Quantitative DSB sequencing (qDSB-Seq): a method for genome-wide accurate estimation of absolute DNA double-strand break frequencies per cell

Yingjie Zhu[1,9], Norbert Dojer[1,2,9], Anna Biernacka[3,9], Benjamin Pardo[4], Romain Forey[4], Magdalena Skrzypczak[3], Bernard Fongang[1], Jules Nde[1], Raziyeh Yousefi[1], Philippe Pasero[4], Krzysztof Ginalski[3,8] and Maga Rowicka[1,6,7,8]

[1] Department of Biochemistry and Molecular Biology, University of Texas Medical Branch at Galveston, Galveston, Texas, USA.

[2] Institute of Informatics, University of Warsaw, Warsaw, Poland.

[3] Laboratory of Bioinformatics and Systems Biology, Centre of New Technologies, University of Warsaw, Warsaw, Poland.

[4] Institute of Human Genetics, Montpellier, France.

[5] Institute for Translational Sciences, University of Texas Medical Branch at Galveston, Galveston, Texas, USA.

[6] Sealy Center for Molecular Medicine, University of Texas Medical Branch at Galveston, Galveston, Texas, USA.

[7] Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch at Galveston, Galveston, Texas, USA.

[8] Correspondence should be addressed to M.R. (Maga.Rowicka@utmb.edu) or K.G. (K.Ginalski@cent.uw.edu.pl)

[9] These authors contributed equally

Short title: Precise genome-wide estimation of DNA break frequencies per cell

Keywords: qDSB, quantitative DSB sequencing, DNA double-strand break, DSB, sequencing, BLESS, computational analysis, genome instability, Mec1, hydroxyurea, DNA replication, replication stress, Zeocin, NotI, I-SceI, fork barriers, rDNA

Abstract

Emerging genome-wide methods for mapping DNA double-strand breaks (DSBs) by sequencing (e.g. BLESS) are limited to measuring relative frequencies of breaks between loci. Knowing the absolute DSB frequency per cell, however, is key to understanding their physiological relevance. Here, we propose quantitative DSB sequencing (qDSB-Seq), a method to infer the absolute DSB frequency genome-wide. qDSB-Seq relies on inducing spike-in DSBs by a site-specific endonuclease and estimating the efficiency of the endonuclease cleavage by sequencing or PCR. This spike-in frequency is used to quantify DSB sequencing data. We present validation of the qDSB-Seq method and results of its application. We quantify DSBs resulting from replication stress and the collapse of replication forks on natural fork barriers in the ribosomal DNA. The qDSB-Seq approach can be used with any DSB sequencing method and allows accurate comparisons of absolute DSB frequencies across samples and precise quantification of the impact of various DSB-causing agents.

Introduction

DNA double-strand breaks (DSBs) underlie genomic instability that drives cancer development[1] and are also induced during physiological processes[2]. Moreover, DSBs are created by CRISPR/Cas system, the most promising method for genome editing[3]. There is consequently a tremendous interest in measuring DSBs precisely genome-wide. Starting with the BLESS method[4] developed in 2013, several high-resolution and direct methods to label DSBs have been developed[5-7]. To judge the physiological relevance of observed DSBs, though, knowing their frequency (DSB number per cell) is crucial. So far, there is no method to assess such absolute break frequency, other than the inaccurate marking using the phosphorylated histone variant H2A.X (γH2A.X). γH2A.X labeling is known to lead to false positives since γH2A.X also labels single-strand DNA breaks and inactive chromosome X[8-10]; sometimes even pan-nuclear γH2A.X is observed[11]. The lack of an accurate method to assess DSB frequency limits our understanding of the physiological relevance of observed DSBs and hinders between-sample comparisons. Here, we propose a general, yet simple method for accurate quantification of average DSB number per cell using sequencing.

Quantitative DSB sequencing (qDSB-Seq) method relies on inducing spike-in DSBs with known or easy-to-determine frequency, which is later used to normalize and quantify DSB sequencing data. Such spike-in DSBs can be variously implemented. We recommend supplementing a DSB-labeling protocol with a gentle restriction digestion (**Fig. 1**), thus introducing spike-in breaks in predefined genomic loci. A cell system in which a restriction enzyme is expressed endogenously, such as DivA cells[12], can also be used or, cells engineered to express a restriction enzyme and to contain a cassette encompassing its recognition site, such as I-SceI strain, discussed below. Irrespective of the system used and of the manner of inducing spike-in DSBs (*in vivo* or *in vitro)*, the next step is paired-end (PE) sequencing of genomic DNA (gDNA) or quantitative PCR (qPCR). These methods are used to calculate absolute frequency (DSB per cell) of spike-in DSBs which is then used to normalize the DSB sequencing signal (e.g. from BLESS or i-BLESS data[4, 13]) (**Fig. 1** and **Methods**). For a restriction enzyme with a single recognition site in the genome, after mapping PE gDNA fragments, spike-in break frequency is estimated by calculating the ratios of fragments covering the cut site to those ending at that location (**Methods**). Next, background related to fragmentation preceding sequencing is estimated and subtracted (**Methods**). For a restriction enzyme with multiple recognition sites in a given genome, all those sites are aligned, and then the same estimation of overall cutting frequency is performed and aggregate background frequency subtracted. We also discuss correcting for sequencing bias in **Methods**.

Results

Validation of qDSB-Seq approach

Estimation of an enzyme cutting frequency. First, we validated our estimation of cutting frequency of the I-SceI restriction enzyme (calculated from gDNA data using the qDSB-Seq method (**Eq. 1**)), with the cutting frequency obtained from a semi-quantitative polymerase chain reaction (sqPCR) (**Fig. 2a)**. To this end, we used a strain engineered to

carry a single I-SceI recognition site. The cutting efficiency calculated by qDSB-Seq was 0.74, while sqPCR estimation was 0.79 (σ=0.03, 95% confidence interval 0.73-0.85) (**Fig. S1)**. We also performed another biological replicate with lower enzyme cutting frequency and validation using Southern blotting. In this experiment, cutting efficiency estimated by qDSB-Seq and Southern blotting were consistent: 0.38±0.03 and 0.3±0.1, respectively. We also tested the repeatability of the results in NotI-treated samples (39 cutting sites in the yeast genome). The standard deviation of the cutting frequencies of the enzyme tested was ~0.01 for technical replicates, and 0.02-0.03 for samples from different biological conditions that could affect local sequencing yield, including cell cycle phase. Based on these validation experiments, we conclude that qDSB-Seq yields accurate estimates of cutting frequency (**Methods**).

qDSB-Seq based on spike-in (Fig. 2b). The principle of spike-in is that a sample with known parameters (spike-in) is mixed with the sample being measured. In our case, a spike-in with known DSB frequency per cell is needed. Therefore, we mixed wild type cells with cells engineered to express I-SceI endonuclease and carry a cassette with a single I-SceI cutting site per genome (**Fig. 2b**). We employed a CASY Cell Counter (Roche Applied Science) for accurate cell counting to ensure that 2% of the mixed cells carried the I-SceI cassette. The frequency of I-SceI cutting in the engineered cells was determined by our formula (**Methods Eq. 1)** and independently by qPCR, as described above, both yielding similar results (qPCR: 0.79 and qDSB-Seq: 0.74). The 74% cutting efficiency obtained from qDSB-Seq was used for further calculations (**Methods Eq. 2-4**). After accounting for the 2% dilution, the final result was that I-SceI induced 0.015 DSB per cell. This estimated cutting frequency of I-SceI and the total number of I-SceI sequencing reads were used to convert the number of DSB-related sequencing reads into the number of DSBs per cell (**Methods Eq. 1-4**).

qDSB-Seq based on restriction digestion (Fig. 2c). Here, we used NotI restriction enzyme, but any other restriction enzyme that would cut the genome of the chosen organism in at least one place can be utilized (see **Methods** for a discussion of optimal restriction enzyme selection). The sequence recognized by NotI restriction enzyme, 5'-GCGGCCGC-3', occurs 39 times in the genome of the W303 budding yeast strain we used. We estimated NotI cutting frequency using the formula in **Eq. 1,** validated experimentally, as described above. In the WT HU sample, NotI cutting frequency was estimated to be 0.63±0.02. Knowing the cutting frequency of NotI in the sample, we calculated the read/DSB ratio in the same manner as was done for I-SceI. Next, as for I-SceI, we counted the total number of DSB-related sequencing reads and estimated the total number of breaks in the NotI experiment (**Methods**).

Comparison of qDSB-Seq results based on I-SceI spike-in and NotI restriction digestion. Next, we tested our entire approach. To this end, we compared results of qDSB-Seq performed in the same sample (wild-type cells treated with hydroxyurea, WT HU), using two different qDSB-Seq variants described above, based on I-SceI spike-in (**Fig. 2b**) or NotI restriction digestion (**Fig. 2c)**. The I-SceI spike-in method, validated as described above, can serve as validation for the restriction enzyme NotI-based qDSB-Seq quantification. qDSB-Seq using the I-SceI spike-in and the NotI restriction digestion gave

very similar results: 213±9 DSB/cell and 219±7 total breaks per cell, respectively (both numbers after subtracting breaks induced by NotI and labeled telomere ends) (**Fig. 2d**). Moreover, comparison of qDSB-Seq and sqPCR also gives consistent results (**Fig. S1**). Thus, we conclude the qDSB-Seq method has been validated as accurate.

Applications of qDSB-Seq

Quantification of DSBs induced by chemotherapeutic compounds. Many of the drugs used for cancer chemotherapy generate DSB levels that are dangerous to normal proliferating cells[14, 15]. To decrease the adverse effects of chemotherapy on normal cells, it is important to monitor this damage. Here, we studied effects of Zeocin, a member of the bleomycin drug family, which can induce radiomimetic DSBs in cells, similar to ionizing radiation. To quantify the number of breaks caused by Zeocin, we treated cells synchronized in G_1 phase with 100 μg/ml Zeocin for 1h and performed qDSB-Seq with gentle NotI digestion on treated (ZEO) and control (G1) samples. Based on the gDNA sequencing, we estimated NotI cutting frequency in ZEO and G1 samples and used qDSB-Seq to calculate DSB numbers. The number of DSBs estimated in Zeocin-treated cells and control was 16.5±1.7 DSBs/cell and 3.0±0.3 DSBs/cell, respectively, meaning that Zeocin induces on average 13.2±1.9 DSBs per cell (**Fig. 3a** and **Table S1**).

Quantification of DSBs induced by replication stress. Although cells have evolved complex mechanisms, called checkpoints, to stabilize or restart stalled replication forks, such forks may still collapse, particularly if replication stress persists or if components of the response to replication stress are disabled[16, 17]. Here, we studied five samples with induced replication stress. As a control, we used wt S phase cells. The number of DSBs in S phase cells detected by qDSB-Seq was much higher than in G_1 phase cells (**Fig. 3b**), showing that even normal replication causes considerable stress. Next, we studied wt cells treated with camptothecin (CPT), a topoisomerase I (Top1) inhibitor, known to cause fork stalling and reversal, eventually leading to DSBs. Indeed, we saw an increased DSB level in the CPT sample, as compared with the control (**Fig. 3c**). The second sample consisted of wt cells released synchronously into S phase from a G_1 arrest and then treated with 200mM hydroxyurea for one hour (WT HU). In this sample, all the replication forks were severely slowed down or stopped due to dNTP depletion, which is known to induce DSBs. Here, we saw that such a massive replication arrest causes a substantial increase in the observed DSBs, unlike treatment with low doses of CPT (**Fig. 3c**). In the last three analyzed samples, we studied the effects of *mec1-1* and *mus81*Δ mutations on HU-induced DSBs. Mec1 is an activator of checkpoint upon genotoxic stress and is essential for the maintenance of stable replication forks. In eukaryotes, Mus81 functions as a structure-selective endonuclease in the resolution of branched DNA intermediates, such as stalled replication forks[18-20]. Once a stalled replication fork is cleaved, it produces a 1-ended DSB that would be detectable by i-BLESS, a DSB-labeling method we used (**Methods**). To examine the role of the checkpoint kinase Mec1 and the nuclease Mus81 in the cleavage of replication forks, we quantified DSBs in *mec1-1*, *mus81*Δ and *mec1-1 mus81*Δ cells treated with HU. The results (**Fig. 3d** and **Table S2**) show that disabling Mec1, which prevents fork collapse in response to HU, increases DSBs by 32%. In contrast, inactivation of Mus81 did not significantly affect

DSB frequency. Moreover, inactivation of Mus81 in *mec1–1* cells did not reduce DSB levels, indicating that deregulation of Mus81 cleavage is not increasing DSBs in the absence of Mec1. In summary, we show that qDSB-Seq allows precise quantification of DBS levels, spanning two orders of magnitude, and in a variety of samples. It also gives insights into mechanisms of DSB creation.

Quantification of DSBs caused by natural replication fork barriers. Sites promoting pausing of the replication forks have been identified in many genomes[21]. In the budding yeast, the replication fork barriers (RFBs) located within the ribosomal DNA (rDNA), protect the nearby, highly expressed rRNA genes from head-on collisions with incoming replication forks[21]. It is not known what percentage of forks pauses at the barriers, and whether forks only stop or are also disassembled and later repaired by recombination[22]. Here, we used qDSB-Seq to quantify DSBs associated with fork arrest at the RFB1-2 in samples discussed above (**Fig. 3**). As expected, in G_1 arrested cells (samples G1 and ZEO), there was no increased DSB signal in the vicinity of replication fork barriers, but we observed such an increased signal in all six samples collected during S phase (**Fig. 4a**). Interestingly, in these latter samples, the frequency of DBSs at the fork barriers was not proportional to the total number of DSBs in the sample. The lower proportion of RFB-related DSBs in HU-treated samples (**Fig. 4b**) may reflect the fact that most of the forks do not reach the RFBs in the presence of HU due to dNTP depletion[23]. Moreover, the signal is higher in HU-arrested *mec1-1* and *mec1-1 mus81Δ* cells (1.53 and 1.50 DSBs/cell, respectively) because forks progress further in HU-treated *mec1* mutants due to the deletion of the *SML1* gene[24]. One should also note that RFBs lie within an rDNA array, ~150 copies of which are present in the W303 strain genome. Therefore, the frequency of breaks we calculate, e.g. 1.53 DSBs/cell for *mec1-1* HU sample, corresponds to only 0.01 DSBs/cells on average at each RFB in an individual rDNA array. The results indicate that inhibiting the Mec1 checkpoint has adverse effects on the stability of forks stalled at RFBs, causing their collapse more often than other treatments. We also reveal an order of magnitude variance of stability of replication forks stalled at RFBs. For example, forks are relatively stable (0.0007 DSB per cell and rDNA array) in wt cells treated with HU for 1h, but ~14 times less stable in HU-arrested Mec1 deletion mutants (0.01 DSB per cell and rDNA array). These breaks seem to be caused by Mus81-mediated resection since they disappear in *mus81Δ* (samples taken at the same stage of the cell cycle progression).

Also, we uncovered subtle differences in the positions of DSBs near RFBs, that cannot be explained by different fork progression in HU-treated cells (**Fig. 4c)**. This result may indicate that HU affects mechanisms of maintenance of fork stability at RFB in a manner deeper than changing the percentage of collapsed forks (**Fig. 4**). Indeed, HU treatment decreases the number of DSBs at RFBs observed in wild type cells by 6-fold (**Fig. 4a**), which may be related to a indirect role HU plays in inhibiting homologous recombination[25], which is believed to be the part of the fork disassembly process of forks stalled at RFBs.

Other potential approaches to quantifying DSBs.

Quantifying DSBs using γH2A.X. Phosphorylating H2A.X (H2A in yeast) histone to γH2A.X is one of the earliest steps in a cell's response to DSBs. Therefore, antibodies against γH2A.X are often used to visualize DSBs in cells and, in spite of known drawbacks (**Introduction**), quantify them. Now that we are able to quantify DSBs accurately using qDSB-Seq, we will test the accuracy of γH2A.X estimation. To this end, we employed DivA cells[12]; human U2OS cells engineered to express the AsiSI restriction enzyme endogenously. 4-hydroxytamoxifen (4-OHT) treatment acts as an on-switch for AsiSI transport to nuclei, leading to cutting the human genome at 1,219 recognition sites, with varying efficiency. We used the qDSB-Seq method to calculate the number of DSBs induced in DivA cells by tamoxifen treatment, using gDNA sequencing, as described above and in **Methods**. Even though in human data gDNA sequencing coverage is low, aligning all 1,219 AsiSI cutting sites allows a confident estimation of the total number of DSBs induced by AsiSI (**Supplementary File**). Since a low number of AsiSI induced breaks is also present in untreated DivA cells, we subtracted the number of AsiSI breaks in untreated cells from the number of DSBs produced by AsiSI in the +4-OHT cell. The BLESS data was used to define the extent of homologous repair of AsiSI induced breaks in the studied conditions, which was minimal. Our qDSB-Seq estimate of the number of 4-OHT-induced DSBs to be 9.7±3.9 DSBs/cell. Legube and colleagues employed counting γH2A.X foci and obtained ~80 DSBs induced upon 4-OHT in DivA cells in the same conditions[26], that is 8 times higher number than we did. Such discrepancy is not surprising, considering that γH2A.X signal spreads between 400 kb and 2 Mb from the site of a break[12], and thus it is unlikely that there is a mechanism allowing for its immediate removal upon break repair. Thus, the number of concurrent DSBs may be overestimated due to a lingering signal from already repaired DSBs. We propose that such persistence of γH2A.X signal after DSBs were repaired causes the γH2A.X-based estimate of AsiSI-induced DSBs[26] to be 8 times higher than the qDSB-Seq quantification. Indeed, such signal persistence seems to be a general phenomenon. When we performed timecourse i-BLESS DSBs labeling and γH2A ChIP-Seq in yeast, the time delay between the γH2A signal and i-BLESS was clear (**Fig. S2)**. It does not mean that counting γH2A.X or γH2Afoci would always lead to overestimating a number of DSBs. As **Fig. S2** shows, early on γH2A signal is relatively low compared with i-BLESS, but the proportion change later, with last time point showing still increased γH2A near origin, while the i-BLESS signal already disappeared. Counting γH2A.X foci may also lead to underestimating the number of DSB by detecting adjacent foci as a single one. In summary, counting γH2A.X or γH2A foci is not a reliable method of estimating DSB numbers per cell, it can lead to both substantial over- and under-estimation of the DSB numbers.

Purely computational DSB quantification

Telomere-based normalization. One may think of telomeres as natural "spike-ins" present in the genome since in non-tumoral cancer cells they are present in constant quantities per cell. Therefore, we attempted telomere-based normalization in samples for which DSB levels were also quantified using qDSB-Seq (**Table S3**). We discovered that telomere-based normalization led to massive errors (up to 2 orders of magnitude in G_1 sample, **Table S3**). Moreover, even relative DSB levels were not correctly predicted. This result shows that problems with telomere-based normalization are not caused by

partial deproteinization of telomere ends but to a variable degree of telomere protection across conditions. Reason of this phenomenon is not clear; it may be related to T-loop or D-loop formation. Moreover, even small changes in the sample preparation procedure (no fixation versus 5 min 2% formaldehyde fixation) can change telomere-based normalization by 2-fold, while different doses of proteinase K can change results by order of magnitude (**Table S4**). Based on these data, it is evident that telomere-based normalization is generally not feasible.

Normalization to the background. Another possible approach is normalization to background level, relying on the assumption that signal varies in a small fraction of genomic locations, similar to the principle employed e.g. in differential gene expression analysis. However, in our experience, this is not the case for DSB-sequencing data (see also[27]). For example, breaks may be induced evenly, e.g. by ionizing radiation. By definition, as long as DSBs are induced uniformly, normalization to the background would miss them, irrespective of the magnitude of the induced breaks. Analysis based on background normalization would, therefore, reach a false conclusion in any sample in which DSBs are induced uniformly, e.g. with irradiation. In other samples, the results of background-based normalization are no better (**Table S5**). Such analysis tends to show only a small difference in predicted DSB levels between samples, and relative break levels are often clearly predicted incorrectly, e.g. less DSBs in HU treated S phase samples than in the untreated G_1 sample. The reason background correction does not work well is that it is based on the assumption that factors causing background DSBs are virtually constant between the samples (e.g. handling artifacts, spontaneous breaks) and not by condition-specific circumstances (e.g. radiation or breaks induced in un-replicated regions during chromosome replication). However, in our experience background in DSB-sequencing data depends substantially on experimental treatment[27] and that is why background normalization is not appropriate.

Discussion

Several genome-wide methods to detect DSBs with single-nucleotide resolution using sequencing have been developed recently, but their usefulness is limited by the fact that they only allow to compare DSB level between genomic loci in the same sample, but not between samples. Here, we proposed a general approach that solves this problem, by complementing a DSB-labeling method by the qDSB-Seq approach, allowing us to estimate absolute DSB frequencies (per cell) genome-wide. We tested qDSB-Seq approach by combining it with DSB-labeling using BLESS and i-BLESS, but it can also be used with any other DSB-labeling method[4-7]. We used several independent methods for qDSB-Seq validation: qPCR, Southern blotting, and cell mixing. We have shown that the qDSB-Seq allows accurate sample-to-sample comparisons of DSB frequencies. qDSB-Seq is easy to use and to make it even more practical we proposed several implementations.

We most recommend implementing qDSB-Seq by a partial restriction digestion *in vitro*. Such approach does not require organism-specific constructs, such as DivA cells or the I-SceI strain discussed above. Moreover, since digestion is performed *in vitro*, induced breaks cannot be repaired by homologous recombination, which simplifies the calculation of the number of spike-in breaks caused by a restriction enzyme (**Methods**). On the other hand, *in vitro* restriction digestion may require a pilot experiment to select an optimal

concentration of a restriction enzyme. Choosing too small enzyme concentration may result in cutting efficiency too low to be estimated accurately, while too high concentration may cause too many reads arising from spike-in digestion, lowering effective coverage. We have shown that qDSB-Seq works with more than two orders of magnitude of numbers of induced DSBs (0.02 to 28 DSBs/cell), so it is not necessary to select a near optimal cutting frequency, it is enough to avoid extremely low or high spike-in frequencies (see also **Methods**). The qDSB-Seq using cells engineered to carry spike-in DSBs is easier in this respect, as it is straightforward to control a level of spike-in by determining mixing ratio (**Methods**). However, thus engineered cells should be from the same organism in which DSB sequencing is performed.

Another interesting aspect of our results is the high number (>40 DSBs/cell) of breaks observed in all yeast S phase samples (**Fig. 3**). Since in untreated G_1-phase cells we only detected 3 DSBs/cell, high levels of DSBs in S-phase samples should result from S phase-specific DNA damage rather than from noise in DSB sequencing. These results raise possibility that reversed, but not cleaved, replication forks are labeled together with DSBs by methods such as BLESS and i-BLESS. More research beyond the scope of this paper, especially cryoEM studies, are necessary to clarify. Moreover, since sequencing data reports population means, a small fraction of cells with a very high number of breaks may contaminate the whole sample. For example, such small population of cells can undergo substantial DNA fragmentation caused by unresolved replication stress or apoptosis. The above example shows that knowing precise levels of DSBs per cell leads to a deeper reflection on the meaning of the DSB sequencing results and indicates that they should not be interpreted as a typical number of DSBs per cell. More complex mathematical framework and complementary data are needed to estimate a typical number of DSBs per cell, as we discuss elsewhere[27].

We also evaluated the performance of other potential DSB quantification methods, such as estimating a number of DSBs per cell by counting γH2A.X foci. Unlike direct DSB-labeling, which is instantaneous, γH2A.X signal accumulates over time. Such time delay in the appearance of γH2A.X foci can lead both to underestimation and overestimation of the number of DSBs, depending on the time of the γH2A.X labeling in regard to DSBs occurrence and repair. We also showed, that purely computational methods of estimating background are not appropriate, since DSB background is highly variable between samples and conditions. Therefore, there is no substitution to modifying a DSB-sequencing method to allow confident experiment-based normalization, such as our qDSB-Seq method.

In summary, we proposed and tested an easy-to-implement qDSB-sequencing method, which allows accurate comparison between DSB-sequencing data across samples and organisms, thereby allowing to quantify relative contributions of different DSB-causing factors.

Acknowledgements

This research was supported by the NIH grant R01GM112131 to M.R. (Y.Z., N.D., B.F., J.N., R.Y. and M.R.), by the Polish National Science Centre grants

(2011/02/A/NZ2/00014 to K.G. and 2015/17/D/NZ2/03711 to M.S.), Foundation for Polish Science grant (TEAM) to K.G, and Ligue contre le Cancer (Equipe labelisee), Agence Nationale pour la Recherche (ANR) and Institut National du Cancer (INCa) grants to P.P. (B.P., R.F. and P.P.). The authors are grateful to Gaelle Legube for providing DivA cells and fruitful discussions, especially about results of [26] and [12]. The authors are also grateful to Andrzej Kudlicki and Heather Lander for discussions and critical reading of the manuscript.

Authors contributions

M.R. conceived qDSB-Seq and supervised and coordinated the project. M.R. and Y.Z. wrote the manuscript, A.B., K.G. and P.P. and M.S. edited the manuscript. Y.Z. performed data analysis and developed software, N.D. performed initial data analysis and analysis of the human data and developed software. A.B., K.G., B.P., P.P. and M.R. designed experiments. A.B. and M.S. performed BLESS and qDSB-Seq experiments. B.P., R.F. and P.P. performed validation. Y.Z. prepared figures. R.Y., B.F. and J.N. contributed to statistical data analysis. M.S. and K.G. performed paired-end Illumina sequencing. All authors read the manuscript.

Competing Financial interests

The authors declare no competing financial interests.

Methods

Strains and growth conditions. Yeast strains used in this study are listed in **Table S6**. Cells were grown in YPD medium at 25°C until early log phase and were then arrested in G_1 for 170 min with 8 µg/ml α-factor. YBP-275 strain was cultured in YPR medium, galactose was added for 2 h to induce I-SceI cutting. Cells were released from G_1 arrest by addition of 75 µg/ml Pronase (Sigma). HU was added 20 min before pronase release followed by a 40 min (wt) or 1 h incubation (*mec1-1*, *mus81*Δ, *mec1-1 mus81*Δ). After collecting, cells were washed with cold SE buffer (5M NaCl, 500 mM EDTA, pH 7.5) and immediately subjected to DSB labeling.

DSB sequencing. BLESS/i-BLESS DSB labeling was performed as described in[13,28]. Sequencing libraries for BLESS/i-BLESS and respective gDNA samples were prepared using commercially available kits (e.g. TruSeq DNA LT Sample Prep Kit (Illumina) and ThruPLEX DNA-seq Kit (Rubicon Genomics)). BLESS/i-BLESS libraries were prepared without prior fragmentation and further size selection. Quality and quantity of the libraries were assessed on 2100 Bioanalyzer using HS DNA Kit, and on Qubit 2.0 Fluorometer using Qubit dsDNA HS Assay Kit (Life Technologies). The libraries were sequenced (2x75 bp) on Illumina HiSeq2500/HiSeq4000 platforms, according to our modified experimental and software protocols for generation of high-quality data from low-diversity samples, such as resulting from the BLESS or i-BLESS methods[26]. Additionally, qDSB-sequencing was performed, either using NotI restriction digest or I-SceI spike-in, as described below.

qDSB-Seq with I-SceI spike-in. Here, we used a yeast strain engineered to carry a cassette with one I-SceI cutting site, which is not present in any other locations of the yeast genome. We used CASY Cell Counter (Roche Applied Science) to mix this spike-in with our sample of interest (wild type cell with replication stress induced by hydroxyurea treatment, HU) in 2:98 proportion. The cutting ratio of the I-SceI endonuclease expressed in the I-SceI strain, was estimated using an unmixed I-SceI strain (see below and **Methods)**.

Sequencing data analysis. We used *InstantSeq*[29] to ensure sequencing data quality before mapping. Next, *InstantSeq* was used to remove BLESS/i-BLESS proximal and distal barcodes (TCGAGGTAGTA and TCGAGACGACG, respectively). Reads labeled with the proximal barcode, which are directly adjacent to DSBs**,** were selected and mapped to the version of the yeast S288C genome (where we manually corrected common polymorphisms) and the human hg19 genome using bowtie[30] v0.12.2 with the alignments parameters '-m1 –v1' (to exclude ambiguous mapping and low-quality reads). The end base pairs of the reads were trimmed using bowtie '-3' parameter. The parameter choice was based on the *InstantSeq* quality report. Hygestat_BLESS v1.2.3 (part of *InstantSeq* software suite[29]) was used to identify genomic regions with a significant difference in normalized read numbers between treatment and control samples.

Selection of a restriction enzyme. Restriction enzymes differ in the number of recognition sites a restriction enzyme must bind to in order to cleave double-strand DNA. Most restriction enzymes (e.g. NotI and BamHI used in this study) cleave one recognition site at a time. Some, however, (e.g. BcgI), cut only when bound to at least two sites at once. For multi-site restriction enzymes, lowering enzyme concentration may sometimes increase cutting efficiency due to site saturation. Therefore, we do not recommend multi-site enzymes for qDSB-Seq, since potential site saturation makes estimating optimal quantities of a restriction enzyme more challenging (below).

Calculation of breaks per cell. We recommend the following procedure:

Calculate cutting frequency (f_{cut}) from genome sequencing:

$$f_{cut} = \frac{\alpha_{cut} N_{cut}}{\alpha_{uncut} N_{uncut} + \alpha_{cut} N_{cut}} - f_{bg} \qquad \text{Equation (1)}$$

where, f_{cut} is cutting frequency; N_{cut} is the number of gDNA fragments on restriction sites cut by enzyme; N_{uncut} is the number of gDNA fragments on restriction sites but without cutting; α is correction factor. We describe below how to compute f_{bg}.

Calculate induced breaks (B_I) on restriction sites:

$$B_I = f_{cut} N_{sites} p \qquad \text{Equation (2)}$$

where, B_I is induced breaks on restriction sites (N_{sites}) and p is the proportion of digested cells (p=1 unless spike-ins are used).

Reads per break in DSB sequencing (r):

$$r = \frac{R_{cut}}{B_I} \qquad \text{Equation (3)}$$

where, r is the number of DSB reads per break estimated from the number of DSB reads on cutting sites (R_{cut}) and induced breaks (B_I).

Breaks per cell (B_{cell}):

$$B_{cell} = \frac{R_{total}}{r} \qquad \text{Equation (4)}$$

Sonication background correction (f_{bg}). To correct for background, we randomly selected genomic windows of the same size as used in the qDSB-Seq method and estimated "cutting frequency" in those intervals as in qDSB-Seq. Thus, the estimated background is subtracted from the computed cutting efficiency.

Background estimation. To estimate background intensity from sequencing data, we used a sliding window method to count the number of reads per kilobase pairs (kb). The windows containing repeats, mitochondrial DNA, and rDNA, were excluded. In addition, we also removed unmapped regions based on gDNA sequencing data. The frequency of reads in windows was calculated and visualized. Then, the number of reads with the highest frequency in the distribution was selected as the background value.

Identification of telomere ends. To estimate the reads from the telomere ends in BLESS/i-BLESS sequencing data, we screened reads labeled with proximal barcode. In the yeast genome, we searched for the telomeric sequence in AC-rich strands using the regular expression CAC{1,10} in the PERL language. In the human genome, we searched for the telomeric sequence in the C-rich strand using regular expression "^(CCCTAA|CCTAA|CTAA|TAA|AA|A)(CCCTAA){1,}(C|CC|CCC|CCCT|CCCTA|CCCTAA)$".

References

1. Khanna, K.K. & Jackson, S.P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nature genetics* **27**, 247-254 (2001).
2. Bassing, C.H., Swat, W. & Alt, F.W. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* **109 Suppl**, S45-55 (2002).
3. Slaymaker, I.M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-88 (2016).
4. Crosetto, N. et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature methods* **10**, 361-365 (2013).
5. Hoffman, E.A., McCulley, A., Haarer, B., Arnak, R. & Feng, W. Break-seq reveals hydroxyurea-induced chromosome fragility as a result of unscheduled conflict between DNA replication and transcription. *Genome Res* **25**, 402-412 (2015).
6. Lensing, S.V. et al. DSBCapture: in situ capture and sequencing of DNA breaks. *Nature methods* **13**, 855-857 (2016).
7. Canela, A. et al. DNA Breaks and End Resection Measured Genome-wide by End Sequencing. *Molecular cell* **63**, 898-911 (2016).
8. Marti, T.M., Hefner, E., Feeney, L., Natale, V. & Cleaver, J.E. H2AX phosphorylation within the G1 phase after UV irradiation depends on nucleotide excision repair and not DNA double-strand breaks. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 9891-9896 (2006).
9. Tuduri, S. et al. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nature cell biology* **11**, 1315-1324 (2009).
10. Chadwick, B.P. & Lane, T.F. BRCA1 associates with the inactive X chromosome in late S-phase, coupled with transient H2AX phosphorylation. *Chromosoma* **114**, 432-439 (2005).
11. Meyer, B. et al. Clustered DNA damage induces pan-nuclear H2AX phosphorylation mediated by ATM and DNA-PK. *Nucleic acids research* **41**, 6109-6118 (2013).
12. Iacovoni, J.S. et al. High-resolution profiling of gammaH2AX around DNA double strand breaks in the mammalian genome. *The EMBO journal* **29**, 1446-1457 (2010).
13. Biernacka, A. et al. in In preparation (2017).
14. Cheung-Ong, K., Giaever, G. & Nislow, C. DNA-damaging agents in cancer chemotherapy: serendipity and chemical biology. *Chemistry & biology* **20**, 648-659 (2013).
15. Jekimovs, C. et al. Chemotherapeutic compounds targeting the DNA double-strand break repair pathways: the good, the bad, and the promising. *Frontiers in oncology* **4**, 86 (2014).
16. Lopes, M. et al. The DNA replication checkpoint response stabilizes stalled replication forks. *Nature* **412**, 557-561 (2001).
17. Sogo, J.M., Lopes, M. & Foiani, M. Fork reversal and ssDNA accumulation at stalled replication forks owing to checkpoint defects. *Science* **297**, 599-602 (2002).

18. Ehmsen, K.T. & Heyer, W.D. Saccharomyces cerevisiae Mus81-Mms4 is a catalytic, DNA structure-selective endonuclease. *Nucleic acids research* **36**, 2182-2195 (2008).
19. Froget, B., Blaisonneau, J., Lambert, S. & Baldacci, G. Cleavage of stalled forks by fission yeast Mus81/Eme1 in absence of DNA replication checkpoint. *Molecular biology of the cell* **19**, 445-456 (2008).
20. Rass, U. Resolving branched DNA intermediates with structure-specific nucleases during replication in eukaryotes. *Chromosoma* **122**, 499-515 (2013).
21. Kobayashi, T. The replication fork barrier site forms a unique structure with Fob1p and inhibits the replication fork. *Molecular and cellular biology* **23**, 9178-9188 (2003).
22. Labib, K. & Hodgson, B. Replication fork barriers: pausing for a break or stalling for time? *EMBO reports* **8**, 346-353 (2007).
23. Pasero, P., Bensimon, A. & Schwob, E. Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus. *Genes & development* **16**, 2479-2484 (2002).
24. Poli, J. et al. dNTP pools determine fork progression and origin usage under replication stress. *The EMBO journal* **31**, 883-894 (2012).
25. Alabert, C., Bianco, J.N. & Pasero, P. Differential regulation of homologous recombination at DNA breaks and replication forks by the Mrc1 branch of the S-phase checkpoint. *The EMBO journal* **28**, 1131-1141 (2009).
26. Caron, P. et al. Non-redundant Functions of ATM and DNA-PKcs in Response to DNA Double-Strand Breaks. *Cell Rep* **13**, 1598-1609 (2015).
27. Zhu, Y. et al. (BioRxiv; 2017).
28. 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).
29. Mitra, A. et al. in submitted (2017).
30. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

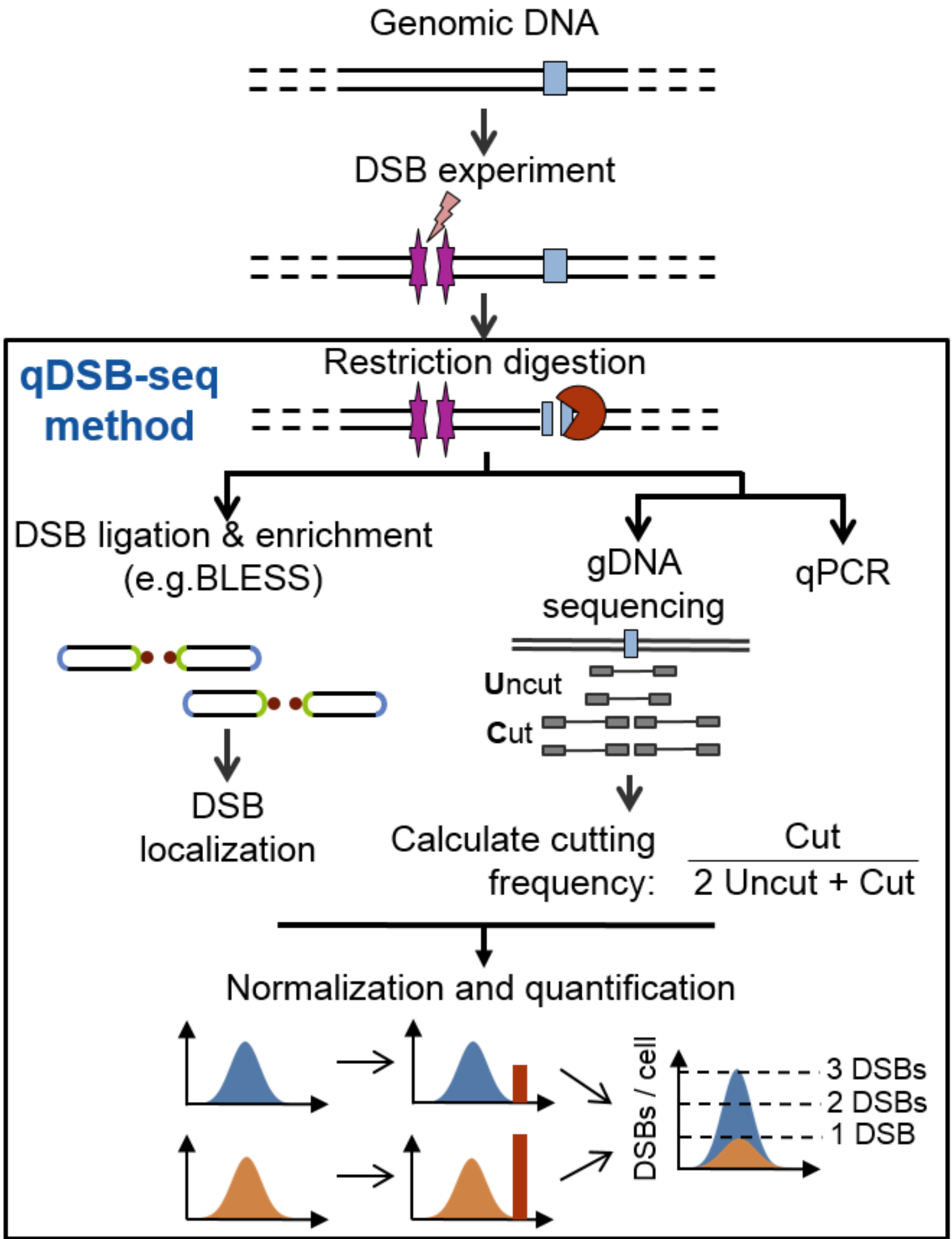
Figure 1. Schema of the qDSB-Seq method. After DSB induction, and before DSB labeling, the sample is gently digested with a restriction enzyme, with the goal to induce occasional breaks, not total fragmentation. Afterward, DSB sequencing and gDNA sequencing or qPCR are performed and used to estimate the cutting frequency of the enzyme. This cutting frequency is used to quantify the absolute frequency (i.e. per cell) of DSBs in the sample. The method also allows for quantitative comparison of break frequencies and distributions between samples.

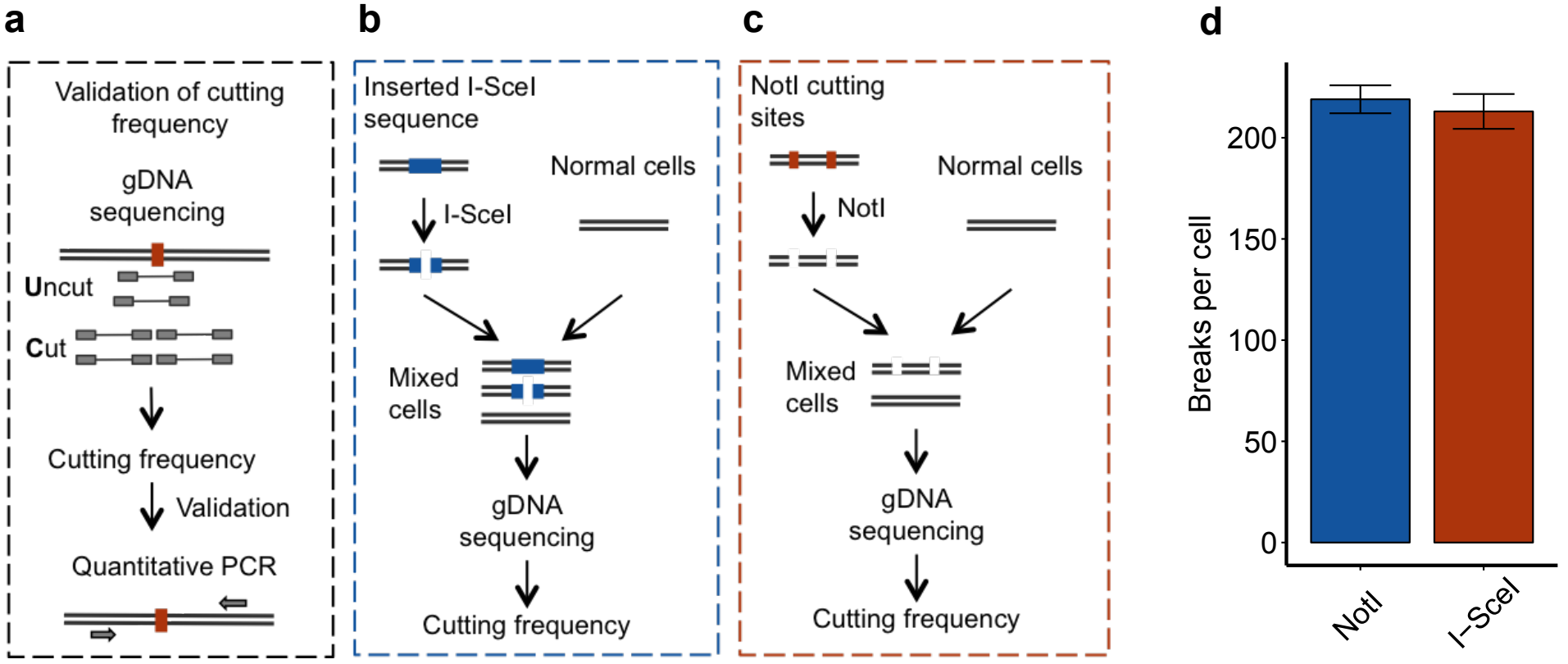
Figure 2. Validation of the qDSB-Seq method. (a) The simplified computational method to calculate restriction enzyme cutting frequency was validated by qPCR. **(b)** A spike-in with known DSB frequency (determined by qPCR) is mixed with experimental

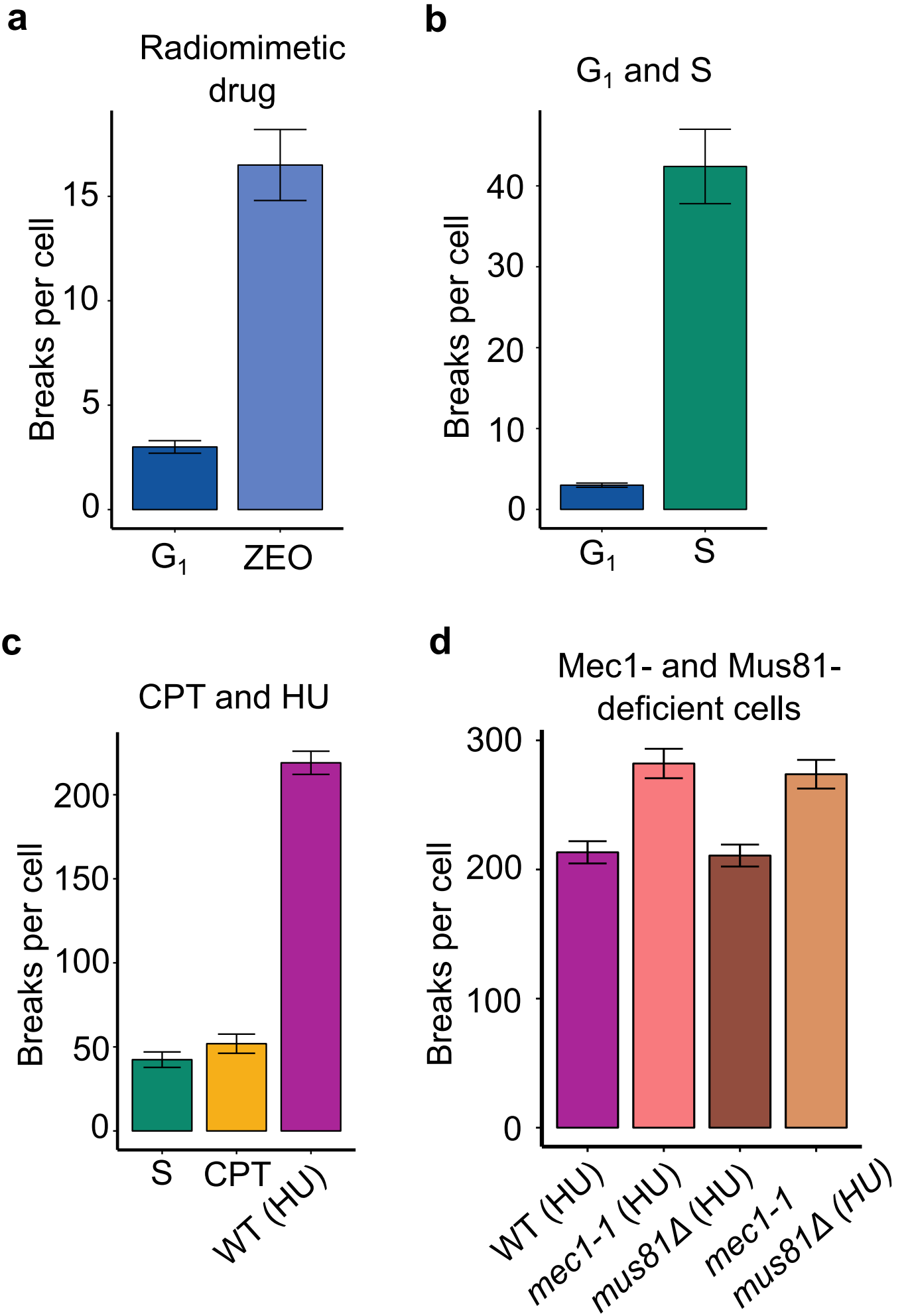
cells in controlled proportion. **(c)** Quantifying DSBs using restriction enzyme NotI in *vitro*, which cut yeast genome at multi-loci. **(d)** Comparison of quantifications using qDSB-Seq with validated I-SceI spike-in and NotI digestion.

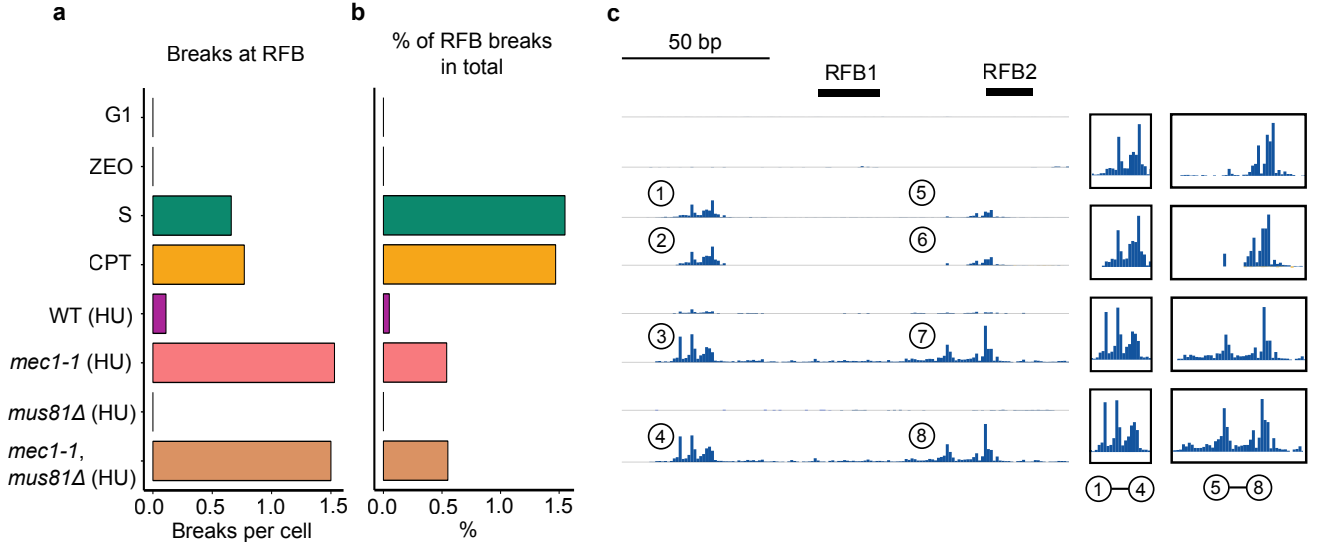
Figure 3. Quantitative DSB-sequencing for various samples in different cell-cycle phases and conditions. (a) Quantification of DSBs induced by a radiomimetic drug, Zeocin (100 µg/ml). **(b)** Comparison of DSBs levels in G1 and S phase. The wild-type cells were arrested in G1 and collected or released into S phase for 45 min. **(c)** Quantification of DSBs in the wild-type cells during unperturbed S phase (S), camptothecin (CPT, 100 µM) or hydroxyurea treatment (HU, 200 nM). CPT and HU trigger formation of stalled forks and paused forks, respectively. **(d)** Quantification of DSBs in wild-type and Mec1- and Mus81-deficient cells under HU treatment.

Figure 4. Quantification of DSBs resulting from replication forks (incoming from the left) collapsed at RFBs. (a) DSBs per cell combined for RFB1 and RFB2. **(b)** Percentage of breaks related to RFBs in all breaks. **(c)** Zoom-in of DSB-sequencing reads near RFBs, single nucleotide resolution. Note differences in DSB positions at RFB2 between HU-treated *mec1-1* samples (7 & 8) and CPT and S samples (5 & 6).


Genomic DNA
DSB experiment
qDSB-seq method
Restriction digestion
DSB ligation & enrichment (e.g.BLESS)
gDNA sequencing
qPCR
Uncut
Cut
DSB localization
Calculate cutting frequency:
Cut
2 Uncut + Cut
Normalization and quantification
DSBs / cell
3 DSBs
2 DSBs
1 DSB


a
Validation of cutting frequency
gDNA sequencing
Uncut
Cut
Cutting frequency
Validation
Quantitative PCR
b
Inserted I-SceI sequence
Normal cells
I-SceI
Mixed cells
gDNA sequencing
Cutting frequency
c
NotI cutting sites
Normal cells
NotI
Mixed cells
gDNA sequencing
Cutting frequency
d
Breaks per cell
200
150
100
50
0
NotI
I-SceI


a
Radiomimetic drug
Breaks per cell
0
5
10
15
G1
ZEO
b
G1 and S
Breaks per cell
0
10
20
30
40
G1
S
c
CPT and HU
Breaks per cell
0
50
100
150
200
S
CPT
WT (HU)
d
Mec1- and Mus81-deficient cells
Breaks per cell
0
100
200
300
WT (HU)
mec1-1 (HU)
mus81Δ (HU)
mec1-1
mus81Δ (HU)


a
Breaks at RFB
G1
ZEO
S
CPT
WT (HU)
mec1-1 (HU)
mus81Δ (HU)
mec1-1, mus81Δ (HU)
0.0
0.5
1.0
1.5
Breaks per cell
b
% of RFB breaks in total
0.0
0.5
1.0
1.5
%
c
50 bp
RFB1
RFB2
1
2
3
4
5
6
7
8
1—4
5—8