Integrated analysis of DSB patterns reveals precisely DSB formation mechanisms following replication fork collapse

Yingjie Zhu[1,9], Anna Biernacka[2,9], Benjamin Pardo[3], Romain Forey[3], Norbert Dojer[1,7], Raziyeh Yousefi[1], Jules Nde[1], Bernard Fongang[1], Abhishek Mitra[1], Ji Li[1], Magdalena Skrzypczak[2], Andrzej Kudlicki[1], Philippe Pasero[3], Krzysztof Ginalski[2], Maga Rowicka[1,4,5,6]

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

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

[3] IGH Institute of Human Genetics, Montpellier, France.

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

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

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

[7] Current address: Institute of Informatics, University of Warsaw, Warsaw, Poland

[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: Replication fork fate elucidated from DSB sequencing

keywords: replication fork, replication stress, DNA double-stranded break, sequencing, replication origin, BLESS, PFGE, Poisson distribution, model based, population heterogeneity

Abstract

DNA double-stranded breaks (DSBs) are rare in healthy cells, and therefore few cells with highly-fragmented DNA can obscure DSB sequencing results, but not pulsed-field gel electrophoresis data. Therefore, to infer mechanism of DSB formation and DSB population distribution, we computationally integrate both data types. We found that after hydroxyurea-induced replication stress, 1-ended DSBs, resulting from collapsed replication forks, are typically population representative, while majority of 2-ended DSBs (79-100%) arise from small subpopulations of highly-damaged cells. To study replication fork collapse, we inhibited the checkpoint protein Mec1 and endonuclease Mus81 and quantified that 19% and 13% of hydroxyurea-induced DSBs are Mec1- and Mus81-dependent, respectively. We also clarified that Mus81 is preferentially cleaving reversed replication forks and Mus81-induced DSBs are completely Mec1-dependent.

Introduction

DNA double-strand breaks (DSBs) are the most deleterious form of DNA damage. DSBs can either arise spontaneously, e.g. during replication or be induced, e.g. by ionizing radiation or chemicals, including many chemotherapy drugs (*1*). DSBs are also induced during genome editing by the CRISPR/Cas system (*2*). A better understanding of mechanisms of DSB formation is therefore of great interest and crucial to effective prevention and treatment of the associated diseases, including cancer. Nevertheless, the quantitative knowledge of how genomic DNA breaks in response to various stressors is still lacking. Starting with our BLESS method in 2013 (*3*), several methods for direct DSB-labeling genome-wide have been developed (*4-6*), including i-BLESS (*7*) used here. These dramatic improvements in DSB detection techniques yielded only modest progress in understanding mechanisms of DSB formation, due to the challenging nature of DSB sequencing data and lack of computational approaches for its more advanced analysis. DSBs are very rare events, even under DSB induction, since cells can only tolerate few DNA breaks. Therefore, even a fraction of a percent of e.g. senescent or apoptotic cells exhibiting massive DNA fragmentation can obscure the DSB sequencing data.

Here, we address this challenge by proposing a computational framework to analyze subpopulations of cells with different DSB patterns (**Fig. 1a**). The method relies on DSB sequencing and pulsed-field gel electrophoresis (PFGE) (**Fig. 1b**) as main input data, which we integrate computationally. The PFGE is less sensitive than sequencing, but, unlike DSB sequencing data, it cannot be obscured by a small subpopulation of highly-damaged cells (**Fig. 1c**). Our newly introduced quantitative DSB-sequencing (qDSB-Seq) (*8*) method allows us to quantify the number of DSBs per cell (**Fig. 1a**), which is key to understanding their physiological relevance. To further study DSB formation, we classify DSBs into patterns related to their generation mechanisms. Such classification allows to quantify the contributions of different sources of DNA damage in a given condition.

We show how our framework is applied to studying a complex repertoire of 1-ended and 2-ended DSBs induced by replication stress. 1-ended DSBs (1-DSBs, **Fig. 1d**), are a unique type of DSBs resulting from collapsed forks, while all other DSBs are 2-ended (2-DSBs). 1-DSBs manifest as an asymmetric sequencing read pattern (**Fig. 1e**), which can be utilized to detect active replication origins. Learning DSB patterns related to specific DSB inducing mechanisms allows classifying breaks according to mechanisms of their formation and quantify sources of DNA damage.

To evaluate DSB patterns caused by replication stress, we treated diploid S phase yeast cells with hydroxyurea (HU), which slows replication by depleting dNTPs (**Fig. 2a**). To study the mechanisms of stabilization and repair of collapsed forks, we used strains lacking functional checkpoint protein Mec1 and endonuclease Mus81 that was proposed to cleave stalled forks (**Fig. 2a**). qDSB-Seq, a method for quantitative DSB sequencing (*8*), and i-BLESS DSB labeling (*7*) were used in all eight samples to estimate an average number of DSBs per cell (**Fig. 2a**).

To better understand the DNA damage induced by replication stress, we analyzed the resulting DSB read patterns to classify DSBs according to their creation mechanisms.

Results

Identification of replication-dependent 1-DSBs

For an individual DSB-sequencing read, it is impossible to distinguish whether it originates from a 1-DSB or 2-DSB. Therefore, to identify true 1-DSBs, originating from fork collapse, we first determined active origins and their replicated vicinity (fork range), since forks collapse only in replicated regions. Briefly, we mapped the i-BLESS-labeled (*7*), sequenced DSB reads separately into the + and - strands of the yeast genome and score the DSB distribution on its resemblance to the origin pattern (**Fig. 1e, Methods)**. To facilitate pattern detection, we removed the high-frequency noise (**Fig. 2b**), e.g. related to nucleosome spacing (**Fig. S1)**, while preserving lower-frequency components related to replication. This computational method gives results very similar to optimizing experimental DSB detection protocol (**Fig. S2**). In thus filtered data (**Fig. 2b**), we used a sliding window to find regions with the highest 1-DSB score ($R_{1\text{-DSB}}$) (**Methods Eq. 2-4**). Finally, the hypergeometric test is used to estimate the significance of the $R_{1\text{-DSB}}$ scores and thus select predicted origins (**Fig. S3)**.

This method yields highly precise and sensitive origin predictions (**Table S1**, **Fig. S4**). For example, for *mec1-1 mus81Δ* HU1h sample, we detect 285 origins (at 95% precision, 85% sensitivity) or 255 origins (at 100% precision, 83% sensitivity). Hereafter, 100% precision origin predictions are used (**Fig. 2c**).

Prediction of active replication origins allowed us to distinguish between breaks resulting directly from replication fork collapse and others. We estimated "fork range", that is the maximal region detected as replicated by a given origin, as a length of an interval exhibiting a statistically significant preference for 1-DSBs caused by collapsed forks from that origin (**Fig. 1e**, **Methods**). Only 1-DSBs detected in these fork ranges were accepted and only those for which a corresponding read on the other strand, which could originate from a putative 2-DSB, cannot be found in the vicinity. Thus, we classified DSBs into 1-DSBs and 2-DSBs.

Inferring DSB distribution across cell population

To interpret the calculated average numbers of DSBs per cell, it is crucial to know if they are representative of the majority of cells in the population (**Fig. 1c**, top vs. bottom). To infer the distribution of DSBs in a cell population, we utilized pulsed-field gel electrophoresis (PFGE), a gel electrophoresis improved to allow separation of >50 kb DNA fragments. During PFGE experiments, replicating chromosomes cannot enter the gel due to replication intermediates and therefore remain in the well (**Fig. 1b**, and **Fig. 3a**, structures 3-4). Non-replicating chromosomes would enter the gel and form a distinct band, corresponding to intact chromosomes, while the fragmented ones are detected as a smear (**Fig. 1b**). We performed PFGE with fluorescent probes to chromosome III for each four HU and Recovery samples (**Fig. 2a**). PFGE data thus provide information on percentage of replicating, non-replicating and broken chromosomes III (percentage signal in well-band, chromosome-band and smear, **Fig. 1b**). Each broken, non-replicating chromosome III would contribute equally to the smear signal, irrespective of the number of 2-DSBs. Therefore, the PFGE data is not affected by a small fraction of cells with many DSBs (**Fig. 1c**) and thus perfectly complements DSB sequencing data by providing information on typical distribution of DSBs in cell population.

From DSB-sequencing, we learn the average number of DSBs per cell. PFGE gives information on the fractions of chromosomes: replicating, non-replicating broken, non-replicating intact (**Fig. 3a**). To compare DSB-sequencing and PFGE results one needs to model one data type using the other. First, we modeled 1-DSB distribution in the genome using the Poisson distribution (**Fig. 3b**), i.e. assuming that 1-DSBs occur independently of each other. Both 1-DSBs and replication intermediaries contribute to the well-band signal, so the percentage of DNA in the well band predicted from 1-DSBs alone should be not greater than the observed PFGE value. Nevertheless, our predictions are 4% to 20% higher than observed PFGE values in HU 1h samples (**Fig. 3c**). This small difference may reflect the limited accuracy of PFGE. Alternatively, 1-DSBs may not follow the Poisson distribution, but instead are less likely to occur if some 1-DSBs are already present, e.g. because of a checkpoint or DNA repair mechanism activated by the first DSB.

Nevertheless, it is clear that the average number of 1-DSBs per cell detected by qDSB-Seq (93-115 DSBs/diploid cell, **Fig. 2a**), is either population representative or that even higher number of 1-DSBs are typical for a major subpopulation. The reason is that 1-DSBs are unique, they can be only formed by collapsed replication forks (**Fig. 1d**). Therefore, the theoretical maximum number of 1-DSBs coexisting in a given cell is double the number of active origins (**Fig. 2c**), corresponding to the unrealistic assumption that all origins active in the population fired in each individual cell and that all resulting forks (two per each origin) collapsed and progressed to 1-DSBs. Even in such an extreme case of all active origins firing in every cell and all replication forks collapsing, that would result in 924 1-DSBs/diploid cell *mec1-1* HU 1h sample or 576 in the wt HU 1h sample. That reasoning shows that the minimal size of population affected by 1-DSBs in the described above extreme case would be 20% for wild-type and 10% for mec*1-1* HU 1h samples and that likely these subpopulations are much larger.

In conclusion, in HU 1h samples, 1-DSBs likely occur in virtually all replicating chromosomes III. Similar calculations (**Fig. 3c**) for recovery samples show that 1-DSBs likely occur in a small minority of them, as few as 3% in *mus81Δ*, suggesting an efficient 1-DSBs repair is occurring during 2-hour recovery from the HU treatment. These data also indicate that 1-DSBs are occurring frequently during fork stopping, our data do not show 1-DSBs occurring commonly during fork restart.

Next, we analyzed average detected 2-DSB numbers per cell in our eight samples (**Fig. S5**). These numbers are very high, between 303 and 518 DSBs/cell, unlikely to be tolerated even by the diploid W303 yeast strain we used. Indeed, when we employed the Poisson simulations, as described above, to predict PFGE signal assuming 2-DSBs occur independently, predictions were not consistent with observations (**Fig. 3d**). These results indicate that 2-DSBs are highly preferred in chromosomes III that are already broken, indicating a presence of a small subpopulation of cells with high 2-DSB numbers, overpowering the sequencing data (**Fig. 1c**). We also estimated average 2-DSB numbers per chrIII in majority of the population, using PFGE data and procedure opposite to that shown in **Fig. 3b**. We concluded that virtually all 2-DSBs detected by sequencing in HU 1h samples originate from a subpopulation of cells too small to be detected by PFGE (**Fig. 3e**). For example, a single yeast cell with DNA fragmented into units corresponding to single nucleosomes would give rise to about 70871 DSBs (*9*).

Therefore, as few as 0.3% of cells, undetectable by PFGE, with DNA fragmented into nucleosome-sized units, would cause average 2-DSB number per cell to be ~240, even if no other cells would exhibit 2-DSBs. Indeed, in recovery samples, the majority of 2-DSBs were originating for a small population of cells with highly-fragmented DNA, namely 100% for wild-type, 96% for *mus81Δ*, 95% for *mec1-1 mus81Δ* and 79% for *mec1-1* (**Fig. S6**). The observed high numbers of 2-DSBs are not an artifact of our DSB-labeling method, since they are present only in S phase samples (*10*).

In conclusion, 2-DSBs are induced during recovery in 21% of *mec1-1* cells, that is in five-times as many as in *mus81Δ* sample. Differences between *mec1-1 mus81Δ* and *mec1-1* recovery samples are less significant, since due to high percentage of DNA stuck in the well band in *mec1-1 mus81Δ* sample, there is a bias against detecting high smear signal in that sample. In conclusion, in *mec1-1* cells, there is a substantial number of 2-DSBs induced during recovery, much higher than in *mus81Δ*, number of 2-DSBs in *mec1-1 mus81Δis* likely underestimate by PFGE. Moreover, FACS data (**Fig. S7**) show that in *mec1-1* replication cannot be completed, unlike in other samples. Taken together, these data suggest that in *mec1-1* cells problems with completing replication lead to 2-DSBs. Comparison of these 2-DSB distribution around origins with 1-DSB distribution, indicating stalled fork position, and gDNA read density, indicating replication progress (**Fig. S8**), suggests that these 2-DSBs are distributed homogenously in the genome, consistent with e.g. DNA autolysis in dying cells, rather than occurring during fork-restart in which case they would typically occur further from origins than stalled forks.

Since 1-DSBs detected by sequencing likely are population-representative, unlike 2-DSBs, we used them exclusively in the subsequent analysis.

The role of Mec1 in origin firing

The Mec1 protein was proposed to inhibit firing of late origins during replication stress (*11-14*). Indeed, we detected 60 active late origins in *mec1-1* cells and only 7 in wild-type cells (**Fig. 4f**). We also observed that every mutation we introduced increased the number of origins activated under HU treatment: from 144 in wild-type, 231 in *mec1-1*, and 255 in *mec1-1 mus81Δ* (**Fig. 2c**). On the other hand, number of 1-DSBs per diploid cell directly originating from fork collapse, estimated conservatively as described above, varied less between samples: 115 for wt, 93 for *mec1-1* and 95 for *mec1-1 mus81Δ* (**Fig. 2a**). This result suggests a lower efficiency of origins in Mec1 mutant cells. Indeed, number of DSBs per origin we observed in mutant cells is about half of the number observed in wild-type cells (**Fig. 4a**), both for early and late origins (**Fig. 4b**).

Less efficient origins allow individual forks to travel further, since it takes on average longer before a fork incoming form another fired origin is encountered. Indeed, we observed increased fork ranges for Mec1 mutants (**Fig. 4c**), especially for early origins, for which detection is easier due to more accumulated breaks. Mec1 plays not yet understood role in replication, separate from its checkpoint role (*15*). Based on these data, we hypothesize that Mec1 promotes origin firing, consistent with its known role in contributing to Cdc7 phosphorylation of Mcm2-7 (*15*).

Role of Mec1 in replication stress response

In physiological conditions, replication-induced DSBs are resolved quickly by orchestrated actions of DNA damage response proteins. One of them, Mec1, binds to RPA-coated single-strand DNA (*16*). To better understand the role of Mec1 in DSB response, we analyzed patterns of 1-DSBs in *mec1-1* cells. We observed a distinct pattern of 1-DSBs near origins in untreated *mec1-1* cells, but not in wt cells (**Fig. 4d**). Mec1 is involved in response to DNA damage and activation of DNA repair pathway (*17*), so we interpret the pattern in the cells lacking Mec1 as resulting from impaired DNA damage response, that causes replication-caused 1-DSBs that would be transient in wt cells to persist in *mec1-1* cells. Interestingly, in the HU treated cells the effect is opposite, we observe less 1-DSBs in *mec1-1* cells than in wild-type (**Fig. 2a, 4d, and 4e)**, suggesting multiple roles of Mec1 as a replication stress checkpoint.

Role of Mus81 in resolving stalled forks

Mus81, a catalytic component of a eukaryotic structure-selective endonuclease. Mus81 was proposed to cleave stalled replication forks, although there is no agreement if it only cleaves stalled forks, or only reversed ones, or both with the same efficiency (*18-20*). To clarify, we treated *mus81*Δ yeast cells with replication inhibitor HU for 1 hour. Using qDSB-Seq, we quantified the Mus81-induced DSBs. A total of 100 ± 3 1-DSBs per diploid cell was observed in *mus81Δ* cells, 13% less than 115 ±3 DSBs observed in wild-type cells (**Fig. 5a**), indicating that at least 9-21 DSBs/diploid cell in wild-type (13%) are caused by Mus81-mediated cleavage of forks stalled under HU treatment. Removing Mus81 increases an average fork range from 5.7kb to 6.5kb that is by 0.8kb (**Fig. 5b).** Therefore, we hypothesize that Mus81 frequently cleaves reversed forks, with average reversal range of ~0.8kb (**Fig. 5c**), which is consistent with fork reversal range observed by electron microscopy (*21*). This hypothesis also solves an apparent contradiction of our result that fork range is larger in *mus81Δ* than in wild-type (**Fig. 5c**), with the Aladjem group finding that forks are slower in human *mus81Δ* (*22*). In case of 1-DSBs originated from reversed forks, the total distance travelled by the fork (forward and reverse), exceeds the distance travelled by the fork estimated from 1-DSB location, and therefore fork range cannot be used to estimate fork speed if fork reversal is occurring (**Fig. 5c**).

Interestingly, in *mec1-1* cells, removing Mus81 does not further reduce the number of DSBs, indicating that Mus81 is not capable of cleaving stalled forks in absence of Mec1, which may be e.g. necessary for recruiting Mus81. Indeed, fork range in *mec1-1* and *mec1-1 mus81Δ* cells is the same (**Fig. 5d**), in contrast with wild type, where removing Mus81 is increasing fork range significantly (**Fig. 5c**). That result support our hypothesis that Mec1 is needed for Mus81 to cleave stalled forks and that is why both reduction in DSBs and the alteration of fork range is not observed in *mec1-1* cells upon removing Mus81, since Mus81 is anyway incapable of cleaving stalled forks in in Mec1-deficient cells.

Discussion

Several methods for direct DSB-labeling genome-wide that have been developed recently DSB affordable DSB detection with unprecedented single nucleotide resolution. Since DSBs are rare in healthy cells, DSB sequencing data from such cells can be easily overshadowed by a small fraction of the cell population with substantial DNA damage. This challenge is unique to DSB sequencing, as noise orders of magnitude higher than signal of interest is not normally encountered in any other genomic or omics data.

Therefore, to draw correct conclusions from DSB sequencing data, it is crucial to consider not only directly measured average DSB numbers per cell, but also how they are distributed in the population. The best solution is to complement DSB sequencing with additional experiments allowing estimating the distribution of DSB counts across the population of cells. Here, we used pulse-field gel electrophoresis and provided conceptual and mathematical framework for integrating these data. Such analysis is crucial for understanding the physiological relevance of measured data and origins of the observed breaks.

In our case, 1-ended breaks that we can directly link to fork collapse tend to be representative of the major population. This is likely to be a typical situation, if care is exercised to identify only genuine 1-DSBs. Due to their unique nature, originating from collapsed replication forks, the maximal theoretically achievable number of 1-DSBs is twice the number of replication origins, corresponding to 1-DSBs every ~25kb in yeast. On the other hand, 2-DSBs can be also caused by any mechanism, such as irradiation, but also replication stress (e.g. breaking of non-replicated DNA). Some mechanisms, such as radiation, may affect all cells to a similar extent in terms of number of DSBs caused. On the other hand, certain other phenomena, such as replication stress, may lead to a low number of 2-DSBs in majority of cells, but very high number in few cells that likely cannot resolve replication stress and may therefore enter senescence, like in the examples discussed above. 2-DSBs from such cells have high potential to obscure the data. For example, in cells undergoing apoptosis, DNA is fragmented into nucleosome-sized units, thus creating several orders of magnitude more breaks than in a typical cell in a population.

This example shows that careful analysis is needed to interpret DSB sequencing data. Simplistic solutions such as rejecting all 2-DSBs would not be correct, as we showed that some of them are genuine. We also showed that analysis of the genome-wide distribution of DSB patterns allows drawing inference and state hypotheses about mechanisms of their formation. In our case, 2-DSBs in the fork regions are not distributed as we expected them to be if they were caused by fork re-start, but rather are proportional to gDNA density, indicating their likely random distribution in the genome, consistent with the hypothesis they originate from cells with massively fragmented DNA.

Moreover, analysis of DSBs originating from the collapsed forks gives also insights into the dynamics of unperturbed replication forks and cell replication program. We developed methods for highly precise detection of location and positions of the active replication origins using DSB sequencing data, which can be interesting on its own. For example, we discover that in addition to its known role in inhibiting late origins, Mec1 seems to increase significantly, on average twice, the origin efficiencies.

We also show that careful DSB quantification and classification of breaks according to their mechanisms allows new mechanistic insights. Here, we clarified the role of Mus81 in replication stress, showing that Mus81 is preferentially cleaving reversed forks (with average reversal of 0.8kb) and all that this cleavage is completely Mec1-dependent. We also showed that Mus81 is responsible for 13% of DSBs induced by HU and that 19% of such DSBs are Mec1-dependent **(Fig. 5e)**. Mec1-dependent DSBs are also present in absence of Mus81, that is they likely do not require cleaving of reversed replication forks.

In conclusion, we showed how even typically undetectable population heterogeneity can severely contaminate DSB sequencing data and how this problem can be solved by a computational integration of complementary data types, here by including PFGE in the analysis. Due to DSB sequencing data susceptibility to noise from few highly damage cells, to ensure correct interpretation of DSB-sequencing data it is critical that distribution of DSBs across cell population will be interrogated.

Methods

Strains and growth conditions

Yeast strains used in this study (W303), were provided by the Pasero lab. Cells were grown in YPD medium at 25°C until early log phase and were then arrested in G1 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 G1 arrest by addition of 75 μg/ml Pronase (Sigma), 200 mM HU (Abcam) was added 20 min before Pronase release followed by 1 h incubation. The cells were subsequently centrifuged and resuspended in fresh medium to recover for 2 h. After collecting, cells were washed with cold SE buffer (5M NaCl, 500 mM EDTA, pH 7,5) and immediately subjected to DSB labelling.

DSB sequencing. i-BLESS DSB labeling was performed as described in (*7*),(*23*). Sequencing libraries for 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)). 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 i-BLESS (*23*). Additionally, qDSB-sequencing was performed, either using NotI restriction digestion or I-SceI spike-in, as described in (*8*).

Quantitative DSB sequencing

We used qDSB-seq method to normalize and quantify DSBs, as described previously (*8*). To implement qDSB-Seq, we used two approaches. First, we used YBP-275 strain with Gal-inducible I-SceI endonuclease and I-SceI recognition site introduced at ADH4 locus on chromosome VII. Next, we performed i-BLESS and genomic DNA sequencing.

Genomic DNA was used to quantify DSBs by estimating cutting frequency of the enzyme used (*8*).

Libraries preparation and Illumina sequencing

Sequencing libraries for 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)). 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 by 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 i-BLESS (*23*).

Sequencing data analysis

We used *InstantSeq (24)* to ensure sequencing data quality before mapping. Next, *InstantSeq* was used to remove 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 yeast S288C genome and the human hg19 genome using bowtie (*25*) v0.12.2 with the alignments parameters '-m1 –v1' (to exclude ambiguous mapping and low-quality reads). To correct single nucleotide variants of yeast strains, we mapped genomic reads of yeast strain used in our experiments to S288C genome using bowtie v0.12.2 allowing 2 mismatches. The reference base pairs were corrected if the variant was present in > 95% cases. 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 (*24*)) was used to identify genomic regions with a significant difference in normalized read numbers between treatment and control samples.

Identifying DSBs originating from fork collapse

First, we map the i-BLESS-labeled (*7*), sequenced DSB reads separately into the Watson (+) and Crick (-) strands of the reference genome. To predict active origins, we employ a model learned from an asymmetric pattern that centers upon origins (**Fig. 1e)**. To facilitate detecting such pattern, we first filter the sequencing data to remove the high-frequency noise (**Fig. 2b**). In our experience, the majority of noise in DSB sequencing data is high-frequency, often related to nucleosome spacing (**Fig. S1)**. On the other hand, replication related signal shows much lower frequency (at least several kb period). Once the data is transformed into the Fourier transform space (frequency space), the high-frequency signal is removed from thus transformed DSB sequencing data using a Fourier-based smoothed low-pass filter (**Fig. 2b**, center box), a type of filtering often employed in acoustic signal processing. The Fourier-based low-pass filter, which by definition removes high-frequency data, and has thus increased signal-to-noise ratio in our DSB sequencing data. Low-pass filter noise removal method proved very effective, giving results very similar to those obtained by optimizing DSB detection procedure aimed to

minimize noise (**Fig. S2**). The low-pass filtering is especially crucial in case of a weaker signal or noisier data (**Fig. S2**, top panel). Once the data is thus filtered and transformed back to chromosomal coordinates using the inverse Fourier transform (**Fig. 2b**), we applied a sliding window to calculate the 1-DSB ratio ($R_{1\text{-DSB}}$) using the filtered DSB signal (**Eq. 1**) from + and – strands (**Eq. 2-4**). To accommodate variance in origin firing times, and consequently in length of replicated regions, we calculated $R_{1\text{-DSB}}$ score for the multiple consequently in length of replicated regions, we calculated $R_{1\text{-DSB}}$ score for the multiple window sizes and the maximal region replicated by forks originating from a given origin (fork range), is defined. Candidate origins are expected to have a positive $R_{1\text{-DSB}}$ score (the higher score the better, with 1 being maximal), while candidate fork meeting locations should have a negative $R_{1\text{-DSB}}$ score (the lower the better, minimal -1). Finally, to predict high-confidence origins and fork meeting locations, we used hypergeometric test to calculate significance of the $R_{1\text{-DSB}}$ scores for each origin. Accepted candidate origins are selected using this p-value and desired FDR threshold (**Fig. S3)**, $R_{1\text{-DSB}}$ scores are only used to prepare a short list of candidate origins. After identifying the region in which predicted origin is located, smaller steps are used to predict accurate origin location.

To validate our fork pattern and origin predictions, we performed several replication stress experiments in the budding yeast (**Fig. 2a**). We performed fork range and candidate origin identification as above, considering intervals from 1 to 10 kilobases for individual fork ranges. We assumed that origins listed in OriDB (*26*) and confirmed by BrdU staining in our conditions were true positives, while others were false positives, with exception of a few predicted origins that were in OriDB, but not stained by BrdU, which were also assumed true positives. Sensitivity and precision curves for predictions are shown for hydroxyurea (HU) treated cells (**Fig. S4**), that is in a condition when replication is stopped early, allowing only for detection of early origins. Depending on desired FDR, we predicted 144 origins with 100% precision, and 168 origins with 95% precision in wt HU-treated cells and >200 origins in *mus81Δ* cells and *mec1-1* cells. Interestingly, at 95% precision we predicted 22 origins that were not detected by BrdU staining in the same conditions, but are listed as origins in OriDB, so we consider them genuine. Not all of origins we currently label as false positives must be errors, they may be yet-unknown weak/dormant origins we discovered due to high-sensitivity of our methods. In further analysis, we used *p*-value thresholds not allowing any false positives (FDR=0, precision 100%).

Being able to predict active replication origins in a given condition allowed us to distinguish between breaks resulting directly from replication fork collapse and others. To this end, first "fork range" was estimated (see below). Briefly, fork range is an interval ranging from 1 kb to 10 kb on each side of an origin, where we observe a statistically significant preference for 1-DSBs consistent with fork collapse in that region. Next, 1-DSBs assumed to originate from collapsed forks are counted by setting strict criteria: only 1-DSBs in fork range are counted, and only those that cannot be a part of a 2-DSB, that is the number of 1-DSBs with the orientation opposite to expected fork direction is subtracted from number of the 1-DSB with the consistent direction.

Fourier-based low-pass filter

The i-BLESS data contain high frequency quasi-periodic components related to nucleosome spacing and potential sequencing artifacts (apparent in the power spectrum of the read count distribution, **Fig. S1**). To remove the spurious signal from further analysis, we implemented and applied a low-pass filtering approach. The low-pass filter was performed by multiplying the Fourier Transform, *D(f)*, of the read density *d(x)*, by a kernel *k(f)*, defined as a smoothed step function

$$k(f) = \exp(-(c/f)^6) \qquad \text{Equation (1),}$$

where *f* is frequency and *c* is the cut-off frequency (here we used $c=2\pi/5000\ nt^{-1}$). After filtering in the frequency domain, the data are transformed back to position in the real domain (chromosomal coordinate) by inverse Fourier transform, thus the filtered read density $d_{filtered}(x)$ is calculated as the inverse Fourier transform of

$$D_{filtered}(f) = D(f)\ k(f).$$

The exponent 6 in the kernel (**Eq. 1**) has been chosen empirically to assure a smooth transition while maintaining an efficient separation of sequencing and nucleosome-related noise from the low-frequency components related to the replication domains. The Fourier transforms were implemented by real-to-complex routines from the FFTW library.

This Fourier-based low-pass filter, which by definition removes high-frequency data, increases signal-to-noise ratio in our DSB sequencing data. Low-pass filter noise removal method proved very effective, giving results very similar to those obtained by optimizing DSB detection procedure aimed to minimize noise (**Fig. S2**). The low-pass filtering is especially crucial in case of a weaker signal or noisier data (**Fig. S2**, top panel).

Predicting replication-induced 1-DSBs.

We first calculated the numbers of the DSB reads mapped to the left of a candidate origin, mapped to the Watson (WL) and Crick strands (CL), respectively. Analogously, CR and WR were calculated. To eliminate 2-DSBs, we use the difference of read depth between both strands as an estimate of number of 1-DSBs in a region:

$$S_{1-DSB} = (WL - CL) + (CR - WR), \qquad \text{Equation (2)}$$

Analogously, the estimated number of 2-DSBs is:

$$S_{2-DSB} = CL + WR, \qquad \text{Equation (3)}$$

These are conservative estimates; they may lead to underestimating DSBs caused by fork collapse in regions where breaks from forks travelling in both directions are present.

To identify replication origins and fork meeting locations, we defined a ratio for 1-ended DSB ($R_{1\text{-}DSB}$), based on the model of replication-induced 1-DSBs (**Fig. 1c**):

$$R_{1-DSB} = \frac{S_{1-DSB}}{S_{total}}, \qquad \text{Equation (4)}$$

$$R_{2-DSB} = \frac{S_{2-DSB}}{S_{total}}, \qquad \text{Equation (5)}$$

$R_{1\text{-DSB}}$ ranges from -1 to 1. $R_{2\text{-DSB}}$ ranges from 0 to 1. In the case of an ideal candidate origin with only replication-induced 1-ended DSB, $R_{1\text{-DSB}} = 1$, $R_{2\text{-DSB}} = 0$. In the case of an ideal fork meeting region, $R_{1\text{-DSB}} = -1$, $R_{2\text{-DSB}} = 1$. In the case of an ideal 2-DSB only region, $R_{1\text{-DSB}} = 0$, $R_{2\text{-DSB}} = 0.5$ (**Fig. 2b**).

$R_{1\text{-DSB}}$ score encodes a simplified origin model. More specific models can be constructed, but since they would depend on origin firing time and progress of replication at the time when sample was taken. Therefore, we prefer a more general $R_{1\text{-DSB}}$ score, which proved very efficient in identifying candidate origins. We used sliding windows ranging from 2 kb to 20 kb to identify candidate origins (local maxima of $R_{1\text{-DSB}}$ score). Once candidate origins were identified, the fork range was defined as the longest range (in multiples of 1 kb), where the substantial (>0.5%) enrichment of 1-DSBs was observed in each consecutive 1 kb interval. Then, significance of the enrichment of 1-ended DSB in thus defined fork range was calculated using hypergeometric test, only candidate origins with significant p-values for 1-DSB enrichment were selected as predicted origins.

Pulsed-field gel electrophoresis

Pulsed-field gel electrophoresis (PFGE) was performed as described previously (*27*). Gel electrophoresis image was analyzed using software GelAnalyzer (http://www.gelanalyzer.com/) and ImageJ (https://imagej.nih.gov/ij/index.html). Probes of ARS305, 307, 315 on chrIII were used.

References

1. A. Aguilera, B. Gomez-Gonzalez, Genome instability: a mechanistic view of its causes and consequences. *Nature reviews. Genetics* **9**, 204-217 (2008).
2. I. M. Slaymaker *et al.*, Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-88 (2016).
3. N. Crosetto *et al.*, Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature methods* **10**, 361-365 (2013).
4. E. A. Hoffman, A. McCulley, B. Haarer, R. Arnak, W. Feng, Break-seq reveals hydroxyurea-induced chromosome fragility as a result of unscheduled conflict between DNA replication and transcription. *Genome research* **25**, 402-412 (2015).
5. S. V. Lensing *et al.*, DSBCapture: in situ capture and sequencing of DNA breaks. *Nature methods* **13**, 855-857 (2016).

6. A. Canela *et al.*, DNA Breaks and End Resection Measured Genome-wide by End Sequencing. *Molecular cell* **63**, 898-911 (2016).
7. A. Biernacka *et al.*, i-BLESS method: ultra-sensitive method for detection of DNA double-strand breaks. *In preparation*, (2017).
8. Y. Zhu *et al.*, Quantitative DSB sequencing (qDSB-Seq): a method for genome-wide accurate estimation of absolute DNA double-strand break frequencies per cell. *bioRxiv*, (2017).
9. W. Lee *et al.*, A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics* **39**, 1235-1244 (2007).
10. Y. Zhu *et al.*, Quantative DSB sequencing (qDSB-Seq): a method for genome-wide accurate estimation of absolute DNA double-strand break frequencies per cell. (2017).
11. M. Gauster *et al.*, Dysregulation of placental endothelial lipase and lipoprotein lipase in intrauterine growth-restricted pregnancies. *The Journal of clinical endocrinology and metabolism* **92**, 2256-2263 (2007).
12. C. Santocanale, J. F. Diffley, A Mec1- and Rad53-dependent checkpoint controls late-firing origins of DNA replication. *Nature* **395**, 615-618 (1998).
13. K. Shirahige *et al.*, Regulation of DNA-replication origins during cell-cycle progression. *Nature* **395**, 618-621 (1998).
14. J. A. Tercero, J. F. Diffley, Regulation of DNA replication fork progression through damaged DNA by the Mec1/Rad53 checkpoint. *Nature* **412**, 553-557 (2001).
15. J. C. Randell *et al.*, Mec1 is one of multiple kinases that prime the Mcm2-7 helicase for phosphorylation by Cdc7. *Molecular cell* **40**, 353-363 (2010).
16. M. Clerici, C. Trovesi, A. Galbiati, G. Lucchini, M. P. Longhese, Mec1/ATR regulates the generation of single-stranded DNA that attenuates Tel1/ATM signaling at DNA ends. *The EMBO journal* **33**, 198-216 (2014).
17. A. Marechal, L. Zou, DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harb Perspect Biol* **5**, (2013).
18. K. T. Ehmsen, W. D. Heyer, Saccharomyces cerevisiae Mus81-Mms4 is a catalytic, DNA structure-selective endonuclease. *Nucleic acids research* **36**, 2182-2195 (2008).
19. B. Froget, J. Blaisonneau, S. Lambert, G. Baldacci, 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. K. Hanada *et al.*, The structure-specific endonuclease Mus81 contributes to replication restart by generating double-strand DNA breaks. *Nature structural & molecular biology* **14**, 1096-1104 (2007).
21. J. M. Sogo, M. Lopes, M. Foiani, Fork reversal and ssDNA accumulation at stalled replication forks owing to checkpoint defects. *Science* **297**, 599-602 (2002).
22. H. Fu *et al.*, The DNA repair endonuclease Mus81 facilitates fast DNA replication in the absence of exogenous damage. *Nature communications* **6**, 6746 (2015).

23. A. Mitra, M. Skrzypczak, K. Ginalski, M. Rowicka, Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. *PloS one* **10**, e0120520 (2015).
24. A. Mitra *et al.*, in *submitted*. (2017).
25. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).
26. C. C. Siow, S. R. Nieduszynska, C. A. Muller, C. A. Nieduszynski, OriDB, the DNA replication origin database updated and extended. *Nucleic acids research* **40**, D682-686 (2012).
27. A. Lengronne, P. Pasero, A. Bensimon, E. Schwob, Monitoring S phase progression globally and locally using BrdU incorporation in TK(+) yeast strains. *Nucleic acids research* **29**, 1433-1442 (2001).

Figures and Tables

Figure 1. Integrative analysis of DSB formation. (a) Overall approach. Different data sources are integrated computationally, DSB mechanisms and their prevalence in population are inferred. **(b) Main data sources:** DSB-sequencing and pulsed-field gel electrophoresis. **(c) Differences between DSB-sequencing and PFGE.** DSB sequencing reports and average number if DSBs in population, so the result may not reflect typical DSB number in population, if even small subpopulation of cells with high number of breaks are present (e.g. dead cells). PFGE detects lengths of DNA fragments, so it does not have a problem between distinguishing if breaks occurring in a small population of cells or are representative of majority of population. **(d) 2-ended and 1-ended DSBs. (e) Expected DSB pattern resulting from fork collapse.** The double-stranded breaks are generated when replication fork passes through the single-stranded breaks (nicks). The collapse of forks travelling to left and right will result in reads mapped to the + and - strands, respectively. The expected total pattern is shown below.

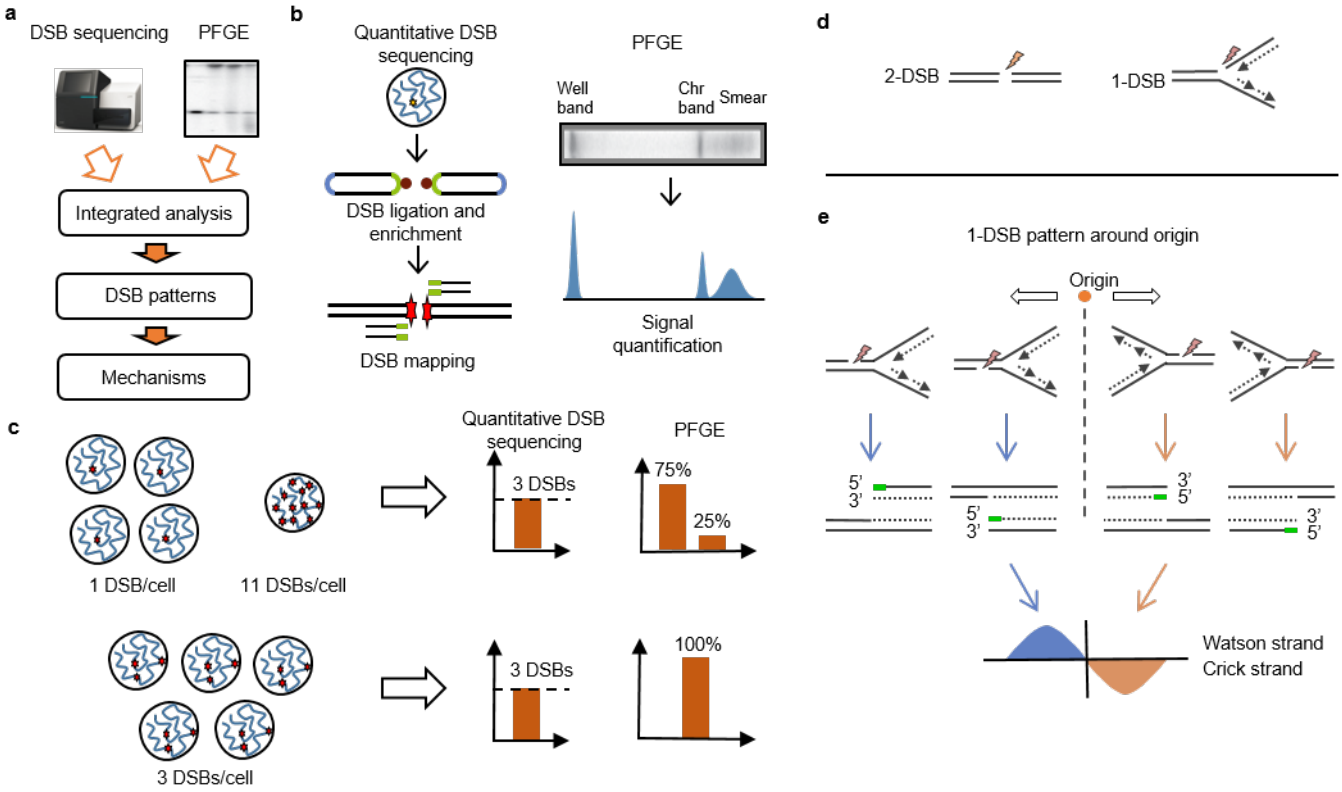
Figure 2. Quantification of 1-DSBs in replication stress and release conditions and active origin inference from 1-DSB signal. (a) Experimental conditions and results of 1- and 2-DSB quantification. Yeast cells were synchronously released from α-factor G_1 phase arrest and treated with, and recovered from HU. Quantification of 1- and 2-DSBs is explained below. **(b) Active origin inference after noise filtering**. DSB sequencing data were transformed to frequency signal using Fourier transform, then low pass filter and inverse Fourier transform was executed to remove high frequency signal. $R_{1\text{-}DSB}$ score was calculated to identify origin and range of replication. **(c) Numbers of origins detected in HU-treated samples**.

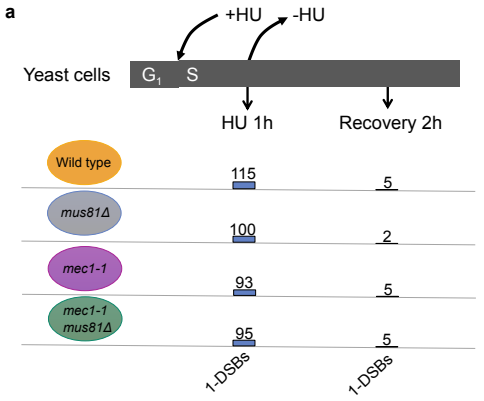
Figure 3. Comparing qDSB-Seq and PFGE results. (a) Results of DSB sequencing and PFGE for different chromosomal fragments: (1) Unbroken linear DNA; (2) Linear DNA with 2-DSB; (3) Replicated DNA but unbroken; (4) Replicated DNA with 1-DSB. **(b)** An example of use of the Poisson distribution to simulate population-wide DSB distributions using they measured number per chromosome III (here 1.13). **(c)-(d)** Comparison of predicted (Poisson model) and observed (PFGE) well band and smear

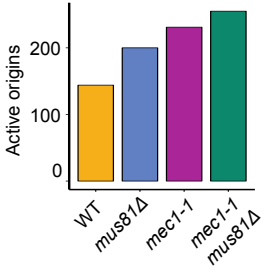
band on chrIII: **(c)** Well band simulation based on qDSB-seq 1-DSB data; **(d)** Smear simulation based on qDSB-seq fork 2-DSBs. **(e)** Cell population of 1-DSBs and 2-DSBs. Small subpopulation with high number of 2-DSBs cannot be detected by PFGE.

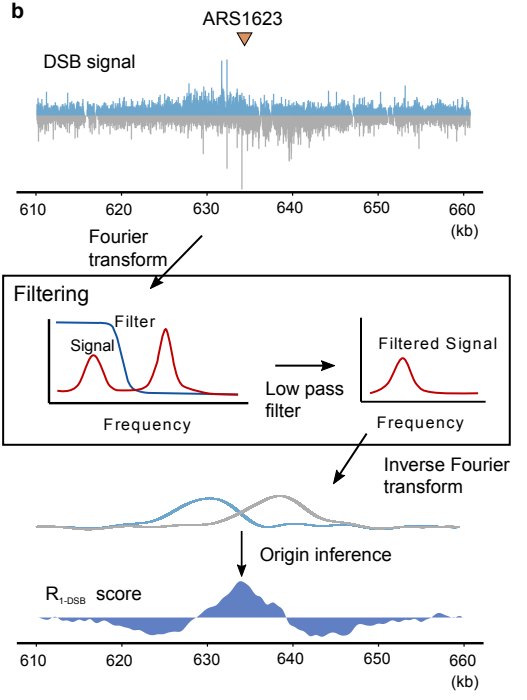
Figure 4. The roles of Mec1 in origin firing and response to replication stress. (a) Correlation between replication-dependent 1-DSBs and firing time. Lower 1-DSB level was observed in *mec1-1* mutant cells compared with wild-type. **(b)** Meta-profile of 1-DSBs per diploid cell for early and late origins. The average of 1-DSBs in aligned origins was presented. Early origins show faster fork progression than late origins. **(c)** Increase in fork range length in Mec1-deficient cells in comparison to wild-type cells for active and early origins. **(d-e)** Meta-profile of 1-DSBs near origins for wild-type and Mec1-deficient cells, untreated **(d)** and HU-treated **(e). (f)** Visualization of fork ranges and 1-DSBs for selected origins. Mec1 show temporal regulation for both early and late origins in chrXVI.

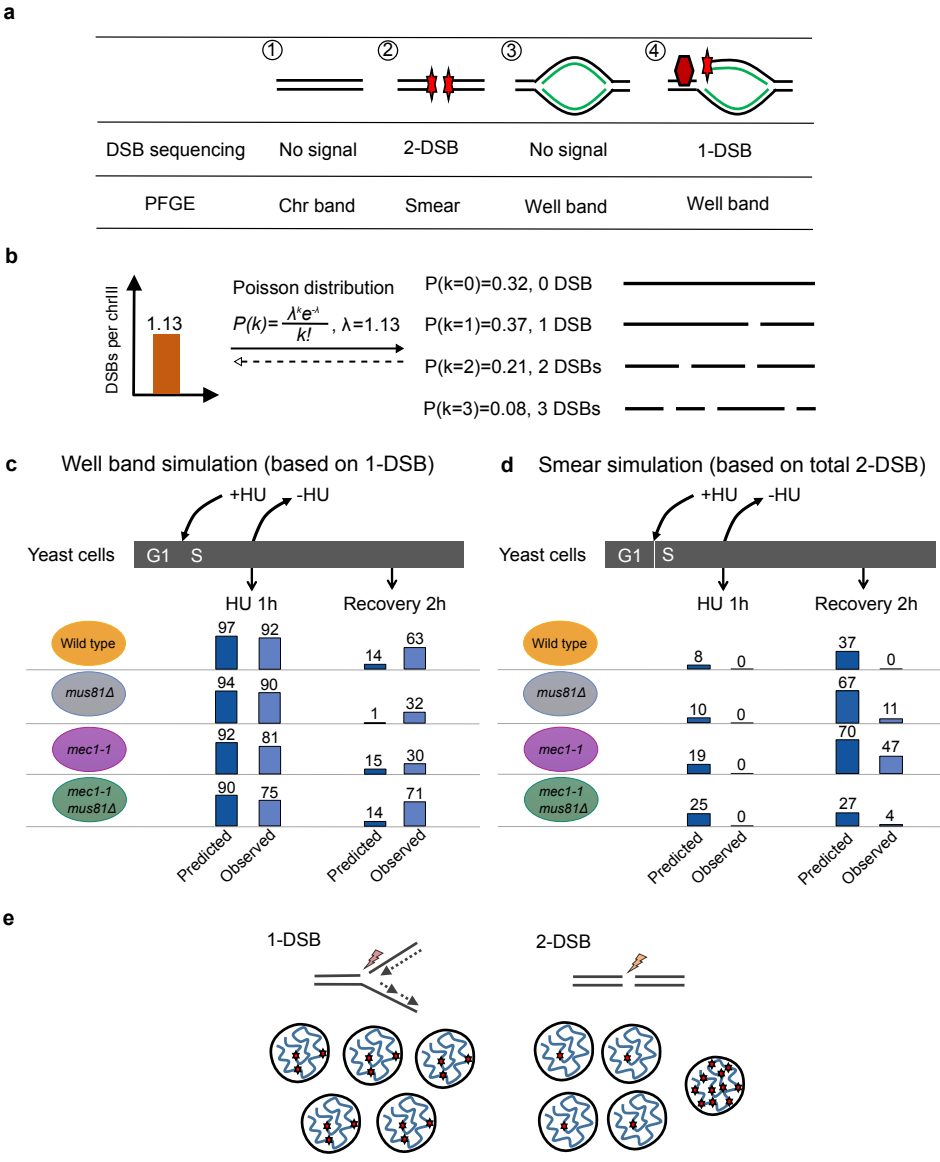
Figure 5. Mus81 cleaves reversed replication fork causing 1-ended DSBs. (a) Quantification of 1-DSBs per cell. **(b)** Longer fork range in *mus81Δ* than WT, indicating that in wild-type cells Mus81 protein is cleaving predominantly reversed forks. **(c)** A model of how Mus81 cleaving reversed replication fork would impact observed 1-DSB patterns. In *mus81Δ*, fork collapses due to other factors, such as a DNA nick on single strand DNA. **(d)** Similar fork range length in *mec1-1* and *mec1-1 mus81Δ* cells. **(e)** Mec1- and Mus81-dependent DSBs in wild-type HU 1h sample.

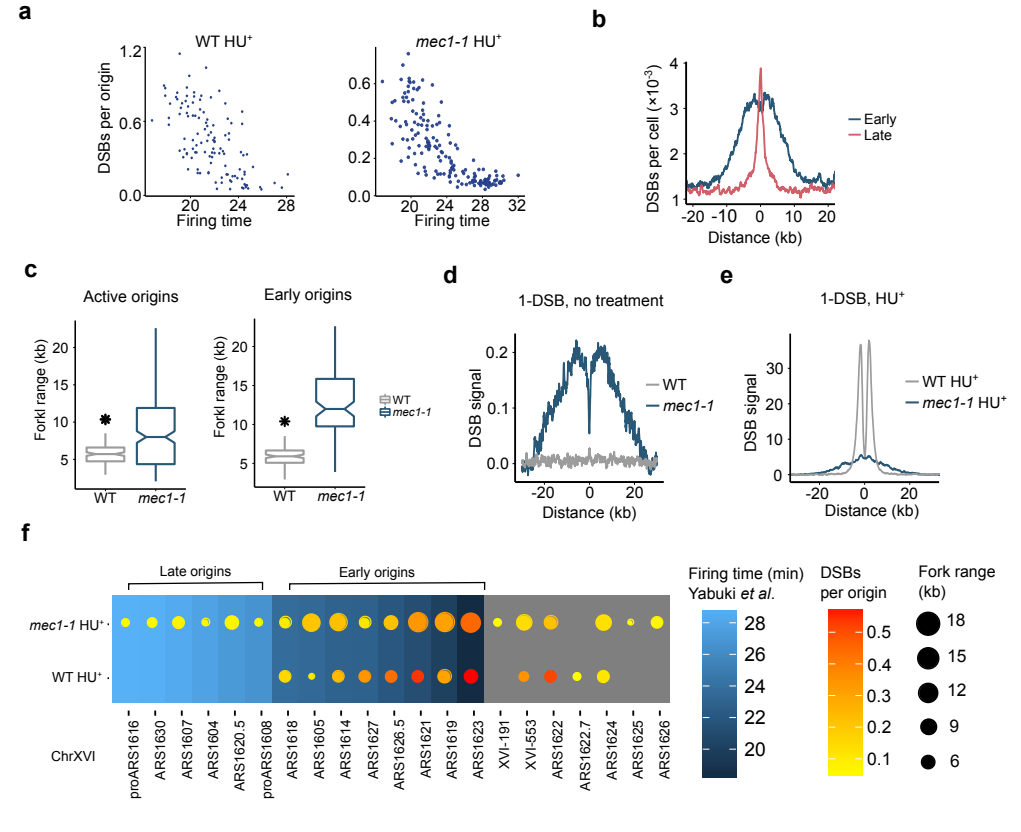

a
DSB sequencing
PFGE
Integrated analysis
DSB patterns
Mechanisms
b
Quantitative DSB sequencing
DSB ligation and enrichment
DSB mapping
PFGE
Well band
Chr band
Smear
Signal quantification
c
1 DSB/cell
11 DSBs/cell
Quantitative DSB sequencing
3 DSBs
PFGE
75%
25%
3 DSBs/cell
3 DSBs
100%
d
2-DSB
1-DSB
e
1-DSB pattern around origin
Origin
5'
3'
Watson strand
Crick strand

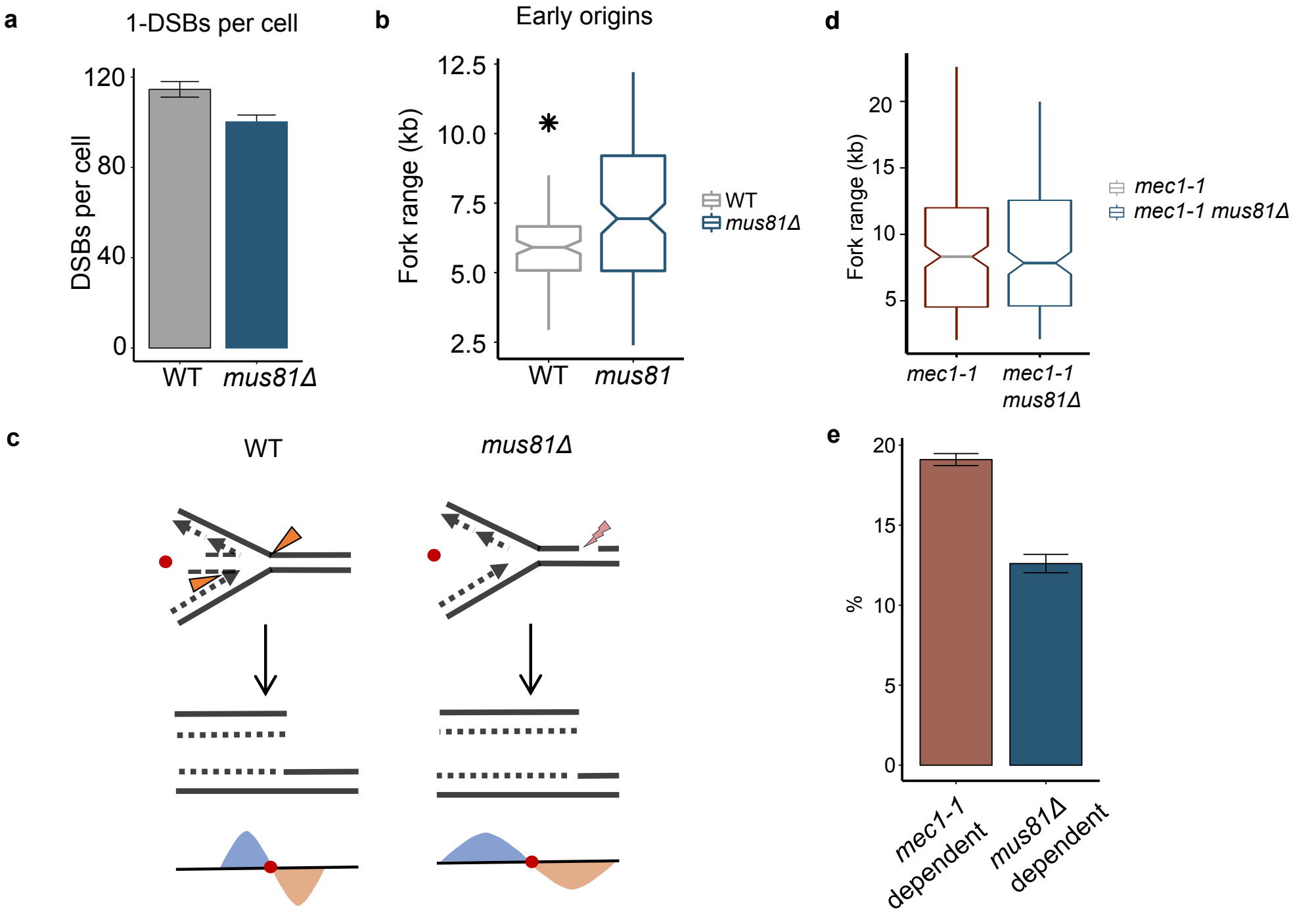

a
+HU
-HU
Yeast cells
G1
S
HU 1h
Recovery 2h
Wild type
115
5
mus81Δ
100
2
mec1-1
93
5
mec1-1
mus81Δ
95
5
1-DSBs
1-DSBs


c
Active origins
200
100
0
WT
mus81Δ
mec1-1
mec1-1
mus81Δ


b
ARS1623
DSB signal
610
620
630
640
650
660
(kb)
Fourier transform
Filtering
Filter
Signal
Frequency
Low pass filter
Filtered Signal
Frequency
Inverse Fourier transform
Origin inference
R1-DSB score
610
620
630
640
650
660
(kb)


a
①
②
③
④
DSB sequencing
No signal
2-DSB
No signal
1-DSB
PFGE
Chr band
Smear
Well band
Well band
b
DSBs per chrIII
1.13
Poisson distribution
P(k)= λ^k e^-λ / k!, λ=1.13
P(k=0)=0.32, 0 DSB
P(k=1)=0.37, 1 DSB
P(k=2)=0.21, 2 DSBs
P(k=3)=0.08, 3 DSBs
c
Well band simulation (based on 1-DSB)
+HU
-HU
Yeast cells
G1
S
HU 1h
Recovery 2h
Wild type
mus81Δ
mec1-1
mec1-1 mus81Δ
97
92
14
63
94
90
1
32
92
81
15
30
90
75
14
71
Predicted
Observed
Predicted
Observed
d
Smear simulation (based on total 2-DSB)
+HU
-HU
Yeast cells
G1
S
HU 1h
Recovery 2h
Wild type
mus81Δ
mec1-1
mec1-1 mus81Δ
8
0
37
0
10
0
67
11
19
0
70
47
25
0
27
4
Predicted
Observed
Predicted
Observed
e
1-DSB
2-DSB


a
WT HU+
mec1-1 HU+
DSBs per origin
Firing time
b
DSBs per cell (×10-3)
Early
Late
Distance (kb)
c
Active origins
Early origins
Forkl range (kb)
WT
mec1-1
d
1-DSB, no treatment
DSB signal
Distance (kb)
e
1-DSB, HU+
WT HU+
mec1-1 HU+
f
Late origins
Early origins
mec1-1 HU+
WT HU+
ChrXVI
proARS1616
ARS1630
ARS1607
ARS1604
ARS1620.5
proARS1608
ARS1618
ARS1605
ARS1614
ARS1627
ARS1626.5
ARS1621
ARS1619
ARS1623
XVI-191
XVI-553
ARS1622
ARS1622.7
ARS1624
ARS1625
ARS1626
Firing time (min)
Yabuki et al.
DSBs per origin
Fork range (kb)


a
1-DSBs per cell
DSBs per cell
120
80
40
0
WT
mus81Δ
b
Early origins
Fork range (kb)
12.5
10.0
7.5
5.0
2.5
*
WT
mus81
WT
mus81Δ
d
Fork range (kb)
20
15
10
5
mec1-1
mec1-1 mus81Δ
mec1-1
mec1-1 mus81Δ
c
WT
mus81Δ
e
%
20
15
10
5
0
mec1-1 dependent
mus81Δ dependent