Published online 13 June 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 13* ***7841–7854***
doi: 10.1093/nar/gkx522

Genome-wide profiling of S/MAR-based replicon contact sites

Claudia Hagedorn[1,*], Andreas Gogol-Döring[2], Sabrina Schreiber[3], Jörg T. Epplen[1,3] and Hans J. Lipps[1]

[1]University of Witten/Herdecke, ZBAF, Institute of Cell Biology, Stockumer Strasse 10, 58453 Witten, Germany, [2]Technische Hochschule Mittelhessen (University of Applied Sciences), Department of Bioinformatics, Wiesenstrasse 14, 35390 Gießen, Germany and [3]Department of Human Genetics, Ruhr-University, Universitätsstraße 150, 44801 Bochum, Germany

Received August 23, 2016; Revised May 29, 2017; Editorial Decision May 31, 2017; Accepted June 05, 2017

ABSTRACT

Autonomously replicating vectors represent a simple and versatile model system for genetic modifications, but their localization in the nucleus and effect on endogenous gene expression is largely unknown. Using circular chromosome conformation capture we mapped genomic contact sites of S/MAR-based replicons in HeLa cells. The influence of *cis*-active sequences on genomic localization was assessed using replicons containing either an insulator sequence or an intron. While the original and the insulator-containing replicons displayed distinct contact sites, the intron-containing replicon showed a rather broad genomic contact pattern. Our results indicate a preference for certain chromatin structures and a rather non-dynamic behaviour during mitosis. Independent of inserted *cis*-active elements established vector molecules reside preferentially within actively transcribed regions, especially within promoter sequences and transcription start sites. However, transcriptome analyses revealed that established S/MAR-based replicons do not alter gene expression profiles of host genome. Knowledge of preferred contact sites of exogenous DNA, e.g. viral or non-viral episomes, contribute to our understanding of episome behaviour in the nucleus and can be used for vector improvement and guiding of DNA sequences to specific subnuclear sites.

INTRODUCTION

A characteristic of the eukaryotic cell nucleus is the spatial separation of different genome compartments involved in essential biological processes, such as transcription and replication. It is well accepted that interphase chromosomes are organized in territories within the nucleus, with active and repressive regions occupying different subnuclear compartments (1). Interchromosomal interactions occur preferentially within active chromatin compartments and between inactive chromatin compartments, but are rarely found between both chromatin compartments. This spatial segregation is probably due to the fact that active and inactive compartments interact with specific subnuclear sites. Active genes often associate with subnuclear structures enriched with RNA/DNA polymerases and transcription or replication factors (also referred to as transcription/replication factories) (2), while non-transcribed chromatin regions often overlap with lamin-associated domains and are found near the nuclear envelope (3). According to a model proposed by Cook in 1999 polymerases and other proteins (e.g. transcription factors) are clustered in 'factories' and attached to a subnuclear structure (2). The term factory is applied to sites where transcription (4), replication (5) and repair (6) of endogenous DNA occurs. Recently published data obtained by chromosomal conformation capturing (3C) demonstrate that distal transcribed or transcription factor associated sequences are often in close proximity to each other in the 3D nucleus (7–9) supporting the model of transcription factories. These factories seem to be specialized: each RNA polymerase (I, II, III) accumulates in certain factories (10), whereas RNA polymerase II factories are further specialized in terms of promoter type (11), regulating pathway (12) or gene families (13–15).

For a deeper understanding of the relevance of 3D nuclear structure for the regulation of replication and transcription of exogenous DNA, a non-viral autonomous replicon which can easily be genetically modified, and understanding its behaviour within the nucleus would be most desirable. Based on the observation that an origin of replication binds to a subnuclear structure, most probably the

[*]To whom correspondence should be addressed. Tel: +49 2302 926270; Fax +49 2302 926220; Email: claudia.hagedorn@uni-wh.de
Present address: Claudia Hagedorn, University of Witten/Herdecke, ZBAF, Chair for Biochemistry and Molecular Medicine, Stockumer Strasse 10, 58453 Witten, Germany.

nuclear matrix, at the onset of DNA replication a minimal replication and expression system was constructed in our lab (16). This expression vector carrying a scaffold/matrix attachment region (S/MAR) was shown to replicate autonomously in numerous cell lines, including primary cells (17–20). In the interphase nucleus, the replicon binds to the nuclear matrix by an interaction of the S/MAR with the prominent matrix protein SAF-A (21). S/MAR-based episomes replicate once per cell cycle during early S-phase and the origin recognition complex can assemble at various regions of the episome (22). Establishment efficiency of S/MAR-based replicons is a stochastic and infrequent event and strongly depends on the nuclear compartment the vector reaches after transfection, a phenomenon probably characteristic of all episomal replicons as e.g. EBV (23). Using fluorescence *in situ* hybridization (FISH) it was shown on a single cell level that established replicons co-localize with early replicating foci and post-translational histone modifications associated with transcriptional activity (24). Since a preference of S/MAR-based replicons for specific chromosomes or chromosomal regions could not be observed (24), it can be suggested that specific chromatin pattern assembled over specific sequence elements (e.g. genes, promoters) are favoured for association.

Using circular chromosome conformation capture (4C), we mapped genomic contact sites of three different S/MAR-based replicons stably established in HeLa cells: pEPI-EGFP, pEPI-HS4 (containing an insulator downstream of S/MAR element) and pEPI-Intron (containing an intron between CMV promoter and EGFP). Independent of the inserted *cis*-active elements (HS4-insulator and intron), all three replicons reside preferentially within actively transcribed regions. However, pEPI-EGFP and pEPI-HS4 displayed distinct contact sites, whereas pEPI-Intron showed a rather broad genomic contact pattern. Transcriptome analyses revealed that established S/MAR-based replicons do not alter gene expression profiles of host genome, an observation most important for the use of these replicons in applied biotechnology.

Here, we describe for the first time on a global level the localization of episomal S/MAR-based replicons in the nucleus and the epigenetic features of preferred contact sites. The knowledge of preferred contact sites of S/MAR-based replicons within the genome provides new insights on episome establishment that can be used to improve establishment efficiencies and guiding of specific DNA sequences into specific nuclear compartments.

MATERIALS AND METHODS

Replicons, transfection and cell culture conditions

Within this study the S/MAR-based replicons pEPI-EGFP, pEPI-HS4 and pEPI-Intron were used. The chicken hypersensitive site 4 (HS4)-insulator sequence downstream of and in opposite orientation to the S/MAR element (pEPI-HS4) was shown to increase establishment efficiencies of S/MAR-based replicons (25,26). An intron inserted between promoter and transgene was shown to affect localization of minichromsomes (11). Therefore, intron1 of human beta-globin gene (HBB) was cloned between CMV and EGFP resulting in pEPI-Intron. Replicons were transfected into HeLa cells (German Resource Centre for Biological Material, DSMZ; Braunschweig, Germany) using FuGene HD transfection reagent (Roche; Basel, Switzerland) or Amaxa® Cell Line Nucleofector® Kit T (Lonza; Basel, Switzerland). Cells were maintained in DMEM Medium (PAN Biotech; Aidenbach, Germany) supplemented with 10% foetal bovine serum (PAA; Cölde, Germany), penicillin (10 000 units/ ml)/ streptomycin (10 mg/ml) (PAA; Cölde, Germany), 50 μg/ml partricin (Biochrom; Berlin, Germany) and non-essential amino acids (PAA Cölde, Germany). Twenty-four hours post-transfection, transient transfection efficiency was determined by fluorescence microscopy and flow cytometric analysis (FACS). Generally, transfection efficiency was 50–70%. Cells were selected in the presence of 400 μg/ml G418 for 14 days and maintained in the absence of selection thereafter. Long-term EGFP expression of the mixed populations and single cell clones grown in the absence of selection was verified by FACS analyses.

Copy number analyses of HeLa genome

Genomic DNA was isolated from HeLa mixed populations carrying either pEPI-EGFP, pEPI-HS4, or pEPI-Intron using a standard protocol for salting-out (27) and 250 ng DNA were analyzed for copy number changes using CytoScan® HD Arrays (Affymetrix (Santa Clara, California, USA); >2.6 million markers, resolution limit of 25–50 kb) and respective reagents and instruments according to the manufacturer's instructions. Data were further processed using the Chromosome Analysis Suite version 3.0.0.42 (*NetAffx Library* 33.1 (UCSC genome assembly hg19), Affymetrix). Copy numbers and chromosomal positions were exported and plotted into Circos diagrams.

Circular chromosome conformation capture (4C)

Circular chromosome conformation capture on pEPI was performed as described previously (28). Briefly, interacting DNA segments were cross-linked with 1% paraformaldehyde and nuclei were isolated using a lysis buffer containing 0.2% NP-40 followed by incubation at 37°C in presence of 0.3% SDS. Prior digestion, SDS was sequestered with 2% Triton X-100 and subsequently digested with EcoRI (400 U, 24 h). Digestion efficiencies were quantified in qPCR using qTower Light Cycler (Analytic Jena; Jena, Germany) with FastStart DNA Master^{PLUS} SYBER Green I reaction mix (Thermo Fisher Scientific; Waltham, Massachusetts) and primers covering the EcoRI restriction site within the pEPI genome normalized to a non-digested region (primer sequences see Supplementary Table S1). Digested samples were purified followed by diluted proximity ligation at 16°C, resulting in a 3C-library. For construction of the 4C library, the cross-link was reversed and samples were subjected to digestion with NlaIII followed by a second ligation. Using specific primers (Supplementary Table S1) close to the respective ligation sites, genomic sequences were amplified and subjected to deep sequencing (Illumina HiSeq2000; 2 × 125 bp (mixed populations), 2 × 100 bp (single-cell derived populations); >10 million reads/ sample). For verification of obtained results, independently constructed 4C libraries were cloned in pGEM-Teasy (Promega, Madison, USA) and 96 clones per population were sequenced and analysed (see Supplementary Material).

Read mapping and statistical analyses

The sequencing read pairs were processed as follows: First, we *in-silico* generated all possible self-ligation junction of the pEPI, i.e. all sequences that could be generated by completely digesting pEPI with either NlaIII or EcoRI and then ligating two of the fragments. If any one of the two reads belonging to a read pair mapped completely to one of these self-ligation junctions or to pEPI itself, then the read pair was discarded. In the remaining reads, we searched for pEPI marker sequences, i.e. sequences starting with the PCR primer and ending with the first downstream occurrence of the corresponding enzyme restriction site in pEPI. Read pairs containing the last 10 bp of a pEPI marker sequence followed by a sequence of length $\geq$20 bp not part of pEPI (i.e. possibly originating from the human genome), were selected for the subsequent analysis. All pEPI parts of the reads were stripped and the rest was mapped against the human reference genome (hg19) using the Bowtie 2 read mapping tool (29). A read pair was treated as a 'hit', if both reads in the pair could be uniquely and coherently mapped to genomic positions with distance of $\leq$1 kb between each other. Redundant hits originating from identical read pairs were treated as a single hit. The hits were then assigned to the closest EcoRI site. 4C data sets are available on GEO database (GSE97858). The statistical analyses for calculating the enrichments/depletions of detected EcoRI sites in different genomic features was done as described before (30). As a background model, we used the set of all genomic EcoRI restriction sites in the human genome.

DNA FISH

Episomal DNA and respective contact sites were detected using DNA FISH (31). Briefly, cells grown to 90% confluence on chamber slides were fixed with 4% paraformaldehyde (10 min at 20°C). To make nuclear DNA accessible for probes without affecting 3D chromatin architecture, cells were treated with 0.5% Triton-X100/PBS (15 min at 20°C), incubated overnight in 20% glycerol and subjected to repeated freezing in liquid nitrogen (five cycles). For deproteinization, slides were incubated in 0.1 N HCl (5 min at 20°C) and stored at 4°C in 50% formamide (pH 7.0)/2× SSC (0.3 mol/l NaCl and 0.03 mol/l sodium citrate, pH 7.0) prior hybridization (48 h at 37°C). Hybridization mix contained 50 ng/μl of DIG and/or biotin labelled probe, 50% formamide, 2× SSC, 10% dextran sulphate, 40 mmol/l phosphate buffer (23 mmol/l Na_2HPO_4, 17 mmol/l NaH_2PO_4, pH 7.0), 0.1% SDS, 1× 'Denhardt's' buffer (0.02% Ficoll 400, 0.02% polyvinylpyrolidone, and 0.02% bovine serum albumin) and 2.5 μg/μl sheared salmon sperm DNA. Probes were prepared using DNA polymerase, DIG-11-dUTP or biotin-16-dUTP, and specific primers amplifying EGFP gene or chromosomal contact sites (see Supplementary Table S1). Nuclear and probe DNA was denatured simultaneously (2 min at 75°C). Post hybridization, slides were washed three times each in 2× SSC (5 min at 37°C) and 0.1× SSC (5 min at 65°C) and blocked in 4% BSA/4× SSC/0.2% Tween-20. Immunodetection was performed using the following antibodies in indicated order: rabbit anti-DIG (45 min at 37°C, 1:80; Sigma-Aldrich, St. Louis, USA), goat anti-rabbit-Alexa488 (45 min at 37°C, 1:200; Invitrogen, Carlsbad, California), mouse anti-biotin (45 min at 37°C, 1:80; Sigma-Aldrich, St. Louis, USA), and goat anti-mouse-Alexa555 (45 min at 37°C, 1:200; Invitrogen, Carlsbad, CA, USA). Slides were washed three times (3 min at 20°C) after each decoration step and mounted in Vectashield (Biozol, Eching, Germany).

3C-qPCR

To confirm detected contact sites we performed qPCR on 3C-libraries that were constructed as described above. After diluted proximity ligation, 3C libraries were precipitated and further purified using NucleoSpin® gDNA Clean-up kit (Macheray & Nagel; Düren, Germany). For 3C-qPCR we used replicon-specific primers close to the EcoRI site and primers specific for the respective chromosomal contact site. Obtained data were quantified using the $\Delta\Delta C_t$ method (32). First, C_t-values of ligation products (S/MAR-based replicon with chromosomal site) were normalized to C_t values of overall S/MAR-based replicons in each sample (EGFP) and then compared to C_t values in an undigested und unligated sample. Primers used for quantitative PCR are listed in Supplementary Table S1.

Nuclear fractionation

Nuclear fractionation was performed as described before (25,33). Briefly, cells were detached by treatment with trypsin. An aliquot of 10^7 cells per reaction was washed once with cold PBS followed by incubation in cytoskeleton buffer (10 mmol/l PIPES, 300 mmol/l saccharose, 100 mmol/l NaCl, 3 mmol/l $MgCl_2$, 1 mol/l EGTA; 4 min on ice) and centrifugation (1000 × g, 4°C, 3 min); the supernatant contained soluble cyto- and nucleoplasmic proteins. Nuclei were then incubated in extraction buffer (10 mmol/l PIPES, 300 mmol/l saccharose, 250 mmol/l ammonium sulphate, 3 mmol/l $MgCl_2$, 1 mol/l EGTA; 4 min on ice) and collected by centrifugation as above; the supernatant contained soluble nuclear components including histone H1. Digestion was performed in digestion buffer (10 mmol/l PIPES, 300 mmol/l saccharose, 50 mmol/l NaCl, 3 mmol/l MgCl2, 1 mol/l EGTA; 3 h, 37°C) using four enzymes that do not cut within the replicons (NotI, EcoRV, XhoI and PvuI) together with either MlsI (linearising pEPI-EGFP) or HindIII (linearising pEPI-Intron and pEPI-HS4). All used enzymes do not cut within the chromosomal regions. Subsequent centrifugation left the complete matrix in the pellet fraction while supernatant contained DNA and histones. After extraction with 2 mol/l NaCl buffer (10 mmol/l PIPES, 300 mmol/l saccharose, 2 mol/l NaCl, 3 mmol/l $MgCl_2$, 1 mol/l EGTA; 4 min on ice) matrix-associated proteins, which are not part of the core filament were located in the supernatant; matrix-associated DNA and the core filament network (matrix skeleton) were located in the pellet. All fractions were subjected to Proteinase K digestion, DNA was precipitated and further purified using NucleoSpin® gDNA Clean-up kit (Macheray & Nagel; Düren, Germany). Whole genomic DNA served as input control; primers used for quantitative PCR are listed in Supplementary Table S1.

Motif discovery, GO term analysis and *in silico* prediction of putative S/MARs

Web-based tool MEME was used to discover common motifs within frequently occurring contact sites (pEPI-EGFP, pEPI-Intron score $\geq$ 1000; pEPI-Intron score $\geq$ 2000) applying default parameters, maximum number of motifs was set to 10. Identified motifs were further analyzed for significantly association with genes linked to one or more Genome Ontology (GO) terms (*P*-value cut-off 0.01; http://meme-suite.org). GO terms were then tested for enrichment (*P*-value cut-off 10^{-3}) using 'GOrilla' (34) and visualised as a network model using REViGO (35) (node size represents fold-change enrichment, colour represents associated *P*-value). Putative S/MARs were predicted *in silico* using the web-based tool WebSIDD (http://benham.genomecenter.ucdavis.edu/sibz/) (36) applying default parameters.

RNAseq/ transcriptome analyses

To determine the influence of episomally maintained S/MAR contained replicons on gene expression, RNAseq was performed. Total RNA was extracted in Trizol (Invitrogen; Carlsbad, California) from both, HeLa wildtype cells and a mixed populations stably maintaining pEPI-EGFP (HeLa T55E), used for library preparation (NEBNext® mRNA Library Prep Reagent Set for Illumina; NEB, Frankfurt, Germany) and subsequently subjected to deep sequencing (Illumina NextSeq500, single-end, 75bp). Obtained raw data were processed using TopHat-Cufflinks pipeline on Galaxy server (http://usegalaxy.org) as described before (37–39). Briefly, reads were mapped against the human genome (GRCh19) using TopHat, followed by assembling and FPKM value estimation with CufLinks using default parameters for single-end reads. Significant changes in transcript expression were calculated using CuffDiff. Data visualization was performed with CummeRbund (37–39). Data sets are available on GEO database (GSE97725).

For verification in qPCR 1μg DNase-treated RNA was reverse transcribed into cDNA using First Strand Synthesis Kit (Thermo Fisher Scientific; Waltham, Massachusetts) and subsequently used in quantitative PCR (genes and primer sequences are listed in Supplementary Table S2).

RESULTS

S/MAR-based replicons do not show chromosomal preferences

Stable establishment of S/MAR-based replicons is a stochastic event and seems to depend on the nuclear compartment the vector reaches after transfection (24). To identify DNA sequences to which S/MAR-based replicons, containing different functional genomic elements, are in close proximity, circular chromosome conformation capturing (4C) coupled to high-throughput sequencing was performed. Either mixed populations or single cell clones of HeLa cells stably transfected with pEPI-EGFP, pEPI-HS4 (HS4 insulator downstream of S/MAR), and pEPI-Intron (HBB intron1 between CMV promoter and EGFP) (Figure 1A) were crosslinked, digested with EcoRI and NlaIII and ligated (Figure 1B). Subsequently, genomic sequences were amplified (Figure 1C) and sequenced (100 bp/125 bp, paired end). After stringent filtering for reads that (i) contain bait sequence (pEPI DNA) ligated to genomic DNA and (ii) could be uniquely mapped to the human genome, we obtained 123.895 (pEPI-EGFP), 533.221 (pEPI-HS4), and 369.896 (pEPI-Intron) reads of mixed populations. Here, the number of mapped reads per chromosome correlated with chromosome size (pEPI-EGFP $r = 0.788$, $P < 0.0001$; pEPI-HS4 $r = 0.532$, $P = 0.009$; pEPI-Intron $r = 0.848$, $P < 0.0001$; Figure 1D), with an accumulation of pEPI-HS4 reads on chromosomes 7 and 12 (20.94% and 17.17% of all reads mapped to chromosomes 12 and 7, respectively). No significant correlation with gene density of chromosomes could be observed (pEPI-EGFP $r = –0.165$, $P = 0.45$; pEPI-HS4 $r = 0.341$, $P = 0.11$; pEPI-Intron $r = –0.311$, $P = 0.148$; Figure 1E).

Using 10^6 cells per experiment and assuming two vector copies per cell on average, contact sites of 2×10^6 vector copies were examined. Out of the 123 895 reads obtained for pEPI-EGFP 1215 different contact sites were detected, 5276 contact sites out of 533.221 reads for pEPI-HS4, and 1136 contact sites out of 369 896 for pEPI-Intron. Regions of up to 100 kb in length containing reads with a minimum score (number of reads mapped to a unique locus) of 100 were summarized to HotSpots, overlapping HotSpots were further summarized to one Hotspot, resulting in 295 (pEPI-EGFP), 1109 (pEPI-HS4) and 473 (pEPI-Intron) HotSpots (Figure 2A–C). HotSpots and contact sites with a frequency $\geq$1000 were considered as significant and are termed *frequent HotSpots/ contact sites* below. Whereas only few and non-clustered, frequent HotSpots (score $\geq$1000) were detected for pEPI-EGFP (16 HotSpots, Figure 2D), those for pEPI-HS4 (34 HotSpots, Figure 2E) appear clustered to distinct chromosomal loci (chromosomes 1, 2, 7 and 12). In contrast, HotSpots of pEPI-Intron (135 HotSpots, Figure 2F) appear rather evenly distributed throughout the genome. Remarkably, frequent HotSpots of pEPI-Intron contained significantly fewer intron-less genes when compared to genes within frequent HotSpots of pEPI-EGFP (O/E 0.26, $P = 0.0001$) and pEPI-HS4 (O/E 0.21, $P < 0.0001$) (Supplementary Figure S1). Recent studies revealed that HeLa cells display a remarkably high level of aneuploidy (40). To exclude that detected HotSpots simply reflect chromosomal amplifications we determined copy numbers (CN) throughout the genome. As shown in Figure 2, HeLa cells used in this study are predominantly triploid (orange dots, CN = 3) with some highly amplified chromosomal regions (e.g. p-arm of chromosome 5; dark red dots, CN > 5). However, no correlation between frequently occurring HotSpots (score $\geq$1000) and amplified regions (CN > 5) was found, demonstrating that there are indeed specific chromosomal loci to which S/MAR-based replicons

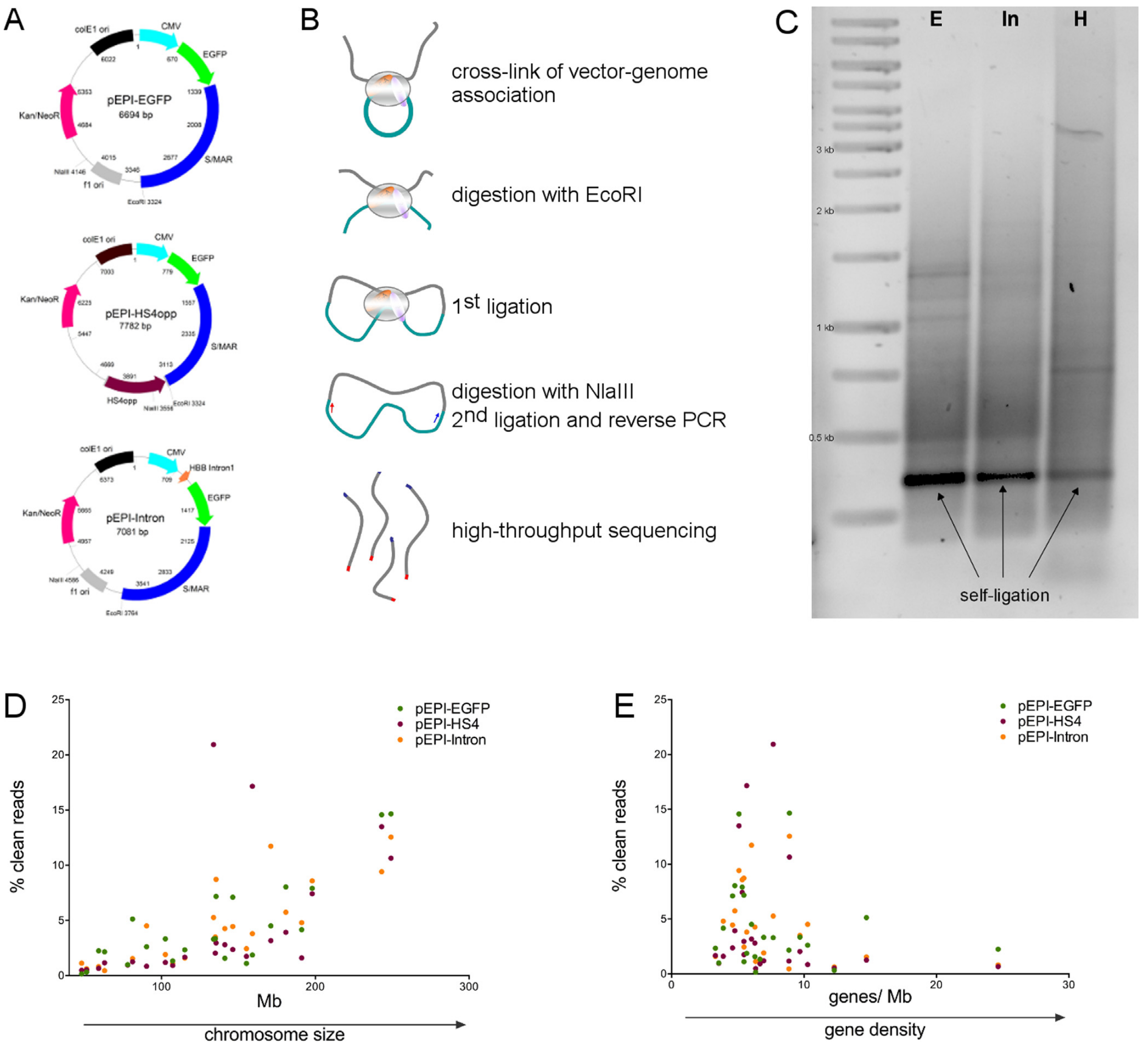

Figure 1. Circular chromosome conformation capture (4C) on S/MAR-based replicons. (**A**) To identify genomic contact partners of S/MAR-based replicons pEPI-EGFP, pEPI-HS4 (chicken hypersensitive site 4 (HS4)-insulator sequence downstream of S/MAR) and pEPI-Intron (human beta-globin (HBB) intron1 between CMV and EGFP), cells stably maintain these replicons were subjected to 4C. (**B**) Vector-genome interactions were cross-linked, followed by two separate digestion and ligations steps. Replicon-specific primers were used to amplify unknown genomic sequences. Obtained amplicons (**C**) were subjected to deep sequencing and mapped to the human genome. *E, pEPI-EGFP; H, pEPI-HS4; In, pEPI-Intron*. Number of detected contact sites correlated with (**D**) chromosome size, but not with (**E**) gene density.

are in close proximity. To verify that S/MAR based vectors are not randomly distributed in the genome we repeated the 4C experiments with two independent mixed populations maintaining pEPI-EGFP or pEPI-HS4, respectively. Here, 4C libraries were introduced into a cloning vector and 96 clones of each library were sequenced and analyzed. If S/MAR based replicons indeed have preferential contact sites, an accumulation of contact sites should also be seen when analyzing a small set of ligated episome-chromosome contacts. A similar distribution of contact sites as in the global analysis was detected and many of these contact sites were identical to those found before (Supplementary Figure S2 A-D and Supplementary Material). The three most-frequently detected contact sites of pEPI-EGFP and pEPI-HS4 were also verified *in situ* using DNA FISH (Figure 2G and H and Supplementary Figure S2 F) and the three most-frequent contact sites of each replicon were additionally verified in an independent 3C library using PCR analyses (Supplementary Figure S2 E).

Assuming that S/MAR-based replicons are positioned randomly in the nucleus and/ or behave dynamic during mitosis, in a mixed population we would expect that each read obtained for the different replicons should be associated with a different genomic EcoRI site. The observation of considerably fewer contact sites for each construct suggests both, preferred contact sites of S/MAR-based replicons with host genomes and a rather non-dynamic behaviour during mitosis. Therefore a very distinct contact pattern should be expected in individual clones. For this reason, we analyzed the contact sites of S/MAR-based replicons in individually established single-cell derived populations. Clones #1 and #2 of pEPI-EGFP displayed 2261 and 1434 HotSpots, for pEPI-HS4 clone#1 742 HotSpots were detected (Supplementary Figure S3 A–C). Remark-

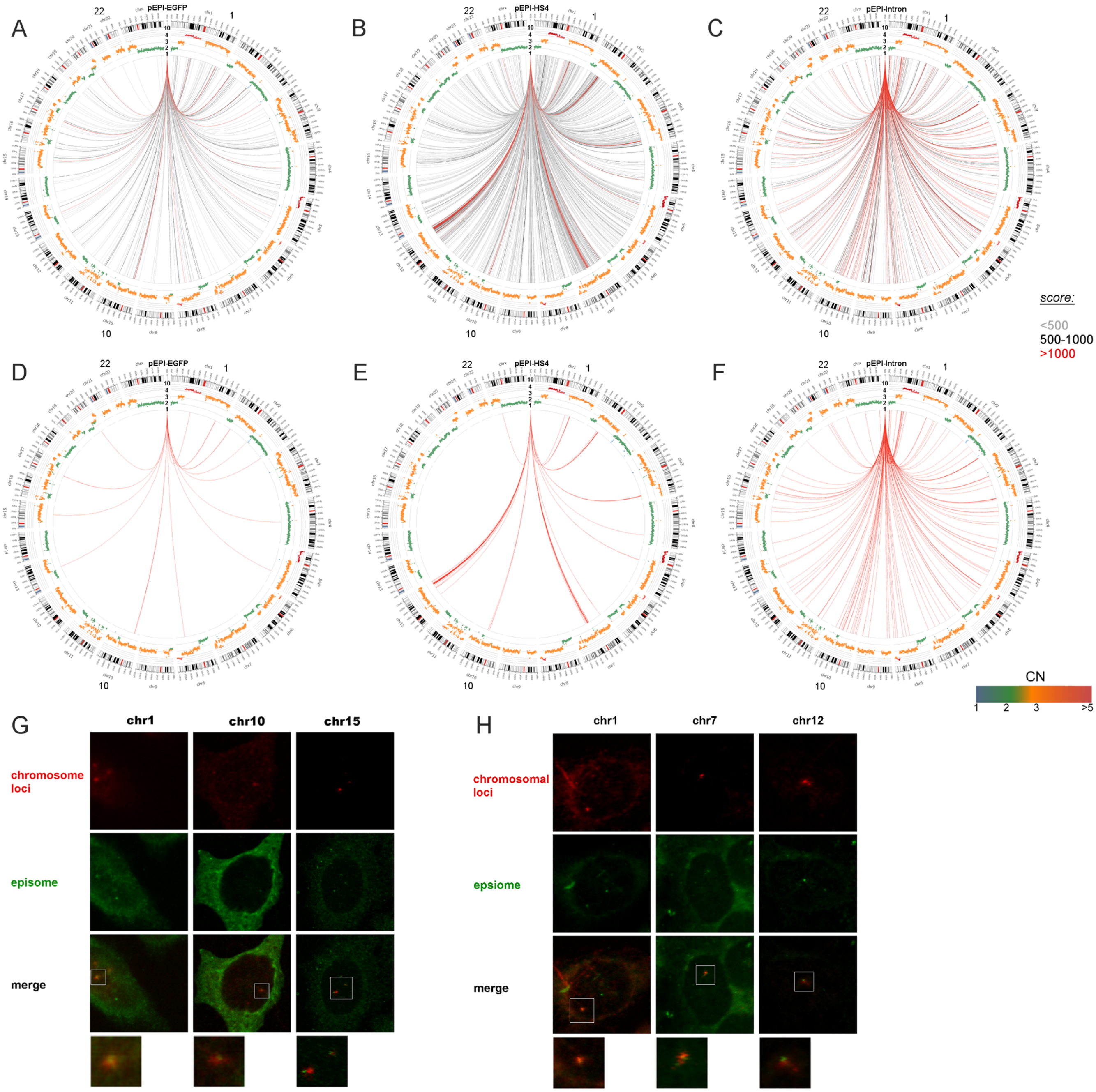

Figure 2. Circos diagrams of detected interaction HotSpots in mixed populations. Chromosomes are arranged clockwise starting with chromosome 1. Each line starting from pEPI (centre) to a chromosomal locus represents a contact HotSpot with a minimum score (detection frequency) of 100. Copy numbers (CN) of HeLa genome are indicated with colour gradient from blue for CN = 1 to red for CN = 10 (**A**) 295 HotSpots were detected for pEPI-EGFP, (**B**) 1109 HotSpots for pEPI-HS4 and (**C**) 473 HotSpots for pEPI-Intron. HotSpots with a score ≥1000 (frequent HotSpots) were considered as significant and are shown in red. (**D**) Sixteen frequent HotSpots were detected for pEPI-EGFP, (**E**) 34 frequent HotSpots for pEPI-HS4, and (**F**) 135 frequent HotSpots for pEPI-Intron. Light gray lines, interaction HotSpots with score 100–500; black lines, interaction HotSpots with score 500–1000; red lines, interaction HotSpots with score >1000. Using DNA FISH, the three most frequently detected contact sites of (**G**) pEPI-EGFP and (**H**) pEPI-HS4 were verified in situ. Episomes and chromosomal loci were detected using specific probes tagged with DIG (episome, green) and biotin (chromosomal contact site, red), respectively.

ably, frequently occurring HotSpots (score $\geq$ 1000) were highly clustered to a specific chromosomal locus. For pEPI-EGFP clone#1 four frequent HotSpots cluster on chromosome 20 (Supplementary Figure S3 E). Twelve of 16 frequent HotSpots of pEPI-EGFP clone#2 cluster on chromosome 1 (Supplementary Figure S3 F), and five of six frequent HotSpots of pEPI-HS4 were found to cluster on chromosome 9 (Supplementary Figure S3 G). These frequently occurring Hot Spots were also identified in mixed populations, albeit not as frequently occurring contact sites (score < 1000). The number of frequently occurring HotSpots may correlate with the number of S/MAR-based replicons per individual cell. As observed in mixed population, pEPI-Intron clone#1 displayed a comparable broad contact pattern with 182 HotSpots of which 94 occurred frequently (score $\geq$ 1000) and were not restricted to a certain locus (Supplementary Figure S3 D and H). Again, the observed distinct HotSpots of pEPI-EGFP and pEPI-HS4 in clonal populations indicates that genomic contact sites are 'inherited' during mitosis. In contrast, the high number of frequently occurring HotSpots for pEPI-Intron in a single-cell derived population indicates either a dynamic behaviour during mitosis or a high variability of co-transcribed loci. However, comparison of contact patterns of all three replicons in mixed populations, as well as in single-cell derived populations, implies an impact of genomic elements (e.g. insulator, intron) on the behaviour of episomes.

Contact sites are enriched for markers of open chromatin

As outlined above, detected contact pattern indicate the existence of preferred sites of localization. Indeed, *in situ* hybridization and 3D microscopy on single cells revealed that S/MAR-based replicons co-localize with epigenetic markers associated with active transcription (24). These observations prompted us to analyze epigenetic modifications and chromatin status of the identified contact sites to obtain a global picture of common features that are linked to vector-genome contact sites. The association of episomal contact sites with specific genomic and epigenetic features was evaluated using random controls generated computationally. ChIP-seq, DNase-Seq and chromatin segmentation data from ENCODE (41) were used to define chromatin segments, chromatin accessibility and transcriptional activity. Features with $P \leq 0.001$ were considered as significant and are shown in Figure 3 (dashed bars). Genomic contact sites in mixed populations of pEPI-EGFP, pEPI-HS4 and pEPI-Intron were significantly enriched for transcribed genome segments and histone modifications associated with active transcription and open chromatin structure (H3K27ac, H3K36me3, H3K4me1/2/3, H3K79me2, H3K9ac, H4K20me1). Notably, an up to 3.48-fold enrichment of transcription start sites (TSS) within the genomic contact sites was observed for all three replicons (Figure 3A, $P < 0.001$). Only pEPI-HS4 displayed a moderate enrichment of enhancer elements within its genomic contact sites (1.77-fold, $P < 0.0001$; pEPI-EGFP 1,23-fold, $P = 0.454$; pEPI-Intron 1,79-fold, $P = 0.019$), whereas a moderate but significant enrichment for (CCCTC-binding factor (CTCF) binding was detected for pEPI-EGFP (1.63-fold, $P < 0.0001$) and pEPI-Intron (1.72-fold, $P < 0.0001$), but not for pEPI-HS4 (1.14-fold, $P = 0.006$). Episomal contact sites appeared to be less frequent within transcriptional silent regions (repressed), lamina associated domains (LAD), and histone modification marks associated with transcriptional silencing and heterochromatin (H3K27me3, H3K9me3) (Figure 3A, $P < 0.001$). Focusing on gene coding regions, a preference of episomal replicons to interact with gene bodies and intronic sequences, as well as with TSS spanning sequences was observed. Contact sites of all three replicons were especially enriched for promoter sequences (Figure 3B; pEPI-EGFP 2.14-fold, $P < 0.0001$; pEPI-HS4 2.0-fold, $P < 0.0001$; pEPI-Intron 1.82-fold, $P = 0.0001$). Episomes pEPI-EGFP and pEPI-Intron did not show a significant preference to interact with transcription end sites (TES), whereas in contrast contact sites of pEPI-HS4 were slightly enriched for TES (1.26-fold, $P < 0.0001$; Figure 3B). All three replicons showed a moderate tendency to interact with genomic sites near CpG islands (±5–20 kb; Figure 3B).

Again, a similar enrichment pattern was observed in an independent 4C analysis of a mixed cell population with up to 69% of contact sites being associated with active regions (Supplementary Results and Supplementary Figure S2 A–D) and in single-cell derived populations (Supplementary Figure S4). These data indicate that S/MAR-based episomal replicons favour open chromatin regions, especially sequences spanning promoters and transcription start sites for association.

Contact sites are enriched for polymerase II binding and located in proximity to potential origins of replication

Based on the observation that transcription factories are specialized in terms of regulating pathway (12) or gene families (13–15), we tested genes within preferred contact sites of the mixed populations for an overrepresentation of certain gene families or pathways, using Panther Classification System (42). Within frequently occurring HotSpots (score $\geq$ 1000), none of the replicons showed an enrichment for neither gene families nor pathways. Considering all detected HotSpots, those of pEPI-HS4 were slightly and non-significantly ($P \geq 0.01$) enriched for Gene ontology (GO) terms *phosphatidylinositol binding* (3-fold enrichment, $P = 0.021$) and *ATP binding* (1.6-fold enrichment, $P = 0.038$), whereas genes within pEPI-Intron HotSpots were found to be enriched for GO term *cell adhesion* (2.9-fold enrichment, $P = 0.0097$).

We next searched for common protein binding motifs within the frequent contact sites (pEPI-EGFP and pEPI-HS4 score $\geq$ 1000; pEPI-Intron score $\geq$ 2000). Sequences spanning these contact sites (3000 bp) were subjected to MEME, a web-based motif discovery tool. Within these regions common protein binding motifs with (G)GAGG or stretches of $(T)_{4–11}$ occurred frequently (Supplementary Figure S5). These motifs were further analyzed for enrichment of associated GO terms using the web-tool *Gene Ontology for Motifs* (GOMo). Detected protein binding motifs within frequent contact sites of all three replicons were associated with GO terms *binding*, *sequence-specific DNA binding*, *transcription regulator activity*, and *histone binding* (Figure 4A), indicating an enrichment of transcription factor binding sites (TFBS) within frequent contact

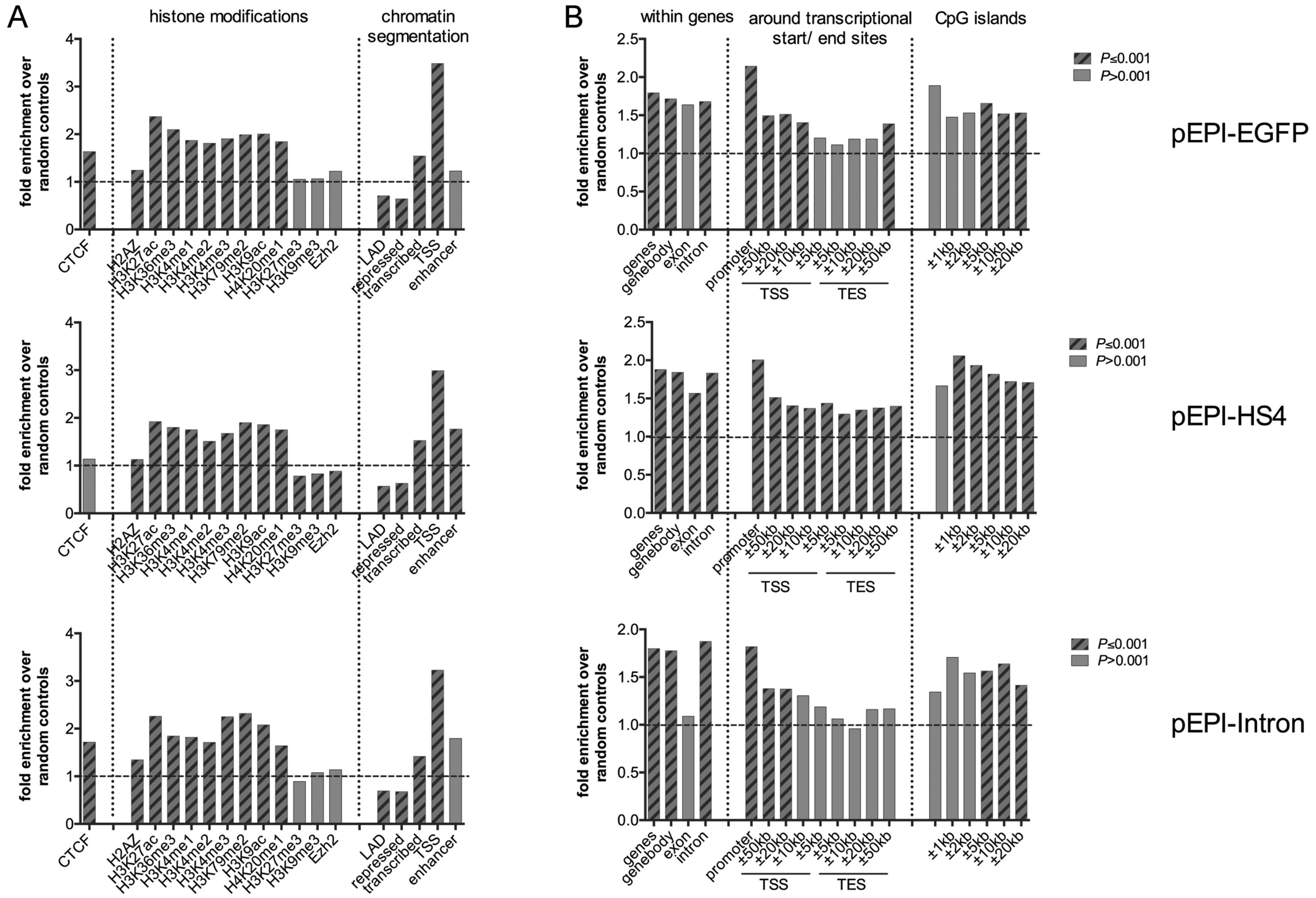

Figure 3. Genomic and epigenetic features of detected contact sites in mixed populations. Enrichment of contact sites of pEPI-EGFP, pEPI-HS4, and pEPI-Intron compared to random sites in the vicinity of specific (**A**) epigenetic and (**B**) genomic features. Values are given as the proportion of contact events divided by the proportion of random events. Enrichment with $P \leq 0.001$ was considered as significant (dashed bars).

sites. Consequently, we tested the enrichment of RNA polymerase II (RNAP2) binding from ChIP-seq data (*SYDH*, ENCODE/Stanford/Yale/USC/Harvard) within all detected contact sites. Consistent with the observation that S/MAR-based replicons preferentially associate with actively transcribed chromatin (Figure 3), a significant enrichment of RNAP2 binding was detected for overall RNAP2 (pEPI-EGFP 2.15-fold, $P = 0.012$; pEPI-HS4 1.83-fold, $P < 0.0001$; pEPI-Intron 3.74-fold, $P < 0.0001$), as well for active (phosphorylated) RNAP2-Ser2 (pEPI-EGFP 3.02-fold, $P < 0.0001$; pEPI-HS4 2.69-fold, $P < 0.0001$; pEPI-Intron 1.95-fold, $P = 0.003$; Figure 4B). However, beyond an enrichment of putative albeit different TFBS within frequent contact sites we could not identify other sequence motifs or regulatory pathways being enriched for one replicon.

A co-localization of S/MAR-based replicons with early-replicating foci (24) and an association with the nuclear matrix has been described before (33). Based on these previous observations, we speculated that contact sites of S/MAR-based replicons may also function as or are at least in close proximity to origins of replication. Recently published studies connected dimethylated lysine 79 of histone 3 (H3K79me2) with replication initiation (43) and found shared replication origins strongly associated with chromatin modifications H3K4me3, H3K9ac and unmethylated CpG islands (44). When considering all detected contact sites, an enrichment of H3K4me3, H3K79me2 and H3K9ac was detectable for all three replicons (Figure 3A), whereas no significant association with early replicating sequences (pEPI-EGFP 0.52-fold, $P = 1$; pEPI-HS4 0.96-fold, $P = 1$, pEPI-Intron 2.2-fold, $P = 0.11$) was shown. Yet, the three most frequently occurring contact sites of each vector were located within early replicating regions containing (unmethylated) CpG islands, associated with active histone modification marks (H3K4me3, H3K9ac, H3K79me2) coupled with a lack of repressive histone modification marks (H3K27me3) (Supplementary Figure S6A–C). As outlined above, a moderate association with CpG shores (±5–20 kb from CpG island) was detected for all three replicons, while pEPI-HS4 contact sites were associated with CpG shores ±2 kb from CpG islands (1.9-fold, $P < 0.0001$; Figure 3B). However, enrichment within CpG shores does not seem to depend on methylation status of the respective CpG island: pEPI-EGFP and pEPI-Intron showed either only low or no significant association with methylation status, while pEPI-HS4 was significantly ($P < 0.001$) associated with both, methylated and unmethylated

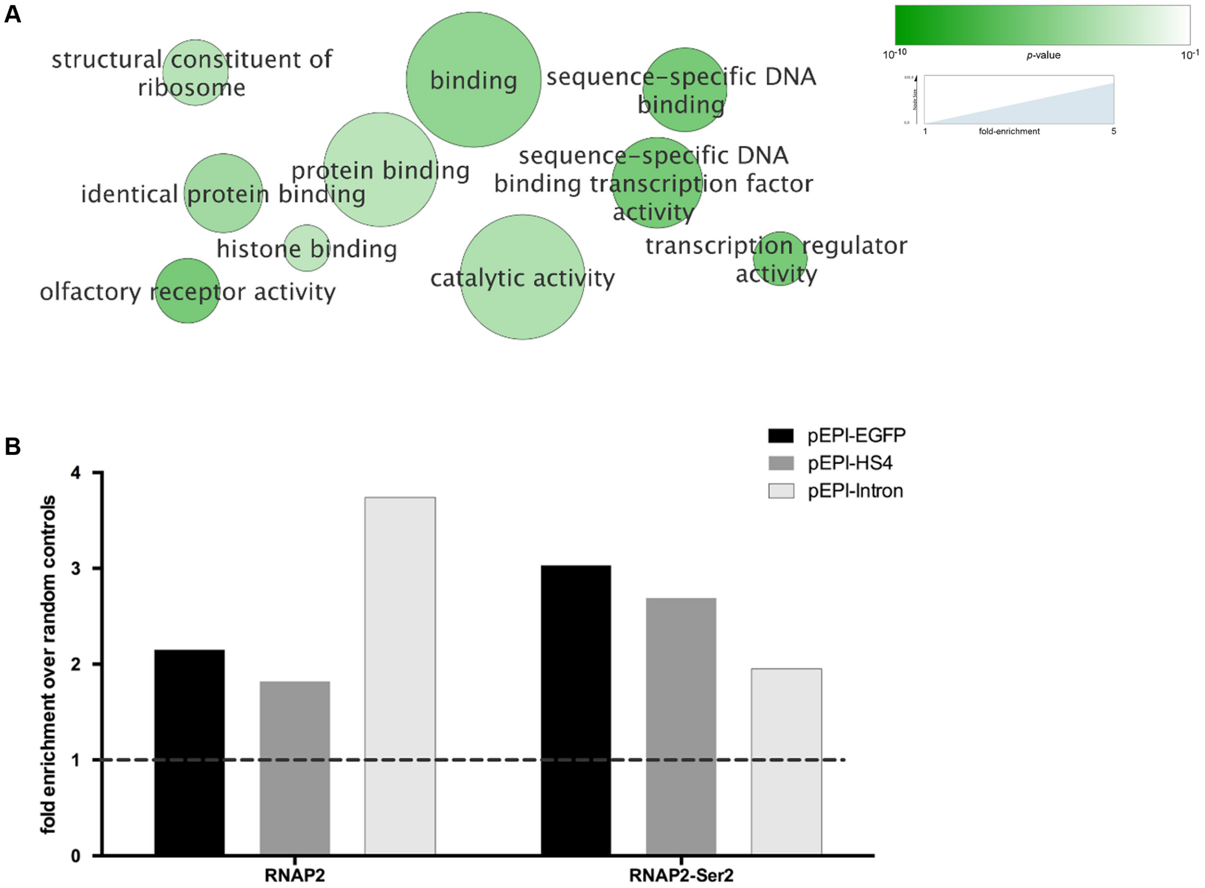

Figure 4. Frequent contact sites are contain transcription factor binding motifs and are enriched for RNAP2. (**A**) Common protein binding motifs identified with web-based tool MEME within the frequent contact sites (pEPI-EGFP and pEPI-HS4 score ≥ 1000; pEPI-Intron score ≥ 2000) were associated with GO terms *binding*, *histone binding*, *sequence-specific DNA binding*, and *transcription regulator activity.* Colour intensity increases with significance, node size represents enrichment of term in dataset. (**B**) Contact sites of all three S/MAR-based replicons are significantly enriched for binding of overall RNAP2 as well for active (phosphorylated) RNAP2-Ser2.

CpG islands with a tendency towards unmethylated CpGs (Supplementary Figure S6 D).

Since S/MAR-based replicons associate with the nuclear matrix it should be expected that sequences in close proximity are also matrix-associated and should include other S/MARs. Using an web-based *in silico* prediction tool for the three most-frequent contact sites of each replicon (36), we found at least one putative S/MAR in close proximity (Figure 5A). We then isolated nuclear matrix associated sequences as described before (25,33) and amplified the three most-frequent contact sites of each replicon. Episomal DNA as well as these contact sites were found to be enriched in the nuclear matrix/ core-filament fraction. Since transcriptionally active chromatin is described to be associated with the nuclear matrix (45), the housekeeping gene *plasminogen activator, tissue* (PLAT) served as positive control, whereas a non-coding transcriptionally inactive region of chromosome 9 found to be present in both, free DNA and matrix/ core-filament fraction (Figure 5B).

Episomal S/MAR-based replicons do not alter expression profile of host genome

As described above, S/MAR-based replicons tend to interact with chromosomal sites of active transcription, favouring promoter sequences and transcription start sites. Genes displaying promoter-promoter interactions have been shown to be not only transcribed cooperatively but are also capable of co-activating other promoters within such an interacting cluster (8). For this reason, it cannot be excluded that S/MAR-based replicons do have an influence on endogenous gene expression. We analyzed the transcriptome of untransfected HeLa cells (wild-type) and compared with HeLa cells stably established pEPI-EGFP. Obtained reads were mapped and analyzed using the CufFLinks pipeline and visualized using cummeRbund (37–39). In Figure 6A, *Fragments per Kilobase Million* (FPKM) values of untransfected HeLa (control; X-axis) have been plotted against FPKM values of HeLa cells stably maintaining pEPI-EGFP (Y-axis) and no global changes in gene expression in pEPI-EGFP maintaining cells could be observed. To identify differentially expressed genes, false discovery rate (FDR) cut off was set to 0.01 and –$\log_{10}$ of FDR ad-

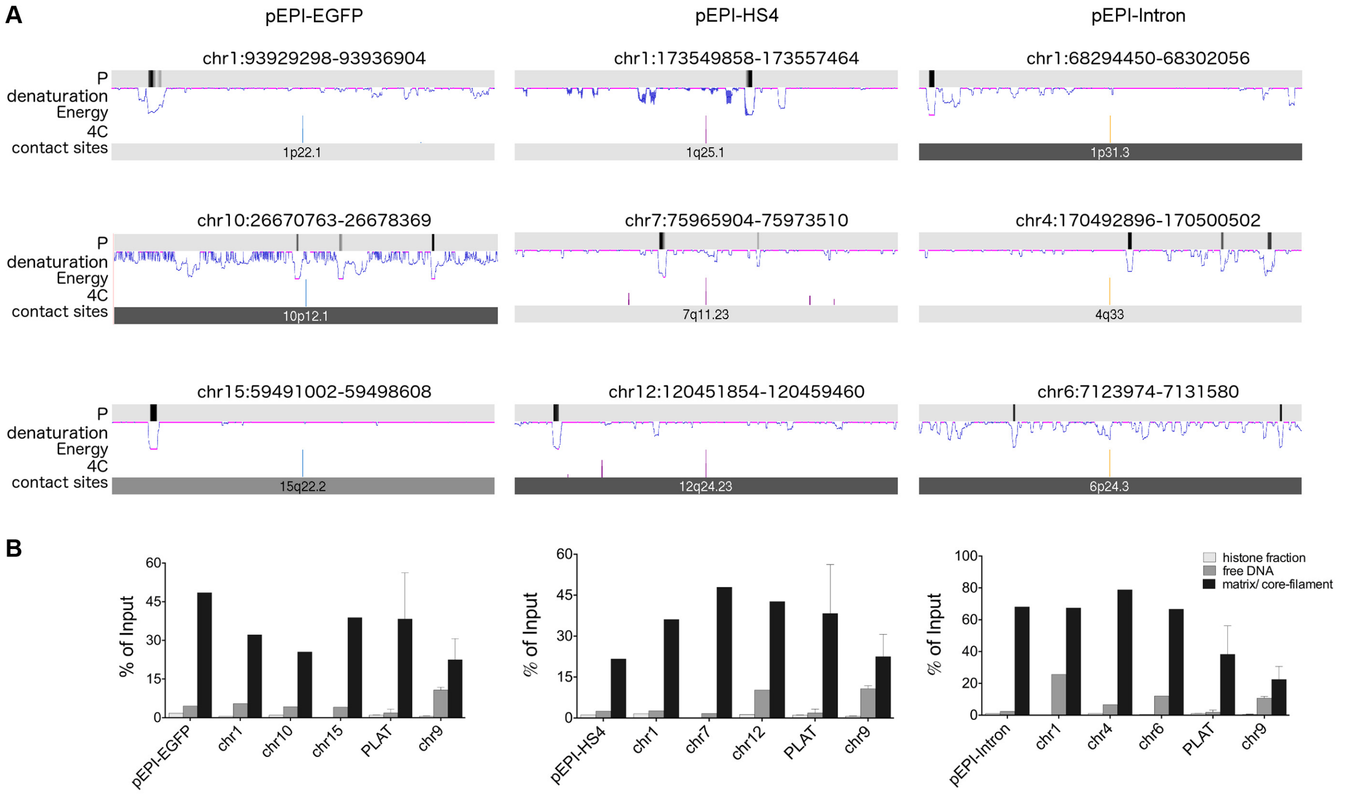

Figure 5. Frequent contact sites are in close proximity to putative S/MARs and associated with the nuclear matrix. (**A**) Putative S/MARs in close proximity to the three most frequent contact sites were identified with WebSIDD prediction tool. *denaturation Energy, energy needed to force a base pair at a respective position open; P, probability of strand separation (black bars indicate high probabilities)*. (**B**) Three most-frequent contact sites of S/MAR-based replicons pEPI-EGFP, pEPI-HS4 and pEPI-Intron were found to be associated with the nuclear matrix. Housekeeping gene PLAT and a transcriptionally silent portion of chromosome 9 served as positive and negative control, respectively. *PLAT, plasminogen activator (tissue)*.

justed *P*-value (*q*-value) has been plotted against $\log_2$ of fold change. Significantly differentially expressed genes with $q <$ 0.01 and an absolute fold-change of 2 would appear red and were not detected when comparing expression levels of untransfected HeLa cells with cells stably maintaining pEPI-EGFP (Figure 6B). Genes that displayed high fold change and low *q*-values ($q < 0.05$) were not located within replicon contact sites. For verification we chose five of these genes (Supplementary Table S1) and compared expression levels between untransfected HeLa and HeLa stably maintaining pEPI-EGFP within three replicates; two genes that did not show altered expression in RNAseq analyses served as controls. As shown in Figure 6C no significant changes in genes expression ($P < 0.01$) were detected. Comparing the contact patterns of pEPI-EGFP, pEPI-HS4 and pEPI-Intron with our transcriptome data, it becomes obvious that contact sites of S/MAR-based replicons are located within actively transcribed regions, although no correlation between contact frequency and transcription level could be observed (Figure 6D).

DISCUSSION

Understanding the behaviour of autonomously replicating episomes may not only contribute to our understanding of genome organisation in the nucleus but is also vitally important for their use in biotechnology and gene therapy. The development of the chromosome conformation capturing (3C) technique, based on the strikingly simple idea that digestion and religation of fixed chromatin ends allows for detection of DNA contact sites and frequencies (46), enabled researchers to study topological properties and spatial organization of chromosomes in the nucleus. In this study, we used the 3C-derived circular chromosome conformation capture (4C) to study the chromosomal localization of non-viral S/MAR-based episomal replicons. For the first time, contact sites of exogenous episomal DNA were mapped and characterized. Using the 4C technique we identified genomic contact sites of S/MAR-based replicons in two independently established HeLa mixed populations and independently established single-cell derived populations. Mapped contact sites do not correlate with amplified regions of the HeLa cell line used and therefore can be regarded as being specific. However before using S/MAR-based replicons in applied biotechnology results of this proof-of-principle study should be confirmed in each cell type used for application.

To address the impact of *cis*-acting elements on genomic localization, 4C was also performed with cell populations maintaining S/MAR-based replicons harbouring either an insulator sequence downstream the S/MAR (pEPI-HS4) or an intron between CMV promoter and EGFP transgene (pEPI-Intron). Compared to pEPI-EGFP (16 frequent contact sites on 12 chromosomes), the insulator contain-

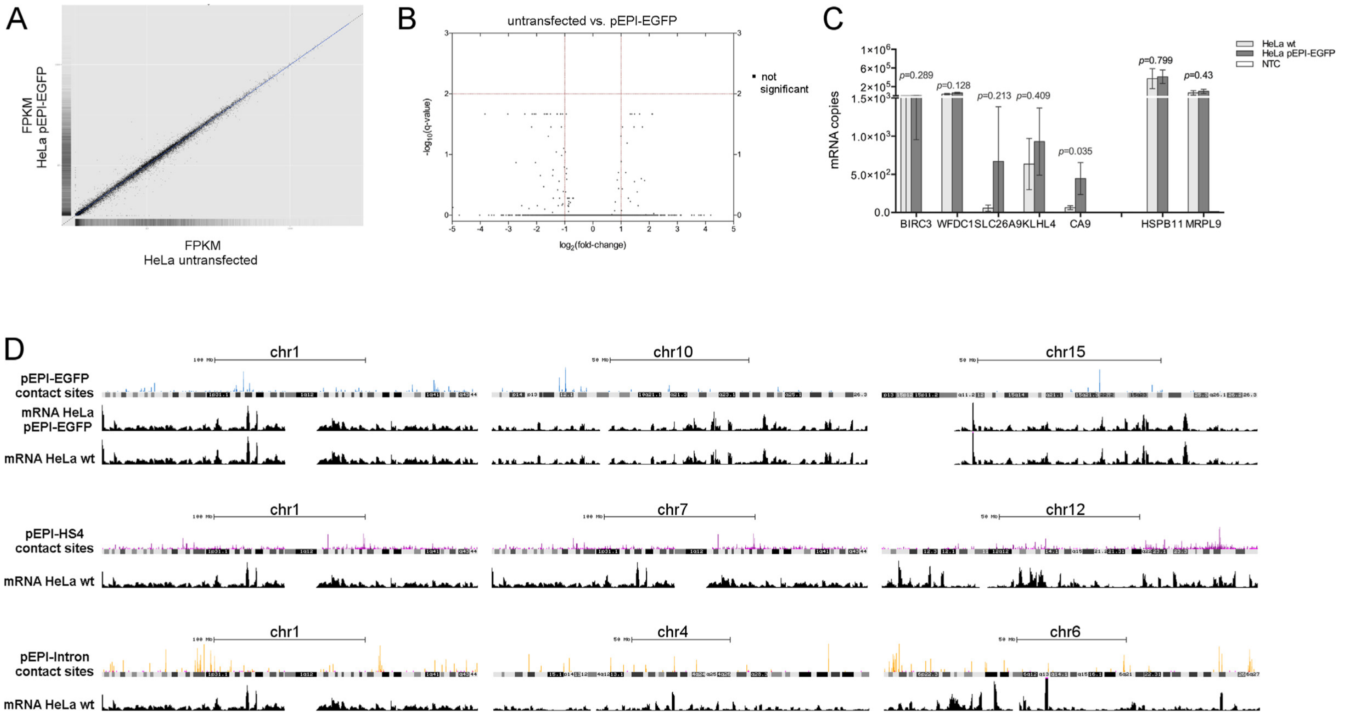

Figure 6. S/MAR based replicons do not alter gene expression. (**A**) FPKM values of untransfected HeLa cells plotted against FPKM values of HeLa cells stably maintaining pEPI-EGFP. No global changes in gene expression were observed. *FPKM, Fragments per Kilobase Million*. (**B**) Genes are ranked in a volcano plot according to their statistical *P*-value (y-axis) and their relative abundance ratio ($\log_2$ fold-change) between untransfected HeLa cells (wild-type) and HeLa cells stably maintaining pEPI-EGFP. Significantly differentially expressed genes with an corrected *P*-value $q < 0.01$ and an absolute fold-change of 2 would appear red and were not detected. (**C**) Expression Level of five potential differentially expressed genes were analyzed in qPCR in three independent replicates. Genes *HSPB1* and *MRPL9* served as controls (see also Supplementary Table S1). (**D**) Contact sites of pEPI-EGFP (upper panel), pEPI-HS4 (middle), and pEPI-Intron (lower panel) are located within actively transcribed regions, but independent of transcription level.

ing replicon pEPI-HS4 possessed a clustered contact pattern with 34 sites of frequent contacts distributed to loci on 7 chromosomes, especially on chromosomes 7 and 12. In contrast, frequent HotSpots of pEPI-Intron (score ≥ 1000) were evenly distributed throughout the genome (135 HotSpots on 21 chromosomes). Replicon-specific contact pattern were detected in two different mixed populations, but were also found in individual clones. Moreover, identified contact sites were also verified by *in situ* experiments. It therefore seems reasonable to assume that results obtained with mixed populations are characteristic for the respective S/MAR-based replicon. Obviously, different genomic elements cloned in a S/MAR-based vector result in different contact pattern of the respective replicons, indicating a function-dependent influence of *cis*-acting genomic elements (e.g. insulator, introns) on replicon localization in the nucleus. We interpret the detected genomic contact sites of S/MAR-based replicons as co-transcribed sequences in transcription factories. According to the model of transcription taking place in specialized factories (2,11), we suppose that S/MAR-based replicons are also transcribed in certain factories with respect to inserted genomic elements. The wide-spread contact pattern and increased transgene expression of pEPI-Intron, observed in both, mixed populations and single-cell derived populations, may therefore result from its transcription in a variety of transcription factories that are specialized for genes with introns as containing splicing factors (11). Since ∼91% of all genes contain introns, the majority of transcription factories should contain splicing factors. This also fits with our observation that contact sites of pEPI-Intron contain significantly fewer intron-less genes than contact sites of pEPI-EGFP and pEPI-HS4. We searched for further common features of co-transcribed genes for each replicon, e.g. common transcription factors and regulatory pathways, but except an enrichment of transcription factor binding sites and RNA polymerase II binding within the co-transcribed sequences no replicon-specific features could be detected (Figure 4 and Supplementary Figure S5). Extensive clustering of pEPI-HS4 contact sites to specific chromosomal loci may therefore either result from an efficient establishment process, mediated by an HS4-nuclear matrix interaction (25,47) or from locally highly intermixing chromatin domains (48).

Despite the genomic element-specific contact pattern, the epigenetic signature of contact sites is replicon-independent and enriched for active histone marks (H3K4me3, H3K79me2) in the absence of repressive histone marks and sequences close to the nuclear envelope and lamina-associated domains (LAD). Specific contact pattern of the used replicons indicate that *cis*-acting genomic sequences associate with specific subnuclear structures. However, it is not fully clear whether *cis*-acting genomic elements guide the replicon in specific transcription factories thus determining number and variety of co-transcribed sequences (pEPI-Intron) or induce changes in spatial chromatin organization resulting in replicon specific contact pattern (pEPI-HS4) (49). We show here that S/MAR-based replicons preferentially associate with actively transcribed chromatin and are in close proximity to putative endogenous S/MAR sequences. A co-localization with active histone modifications, early replicating foci, and splicing speckles in the absence of repressive chromatin markers has been shown before (24) and could now be globally confirmed. Since we have described before that an active transcription running into or over the S/MAR is essential for episomal replication and maintenance this observation supports our previous data (50). At this point, we can only speculate about cell-type specificity of detected contact pattern. Since contact pattern are not restricted to specific chromosomes but to certain epigenetic signatures and subnuclear sites dependent on inserted genomic elements, we assume that contact pattern will vary from cell-type to cell-type as transcriptional landscapes vary in a cell-type specific manner. It is very likely that also in other cell types only a very limited number of contact sites will be observed.

The observation of only few and clustered contact sites strongly suggests the existence of preferred contact sites and a rather non-dynamic behaviour during mitosis. This hypothesis is supported by verification of the three most-frequent contact sites of S/MAR-based replicons pEPI-EGFP and pEPI-HS4 *in situ* (Figure 2G and H and Supplementary Figure S2 F) and in an independent 3C library (Supplementary Figure S2 E). However, transcribed genes are known to often oscillate between active and inactive states for short periods of time and the chromatin contacts that define gene expression are very complex (51). Therefore, it might be that S/MAR-based replicons do show some dynamics throughout a cell cycle with respect to their transcriptional state. S/MAR-based as well as other episomal replicons establish with on average 5–10 copies/ cell (52). While in single-cell derived populations of pEPI-EGFP and pEPI-HS4, the 4–16 frequently occurring contact HotSpots probably represent the number of S/MAR-replicons per cell, this assumption is unlikely for the observed 94 frequently occurring HotSpots of pEPI-Intron. It might be that, unlike pEPI-EGFP and pEPI-HS4, pEPI-Intron behaves rather dynamic during mitosis, resulting in new contact sites after each cell division. Since we could verify the three most-frequent contact sites of pEPI-Intron in an independent 3C-library (Supplementary Figure S2 E), it is conceivable that the high number of frequently occurring contact sites reflects the variety of genomic loci that are co-transcribed in factories in which pEPI-Intron is located and transcribed.

Topologically associating domains (TADs) are considered as the fundamental structural building blocks of chromosomes (53,54) and seem to be tissue-invariant (55). TAD boundaries are enriched for various genomic features, e.g. CTCF and promoter-associated histone marks (H3K4me3) (55,56) that were also found to be enriched in detected contact sites. We therefore speculate that S/MAR-based replicons tend to co-localize with TAD boundaries. During mitosis major features of TADs that are linked to gene expression are lost. It is suggested that cell-type specific DNA elements like enhancers and promoters as well as TAD boundaries are bookmarked by remaining nucleosome free and thus are accessible for proteins such as transcription factors or RNA polymerase II to re-associate, inducing correct chromosome folding and gene expression (57). Since the epigenetic signature of S/MAR-based replicons dynamically changes in a cell-cycle dependent manner with a specific removal of histone modifications during mitosis (58), we hypothesize that genomic localization and active transcription of S/MAR-based replicons is memorized during mitosis the same way as it is for endogenous genes and TADs.

S/MAR-based replicons used in this study do not code for viral proteins. However, their observed preference to co-localize with promoter sequences and transcription start sites, and recent observations that cooperatively transcribed promoters can influence each other (8) rise concerns that S/MAR-based replicons have the potential to alter endogenous gene expression. Therefore, we compared the transcriptome of untransfected HeLa cells with HeLa cells stably maintaining pEPI-EGFP. Setting the FDR to <0.01 we found no significantly differentially expressed genes. This finding is of utmost importance for potential gene therapeutic application of S/MAR-based replicons. This study is the first comprehensive analysis of genomic contact sites of a non-viral, autonomously replicating episome. It preferentially associates with a subset of actively transcribed genes but within this study we were not able to detect common characteristics of these loci. The genomic contact sites are very non-dynamic, but are influenced by genomic *cis*-acting sequences incorporated into the replicon. The present work provides not only the basis for a systematic search of sequences determining nuclear localization, but also a proof-of-principle that 3C-based techniques are versatile tools to localize exogenous DNA within the 3D nucleus. This may lay the foundation for a routine application of 3C/4C to localize exogenous DNA like viral genomes or non-viral episomes in the nucleus.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

C.H. thanks Prof. P. Cook (University of Oxford) and Dr A. Papantonis (CMMC Cologne) very much for introducing her into the secrets of 4C technology. C.H. was supported by the Peter and Traudl Engelhorn foundation. Circos diagrams were created by Stefan Hagedorn (Ilmenau University of Technology).

FUNDING

Peter and Traudl Engelhorn Foundation. Funding for open access charge: University Witten/ Herdecke, Centre for Biomedical Education and Research (ZBAF).

Conflict of interest statement. None declared.

REFERENCES

1. Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev.*, **2**, 292–301.

2. Cook,P.R. (1999) The organization of replication and transcription. *Science (New York, N.Y.)*, **284**, 1790–1795.
3. Guelen,L., Pagie,L., Brasset,E., Meuleman,W., Faza,M.B., Talhout,W., Eussen,B.H., de Klein,A., Wessels,L., de Laat,W. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948–951.
4. Jackson,D.A., Hassan,A.B., Errington,R.J. and Cook,P.R. (1993) Visualization of focal sites of transcription within human nuclei. *EMBO J.*, **12**, 1059–1065.
5. Hozak,P., Hassan,A.B., Jackson,D.A. and Cook,P.R. (1993) Visualization of replication factories attached to nucleoskeleton. *Cell*, **73**, 361–373.
6. Jackson,D.A., Balajee,A.S., Mullenders,L. and Cook,P.R. (1994) Sites in human nuclei where DNA damaged by ultraviolet light is repaired: visualization and localization relative to the nucleoskeleton. *J. Cell Sci.*, **107**, 1745–1752.
7. Gondor,A., Rougier,C. and Ohlsson,R. (2008) High-resolution circular chromosome conformation capture assay. *Nat. Protoc.*, **3**, 303–313.
8. Li,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.
9. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., de Wit,E., van Steensel,B. and de Laat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.
10. Pombo,A., Jackson,D.A., Hollinshead,M., Wang,Z., Roeder,R.G. and Cook,P.R. (1999) Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. *EMBO J.*, **18**, 2241–2253.
11. Xu,M. and Cook,P.R. (2008) Similar active genes cluster in specialized transcription factories. *J. Cell Biol.*, **181**, 615–623.
12. Papantonis,A., Kohro,T., Baboo,S., Larkin,J.D., Deng,B., Short,P., Tsutsumi,S., Taylor,S., Kanki,Y., Kobayashi,M. *et al.* (2012) TNFalpha signals through specialized factories where responsive coding and miRNA genes are transcribed. *EMBO J.*, **31**, 4404–4414.
13. Cai,S., Lee,C.C. and Kohwi-Shigematsu,T. (2006) SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat. Genet.*, **38**, 1278–1288.
14. Noordermeer,D., de Wit,E., Klous,P., van de Werken,H., Simonis,M., Lopez-Jones,M., Eussen,B., de Klein,A., Singer,R.H. and de Laat,W. (2011) Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat. Cell Biol.*, **13**, 944–951.
15. Noordermeer,D., Leleu,M., Splinter,E., Rougemont,J., De Laat,W. and Duboule,D. (2011) The dynamic architecture of Hox gene clusters. *Science*, **334**, 222–225.
16. Piechaczek,C., Fetzer,C., Baiker,A., Bode,J. and Lipps,H.J. (1999) A vector based on the SV40 origin of replication and chromosomal S/MARs replicates episomally in CHO cells. *Nucleic Acids Res.*, **27**, 426–428.
17. Argyros,O., Wong,S.P., Gowers,K. and Harbottle,R.P. (2012) Genetic modification of cancer cells using non-viral, episomal S/MAR vectors for in vivo tumour modelling. *PLoS One*, **7**, e47920.
18. Haase,R., Argyros,O., Wong,S.P., Harbottle,R.P., Lipps,H.J., Ogris,M., Magnusson,T., Vizoso Pinto,M.G., Haas,J. and Baiker,A. (2010) pEPito: a significantly improved non-viral episomal expression vector for mammalian cells. *BMC Biotechnol.*, **10**, 20.
19. Papapetrou,E.P., Ziros,P.G., Micheva,I.D., Zoumbos,N.C. and Athanassiadou,A. (2006) Gene transfer into human hematopoietic progenitor cells with an episomal vector carrying an S/MAR element. *Gene Ther.*, **13**, 40–51.
20. Stehle,I.M., Scinteie,M.F., Baiker,A., Jenke,A.C. and Lipps,H.J. (2003) Exploiting a minimal system to study the epigenetic control of DNA replication: the interplay between transcription and replication. *Chromosome Res.*, **11**, 413–421.
21. Jenke,B.H., Fetzer,C.P., Stehle,I.M., Jonsson,F., Fackelmayer,F.O., Conradt,H., Bode,J. and Lipps,H.J. (2002) An episomally replicating vector binds to the nuclear matrix protein SAF-A in vivo. *EMBO Rep.*, **3**, 349–354.
22. Schaarschmidt,D., Baltin,J., Stehle,I.M., Lipps,H.J. and Knippers,R. (2004) An episomal mammalian replicon: sequence-independent binding of the origin recognition complex. *EMBO J.*, **23**, 191–201.
23. Deutsch,M.J., Ott,E., Papior,P. and Schepers,A. (2010) The latent origin of replication of Epstein-Barr virus directs viral genomes to active regions of the nucleus. *J. Virol.*, **84**, 2533–2546.
24. Stehle,I.M., Postberg,J., Rupprecht,S., Cremer,T., Jackson,D.A. and Lipps,H.J. (2007) Establishment and mitotic stability of an extra-chromosomal mammalian replicon. *BMC Cell Biol.*, **8**, 33.
25. Hagedorn,C., Antoniou,M.N. and Lipps,H.J. (2013) Genomic cis-acting sequences improve expression and establishment of a nonviral vector. *Mol. Ther. Nucleic Acids*, **2**, e118.
26. Moreno,R., Martinez,I., Petriz,J., Gonzalez,J.R., Gratacos,E. and Aran,J.M. (2009) Boundary sequences stabilize transgene expression from subtle position effects in retroviral vectors. *Blood Cells Mol. Dis.*, **43**, 214–220.
27. Miller,S.A., Dykes,D.D. and Polesky,H.F. (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.*, **16**, 1215.
28. Stadhouders,R., Kolovos,P., Brouwer,R., Zuin,J., van den Heuvel,A., Kockx,C., Palstra,R.J., Wendt,K.S., Grosveld,F., van Ijcken,W. *et al.* (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat. Protoc.*, **8**, 509–524.
29. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
30. Gogol-Doring,A., Ammar,I., Gupta,S., Bunse,M., Miskey,C., Chen,W., Uckert,W., Schulz,T.F., Izsvak,Z. and Ivics,Z. (2016) Genome-wide profiling reveals remarkable parallels between insertion site selection properties of the MLV retrovirus and the piggyBac transposon in primary human CD4(+) T cells. *Mol. Ther.*, **24**, 592–606.
31. Cremer,M., Grasser,F., Lanctot,C., Muller,S., Neusser,M., Zinner,R., Solovei,I. and Cremer,T. (2008) Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. *Methods Mol. Biol. (Clifton, NJ)*, **463**, 205–239.
32. Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods (San Diego, Calif)*, **25**, 402–408.
33. Baiker,A., Maercker,C., Piechaczek,C., Schmidt,S.B., Bode,J., Benham,C. and Lipps,H.J. (2000) Mitotic stability of an episomal vector containing a human scaffold/matrix-attached region is provided by association with nuclear matrix. *Nat. Cell Biol.*, **2**, 182–184.
34. Eden,E., Navon,R., Steinfeld,I., Lipson,D. and Yakhini,Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.
35. Supek,F., Bosnjak,M., Skunca,N. and Smuc,T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.
36. Bi,C. and Benham,C.J. (2004) WebSIDD: server for predicting stress-induced duplex destabilized (SIDD) sites in superhelical DNA. *Bioinformatics*, **20**, 1477–1479.
37. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
38. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
39. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
40. Landry,J.J., Pyl,P.T., Rausch,T., Zichner,T., Tekkedil,M.M., Stutz,A.M., Jauch,A., Aiyar,R.S., Pau,G., Delhomme,N. *et al.* (2013) The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda)*, **3**, 1213–1224.
41. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
42. Mi,H., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.*, **8**, 1551–1566.
43. Fu,H., Maunakea,A.K., Martin,M.M., Huang,L., Zhang,Y., Ryan,M., Kim,R., Lin,C.M., Zhao,K. and Aladjem,M.I. (2013)

Methylation of histone H3 on lysine 79 associates with a group of replication origins and helps limit DNA replication once per cell cycle. *PLoS Genet.*, **9**, e1003542.

44. Smith,O.K., Kim,R., Fu,H., Martin,M.M., Lin,C.M., Utani,K., Zhang,Y., Marks,A.B., Lalande,M., Chamberlain,S. *et al.* (2016) Distinct epigenetic features of differentiation-regulated replication origins. *Epigenet. Chromatin*, **9**, 18.
45. Ciejek,E.M., Tsai,M.J. and O'Malley,B.W. (1983) Actively transcribed genes are associated with the nuclear matrix. *Nature*, **306**, 607–609.
46. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science (New York, NY)*, **295**, 1306–1311.
47. Yusufzai,T.M. and Felsenfeld,G. (2004) The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 8620–8624.
48. Boettiger,A.N., Bintu,B., Moffitt,J.R., Wang,S., Beliveau,B.J., Fudenberg,G., Imakaev,M., Mirny,L.A., Wu,C.T. and Zhuang,X. (2016) Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, **529**, 418–422.
49. Yang,J. and Corces,V.G. (2011) Chromatin insulators: a role in nuclear organization and gene expression. *Adv Cancer Res.*, **110**, 43–76.
50. Rupprecht,S., Hagedorn,C., Seruggia,D., Magnusson,T., Wagner,E., Ogris,M. and Lipps,H.J. (2010) Controlled removal of a nonviral episomal vector from transfected cells. *Gene*, **466**, 36–42.
51. Hager,G.L., McNally,J.G. and Misteli,T. (2009) Transcription dynamics. *Mol. Cell*, **35**, 741–753.
52. Jackson,D.A., Juranek,S. and Lipps,H.J. (2006) Designing nonviral vectors for efficient gene transfer and long-term gene expression. *Mol. Ther.*, **14**, 613–626.
53. Gibcus,J.H. and Dekker,J. (2013) The hierarchy of the 3D genome. *Mol. Cell*, **49**, 773–782.
54. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J., Sedat,J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
55. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
56. Huang,J., Marco,E., Pinello,L. and Yuan,G.C. (2015) Predicting chromatin organization using histone marks. *Genome Biol.*, **16**, 162.
57. Dekker,J. (2014) Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture. *Epigenet. Chromatin*, **7**, 25.
58. Rupprecht,S. and Lipps,H.J. (2009) Cell cycle dependent histone dynamics of an episomal non-viral vector. *Gene*, **439**, 95–101.