DSBCapture: *in situ* capture and direct sequencing of dsDNA breaks

Stefanie V. Lensing[1], Giovanni Marsico[1], Robert Hänsel-Hertsch[1], Enid Y. Lam[1,4], David Tannahill[1] & Shankar Balasubramanian[1,2,3]

1. Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK.
2. Department of Chemistry, University of Cambridge, Cambridge, UK.
3. School of Clinical Medicine, University of Cambridge, Cambridge, UK
4. Present address: Cancer Research Division, Peter MacCallum Cancer Centre, East Melbourne, Victoria, Australia
Correspondence should be addressed to S.B.: sb10031@cam.ac.uk

Double-strand DNA breaks (DSBs) continuously arise and cause mutations and chromosomal rearrangements. Here, we present DSBCapture, a sequencing-based method that captures DSBs *in situ* and directly maps these at single nucleotide resolution enabling the study of DSB origin. DSBCapture shows substantially increased sensitivity and data yield compared to other methods. Employing DSBCapture, we uncovered a striking relationship between DSBs and elevated transcription within nucleosome-depleted chromatin.

Due to their mutagenic potential, it is important to identify the precise location of endogenous DSBs and those induced by therapeutic drugs and environmental insults[1,2]. ChIP-seq[3], GUIDE-seq[4] and BLESS[5], are currently used for the genome-wide investigation of DSBs *in situ*. ChIP-seq, however, is indirect and relies on the capture of specific proteins that mark DSBs by proxy, thus compromising both accuracy and resolution. GUIDE-seq profiles off-target cleavage by CRISPR-Cas nucleases and relies on the integration of a double-stranded oligonucleotide at DSBs by non-homologous end joining (NHEJ)[4]. BLESS captures DSBs *via* blunt-ended ligation of a barcoded adapter to broken DNA ends[5], a process less efficient than cohesive-end ligation[6]. Furthermore, sequential addition of the capture and sequencing adapters necessitates two rounds of PCR, a technique known to introduce biases[7]. The resulting libraries suffer from low sequence diversity (i.e. low variation in the initial bases of the sequences) leading to problems during Illumina sequencing. This is normally remedied by diluting the sample DNA with an unrelated library (typically phiX) to artificially increase the sequence diversity however, this reduces data yield of the sample under investigation by over 50 %[8]. Finally, as sequencing is not only initiated from the captured DSB (proximal end) but also from the end generated through fragmentation (distal end), the number of sequencing reads that directly identify the site of DNA damage in single-end sequencing is halved. To overcome these limitations we developed DSBCapture, a substantially improved method derived from BLESS that enables direct *in situ* capture of DSBs using a modified P5 Illumina adapter to facilitate sequencing without additional library preparation steps (**Fig. 1a** and **Supplementary Fig. 1 a,b,c,d**). Ligation of a T-tailed modified P5 Illumina adapter to break sites identifies solely sites of *in situ* DSB formation in single-end sequencing and, as DSBCapture libraries have high sequence diversity, no spike-in of another library is required for sequencing. DSBCapture displays enhanced data yield and quality, higher reproducibility and superior sensitivity compared to BLESS. In a head-to-head comparison DSBCapture identified 4.5-fold

more DSBs (84,946 compared to 18,816) in normal human epidermal keratinocytes (NHEK), providing the most comprehensive DSB landscape in a normal human cell line.

We initially validated DSBCapture using controlled double-strand cleavage by EcoRV in fixed HeLa nuclei. DSBCapture identified 93.7 % of the 430,897 EcoRV restriction sites in the human genome as cleaved (**Supplementary Fig. 2a** and **Supplementary Table 1**). Furthermore, DSB detection is not influenced by the chromatin state or the DNA sequence content. 0.5 % of the detected EcoRV sites lie within open chromatin whilst 99.5 % lie in heterochromatin which corresponds to the distribution of the predicted sites. The average GC content 100 bp either side of cleaved and uncleaved EcoRV sites is comparable (37 % and 36 %, respectively) demonstrating no evident GC bias for DSB detection.

To verify that DSBCapture can detect *in situ* generated DSBs, we performed DSBCapture on AID-DIvA cells (**Supplementary Table 2**) - a U2OS cell line expressing the AsiSI restriction enzyme fused to a modified estrogen receptor ligand-binding domain[9]. Upon 4-hydroxy tamoxifen (4OHT) treatment, the restriction enzyme translocates to the nucleus causing sequence-specific DSBs at GCGATCGC sites. The locations of AsiSI-induced DSBs have been evaluated by ChIP-seq using the DNA damage markers γH2AX, RAD51 and XRCC4[9]. A total of 121 AsiSI sites were identified by DSBCapture, recovering 74 of the 100 'most cleaved' sites previously identified by γH2AX ChIP-seq[9] (**Fig. 1b** and **Supplementary Fig. 2b**). DSBCapture also identified 47 additional AsiSI sites of which 28 (60 %) overlap with XRCC4 or RAD51 ChIP-seq peaks[9], illustrating that DNA repair proteins localize to these sites. The ability of DSBCapture to precisely map nuclease-induced DSBs could make it useful for the study of CRISPR off-target cleavage sites.

In AsiSI-expressing U2OS cells, DSBCapture also recovered 2,372 endogenous DSBs (lacking AsiSI sites). By comparison, GUIDE-seq identified 25 endogenous DSBs in U2OS cells[4] (**Supplementary Fig. 3**). The substantially higher number of DSBs revealed by DSBCapture probably results from the direct mechanism of DSB detection since DSBCapture directly captures DSBs *via* ligation of an oligonucleotide to DNA ends, whereas GUIDE-seq relies on NHEJ to integrate a double-stranded oligonucleotide at DSBs during DNA repair. Overall, this supports improved sensitivity and genome-wide coverage for DSB detection by DSBCapture.

DSBCapture and BLESS were compared side-by-side to study endogenous DSBs in two biological replicates of NHEKs. No DNA was recovered in three negative controls (DSBCapture without T4 DNA ligase during the first or second ligation or using a non-biotinylated modified P5 Illumina adapter) (**Supplementary Fig. 1e**). After artificially enhancing BLESS data yield by repeated sequencing to obtain similar read numbers to DSBCapture, we averaged both replicates and compared the data quality of the two methodologies: 69 % of the DSBCapture data passed filtering versus 26 % of the BLESS data (**Supplementary Table 3**). Overall, 84,946 common high confidence peaks were observed for DSBCapture replicates versus 18,816 for BLESS (**Supplementary Fig. 4a**). **Fig. 1c** exemplifies the increased signal of DSBCapture for two representative genes: *MAP2K3* and *MYC*. Moreover, DSBCapture showed higher reproducibility: 83 % overlapping peaks between two independent experiments versus 63 % in BLESS (**Supplementary Fig. 4a**). Of the DSBs identified using BLESS, 99 % were observed by DSBCapture whereas 78 % of DSBs identified by DSBCapture

were missed by BLESS (**Fig. 1d, Supplementary Fig. 4b**). 73 % of the original peaks were also identified when performing DSBCapture with reduced DNA input (20 μg instead of 50 μg), illustrating good overlap irrespective of input material (**Supplementary Fig. 4c**). These results indicate that DSBCapture has improved sensitivity and higher statistical confidence for DSB detection. On comparing the sequence context of DSBs detected uniquely by DSBCapture to those shared between BLESS and DSBCapture we found that DSBCapture detected a significantly greater proportion of DSBs in genomic regions where the GC content surpassed 70 %. When the GC content of DSBs exceeded 80 %, DSBCapture unique peaks were more than 4.5-fold enriched compared to those shared with the BLESS dataset (**Supplementary Fig. 4d**). G-quadruplex DNA secondary structures form in G-rich DNA in human cells[10]. G-quadruplexes have been implicated as fragile sites during transcription and replication[3,11,12], and associate with somatic copy number alterations in human cells[13]. We found that G-quadruplex-sites, previously mapped in genomic DNA (OQs)[13], were 3-fold enriched over random within high confidence DSBCapture peaks (**Supplementary Fig. 4e**), suggesting that G-quadruplexes are intrinsic sites of DSB formation in NHEKs.

We next analysed high confidence DSBCapture data for NHEKs in the context of chromatin features using ChIP-seq and DNase-seq datasets from ENCODE[14]. Notably, 58 % (11.4-fold enrichment) of DSBs localized with histone H2A.Z, a histone variant transiently incorporated at DSBs and known to have a role in DSB repair[15] (**Fig. 2a, Supplementary Fig. 5a** and **Supplementary Table 4**). Over 76 % (33.3-fold enrichment) of the DSBCapture peaks overlapped with regulatory, nucleosome-depleted regions (DNase-seq), revealing a relationship between regulatory chromatin and genome instability (**Fig. 2a** and **Supplementary Table 4**). This is consistent with the notion that nucleosome density mediates the susceptibility of DNA to genotoxic insults, with euchromatin showing enhanced DNA repair signaling[16,17]. DSBs also correlate with markers of active genes (mono-, di- and tri-methylated H3K4); enhancer regions (H3K27ac and H3K4me1; **Supplementary Fig. 5b**), the architectural protein CTCF and the transcription factor P63 (**Fig. 2a** and **Supplementary Table 4**). 38 % (12.2-fold enrichment) of DSBs overlapped with RNA polymerase II (POL2B) peaks, linking DSBs to transcription (**Fig. 2a** and **Supplementary Table 4**). No enrichment of DSBs was found in heterochromatic regions of the genome (H3K9me3 and H3K27me3), or in regions featuring the transcriptional repressor EZH2 (**Fig. 2a** and **Supplementary Table 4**). Transcription has been linked to DNA damage[18] and DSBs have been found to occur in the proximity of transcription start sites (TSSs) of highly expressed genes[19,20]. Indeed, we found DSBs to be enriched in genic compared to intergenic regions (**Supplementary Fig. 5c**), particularly at 5'UTRs (21.5-fold) and promoters (12.8-fold) (**Fig. 2b**). We analyzed the average number of DSBs present at TSSs (± 1 kb) and within gene bodies of genes at different transcriptional levels. Increased gene expression correlated with increased DNA damage around TSSs, whereas damage within gene bodies showed little association with gene expression (**Fig. 2c**). Furthermore, we found that genes that do not contain DSBs at the TSS are generally not expressed (**Supplementary Fig. 5d**). Four of the five most highly expressed genes in NHEKs are keratin genes (*KRT5*, *KRT14*, *KRT6A* and *KRT17*); DSBCapture identified DSBs at the TSS in all of these genes. In fact, DSBCapture identified DSBs at the TSS in 93 % of the most highly

122 expressed genes (top 5 %). Taken together, these data clearly link elevated gene expression and sites of
123 regulatory, nucleosome-depleted chromatin to DSB formation in NHEKs.
124
125 DSBCapture will be a valuable methodology for the study of DNA damage and repair, where
126 it will enable the elucidation of sites of endogenous and drug-induced DNA damage.

Accession code

The data reported in this paper are available at the NCBI's GEO repository, accession number GSE78172 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78172)

Acknowledgements

We thank G. Legube, LBCMCP, Center for Integrative Biology (CBI), Université de Toulouse, Toulouse, France for providing U2OS AID-DIvA cells. We thank the genomic core facility at Cancer Research UK Cambridge Institute. R.H.H. acknowledges EMBO for support (EMBO Long-Term Fellowship to R.H.H.). We acknowledge support from the University of Cambridge and the Cancer Research UK program. The Balasubramanian laboratory is supported by core funding from Cancer Research UK (C14303/A17197 to S.B.) and by an ERC Advanced Grant (S.B.). S.B. is a Senior Investigator of the Wellcome Trust.

Author Contributions

S.V.L. developed the DSBCapture method, conceived the study, conducted experiments, interpreted results and wrote the manuscript. G.M. conceived the study, performed bioinformatics analyses, interpreted results and wrote the manuscript. R.H.H. contributed to the development of the DSBCapture method, conceived the study, conducted experiments, contributed to bioinformatics analyses, interpreted results and wrote the manuscript. E.Y.L. contributed to the development of the DSBCapture method, conceived the study and conducted experiments. D.T. conceived the study, interpreted results and wrote the manuscript. S.B. conceived the study, interpreted results and wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

1 Srivastava, M. & Raghavan, S. C. DNA double-strand break repair inhibitors as cancer therapeutics. *Chem. Biol.* **22**, 17-29 (2015).

2 Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461**, 1071-1078 (2009).

3 Rodriguez, R. *et al.* Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat. Chem. Biol.* **8**, 301-310 (2012).

4 Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187-197 (2015).

5 Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10**, 361-365 (2013).

6 Marchuk, D., Drumm, M., Saulino, A. & Collins, F.S. Construction of T-vectors, a rapid and general sytem for direct cloning of unmodified PCR products. *Nucleic Acids Res.* **19**, 1154 (1990).

7 Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).

8 Mitra, A., Skrzypczak, M., Ginalski, K. & Rowicka, M. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. *PLoS One* **10**, e0120520 (2015).

9 Aymard, F. *et al.* Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nat. Struct. Mol. Biol.* **21**, 366-374 (2014).

10 Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182-186 (2013).

11 Ribeyre, C. *et al.* The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet.* **5**, e1000475 (2009).

12 Paeschke, K., Capra, J. A. & Zakian, V. A. DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell* **145**, 678-691 (2011).

13 Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 877-881 (2015).

14 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

15 Gursoy-Yuzugullu, O., Ayrapetov, M. K. & Price, B. D. Histone chaperone Anp32e removes H2A.Z from DNA double-strand breaks and promotes nucleosome reorganization and DNA repair. *Proc. Natl. Acad. Sci. USA* **112**, 7507-7512 (2015).

16 Storch, K. *et al.* Three-dimensional cell growth confers radioresistance by chromatin density modification. *Cancer Res.* **70**, 3925-3934 (2010).

17 Misteli, T. & Soutoglou, E. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat. Rev. Mol. Cell. Biol.* **10**, 243-254 (2009).

18 Fong, Y. W., Cattoglio, C. & Tjian, R. The intertwined roles of transcription and repair proteins. *Mol. Cell* **52**, 291-302 (2013).

19 Yang, F., Kemp, C. J. & Henikoff, S. Anthracyclines induce double-strand DNA breaks at active gene promoters. *Mutat. Res.* **773**, 9-15 (2015).

20 Schwer, B. *et al.* Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proc. Natl. Acad. Sci. USA* **113**, 2258-2263 (2016).

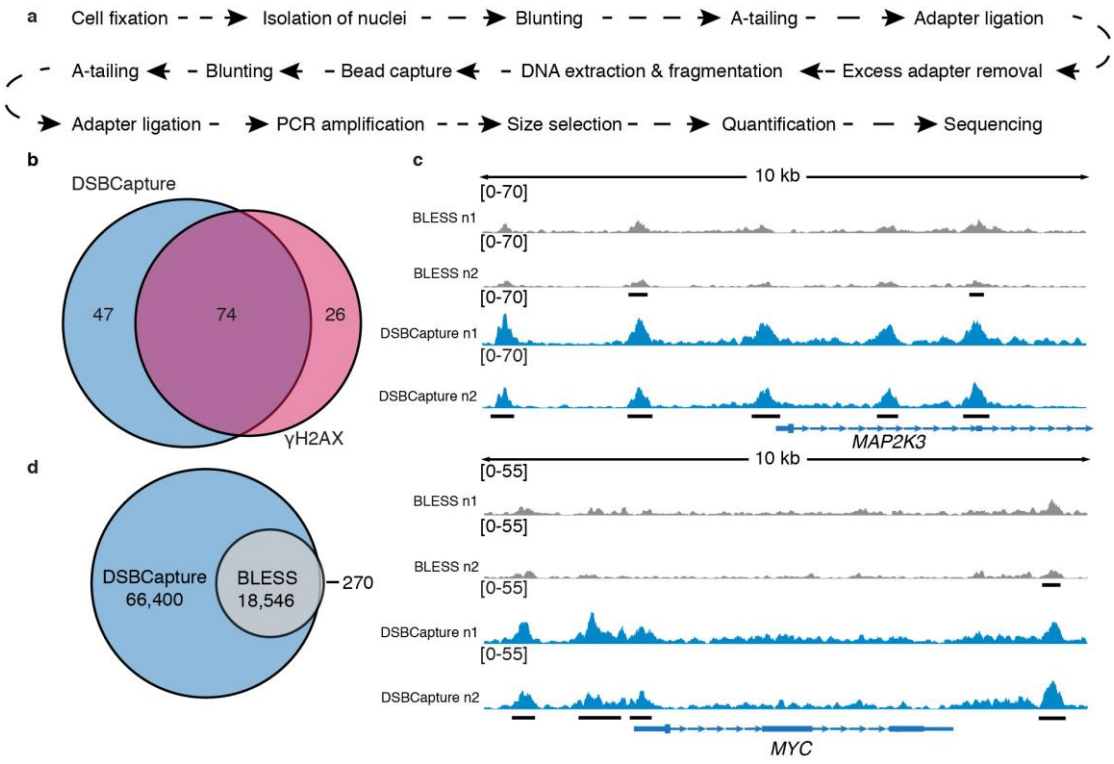


Figure 1 DSBCapture methodology and comparison to BLESS (**a**) DSBCapture workflow. (**b**) Venn diagram illustrating the overlap of DSBs detected at AsiSI sites by DSBCapture and γH2AX ChIP-seq[9]. (**c**) Representative genomic view of DSBs detected by BLESS (grey) and DSBCapture (blue) in two biological replicates. Peaks in common between the two replicates for each method are underlined with black bars; the data range (absolute read counts) is shown in square brackets. Two 10 kb genomic regions in proximity of the *MAP2K3* and *MYC* genes are shown. (**d**) Venn diagram depicting the overlap of high confidence DSBCapture and BLESS peaks.

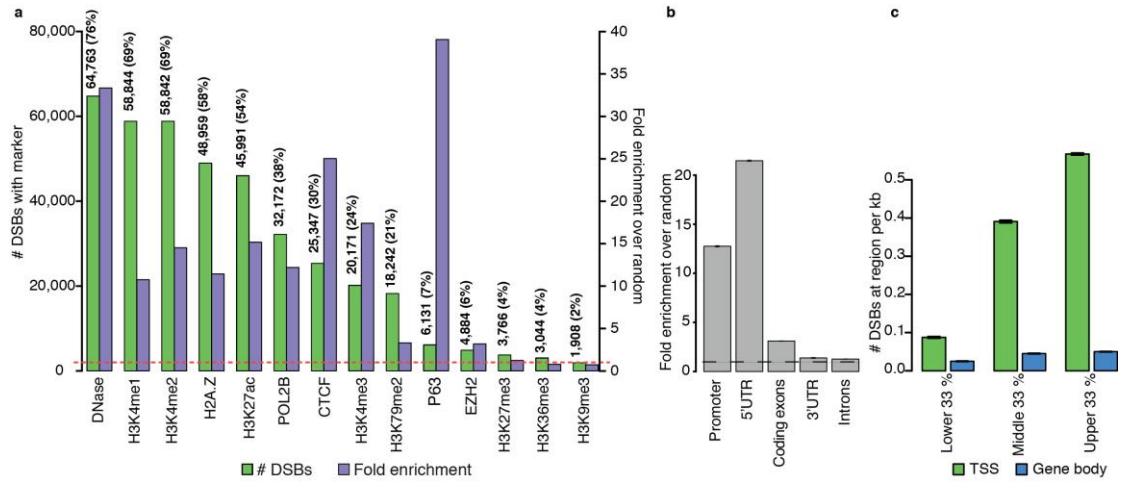


Figure 2 Genomic location and epigenetic context of endogenous DSBs in NHEK cells (**a**) Composite bar plot showing the overlap of high confidence DSBs detected by DSBCapture with epigenetic marks and DNA-binding proteins taken from ENCODE[14]. Green bars: number of DSBs overlapping each mark, ordered from highest (left) to lowest (right). Purple bars: fold change of DSBs overlapping each mark over random, a signal above 1 (dashed red line) is indicative of enrichment. The number and percentages of DSBs that overlap with each mark are shown above their respective bars. (**b**) Fold enrichment of high confidence DSBs in different genomic regions, calculated as the number of peaks divided by the number of randomly shuffled peaks overlapping to each region. Error bars: standard deviation of fold enrichment over random. (**c**) Number of high confidence DSBs detected at TSSs (green bars) or within gene bodies (blue bars) for the lower, middle and upper third of gene expression values, split according to the rpkm values. Error bars: SEM; n = 7,430 for each category.

Online Methods

Cell culture

HeLa cells (CRUK-CI Biorepository and Cell Services Core) were cultured in DMEM (Sigma, D6429) supplemented with 10 % heat inactivated FBS. U2OS AID-DIvA cells were cultured in DMEM supplemented with 10 % heat inactivated FBS and 800 µg/mL G418 Sulfate (Gibco, 10131). To induce nuclear localization of AsiSI, AID-DIvA cells were treated with 300 nM 4OHT (Sigma, H7904) for 4 h. Normal human epidermal keratinocytes (NHEK), pooled from multiple donors, were purchased from Thermofisher (A13401) and cultured in EpiLife medium (Thermofisher, M-EPI-500-CA) supplemented with human keratinocyte growth supplement (HKGS) (Thermofisher, S-001-5). NHEKs were detached using accutase (Sigma, A6964). HeLa and U2OS AID-DIvA cells were STR genotyped and mycoplasma tested. NHEK cells were obtained from Thermofisher and were certified as mycoplasma negative.

BLESS

The method published on http://breakome.utmb.edu was followed; with the exception that DNA was solely fragmented by sonication and amplified using Phusion HF polymerase (NEB, M0530S) as published by Crosetto, *et al*[5]. The TruSeq Nano DNA Library Prep Kit (Illumina, FC-121-4001) was used to generate libraries for sequencing (size selection for 550 bp) and all centrifugation steps were carried out at the speeds described for DSBCapture in the step by step method on the protocol exchange (DOI: 10.1038/protex.2016.52).

Annealing of oligonucleotides

To anneal the modified P5 and P7 Illumina adapters; oligonucleotides (modified P5 and modified P5 complement / modified P7 and modified P7 complement) were made up to 10 µM in 1 × T4 DNA ligase reaction buffer (NEB, B0202S). Oligonucleotides were heated to 95 °C for 10 minutes. Tubes were removed form the heat source and gradually cooled to room temperature.

DSBCapture (EcoRV cleavage in HeLa cells)

During the development of the DSBCapture method, pilot experiments were performed in which permeabilized nuclei were treated with the restriction enzyme EcoRV to generate DSBs. This pilot method was essentially comparable to the final DSBCapture protocol described below with the following significant differences: Cells were fixed in formaldehyde for 10 min. After re-suspension in nucleus break buffer the nuclei were washed 2 × with 1 × NEBuffer 4 (NEB, B7004S) + 1 % TritonX-100 followed by re-suspension in 500 µL 1 × CutSmart buffer (NEB, B7204S). 500 units of EcoRV (NEB, R0195T) were added and the nuclei were incubated over night at 37 °C with shaking at 950 rpm. The 8 min Proteinase K digest was performed at a final concentration of 50 µg/mL. The first blunting step was missed out before A-tailing, as EcoRV is a blunt cutting enzyme. No treatment with Lambda Exonuclease was performed before subsequent DNA isolation. 20 µg of extracted DNA was bound to beads. PCR reactions was performed each using 5 µL beads, 1 µL 20 µM PCR F primer, 1 µL 20 µM

PCR R primer, 10 µL 5 × Phusion HF buffer, 1 µL 10 mM dNTPs and 0.5 µL Phusion HF DNA polymerase (NEB, M0530S) in a final reaction volume of 50 µL with the following cycling parameters: 94 °C 5 min, 94 °C 1 min, 60 °C 45 s, 72 °C 1 min, 72 °C 10 min, 4 °C hold, repeating steps 2-4, for a total of 18 cycles. The DNA was size selected (400-500 bp) by gel electrophoresis and was subsequently extracted using the MinElute gel extraction kit (Qiagen, 28604). The DNA library was subjected to paired-end sequencing on an Illumina MiSeq platform.

DSBCapture (U2OS and NHEK): A detailed protocol can be found on the protocol exchange (DOI: 10.1038/protex.2016.52). Briefly: cells were detached, counted and fixed in complete medium with 2 % formaldehyde (Pierce, 28908) at a density of 1 million cells/1.5 mL for 30 min at room temperature whilst gently rotating. Formaldehyde was quenched by adding glycine to a final concentration of 125 mM. Cells were washed twice in ice cold PBS and lysed in lysis buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.2 % NP40 substitute (Sigma, 74385), cOmplete Roche proteinase inhibitors (REF 11873580001), 1 mM DTT) by gently rotating at 4 °C for 90 min. Nuclei were subsequently re-suspended in nucleus break buffer (10 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.3 % SDS, 1 mM DTT) and incubated at 37 °C for 45 min whilst gently rotating. Nuclei were then re-suspended in 1 × NEBuffer 2 + 0.1 % TritonX-100 (10 mM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9 at 25 °C) and transferred into 2 mL eppendorf tubes (10 million nuclei/tube). Proteinase K (Ambion, AM2546) was added to a final concentration of 100 µg/mL on ice and nuclei were incubated at 37 °C for 8 min before adding an equal volume of 1 × NEBuffer 2 + 0.1 % TritonX-100 + 1:50 PMSF (Sigma, 93482) to kill the reaction. Nuclei were then washed twice in 1 × NEBuffer 2 + 0.1 % TritonX-100 at 4 °C. To repair DNA ends, nuclei were washed once in 1 × Blunting Buffer (NEB, E1201L) + 100 µg/mL BSA (NEB, B9000S) and were blunt-end repaired using the NEB Quick Blunting Kit (NEB, E1201L) + 100 µg/mL BSA in a final volume of 100 µL at 25 °C for 45 min, shaking for 10 s at 800 rpm, every 5 min. Nuclei were then washed three times with 1 × NEBuffer 2 + 0.1 % TritonX-100 at 4 °C and were subsequently A-tailed using Klenow Fragment 3'-5' exo- (NEB, M0212L) and dATP (Promega, U120D) in a final volume of 50 µL at 37 °C for 45 min, shaking for 10 s at 800 rpm, every 10 min. Following A-tailing, nuclei were washed 3 × with 1 × NEBuffer 2 + TritonX-100 at 4 °C, once with T4 DNA ligase buffer (NEB, B0202S) + 0.1 % TritonX-100 at 4 °C and once with T4 DNA ligase buffer at 4 °C. The modified P5 Illumina adapter, previously annealed in 1 × T4 DNA ligase buffer, was then ligated to DNA ends using T4 DNA ligase (NEB, M0202M) for 15-20 h at 16 °C in a final reaction volume of 50 µL whilst shaking at 350 rpm for 15 s every 45 min during the ligation. To remove excess adapters, nuclei were washed 2 × in WB buffer + 0.1 % TritonX-100 (5 mM Tris-HCl pH 7.5, 1 mM EDTA, 1 M NaCl) and once in 1 × Lambda Exonuclease Reaction Buffer (NEB, M0262L) and were then treated with 50 units Lambda Exonuclease (NEB, M0262L) in a final reaction volume of 50 µL for 30 min at 37 °C. Following washing in 1 × NEBuffer 2 + 0.1 % TritonX-100 nuclei were then re-suspended in 1 × NEBuffer 2 + 0.5 % TritonX-100 at 4 °C. To extract genomic DNA, nuclei were lysed with Proteinase K (200 µg/ml) for 30 min at 55 °C, shaking at 800 rpm followed by 30 min at 65 °C, shaking at 800 rpm and the DNA was precipitated using isopropanol. The DNA was then re-suspended in 90 µL H_20

and incubated at 55 °C shaking at 800 rpm for 1 h. The isolated DNA was fragmented by sonication to obtain an average fragment size of 200-500 bp and then 50 µg (20 µg only for data depicted in **Supplementary Fig. 4c**) of DNA was bound to 5 µl MyOne strepdavidin C1 Dynabeads (Invitrogen, 65001) at 4 °C for 45 min, whilst gently rotating. Beads were washed 3 × with WB buffer + 0.1 % TritonX-100 and captured DNA was blunt end-repaired using the Quick Blunting Kit in a final reaction volume of 50 µL for 45 minutes at 25 °C, shaking at 800 rpm every 5 min for 10 s. The beads were then washed 3 × in WB buffer + 0.1 % TritonX-100 and DNA ends were once more A-tailed using Klenow Fragment 3'-5' exo- in a final reaction volume of 25 µL at 37 °C for 45 min, shaking at 800 rpm for 10 s every 10 min. After washing beads 3 × with WB buffer + 0.1 % TritonX-100, the modified P7 Illumina adapter, previously annealed in 1 × T4 DNA ligase buffer, was ligated to DNA ends in a final reaction volume of 50 µL for 15-20 h at 16 °C, every 45 minutes samples were mixed for 1 min at 1200 rpm. Beads were then washed 3 × in WB buffer + 0.1 % TritonX-100 before re-suspending in 25 µL nuclease free water. The DNA DSBCapture library was then amplified by PCR with PCR F and PCR R primers and NEB Next PCR mix for 15 cycles, following the manufacturers recommendations (annealing temperature: 65 °C; NEB, M0541L). The amplified DNA library was then purified using a MinElute PCR Purification kit (Qiagen, 28004) and the library was size-selected (250-1200 bp) using a BluePippin (Sage Science). Samples were run on the Bioanalyzer (Agilent, 5067-4626) to determine the average library size, and the library was quantified using the KAPA library quantification kit (Kapa Biosystems, kk4824), following the manufacturers recommendations. Libraries were subsequently sequenced paired-end on an Illumina NextSeq 500 platform.

RNA-seq

Total RNA for RNA-seq experiments was extracted using the RNeasy kit (Qiagen, cat. no. 74104), following the manufacturer's instructions. RNA-seq libraries were generated using the Illumina Truseq RNA HT (stranded mRNA) kit (RS-122-2103).

Epigenome analysis

The ENCODE project NHEK epigenome ChIP-seq data[14] was retrieved from the NCBI's GEO repository as follows: H3K4me3 (GSM945175); H3K4me2 (GSM733686); H3K27ac (GSM733674); H3K4me1 (GSM733698); H3K79me2 (GSM1003527); H3K36me3 (GSM945174); H2A.Z (GSM1003488); POLR2B (GSM733671); H3K9me3 (GSM1003528); H3K27me3 (GSM733701); EZH2 (GSM1003489); CTCF (GSM822271); TP63 (GSE32061), DNase-seq (GSM736556).

Sequencing

For the EcoRV enzyme cleavage experiment, 50 bp paired-end sequencing was conducted on the Illumina MiSeq instrument. For the AID-DIvA U2OS and NHEK cell experiments, 75 bp paired-end sequencing was performed on the Illumina NextSeq 500 instrument. Despite DSB detection with DSBCapture requiring only single-end sequencing, we sequenced all DSBCapture libraries in paired-end mode in order to be able to distinguish DSBs at restriction sites from PCR duplicates and to be able to compare our method to BLESS in NHEK cells.

Sequencing analysis

DSBCapture libraries: fastq files containing sequencing reads were pre-processed to remove the Illumina adapters and to trim low quality tails using trim_galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/); reads were aligned to the human reference genome (*hg19*) using the bwa mem aligner (http://sourceforge.net/projects/bio-bwa/files/) and cleaned for low quality alignments (mapQ < 10) using samtools (clean reads) (http://samtools.sourceforge.net). Duplicates were identified using picard tools (https://github.com/broadinstitute/picard/) and removed (unduplicated reads). Only reads from read 1, which are proximal to the break site, were retained for downstream analysis and are shown in the coverage plots.

BLESS libraries: when performing the Illumina adapter trimming, the BLESS linker sequences are trimmed (TCGAGGTAGTA and TCGAGACGACG for proximal and distal linkers, respectively), and only read pairs containing both linkers are retained. Trimmed fastq files then undergo the same processing pipeline as for DSBCapture libraries. Finally, only sequencing reads that originally contained the proximal linker (originating either from read 1 or read 2) are retained for subsequent analysis and are shown in the coverage plots.

Peak calling analysis

Regions with enriched reads over the background were called as peaks using the MACS2 software (https://github.com/taoliu/MACS/) using default parameters for the statistical threshold (i.e., corrected p-value $q \leq 0.05$) and the options “no-model” and “no control”. For the NHEK cell analysis, peaks where called independently for each replicate on the clean bam files containing only the proximal linker, and the two peaks files were then intersected using the bedtools package (http://bedtools.readthedocs.org/) to generate the high-confidence peak files for both BLESS and DSBCapture. The sample size was constrained by experimental considerations. For the endogenous DSBs mapping in primary NHEK cells, we observed good reproducibility with 2 biological replicates for both BLESS and DSBCapture.

Comparison of DSBs identified by DSBCapture with 50 µg and 20 µg of input material

DSBCapture was performed using 50 µg and 20 µg of input material from the same biological sample ($n = 1$). The two DSBCapture experiments were processed in the same way, as previously described in the paragraphs "Sequencing analysis" and "Peak calling analysis". The 20 µg library was sub-sampled in order to have the same number of unduplicated, proximal reads as the 50 µg library, i.e. 124 million reads. Peak calling was performed on the subsampled library. Peak intersection was calculated by using the bedtools, as previously explained.

GC content analysis of the EcoRV data set

The 430,897 predicted EcoRV sites were split into two subsets: 403,905 (94 % of total) sites displaying a coverage of at least 5 reads, marked as detected, and the remaining 26,992 sites (6 % of the total)

displaying coverage below 5, marked as not detected. The GC content around the EcoRV restriction site was calculated, by considering 100 base pairs either side of the restriction site.

Analysis of EcoRV restriction sites at DNase hypersensitivity sites

DNase hypersensitivity sites in HeLa cells were downloaded from the ENCODE project (https://www.encodeproject.org/; GEO sample accession numbers GSM736564 and GSM736510), high confidence sites were calculated as the intersection of the genomic regions from the two replicates (bedtools intersect), resulting in 96,541 DNase sites. The overlap of the 430,897 predicted EcoRV sites and the cleaved subset (403,905; coverage $\geq$ 5 reads) with the high confidence DNase sites was calculated. The percentage of total and cleaved sites within DNase sites is reported. The non-parametric Chi-squared test for proportions was used for statistical testing.

AsiSI-induced DSB analysis in AID-DIvA U2OS cells

AsiSI restriction sites were identified as DSBs by peak calling as described above. The list of 100 most cleaved AsiSI sites identified by γH2AX ChIP-seq signal, and the RAD51 and XRCC4 ChIP-seq data, were obtained from Aymard, *et al*[9]. Peaks on RAD51 and XRCC4 were called using the MACS2 peak caller with default parameters ($q \leq 0.05$) using the respective input files.

Comparison of DSBCapture to GUIDE-seq

The 25 endogenous DSB hotspots identified in U2OS cells by GUIDE-seq[4] were compared (bedtools intersect) to the 2,372 endogenous DSBCapture peaks identified by MACS2 in AID-DIvA U2OS cells.

GC content and G-quadruplex analysis

The high confidence (~ 85,000) peaks detected in both replicates of DSBCapture were split into two subsets: DSBs uniquely identified by DSBCapture (~ 66,000 peaks) and peaks common to BLESS (~ 19,000). GC content was measured within the peak regions and each peak was assigned to a GC % category. Independently, for both the DSBCapture unique and BLESS-DSBCapture common peak subsets, the number of peaks in each GC % category were divided by the total number of peaks in the subset, resulting in a fraction of peaks within each GC % category for both subsets. The two fractions of each GC % category were divided one by each other to obtain the fold enrichment of DSBCapture unique peaks over the BLESS-DSBCapture common peaks. Values > 1 indicate that a given GC % category is enriched among the DSBCapture unique peaks, or otherwise is depleted in the BLESS-DSBCapture common peaks. Similarly for the G-quadruplex analysis, the fraction of reads from the DSBCapture high confidence peaks overlapping to observed G-quadruplex-forming sequences (OQs[13]) was calculated for each GC % content group and compared to random (fold enrichment). The random sets were generated by random shuffling of DSB genomic intervals throughout the entire genome and calculating the random overlap to OQs (n = 3).

Comparison to the ENCODE data sets in NHEK cells

Data sets for epigenetic marks (histone modifications), DNase hypersensitivity sites (DNase) and DNA binding proteins (CTCF, P63, POL2B) were downloaded from ENCODE (see Epigenome analysis, Methods). The number of high confidence DSBCapture peaks overlapping each mark (command intersectBed of the bedtools package) was calculated for each mark (independently of all other marks) and visualized on a bar plot (green bars in **Fig. 2a** and **Supplementary Fig. 5b**). Additionally, the DSBCapture peak files were randomly shuffled across the genome (command shuffleBed of the bedtools package) three independent times and the overlap of the random sets with each mark was calculated. The ratio of the DSB overlap divided by the average (n = 3) random overlap for each mark was computed (fold enrichment over random) and visualized (purple bars in **Fig. 2a** and **Supplementary Fig. 5b**). Fold enrichment values > 1 indicates that a mark overlaps to DSBs more often than random. Standard deviations from the 3 independent randomizations are not indicated in the bar plot as they lay in the range 0.004-0.12 for all marks, and are therefore negligible when compared to the fold change values.

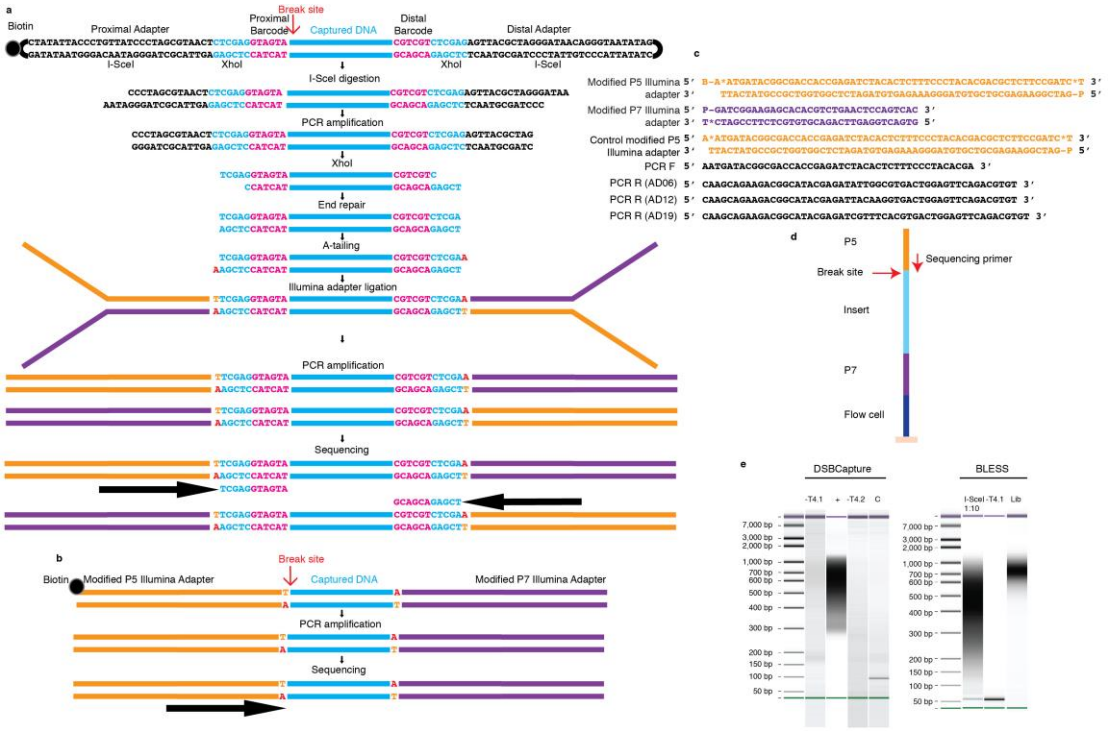
Genomic regions analysis

Gene annotation files for the human genome (*hg19*) were downloaded from the Illumina iGenomes support website (https://support.illumina.com/sequencing/sequencing_software/igenome.html, and different gene features were calculated as follows: Promoter = 1 kb upstream of the transcription start site (TSS); 5'UTR = sequence from the TSS to the annotated translation start codon; Coding exons = all exons that are translated (coding sequences) from the start until the stop codon; 3'UTR = sequence from the annotated end of translation (stop codon) to the end of the last exon; introns = all the regions spliced out during transcript processing. The number of high confidence DSB sites overlapping each region was calculated by intersecting DSBCapture high confidence peaks to each feature separately. The DSBCapture peak files were then randomly shuffled across the genome (command shuffleBed of the bedtools package) three independent times and the overlap of the random sets to each feature was calculated. The ratio of the DSB overlap divided by the average (n = 3) random overlap for each region was calculated (fold enrichment over random). Similarly, the fold enrichment over random for genic regions (i.e. within gene bodies) versus intergenic features was assessed.

Gene expression analysis

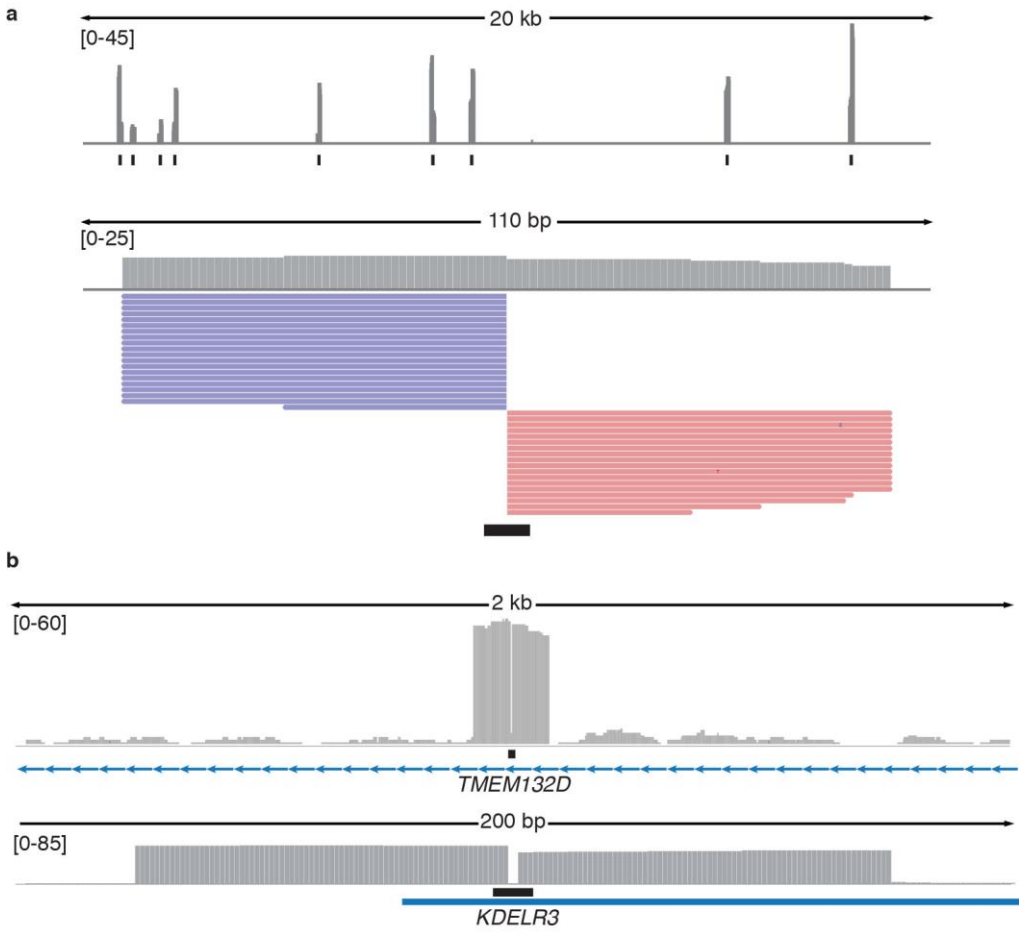
RNA-seq sequencing reads were pre-processed by the trim galore software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) for removal of Illumina sequencing adapters and low quality read tails. Trimmed data were then aligned to the human reference genome (*hg19*) using tophat2 (https://ccb.jhu.edu/software/tophat/index.shtml/). Reads overlapping unambiguously to each gene feature were assessed and counted by the htseq-count (http://www-huber.embl.de/users/anders/HTSeq/). Gene expression values were split into three groups containing the same number of genes (lower, middle and upper third) according to their rpkm gene expression values (rpkm < 0.041; rpkm 0.041-6.08 and rpkm > 6.08), where rpkm is the reads per kilo base (of the total exon length) per million sequenced reads. The average number of high confidence DSBs per kb overlapping ± 1 kb of the TSS (green bars in **Fig. 2c**) and to gene bodies, i.e. all exons and introns

463 excluding ± 1 kb around the TSS (Gene body, blue bars in **Fig. 2c**), were calculated. Genes with length
464 less than 2 kb were excluded from the gene body category. Additionally, all genes were split into two
465 categories according to the presence (n = 12,984) / absence (n = 9,272) of DSBs at ± 1 kb around the
466 TSS and the rpkm for all genes in the two groups was inspected and visualized as box plots.
467
468 **Code availability**
469 The computer code used to analyze all data in this work, including sequencing processing and data
470 comparisons, can be obtained on request by contacting the corresponding author.
471



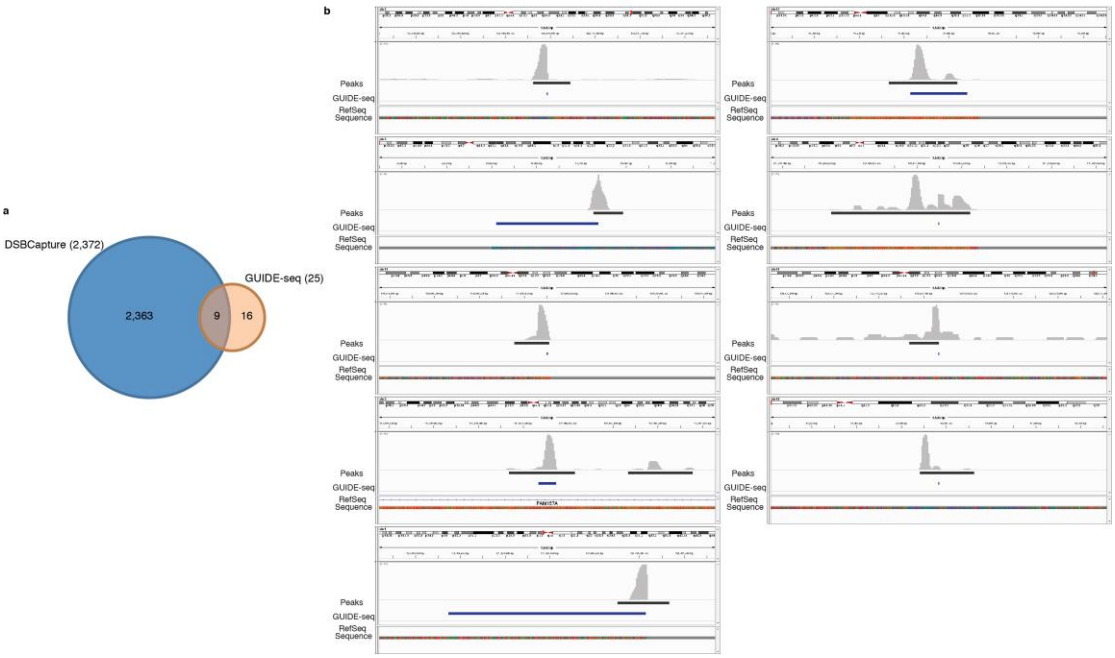
Supplementary Figure 1

DNA processing workflows, adapter sequences and controls. (**a**) BLESS DNA processing workflow following the ligation of both proximal and distal adapters. DNA is digested by I-SceI, PCR amplified, digested by XhoI and subsequently subjected to Illumina library preparation, consisting of end repair, size selection (not shown), A-tailing, Illumina adapter ligation and PCR amplification. Large black arrows indicate the site at which sequencing is initiated; the first 11 bases sequenced are shown. (**b**) DSBCapture DNA processing workflow following the ligation of both modified P5 and P7 Illumina adapters. The DNA is PCR amplified, size selected (not shown), and sequenced. Large black arrow indicates the site at which sequencing is initiated; the first base sequenced is the site of *in situ* DSB formation. (**c**) Sequences of the modified P5, modified P7 and control modified P5 Illumina adapters as well as DSBCapture PCR primers (forward (PCR F) and reverse (PCR R)). AD identifies the Illumina adapter barcode sequence, three example reverse primers are shown; further primers can be created by substituting the barcode sequence. B = biotin; P = phosphorylated; * = phosphorothioate bond. (**d**) Orientation of the DSBCapture library on the Illumina flow cell. The first sequencing primer has complementarity to the P5 Illumina adapter and therefore sequencing is initiated from the P5 end. The ligation of the modified P5 Illumina adapter to the DSB *in situ* enables direct sequencing of the break site in single-end sequencing. The first base sequenced directly identifies the DSB. (**e**) Bioanalyser profiles of the DNA products from DSBCapture and BLESS NHEK libraries. DSBCapture: no product is present in the controls performed without T4 DNA ligase during the first (-T4.1) or second (-T4.2) ligation reactions, or in the control performed with the non-biotinylated control modified P5 lllumina adapter (C). A DSBCapture library was only generated when the complete procedure was carried out (+). BLESS: No product is present in the control performed without T4 DNA ligase during the first ligation reaction (-T4.1). The product of BLESS is shown before Illumina library preparation (I-SceI; diluted 1:10) and after Illumina library preparation (Lib).



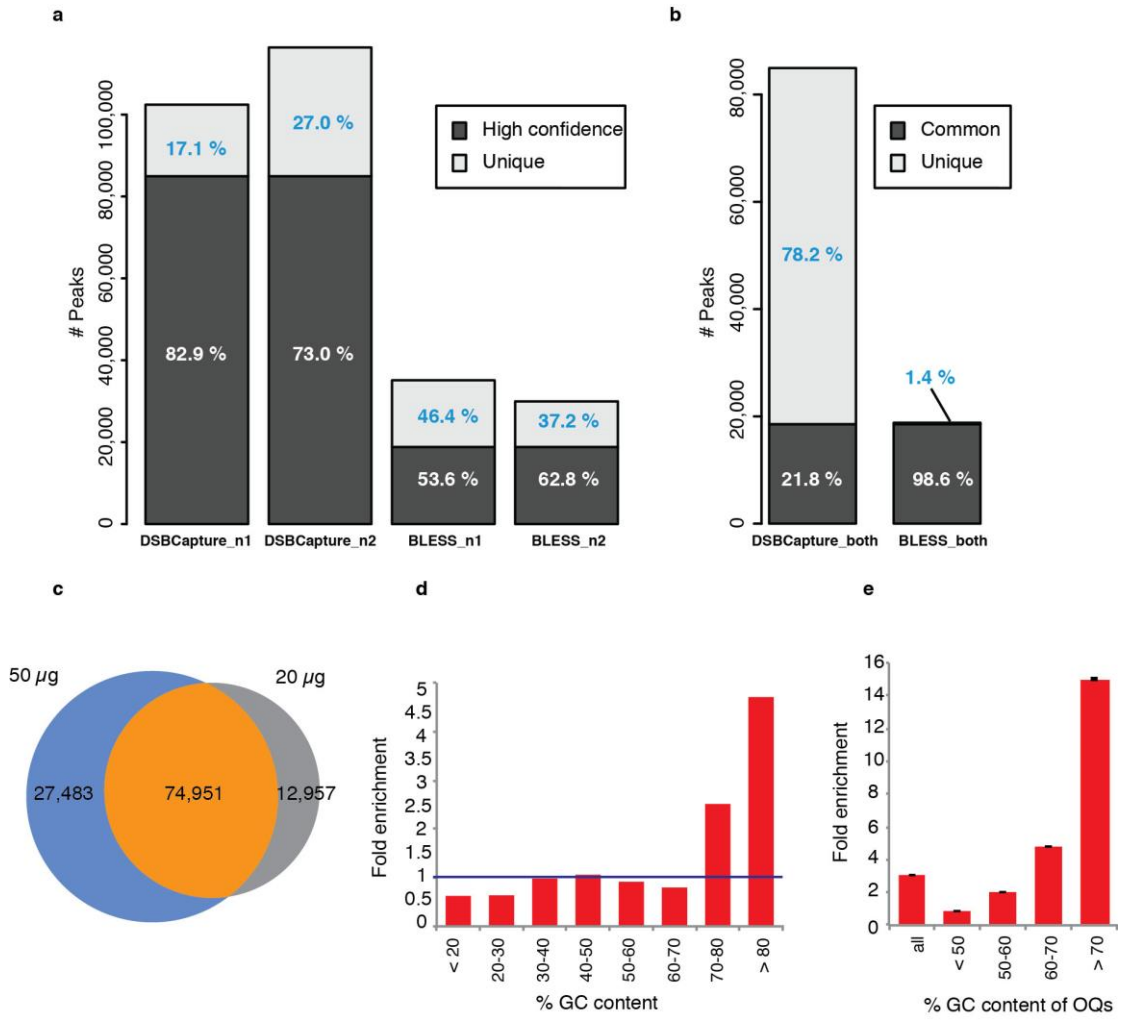
Supplementary Figure 2

DSBs mapped by DSBCapture at EcoRV and AsiSI restriction sites. (**a**) DSBs created by EcoRV cleavage in fixed nuclei, mapped by DSBCapture (n = 1). PCR duplicates have been removed. Data range is shown in square brackets and black boxes illustrate the genomic location of EcoRV sites. A 20 kb region and a 110 bp region are shown. Pink and purple lines: reads from the sense and antisense strand, respectively. As EcoRV is a blunt cutter, reads originate directly from the cleavage site. (**b**) AsiSI cleavage sites (black boxes) detected by DSBCapture (n = 1). Cleavage by AsiSI generates a 2 bp 3' overhang; end processing removes this overhang generating the 2 bp gap in the center of the peak. A 2 kb and a 200 bp region are shown.



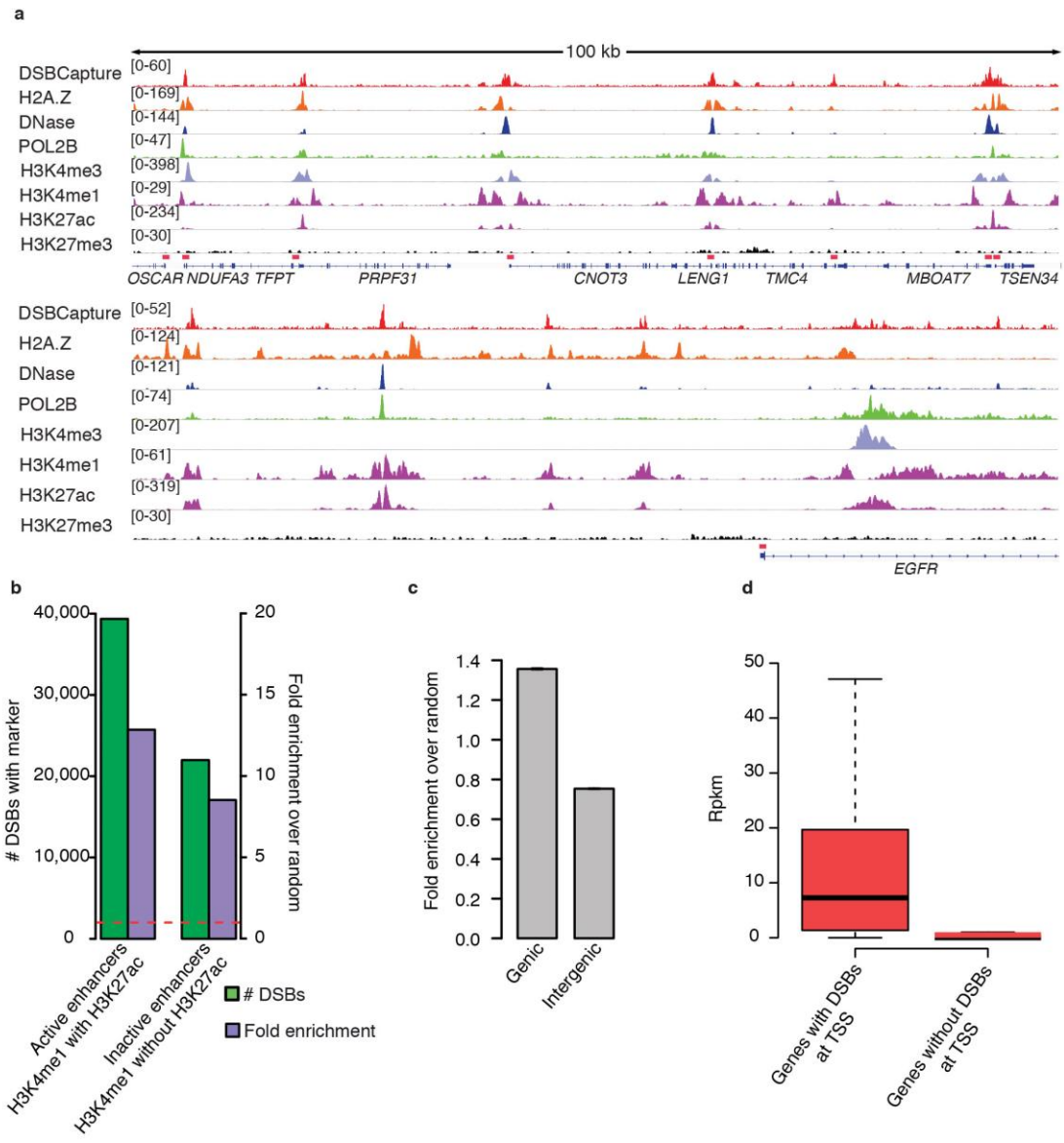
Supplementary Figure 3

Overlap of DSBs detected by DSBCapture and GUIDE-seq in U2OS cells. (**a**) Venn diagram showing the overlap between the DSBCapture peaks and the 25 sites detected by GUIDE-seq[4]. (**b**) Genomic tracts showing the 9 DSBs detected by GUIDE-seq that are also detected by DSBCapture. Each panel shows a genomic view of 2,000 bp around the GUIDE-seq detected DSB hotspot. In each panel, from top to bottom: DSBCapture coverage (grey track); peaks detected in DSBCapture by peak calling (black track); GUIDE-seq sites (dark blue track); RefSeq gene track; reference genome sequence (*hg19*).



Supplementary Figure 4

Analysis of DSBs detected by DSBCapture and BLESS. (**a**) Overlap of peaks between two biological replicate experiments for DSBCapture and BLESS. Peaks called in both replicates (high confidence peaks) were used for data analysis. 84,946 and 18,816 high confidence peaks were identified in DSBCapture and BLESS, respectively. (**b**) Overlap between the high confidence peaks (shown in **a**) from the BLESS and DSBCapture experiments. The vast majority (98.6 %) of the BLESS peaks are also identified by DSBCapture, whereas 78.2 % of DSBCapture peaks are unique to this method. (**c**) Venn diagram showing the overlap of DSBs detected as peaks by DSBCapture performed with 50 µg and 20 µg input material. 74,951 peaks are commonly identified by the two conditions (n = 1). (**d**) Fraction of DSBs with different GC content in the DSBCapture unique peaks divided by the fraction of peaks shared between BLESS and DSBCapture within the same GC content range (fold enrichment). A fold change greater than one represents an increase in DSBs with that particular GC content in the DSBCapture unique peaks. (**e**) Fold enrichment of DSBCapture peaks with OQs[13], calculated as the number of DSBs overlapping to OQs at each indicated % GC sequence content category (x-axis labels) divided by random overlap. All = all 716,311 OQs, irrespectively of GC content; error bars: standard deviation of the fold enrichment over random.



Supplementary Figure 5

Correlation of DSBs with chromatin marks, genic regions and transcription. (**a**) Genomic location of DSBs detected by DSBCapture with respect to histone marks H2A.Z, H3K4me3, H3K4me1, H3K27ac, H3K27me3 as well as DNase and POL2B. Two 100 kb genomic regions are shown; upper: a gene dense region on chromosome 19; lower: a region upstream of the *EGFR* gene. The data range is shown in square brackets, 5'UTRs are highlighted with red boxes. (**b**) Number and fold enrichment of DSBs at active and inactive enhancers. (**c**) Fold enrichment of DSBs in genic and intergenic regions over random. Values > 1 indicate that DSBs are preferentially found within that genomic location. Error bars: standard deviation of the fold enrichment over random. (**d**) Gene expression values measured as rpkm for genes with (left box) or without (right box) DSBs within ± 1 kb of the TSS. Boxes span from the 25th to the 75th percentile with the median marked by a solid bar. All whiskers extend from the 5th to the 95th percentile.

Supplementary Table 1. EcoRV (HeLa) DSBCapture sequencing data. Alignment statistics for a paired-end MiSeq sequencing run of the EcoRV DSBCapture library.

EcoRV (HeLa) DSBCapture experiment	Number
Reads sequenced	43,794,614
Clean reads aligned	31,267,705
Clean reads from read 1	16,366,684
Unduplicated clean reads from read 1	13,449,911
Unduplicated clean reads from read 1 (forward strand)	6,718,566
Unduplicated clean reads from read 1 (reverse strand)	6,731,345

Supplementary Table 2. AID-DIvA U2OS DSBCapture sequencing data. Alignment statistics for a paired-end NextSeq sequencing run of the AID-DIvA U2OS DSBCapture library. The % of proximal reads used was calculated as the number of unduplicated proximal reads divided by the total number of proximal reads sequenced (total number of proximal reads = # Reads/2).

Library	# Reads	# Clean Aligned Reads	# Unduplicated Reads	% Duplication	# Unduplicated Proximal Reads	% Proximal Reads Used
U2OS DSB-Capture	267,536,292	230,074,135	83,734,433	63.6	37,347,231	27.9

Supplementary Table 3. NHEK DSBCapture and BLESS sequencing data. Combined data from multiple paired-end NextSeq sequencing runs of the NHEK BLESS and DSBCapture libraries, where two independent biological replicates were performed (N1 and N2). The % of proximal reads used was calculated as the number of unduplicated proximal reads divided by the total number of proximal reads sequenced (total number of proximal reads = # Reads/2).

Library	# Reads	# Clean Aligned Reads	% Clean Reads	# Unduplicated Reads	% Duplication	# Unduplicated Proximal Reads	% Proximal Reads Used
DSB-Capture N1	353,056,932	314,223,106	89.0	256,152,484	18.5	124,192,457	70.4
DSB-Capture N2	229,658,364	204,047,616	88.8	160,803,701	21.2	77,523,430	67.5
BLESS N1	324,872,812	228,953,681	70.5	147,477,098	35.6	71,305,629	43.9
BLESS N2	585,674,924	317,449,269	54.2	51,725,762	83.7	24,766,932	8.5

Supplementary Table 4. Overlap of high confidence DSBCapture peaks in NHEKs with various marks.

Mark	# DSBs overlapping mark	Fold enrichment of DSB overlapping mark/random	% DSBs overlapping mark
DNase	64,763	33.3	76.2
H3K4me1	58,844	10.7	69.3
H3K4me2	58,842	14.5	69.3
H2AZ	48,959	11.4	57.6
H3K27ac	45,991	15.2	54.1
POL2B	32,172	12.2	37.9
CTCF	25,347	25.0	29.8
H3K4me3	20,171	17.4	23.7
H3K79me2	18,242	3.3	21.5
P63	6,131	39.1	7.2
EZH2	4,884	3.2	5.7
H3K27me3	3,766	1.2	4.4
H3K36me3	3,044	0.8	3.6
H3K9me3	1,908	0.7	2.2