next, we asked whether genes that are marked by cancer-specific nucleosome repositioning (nucleosome loss or gain at their promoters) are characterised by larger NRL changes in BRC. The comparison of NRL in genes marked by differential nucleosome presence at their promoters did not reveal significant differences between those associated with gained versus lost nucleosomes (Figure 3D). Furthermore, the cancer-specific NRL decrease inside genes marked by differential nucleosome occupancy was smaller than the genome-wide change (Figure 3D). This suggests that the change in gene expression *per se* is not the main factor affecting NRL decrease in cancer, and that punctate nucleosome repositioning at regulatory regions is an effect separate from the NRL changes. Interestingly, the genome-wide NRL decrease in cancer was more pronounced in regions with low GC content ([GC] < 39.8%, Figure 3D). On the other hand, most genes have GC content higher than 45%. Thus, additional factors may be at play to explain the global shortening of NRL.

Dependence of nucleosome sensitivity to MNase digestion in tumours on GC content and DNA methylation. To study additional properties of nucleosome repositioning in BRC, we classified nucleosomes into “common nucleosomes” (stable MNase-protected DNA fragments of a given patient found at the same locations in both normal and tumour tissues) and “shifted nucleosomes” (stable MNase-protected DNA fragments of a given patient based on normal tissue that changed their genomic coordinates >20% in tumour). Shifted nucleosomal DNA fragments included a large subnucleosomal fraction (120-140 bp), as well as fragments with other non-canonical sizes and were on average shorter than the stable nucleosomes which did not change their locations in cancer (Figure 4A). Common nucleosomes were characterised by DNA methylation enrichment, while shifted nucleosomes were depleted of DNA methylation (Figure 4B). This is consistent with previous reports that DNA methylation stabilises nucleosomes while de-methylation or hydroxymethylation makes nucleosomes more MNase-sensitive (39,51,52).

The fractioning of DNA fragments strongly depended on GC content, with longer fragments being more GC-rich (Figure 4C and Supplemental Figure S4). This is in line with known effects on DNA fragment lengths by the level of MNase digestion (51,53,54). Interestingly, there was no strong dependence of fragment length on DNA methylation in breast tissues, as opposed to DNA fragment in cfDNA (Figure 4D). Thus, DNA methylation may account for cancer-specific nucleosome repositioning, but probably not for the overall shortening of nucleosome-protected fragments. With respect to the effect of the DNA fragment sizes, the overall distribution of DNA fragments protected from MNase digestion was more enriched with shorter fragments in cancer in comparison with normal samples (Supplemental Figure S2). cfDNA fragment sizes from the same patients were significantly larger than MNase-seq and MNase-H3 DNA fragments, with the peak of the distribution shifting from ~146 bp in MNase-seq to ~165 bp in cfDNA. The size distribution of DNA fragments protected from MNase digestion was shifted towards smaller sizes in tumour versus normal tissues, consistently in different types of genomic regions: inside active and inactive chromatin domains (A- and B-domains), repetitive and non-repetitive regions,

separated into Alu, LINE1, alpha-satellite and micro-satellite repeats (Supplementary Figure S3). The same features of the distributions of DNA fragment sizes were also observed in 10-kb regions with low- and high GC content (Supplementary Figure S3D) and inside heterochromatin regions marked by H3K9me2/H3K9me3 in healthy and breast cancer cell lines (Supplementary Figure S4).

Relationship between nucleosome reorganisation in BRC and linker histones. Linker histones are among the most abundant chromatin proteins which can affect nucleosome organisation. To investigate the effects of linker histones we have used recently reported maps of H1 histone variants determined with ChIP-seq in the BRC cell line T47D (27). Genome-wide, the occupancy of linker histone variants H1.0, H1.2 and H1.5 is negatively correlated with GC content, while histones H1.4 and H1X are preferentially bound to GC-rich regions (Figure 5A). By comparing H1 maps with nucleosome maps determined in this study we clarified that histones H1.0, H1.2 and H1.5 were depleted near nucleosome dyads for protected DNA fragments of all sizes, while H1.4 and H1X occupancies were diminished for sub-nucleosomal and nucleosome-size DNA fragments (<160 bp) but increased for larger-size fragments (Figure 5B). Given this relationship, the redistribution of H1.4 and H1X in BRC can lead to a significant depletion in the fraction of large-size nucleosomal DNA fragments, while the effect of H1.0, H1.2 and H1.5 histones is less pronounced (Figure 5C). A quantitatively similar effect was observed for nucleosome maps determined based on MNase-H3 (Supplementary Figure S6). Based on Figures 5A-C, linker histone H1X may have the largest effect on protection from the digestion of chromatosome particles (>160 bp). Therefore, we selected this histone variant for more advanced analysis. Figure 5D shows that genomic regions that give rise to protected DNA fragments with sizes 160-180 bp are characterised by a flat H1X profile in normal tissues but have a peak of H1X occupancy in the tumours. Furthermore, “common” 160-180 bp fragments protected from digestion both in normal and cancer tissues had ~2-fold H1X enrichment, whereas regions that “lost” or “gained” nucleosomes in tumour had H1X depletion in BRC (Figure 5E). This suggests that H1X may be an important factor affecting the global nucleosome reorganisation in BRC.

NRL shortening in breast cancer is not restricted to specific genomic regions. Genome-wide NRL change could potentially arise due to cancer-specific changes in distinct types of genomic regions. To check this hypothesis, we calculated NRL across all samples from our patient cohort in different types of genomic regions, such as active A-type chromatin compartments (Figure 6A), regions surrounding gene promoters (Figure 6B), Alu repeats (Figure 6C), regions enriched with linker histone H1X (Figure 6D) and differentially methylated DNA (DMRs, Figure 6E). In line with previous studies (40,55,56), more active chromatin regions and gene promoters had a smaller NRL than less active regions (unrelated to the cancer status of the sample). When comparing adjacent tumour versus non-tumour breast tissues for individual patients, all these types of genomic regions were characterised by an NRL decrease in cancer. Alu repeats showed the largest heterogeneity among cancer patients, while DMRs showed less heterogeneity and one of the largest NRL changes between tumour and normal tissues (8 bp). Regions enriched with H1X had a slightly smaller NRL change, followed by gene promoters and active A-compartments, consistent with the smaller cancer-specific NRL change inside the genes reported in Figure 3D. Thus, NRL shortening was characteristic to all studied types of genomic regions, although in a quantitatively different manner.

Discussion

Our analysis revealed systematic nucleosome changes at two levels: (1) Small-scale changes at the level of individual nucleosomes informing about differential gene expression and TF binding (Figures 1 and 2). (2) Global genome-wide decrease in NRL (which holds true in all types of genomic regions that we analysed), as summarised in Figure 6F. Small-scale BRC-associated nucleosome gain was particularly pronounced at CpG islands, promoters of genes involved in BRC signalling, and CTCF binding sites, while nucleosome loss was preferentially associated with L1 repeats and regulators of chromatin organisation. Intriguingly, such nucleosome repositioning was particularly pronounced at promoters of genes encoding DNA-binding proteins, which may be helpful to resolve one of the challenges in cfDNA-based nucleosomics, the determination of the tissue of origin (6). In the current analysis we focused instead on mechanistic effects, leaving the tissue of origin problem for subsequent studies. Another interesting observation is that despite common assumptions in the field of cfDNA nucleosomics (10,57), not all TF binding sites are effective as cancer markers, even if they come from experimental binding maps (Figure 2). We found that only very stringent subsets of such sites, as in the example with ERα in Figure 2B, allow one to stratify patients in terms of their differential TF binding. Furthermore, clear definition of subsets of cancer-sensitive TF sites may require combinations of many experimental TF binding datasets, as in the example with CTCF in Figure 2D.

Our finding that nucleosome reorganisation in cancer is characterised by shorter NRL seemingly resembles the effect of shortening of cfDNA fragments in cancer patients in comparison with healthy people (6,8,58-62). However, it is important to note that NRL is more robust with respect to the experimental setup than the sizes of DNA fragments protected from MNase digestion (the former has been used historically as a biological characteristic of the cells, while the latter may depend on the level of MNase digestion). Indeed, we showed that while the shortening of DNA fragments protected from digestion influences the decrease of NRL, it cannot explain NRL shortening as a sole factor (Figure 3C). Thus, the effect of decreased NRL can be uncoupled from the effect of the shortening of DNA fragments protected from nuclease digestion, although their mechanisms may be related (e.g. linker histone redistribution, changes of DNA methylation, etc).

Many classical studies were devoted to the investigation of NRL changes as a function of cell differentiation or cell state (e.g. see references in (1,63)), and it is well established that NRL can change during differentiation. A hypothesis that NRL can change in cancer was put forward initially in 1976 (21), and has gained traction again in recent publications, in particular relating this to linker histones H1 (64,65). To the best of our knowledge, we provide here the first evidence of systematic NRL decrease in solid tumours based on adjacent tumour/non-tumour tissues from the same patients. It should be noted that the effect of global NRL decrease is different from the local effects of transcription-specific changes of nucleosome spacing at regulatory regions such as gene promoters (6,40) and TF binding sites (9,49,66), as well as from cell cycle-dependent NRL changes (67). The global NRL decrease reported here for breast cancer may also take place in other cancers. Indeed, we have also found a similar genome-wide NRL shortening effect in B-cells from patients with chronic lymphocytic leukaemia, which was more pronounced for more aggressive cancer subtypes (20).

On a practical side, the finding that NRL decreases in tumour tissues may be directly relevant for the clinic, e.g. in relation to liquid biopsies based on cfDNA. Depending on the cancer stage, cfDNA in patients with solid tumours such as BRC may contain 1-30% of DNA from tumour tissues, while the majority of cfDNA usually comes from blood cells. If the nucleosome spacing changes characteristic for the tumour manifest in cfDNA, this may help to classify patients without the need to determine the cells of origin of cfDNA. The use of this effect in patient stratification is beyond the scope of the current publication but will be subject of subsequent work. It is also worth noting that the small patient cohort considered here did not allow us to study systematically the effects of heterogeneity between patients. But even if the heterogeneity between patients is not known, this method may be still applied to monitoring a specific patient, with nucleosome spacing changes as a signal of cancer progression or response to therapy.

We showed here that the NRL decrease in BRC may be associated with several molecular mechanisms, including redistribution of linker histones, in particular the H1X variant, differential DNA methylation, changes in chromatin domains or expression of DNA sequence repeats. Since none of these were found to be *solely* responsible for NRL shortening, it is likely that several mechanisms act synergistically. It is worth mentioning that this aspect of our study was limited by the availability of complementary datasets generated in MCF-7 and T47D BRC cell lines (not in the same patients). The clarification of the molecular mechanisms responsible for these effects will require a separate study, where DNA methylation and linker histones are measured in the same tissues for which the nucleosome maps have been determined. It is also intriguing that nucleosome-level changes may be connected to macroscopic effects, such as changes in the softness of cells/cell nuclei at the periphery of breast cancer tumours (68), as well as changes in the physical properties of chromatin fibers as a feature of cancer progression (69). It is worth noting that cancer cells have many stem cell-like properties. In particular, the chromatin of stem cells is characterised by smaller NRL (9). Thus, both stem cells and cancer have smaller NRL in comparison with their more differentiated counterparts, which may be related to de-differentiation of cancer cells (70). On the mechanistic side, shorter NRL has been shown to change the topology of nucleosome fibres to more deformable and prone to macroscopic self-association (71), consistent with Figure 6F. Thus the 5-10 bp decrease in NRL in breast tissues may be associated with dramatic genome reorganisation, causative mechanistic details of which remain unknown.

Taken together, this study provides the first experimental nucleosome maps in paired tumour/normal breast tissues, which will be a valuable resource for the community, and reported conceptually new effects that deserve subsequent studies analysing the corresponding mechanisms. Furthermore, these new effects allow us to establish some of the missing theoretical foundations for cfDNA-based nucleosomics analysis. These effects are probably not limited to BRC, and future studies are needed to expand this methodology to larger patient cohorts and more cancer types.

Data availability

The processed data generated in this study including mapped nucleosomal fragments have been deposited in the Gene Expression Omnibus (GEO) repository and will be made openly accessible upon the journal publication.

Funding

This work was supported by the Wellcome Trust grant 200733/Z/16/Z and Cancer Research UK grants EDDPMA-Nov21\100044, SEBPCTA-2022/100001.

Acknowledgements

We thank Stuart Newman for the computer cluster support.

References

1. Singh, A.K. and Mueller-Planitz, F. (2021) Nucleosome Positioning and Spacing: From Mechanism to Function. *J Mol Biol*, **433**, 166847.
2. Ahmad, K., Henikoff, S. and Ramachandran, S. (2022) Managing the Steady State Chromatin Landscape by Nucleosome Dynamics. *Annu Rev Biochem*, **91**, 183-195.
3. Teif, V.B. and Clarkson, C.T. (2019) In Ranganathan, S., Gribskov, M., Nakai, K. and Schönbach, C. (eds.), *Encyclopedia of Bioinformatics and Computational Biology*. Academic Press, Oxford, pp. 308-317.
4. Shtumpf, M., Piroeva, K.V., Agrawal, S.P., Jacob, D.R. and Teif, V.B. (2022) NucPosDB: a database of nucleosome positioning in vivo and nucleosomics of cell-free DNA. *Chromosoma*, **131**, 19-28.
5. Im, Y.R., Tsui, D.W.Y., Diaz, L.A., Jr. and Wan, J.C.M. (2021) Next-Generation Liquid Biopsies: Embracing Data Science in Oncology. *Trends Cancer*, **7**, 283-292.
6. Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M. and Shendure, J. (2016) Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, **164**, 57-68.
7. Lo, Y.M.D., Han, D.S.C., Jiang, P. and Chiu, R.W.K. (2021) Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science*, **372**.
8. van der Pol, Y. and Mouliere, F. (2019) Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. *Cancer Cell*, **36**, 350-368.
9. Teif, V.B., Vainshtein, Y., Caudron-Herger, M., Mallm, J.P., Marth, C., Hofer, T. and Rippe, K. (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol*, **19**, 1185-1192.
10. Ulz, P., Perakis, S., Zhou, Q., Moser, T., Belic, J., Lazzeri, I., Wolfler, A., Zebisch, A., Gerger, A., Pristauz, G. *et al.* (2019) Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat Commun*, **10**, 4666.
11. Peneder, P., Stutz, A.M., Surdez, D., Krumbholz, M., Semper, S., Chicard, M., Sheffield, N.C., Pierron, G., Lapouble, E., Totzl, M. *et al.* (2021) Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nat Commun*, **12**, 3230.
12. Rao, S., Han, A.L., Zukowski, A., Kopin, E., Sartorius, C.A., Kabos, P. and Ramachandran, S. (2022) Transcription factor-nucleosome dynamics from plasma cfDNA identifies ER-driven states in breast cancer. *Sci Adv*, **8**, eabm4358.
13. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B. *et al.* (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods*, **14**, 959-962.
14. Cole, L., Kurscheid, S., Nekrasov, M., Domaschenz, R., Vera, D.L., Dennis, J.H. and Tremethick, D.J. (2021) Multiple roles of H2A.Z in regulating promoter chromatin architecture in human cells. *Nat Commun*, **12**, 2524.
15. Takaku, M., Grimm, S.A., Shimbo, T., Perera, L., Menafra, R., Stunnenberg, H.G., Archer, T.K., Machida, S., Kurumizaka, H. and Wade, P.A. (2016) GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. *Genome Biol*, **17**, 36.

16. Bansal, N., Petrie, K., Christova, R., Chung, C.Y., Leibovitch, B.A., Howell, L., Gil, V., Sbirkov, Y., Lee, E., Wexler, J. *et al.* (2015) Targeting the SIN3A-PF1 interaction inhibits epithelial to mesenchymal transition and maintenance of a stem cell phenotype in triple negative breast cancer. *Oncotarget*, **6**, 34087-34105.
17. Shimbo, T., Du, Y., Grimm, S.A., Dhasarathy, A., Mav, D., Shah, R.R., Shi, H. and Wade, P.A. (2013) MBD3 localizes at promoters, gene bodies and enhancers of active genes. *PLoS genetics*, **9**, e1004028.
18. Lidor Nili, E., Field, Y., Lubling, Y., Widom, J., Oren, M. and Segal, E. (2010) p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. *Genome Res*, **20**, 1361-1368.
19. Yang, X., Cai, G.-X., Han, B.-W., Guo, Z.-W., Wu, Y.-S., Lyu, X., Huang, L.-M., Zhang, Y.-B., Li, X., Ye, G.-L. *et al.* (2021) Association between the nucleosome footprint of plasma DNA and neoadjuvant chemotherapy response for breast cancer. *npj Breast Cancer*, **7**, 35.
20. Piroeva, K.V., McDonald, C., Xanthopoulos, C., Fox, C., Clarkson, C.T., Mallm, J.-P., Vainshtein, Y., Ruje, L., Klett, L.C., Stilgenbauer, S. *et al.* (2022) Nucleosome repositioning in chronic lymphocytic leukaemia. *bioRxiv*, 2022.2012.2020.518743.
21. Compton, J.L., Bellard, M. and Chambon, P. (1976) Biochemical evidence of variability in the DNA repeat length in the chromatin of higher eukaryotes. *Proc Natl Acad Sci U S A*, **73**, 4382-4386.
22. Docquier, F., Kita, G.-X., Farrar, D., Jat, P., O'Hare, M., Chernukhin, I., Gretton, S., Mandal, A., Alldridge, L. and Klenova, E. (2009) Decreased Poly(ADP-Ribosyl)ation of CTCF, a Transcription Factor, Is Associated with Breast Cancer Phenotype and Cell Proliferation. *Clinical Cancer Research*, **15**, 5762-5771.
23. Docquier, F., Farrar, D., D'Arcy, V., Chernukhin, I., Robinson, A.F., Loukinov, D., Vatolin, S., Pack, S., Mackay, A., Harris, R.A. *et al.* (2005) Heightened Expression of CTCF in Breast Cancer Cells Is Associated with Resistance to Apoptosis. *Cancer Research*, **65**, 5112-5122.
24. D'Arcy, V., Pore, N., Docquier, F., Abdullaev, Z.K., Chernukhin, I., Kita, G.X., Rai, S., Smart, M., Farrar, D., Pack, S. *et al.* (2008) BORIS, a paralogue of the transcription factor, CTCF, is aberrantly expressed in breast tumours. *British Journal of Cancer*, **98**, 571-579.
25. Brooke, G.N., Powell, S.M., Lavery, D.N., Waxman, J., Buluwela, L., Ali, S. and Bevan, C.L. (2014) Engineered repressors are potent inhibitors of androgen receptor activity. *Oncotarget*, **5**, 959-969.
26. Leach, D.A., Mohr, A., Giotis, E.S., Cil, E., Isac, A.M., Yates, L.L., Barclay, W.S., Zwacka, R.M., Bevan, C.L. and Brooke, G.N. (2021) The antiandrogen enzalutamide downregulates TMPRSS2 and reduces cellular entry of SARS-CoV-2 in human lung cells. *Nat Commun*, **12**, 4068.
27. Serna-Pujol, N., Salinas-Pena, M., Mugianesi, F., Le Dily, F., Marti-Renom, M.A. and Jordan, A. (2022) Coordinated changes in gene expression, H1 variant distribution and genome 3D conformation in response to H1 depletion. *Nucleic Acids Res*, **50**, 3892-3910.
28. Achinger-Kawecka, J., Valdes-Mora, F., Luu, P.-L., Giles, K.A., Caldon, C.E., Qu, W., Nair, S., Soto, S., Locke, W.J. and Yeo-Teh, N.S. (2020) Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin

landscape in endocrine-resistant breast cancer. *Nature communications*, **11**, 1-17.
29. Lee, I., Razaghi, R., Gilpatrick, T., Molnar, M., Gershman, A., Sadowski, N., Sedlazeck, F.J., Hansen, K.D., Simpson, J.T. and Timp, W. (2020) Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods*, **17**, 1191-1199.
30. Lyu, R., Zhu, X., Shen, Y., Xiong, L., Liu, L., Liu, H., Wu, F., Argueta, C. and Tan, L. (2021) Tumour suppressor TET2 safeguards enhancers from aberrant DNA methylation and epigenetic reprogramming in ERα-positive breast cancer cells. *Epigenetics*, 1-15.
31. ENCODE. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
32. Santana-Garcia, W., Castro-Mondragon, J.A., Padilla-Galvez, M., Nguyen, N.T.T., Elizondo-Salas, A., Ksouri, N., Gerbes, F., Thieffry, D., Vincens, P., Contreras-Moreira, B. *et al.* (2022) RSAT 2022: regulatory sequence analysis tools. *Nucleic Acids Res*, **50**, W670-676.
33. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranasic, D. *et al.* (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, **48**, D87-D92.
34. Severson, T.M., Kim, Y., Joosten, S.E.P., Schuurman, K., van der Groep, P., Moelans, C.B., Ter Hoeve, N.D., Manson, Q.F., Martens, J.W., van Deurzen, C.H.M. *et al.* (2018) Characterizing steroid hormone receptor chromatin binding landscapes in male and female breast cancer. *Nat Commun*, **9**, 482.
35. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.
36. Vainshtein, Y., Rippe, K. and Teif, V.B. (2017) NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. *BMC Genomics*, **18**, 158.
37. Quinlan, A.R. (2014) BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics*, **47**, 11 12 11-34.
38. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, **38**, 576-589.
39. Wiehle, L., Thorn, G.J., Raddatz, G., Clarkson, C.T., Rippe, K., Lyko, F., Breiling, A. and Teif, V.B. (2019) DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome research*, **29**, 750-761.
40. Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z. and Sidow, A. (2011) Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516-520.
41. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. and Peterson, H. (2020) gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res*, **9**.
42. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498-2504.

43. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S. *et al.* (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*, **51**, D638-D646.
44. Jansen, M.P.H.M., Knijnenburg, T., Reijm, E.A., Simon, I., Kerkhoven, R., Droog, M., Velds, A., van Laere, S., Dirix, L., Alexi, X. *et al.* (2013) Hallmarks of Aromatase Inhibitor Drug Resistance Revealed by Epigenetic Profiling in Breast Cancer. *Cancer Research*, **73**, 6632-6641.
45. Mueller, E., Sarraf, P., Tontonoz, P., Evans, R.M., Martin, K.J., Zhang, M., Fletcher, C., Singer, S. and Spiegelman, B.M. (1998) Terminal Differentiation of Human Breast Cancer through PPARγ. *Molecular Cell*, **1**, 465-470.
46. Peck, B., Bland, P., Mavrommati, I., Muirhead, G., Cottom, H., Wai, P.T., Maguire, S.L., Barker, H.E., Morrison, E., Kriplani, D. *et al.* (2021) 3D Functional Genomics Screens Identify CREBBP as a Targetable Driver in Aggressive Triple-Negative Breast Cancer. *Cancer Research*, **81**, 847-859.
47. Smith, T.A., Vallis, Y., Neary, D. and Itzhaki, R.F. (1989) Characteristics of lymphocyte chromatin from Alzheimer's disease patients and from young and old normal individuals. *Gerontology*, **35**, 268-274.
48. Ishimi, Y., Kojima, M., Takeuchi, F., Miyamoto, T., Yamada, M. and Hanaoka, F. (1987) Changes in chromatin structure during aging of human skin fibroblasts. *Exp Cell Res*, **169**, 458-467.
49. Teif, V.B., Mallm, J.P., Sharma, T., Mark Welch, D.B., Rippe, K., Eils, R., Langowski, J., Olins, A.L. and Olins, D.E. (2017) Nucleosome repositioning during differentiation of a human myeloid leukemia cell line. *Nucleus*, **8**, 188-204.
50. van Holde, K.E. (1989) *Chromatin*. Springer-Verlag, New York.
51. Teif, V.B., Beshnova, D.A., Vainshtein, Y., Marth, C., Mallm, J.P., Hofer, T. and Rippe, K. (2014) Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res*, **24**, 1285-1295.
52. Teif, V.B., Gould, T.J., Clarkson, C.T., Boyd, L., Antwi, E.B., Ishaque, N., Olins, A.L. and Olins, D.E. (2020) Linker histone epitopes are hidden by in situ higher-order chromatin structure. *Epigenetics Chromatin*, **13**, 26.
53. Schwartz, U., Nemeth, A., Diermeier, S., Exler, J.H., Hansch, S., Maldonado, R., Heizinger, L., Merkl, R. and Langst, G. (2019) Characterizing the nuclease accessibility of DNA in human cells to map higher order structures of chromatin. *Nucleic Acids Res*, **47**, 1239-1254.
54. Chereji, R.V., Kan, T.W., Grudniewska, M.K., Romashchenko, A.V., Berezikov, E., Zhimulev, I.F., Guryev, V., Morozov, A.V. and Moshkin, Y.M. (2016) Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster. *Nucleic Acids Res*, **44**, 1036-1051.
55. Voong, L.N., Xi, L., Sebeson, A.C., Xiong, B., Wang, J.P. and Wang, X. (2016) Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell*, **167**, 1555-1570 e1515.
56. Baldi, S., Jain, D.S., Harpprecht, L., Zabel, A., Scheibe, M., Butter, F., Straub, T. and Becker, P.B. (2018) Genome-wide Rules of Nucleosome Phasing in Drosophila. *Mol Cell*, **72**, 661-672 e664.
57. Doebley, A.L., Ko, M., Liao, H., Cruikshank, A.E., Santos, K., Kikawa, C., Hiatt, J.B., Patton, R.D., De Sarkar, N., Collier, K.A. *et al.* (2022) A framework

for clinical cancer subtyping from nucleosome profiling of cell-free DNA. *Nat Commun*, **13**, 7475.
58. Serpas, L., Chan, R.W.Y., Jiang, P., Ni, M., Sun, K., Rashidfarrokhi, A., Soni, C., Sisirak, V., Lee, W.S., Cheng, S.H. *et al.* (2019) Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc Natl Acad Sci U S A*, **116**, 641-649.
59. Underhill, H.R., Kitzman, J.O., Hellwig, S., Welker, N.C., Daza, R., Baker, D.N., Gligorich, K.M., Rostomily, R.C., Bronner, M.P. and Shendure, J. (2016) Fragment Length of Circulating Tumor DNA. *PLoS genetics*, **12**, e1006162.
60. Guo, J., Ma, K., Bao, H., Ma, X., Xu, Y., Wu, X., Shao, Y.W., Jiang, M. and Huang, J. (2020) Quantitative characterization of tumor cell-free DNA shortening. *BMC Genomics*, **21**, 473.
61. Mouliere, F., Chandrananda, D., Piskorz, A.M., Moore, E.K., Morris, J., Ahlborn, L.B., Mair, R., Goranova, T., Marass, F., Heider, K. *et al.* (2018) Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*, **10**.
62. Ding, S.C. and Lo, Y.M.D. (2022) Cell-Free DNA Fragmentomics in Liquid Biopsy. *Diagnostics (Basel)*, **12**.
63. Beshnova, D.A., Cherstvy, A.G., Vainshtein, Y. and Teif, V.B. (2014) Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions. *PLoS Comput Biol*, **10**, e1003698.
64. Willcockson, M.A., Healton, S.E., Weiss, C.N., Bartholdy, B.A., Botbol, Y., Mishra, L.N., Sidhwani, D.S., Wilson, T.J., Pinto, H.B., Maron, M.I. *et al.* (2021) H1 histones control the epigenetic landscape by local chromatin compaction. *Nature*, **589**, 293-298.
65. Yusufova, N., Kloetgen, A., Teater, M., Osunsade, A., Camarillo, J.M., Chin, C.R., Doane, A.S., Venters, B.J., Portillo-Ledesma, S., Conway, J. *et al.* (2021) Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. *Nature*, **589**, 299-305.
66. Clarkson, C.T., Deeks, E.A., Samarista, R., Mamayusupova, H., Zhurkin, V.B. and Teif, V.B. (2019) CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. *Nucleic Acids Res*, **47**, 11181-11196.
67. Bomber, M.L., Wang, J., Liu, Q., Barnett, K.R., Layden, H.M., Hodges, E., Stengel, K.R. and Hiebert, S.W. (2023) Human SMARCA5 is continuously required to maintain nucleosome spacing. *Mol Cell*, **83**, 507-522 e506.
68. Han, Y.L., Pegoraro, A.F., Li, H., Li, K., Yuan, Y., Xu, G., Gu, Z., Sun, J., Hao, Y., Gupta, S.K. *et al.* (2020) Cell swelling, softening and invasion in a three-dimensional breast cancer model. *Nat Phys*, **16**, 101-108.
69. Stowers, R.S., Shcherbina, A., Israeli, J., Gruber, J.J., Chang, J., Nam, S., Rabiee, A., Teruel, M.N., Snyder, M.P., Kundaje, A. *et al.* (2019) Matrix stiffness induces a tumorigenic phenotype in mammary epithelium through changes in chromatin accessibility. *Nat Biomed Eng*, **3**, 1009-1019.
70. Friedmann-Morvinski, D. and Verma, I.M. (2014) Dedifferentiation and reprogramming: origins of cancer stem cells. *EMBO Rep*, **15**, 244-253.
71. Zhurkin, V.B. and Norouzi, D. (2021) Topological polymorphism of nucleosome fibers and folding of chromatin. *Biophys J*, **120**, 577-585.

Figure legends

Figure 1. The study design and small-scale nucleosome repositioning analysis. A) Scheme of the study: paired tumour/normal breast tissues taken from breast cancer patients numbered P1-P4 were used to determine nucleosome positioning using MNase-seq and MNase-assisted histone H3 ChIP-seq (MNase-H3), complemented by whole-genome sequencing of cell-free DNA extracted from blood plasma from the same patients. The analysis of each sample was performed individually, without pooling. B) Example genomic region (Chr 17: 7,855,262-7,592,516) enclosing gene TP53 and nucleosome occupancy maps for patient P3. Tracks top to bottom: normal tissue MNase-H3 and MNase-seq; tumour tissue MNase-H3 and MNase-seq; cfDNA in two replicates; binding sites of 161 TFs derived from ChIP-seq from ENCODE with Factorbook motifs. Rectangles show regions containing common (brown), gained (green) and shifted nucleosomes (violet). The gained and shifted classifications are not mutually exclusive. C) Fold enrichment of nucleosomes that were gained and lost in tumour versus healthy breast tissues in CpG islands, promoters, CTCF binding sites, L1 repeats and Alu repeats. In all cases except for nucleosomes lost in BRC at CTCF sites, enrichments are statistically significant (Fisher test, $P < 0.005$). D) Network of DNA-binding proteins encoded by genes marked by nucleosome gain at their promoters in tumour tissues.

Figure 2. Nucleosome occupancy profiles around TF binding sites. A) and B) Nucleosome profiles in patient P1 (top) and patient P3 (bottom) around ERα binding sites coinciding in three BRC patients (less stringent, A) and in six BRC patients (more stringent, B) (10-bp smoothing). C) Nucleosome profiles in patient P1 (top) and patient P3 (bottom) around FOSL2 binding sites determined in MCF-7 cells. Black lines – tumour tissues; red – normal tissues; blue – cfDNA from the same patient. D) Nucleosome profiles in patient P3 around CTCF bindings sites common between normal (HMEC and MCF10A) and BRC (MCF-7 and T47D) cell lines (dark yellow); sites that lost CTCF binding in BRC vs normal cell lines (blue), and sites that gained CTCF binding in BRC vs normal cell lines (wine colour).

Figure 3. Nucleosome repeat length shortening in cancer. A) Frequencies of distances between dyads of DNA fragments protected from MNase digestion in two paired normal (N) and tumour (T) samples from patient P2. B) Genome-wide NRL values calculated for each sample reported in this study. The difference of NRLs between normal and tumour tissues is statistically significant (P = 0.014, paired sample t-test). Open squares – average values across all samples in a given condition; horizontal lines – median values for a given condition. The colours of the points corresponding to NRL values in N/T tissues are assigned per patient and per experiment type as follows: MNase-seq: patient P1 (black), P2 (red), P3 (blue), P4 (green); MNase-assisted H3 ChIP-seq: patient P1 (violet), P3 (brown), P4 (cyan). C) NRL in normal (grey violins) and tumour (beige violins) breast tissue calculated inside genomic regions enriched ("long nuc.") and depleted with DNA fragments of sizes 160-180 bp ("short nuc"). The colours of the points are assigned per patient and per experiment type in the same way as in Figure 3A. Paired-sample t-test P values are indicated on the figure. Open squares – average values across all samples in a given condition; horizontal lines – median values for a given condition. D) NRL in normal and tumour breast tissue samples from patient P2 inside genes which are marked by

nucleosome loss or gain at their promoters (black), compared with NRL calculated genome-wide in 10 kb regions with low GC content (<39.8%) (red) and high GC content (>39.8%) (blue).

Figure 4. Interplay between nucleosome stability in BRC, DNA fragment sizes, GC content and DNA methylation. A) The distribution of sizes of common nucleosomes that did not shift in cancer (grey) is skewed towards larger sizes in comparison with shifted nucleosomes whose coordinates changed >20% in cancer (red). B) Aggregate DNA methylation profile based on MCF-7 cells plotted around common and shifted nucleosomes determined in this study. C) GC content of DNA fragments protected from digestion in normal (black) and cancer breast tissues (red) and cfDNA from the same patient cohort (blue). D) DNA methylation based on MCF-7 cells averaged across DNA fragments of different sizes protected from MNase digestion from panel (C).

Figure 5. Effects of linker histones on nucleosome repositioning in breast cancer. A) Correlation of GC content and H1 occupancy in 10-kb regions genome-wide. H1 variants H1.0, H1.2 and H1.5 preferentially bind AT-rich regions, while H1.4 and H1X have a higher affinity for GC-rich regions. B-D) Averaged aggregate profiles of H1 occupancy around different types of DNA fragments protected from MNase digestion. These profiles were first calculated individually for each MNase-seq tumour tissue sample reported in this study, and then averaged across all patients. B) Occupancy of H1 variants around nucleosome dyads calculated for different DNA fragment sizes. C) H1 occupancy in the region [-50, 50] bp around nucleosome dyad calculated for different H1 variants as a function of DNA fragment size. D) Average H1X occupancy in T47D breast cancer cell line calculated around DNA fragments of 160-180 bp in normal (black lines) and tumour breast tissues (red). E) Average H1X occupancy profiles around the consensus set defined based on all tissue samples from this study, of common nucleosomes (black), and nucleosomes which are gained (blue) and lost in cancer (orange).

Figure 6. NRL shortening as a novel breast cancer marker. A-E) NRL values calculated for each individual sample from this study inside different types of genomic regions. A) Transcriptionally active A-compartments of chromatin annotated in breast cancer MCF-7 cells; B) Regions surrounding gene promoters. C) Genomic regions around Alu repeats; D) Regions enriched with H1X linker histones in breast cancer T47D cells; E) Differentially methylated DNA regions annotated based on cell lines MCF-10A, MCF-7 and MDA-MB-231. Open squares – average NRL values for a given condition; horizontal lines – median values. The colours of the points are assigned per sample as in Figure 3A. F) A scheme of the global change of nucleosome compaction in breast cancer tissues. DNA (black) is wrapped around histone octamers (grey) in contact with linker histones (blue). MNase used in this study or apoptotic nucleases present *in situ* (orange) cut DNA not protected by nucleosomes and linker histones. Cancer cells have shorter NRL and contain more nucleosomes which are less protected from MNase digestion due to redistribution of linker histone variant H1X, changes in DNA methylation and other factors.

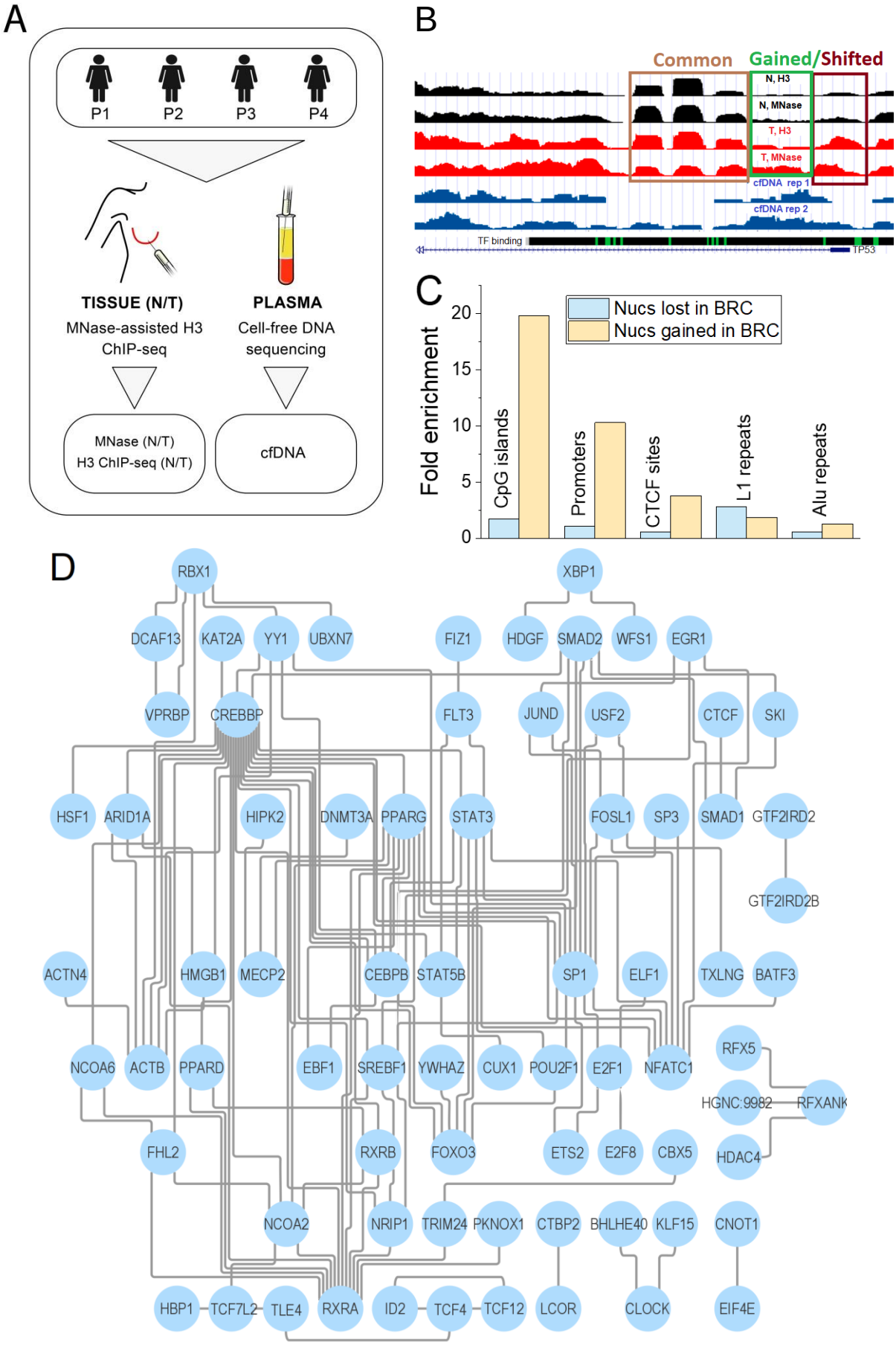
A
P1
P2
P3
P4
TISSUE (N/T)
MNase-assisted H3 ChIP-seq
PLASMA
Cell-free DNA sequencing
MNase (N/T)
H3 ChIP-seq (N/T)
cfDNA
B
Common
Gained/Shifted
N, H3
N, MNase
T, H3
T, MNase
cfDNA rep 1
cfDNA rep 2
TF binding
TP53
C
Fold enrichment
Nucs lost in BRC
Nucs gained in BRC
CpG islands
Promoters
CTCF sites
L1 repeats
Alu repeats
D
RBX1
XBP1
DCAF13
KAT2A
YY1
UBXN7
FIZ1
HDGF
SMAD2
WFS1
EGR1
VPRBP
CREBBP
FLT3
JUND
USF2
CTCF
SKI
HSF1
ARID1A
HIPK2
DNMT3A
PPARG
STAT3
FOSL1
SP3
SMAD1
GTF2IRD2
GTF2IRD2B
ACTN4
HMGB1
MECP2
CEBPB
STAT5B
SP1
ELF1
TXLNG
BATF3
NCOA6
ACTB
PPARD
EBF1
SREBF1
YWHAZ
CUX1
POU2F1
E2F1
NFATC1
RFX5
HGNC:9982
RFXANK
FHL2
RXRB
FOXO3
ETS2
E2F8
CBX5
HDAC4
NCOA2
NRIP1
TRIM24
PKNOX1
CTBP2
BHLHE40
KLF15
CNOT1
HBP1
TCF7L2
TLE4
RXRA
ID2
TCF4
TCF12
LCOR
CLOCK
EIF4E

Figure 1. The study design and small-scale nucleosome repositioning analysis. A) Scheme of the study: paired tumour/normal breast tissues taken from breast cancer patients numbered P1-P4 were used to determine nucleosome positioning using MNase-seq and MNase-assisted histone H3 ChIP-seq (MNase-H3), complemented by whole-genome sequencing of cell-free DNA extracted from blood plasma from the same patients. The analysis of each sample was performed individually, without pooling. B) Example genomic region (Chr 17: 7,855,262-7,592,516) enclosing gene TP53 and nucleosome occupancy maps for patient P3. Tracks top to bottom: normal tissue MNase-H3 and MNase-seq; tumour tissue MNase-H3 and MNase-seq; cfDNA in two replicates; binding sites of 161 TFs derived from ChIP-seq from ENCODE with Factorbook motifs. Rectangles show regions containing common (brown), gained (green) and shifted nucleosomes (violet). The gained and shifted classifications are not mutually exclusive. C) Fold enrichment of nucleosomes that were gained and lost in tumour versus healthy breast tissues in CpG islands, promoters, CTCF binding sites, L1 repeats and Alu repeats. In all cases except for nucleosomes lost in BRC at CTCF sites, enrichments are statistically significant (Fisher test, $P < 0.005$). D) Network of DNA-binding proteins encoded by genes marked by nucleosome gain at their promoters in tumour tissues.

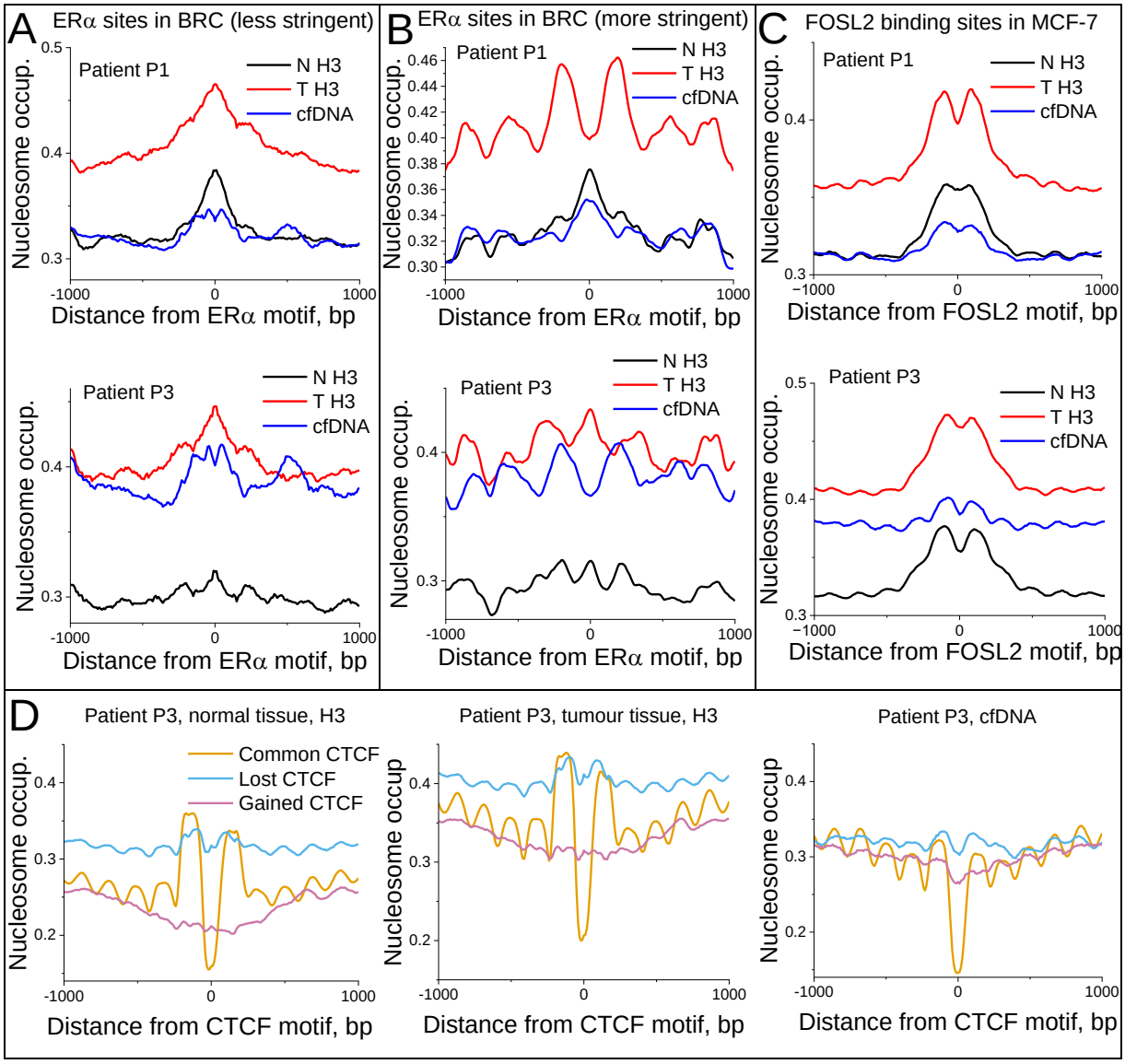

Figure 2. Nucleosome occupancy profiles around TF binding sites. A) and B) Nucleosome profiles in patient P1 (top) and patient P3 (bottom) around ERα binding sites coinciding in three BRC patients (less stringent, A) and in six BRC patients (more stringent, B) (10-bp smoothing). C) Nucleosome profiles in patient P1 (top) and patient P3 (bottom) around FOSL2 binding sites determined in MCF-7 cells. Black lines – tumour tissues; red – normal tissues; blue – cfDNA from the same patient. D) Nucleosome profiles in patient P3 around CTCF bindings sites common between normal (HMEC and MCF10A) and BRC (MCF-7 and T47D) cell lines (dark yellow); sites that lost CTCF binding in BRC vs normal cell lines (blue), and sites that gained CTCF binding in BRC vs normal cell lines (wine colour).

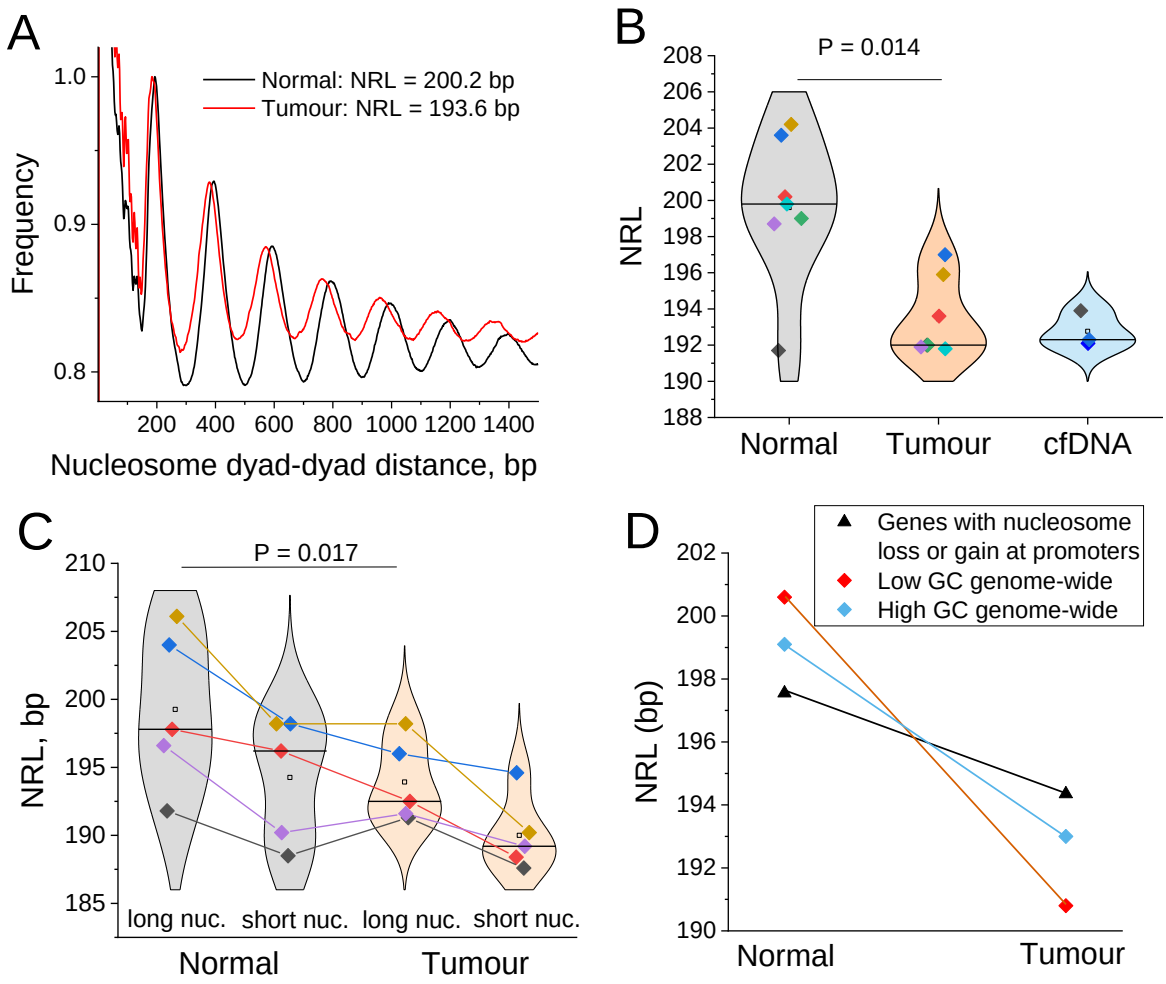

Figure 3. Nucleosome repeat length shortening in cancer. A) Frequencies of distances between dyads of DNA fragments protected from MNase digestion in two paired normal (N) and tumour (T) samples from patient P2. B) Genome-wide NRL values calculated for each sample reported in this study. The difference of NRLs between normal and tumour tissues is statistically significant (P = 0.014, paired sample t-test). Open squares – average values across all samples in a given condition; horizontal lines – median values for a given condition. The colours of the points corresponding to NRL values in N/T tissues are assigned per patient and per experiment type as follows: MNase-seq: patient P1 (black), P2 (red), P3 (blue), P4 (green); MNase-assisted H3 ChIP-seq: patient P1 (violet), P3 (brown), P4 (cyan). C) NRL in normal (grey violins) and tumour (beige violins) breast tissue calculated inside genomic regions enriched ("long nuc.") and depleted with DNA fragments of sizes 160-180 bp ("short nuc"). The colours of the points are assigned per patient and per experiment type in the same way as in Figure 3A. Paired-sample t-test P values are indicated on the figure. Open squares – average values across all samples in a given condition; horizontal lines – median values for a given condition. D) NRL in normal and tumour breast tissue samples from patient P2 inside genes which are marked by nucleosome loss or gain at their promoters (black), compared with NRL calculated genome-wide in 10 kb regions with low GC content (<39.8%) (red) and high GC content (>39.8%) (blue).

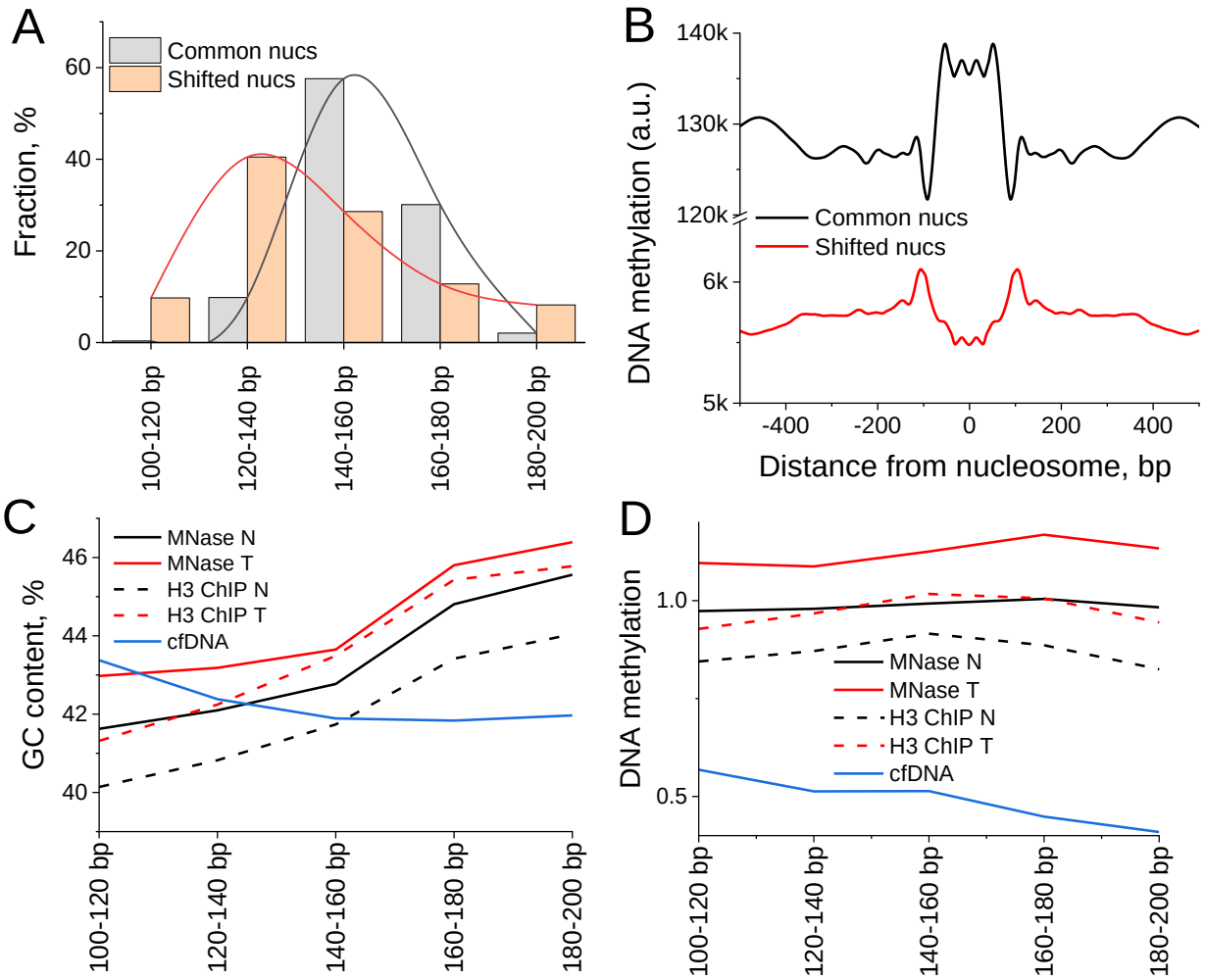

Figure 4. Interplay between nucleosome stability in BRC, DNA fragment sizes, GC content and DNA methylation. A) The distribution of sizes of common nucleosomes that did not shift in cancer (grey) is skewed towards larger sizes in comparison with shifted nucleosomes whose coordinates changed >20% in cancer (red). B) Aggregate DNA methylation profile based on MCF-7 cells plotted around common and shifted nucleosomes determined in this study. C) GC content of DNA fragments protected from digestion in normal (black) and cancer breast tissues (red) and cfDNA from the same patient cohort (blue). D) DNA methylation based on MCF-7 cells averaged across DNA fragments of different sizes protected from MNase digestion from panel (C).

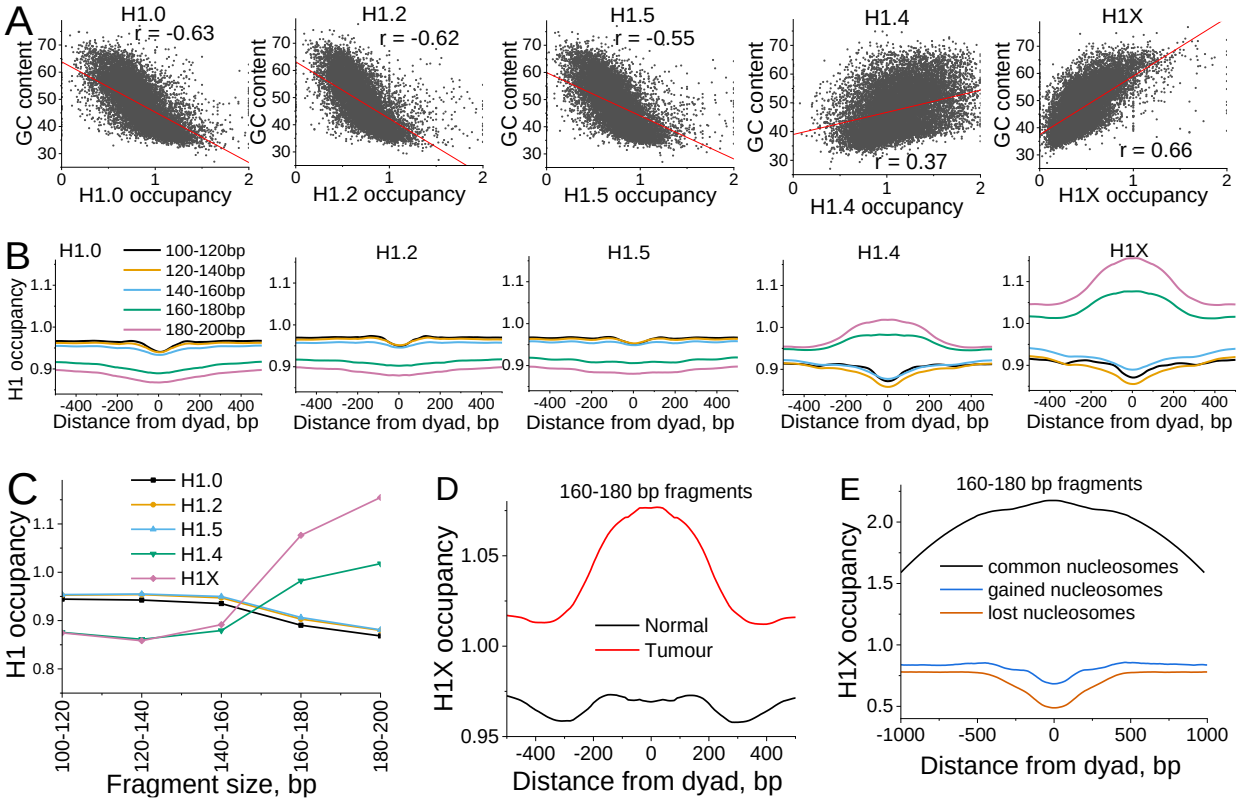

Figure 5. Effects of linker histones on nucleosome repositioning in breast cancer. A) Correlation of GC content and H1 occupancy in 10-kb regions genome-wide. H1 variants H1.0, H1.2 and H1.5 preferentially bind AT-rich regions, while H1.4 and H1X have a higher affinity for GC-rich regions. B-D) Averaged aggregate profiles of H1 occupancy around different types of DNA fragments protected from MNase digestion. These profiles were first calculated individually for each MNase-seq tumour tissue sample reported in this study, and then averaged across all patients. B) Occupancy of H1 variants around nucleosome dyads calculated for different DNA fragment sizes. C) H1 occupancy in the region [-50, 50] bp around nucleosome dyad calculated for different H1 variants as a function of DNA fragment size. D) Average H1X occupancy in T47D breast cancer cell line calculated around DNA fragments of 160-180 bp in normal (black lines) and tumour breast tissues (red). E) Average H1X occupancy profiles around the consensus set defined based on all tissue samples from this study, of common nucleosomes (black), and nucleosomes which are gained (blue) and lost in cancer (orange).

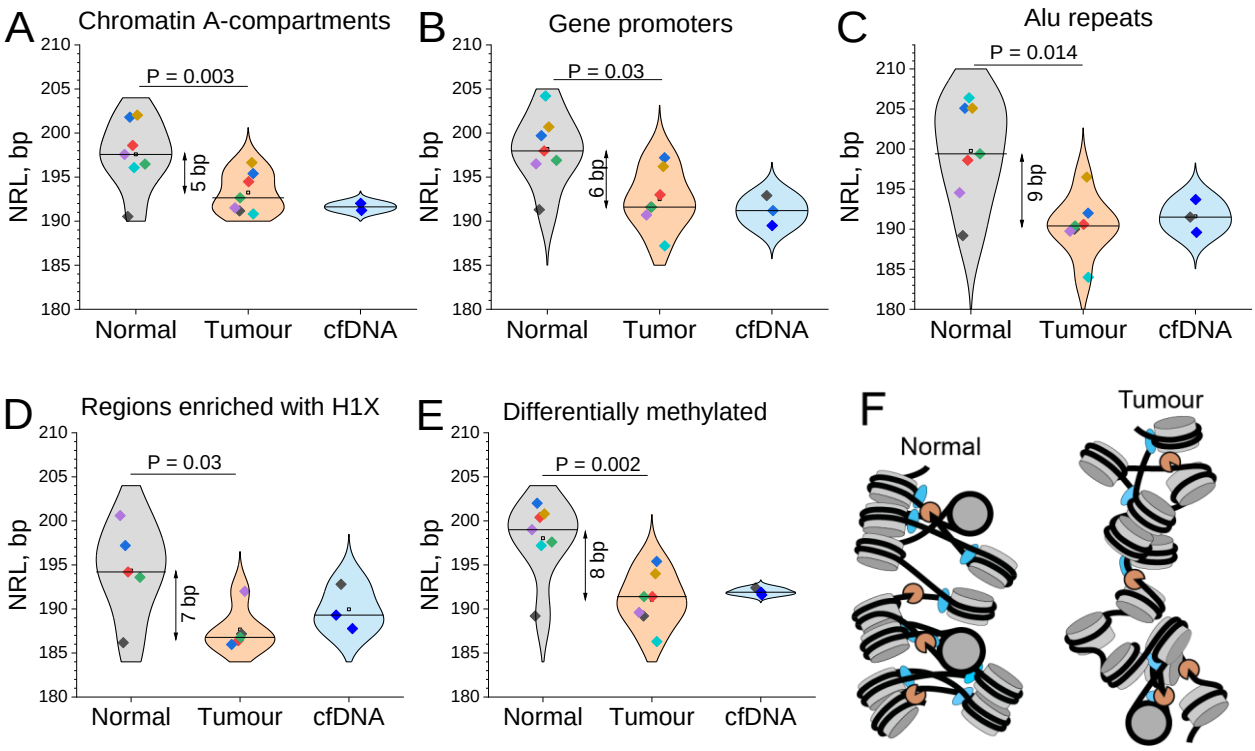

Figure 6. NRL shortening as a novel breast cancer marker. A-E) NRL values calculated for each individual sample from this study inside different types of genomic regions. A) Transcriptionally active A-compartments of chromatin annotated in breast cancer MCF-7 cells; B) Regions surrounding gene promoters. C) Genomic regions around Alu repeats; D) Regions enriched with H1X linker histones in breast cancer T47D cells; E) Differentially methylated DNA regions annotated based on cell lines MCF-10A, MCF-7 and MDA-MB-231. Open squares – average NRL values for a given condition; horizontal lines – median values. The colours of the points are assigned per sample as in Figure 3A. F) A scheme of the global change of nucleosome compaction in breast cancer tissues. DNA (black) is wrapped around histone octamers (grey) in contact with linker histones (blue). MNase used in this study or apoptotic nucleases present *in situ* (orange) cut DNA not protected by nucleosomes and linker histones. Cancer cells have shorter NRL and contain more nucleosomes which are less protected from MNase digestion due to redistribution of linker histone variant H1X, changes in DNA methylation and other factors.