Repeat Expansions Confers WRN Dependence in Microsatellite Unstable Cancers

Niek van Wietmarschen[1,*], Sriram Sridharan[1,*], William J. Nathan[1,2*], Anthony Tubbs[1,*], Edmond M. Chan[3,4*], Elsa Callen[1], Wei Wu[1], Frida Belinky[1], Veenu Tripathi[1], Nancy Wong[1], Kyla Foster[4], Javad Noorbakhsh[4], Kiran Garimella[4], Abimael Cruz-Migoni[2], Joshua A. Sommers[5], Yongqing Huang[4], Ashir A. Borah[4], Jonathan T. Smith[4], Jeremie Kalfon[4], Nikolas G. Kesten[3], Kasper Fugger[6], Robert L. Walker[7], Egor Dolzhenko[8], Michael A. Eberle[8], Bruce E. Hayward[9], Karen Usdin[9], Catherine H. Freudenreich[10], Robert M. Brosh, Jr.[5], Stephen C. West[6], Peter J. McHugh[2], Paul S. Meltzer[7], Adam J. Bass[3,4] & André Nussenzweig[1,‡]

[1] Laboratory of Genome Integrity, National Cancer Institute, NIH, Bethesda, MD, USA

[2] Department of Oncology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom

[3] Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston MA, USA

[4] Broad Institute of Harvard and MIT, Cambridge, MA, USA

[5] Laboratory of Molecular Gerontology, National Institute on Aging, NIH, Baltimore, Maryland, United States of America

[6] DNA Recombination and Repair Laboratory, The Francis Crick Institute, London, U.K.

[7] Genetics Branch, National Cancer Institute, NIH, Bethesda, MD, USA.

[8] Illumina Incorported, San Diego CA, USA

[9] Laboratory of Cell and Molecular Biology, National Institute of Diabetes, Digestive and Kidney Diseases, NIH, Bethesda MD, USA

[10] Department of Biology, Tufts University, Medford, MA, USA.

***Authors contributed equally to this work**

‡Correspondence: ***andre_nussenzweig@nih.gov***

The RecQ DNA helicase WRN is a synthetic lethal target for cancers with microsatellite instability (MSI), a form of genetic hypermutability arising from impaired mismatch repair[1-4]. WRN depletion induces widespread DNA double strand breaks (DSBs) in MSI cells, leading to cell cycle arrest and/or apoptosis. However, the mechanism by which WRN protects MSI cancers from DSBs remains unclear. Here, we demonstrate that TA-dinucleotide repeats are highly unstable in MSI cells and exhibit surprisingly large-scale expansions, distinct from previously described insertion/deletion mutations of a few nucleotides[5]. We show that expanded TA repeats form non-B DNA secondary structures that stall replication forks, activate the ATR checkpoint kinase, and necessitate unwinding by the WRN helicase. In the absence of WRN, the expanded TA-dinucleotide repeats are susceptible to MUS81 nuclease cleavage, leading to massive chromosome shattering. Thus, our study uncovers a distinct biomarker within MSI tumors that underlies the synthetic lethal dependence on WRN, thereby supporting the development of WRN-based therapeutics.

MSI is characterized by hypermutability of short repetitive DNA sequences scattered throughout the genome. MSI arises from deficiency in DNA mismatch repair (MMR) and contributes to the formation of many types of cancers including colorectal cancer (15%), endometrial (20-30%), gastric cancers (15%), and ovarian cancers (12%)[6]. Recent studies demonstrate that multiple cancer types with MSI are reliant on the WRN helicase activity for survival[1-4]. WRN is a member of the RecQ family of DNA helicases including WRN, BLM, and RECQL4, which when mutated manifest as the distinct chromosome instability disorders Werner syndrome, Bloom syndrome, and Rothmund-Thomson syndrome, respectively[7]. RecQ helicases do not display sequence specificity but resolve non-canonical secondary DNA structures such as bubbles, Holliday junctions, and G-quadruplexes that may be encountered during replication and recombination. However, the mechanism by which WRN helicase is required to protect chromosomal integrity of MSI, but not microsatellite stable (MSS), cancers is not understood.

We previously demonstrated that MMR restoration only partially rescued MSI cells from WRN depletion[2]. Recent studies have reported discrepant results on the effect of acute *MLH1* silencing to sensitize MSS cells to WRN depletion[1,3]. We therefore evaluated *WRN* dependency in human primary stomach epithelial cells following knockout of the MMR genes *MLH1* or *MSH2*. After 4 months of culture, *MLH1* or *MSH2* knockout cells failed to develop a dependency upon *WRN* for survival (Extended Data Fig. 1a, b). These data suggest that instead of WRN loss being simply synthetic lethal with impaired MMR, a "genomic scar" may gradually accumulate in MSI cancers that requires WRN's role as a structure-specific helicase.

WRN deficiency generates recurrent DSBs and extensive end-resection in MSI cells

Loss of viability in MSI cells upon *WRN* silencing is associated with a decrease in proliferation and accumulation of DNA double strand breaks (DSBs)[1-4]. Consistently, we found that WRN depletion using a doxycycline-inducible *WRN* shRNA expressed in the MSI KM12 cell line[2] resulted in decreased DNA synthesis and high levels of the DSB marker, KAP1 phosphorylation at Ser 824 (pKAP1)[8], predominantly in the G2/M phase of the cell cycle (Extended Data Fig. 1c-e). In contrast, expression of a seed control WRN.C911 shRNA[2] did not substantially induce DSBs in KM12 (Extended Data Fig. 1f). Analysis of mitotic spreads (n=100) revealed that all chromosomes were shattered in approximately 35% of WRN-depleted KM12 cells (Fig. 1a and Fig. 2a). In contrast, chromosome shattering was not evident in microsatellite stable (MSS) colon cancer line SW837 upon WRN depletion (Fig. 1a and Extended Data Fig. 1e, n=100 metaphases).

To determine whether WRN is necessary to unwind specific regions of the genome to prevent DNA breakage, we queried recurrent sites of DSBs by END-seq[9]. MSI cell lines KM12 and HCT116 show little endogenous DNA breakage; however, upon WRN depletion by either shRNA or siRNA, recurrent DSBs were detected at specific locations throughout the genome (Fig. 1b, Extended Data Fig. 2a). Moreover, END-seq peak intensities were highly reproducible among different experiments (Extended Data Fig. 2b). Using the criteria of 20-fold enrichment over non-treated cells (NT; WRN proficient), END-seq peaks overlapped significantly between HCT116 and KM12 (Extended Data Fig. 2a,c). In contrast, WRN depletion in SW837 did not substantially induce DSBs (Extended Data Fig. 2d). Thus, loss of WRN induces DSBs at recurrent genomic loci that are reproducible across distinct MSI cancer cell lines.

DNA breaks accumulate around TA dinucleotide repeats

END-seq peaks displayed the characteristic pattern of positive and negative strand reads representing the right and left end of DSBs, respectively (Fig. 1c). END-seq reads spread outwards from DSB sites in a pattern consistent with DNA end resection[9,10]. To confirm this, we mapped sites of single-strand DNA (ssDNA) bound by Replication Protein A (RPA)[11,12]. Strikingly, 79% of ssDNA peaks overlapped with END-seq peaks, with resection lengths averaging 500 bp and extending up to 5 kb (Fig. 1b-c, Extended Data Fig. 2e-f). Moreover, polarity of RPA binding was indicative of the accumulation of 3'-overhangs (Extended Data Fig. 2f). Thus, loss of WRN in MSI cells leads to DSBs with extensive 5'-3' end-processing.

The left- and right-end of the DSBs were separated from each other by a variable distance (Fig. 1c, Extended Data Fig. 2g). Interestingly, this "gap" harbored a major reduction in sequencing reads, suggesting that DSBs occur at the borders of these regions. We therefore searched for specific DNA motifs in the gap region in the hg19 reference genome, which revealed a dominant TA dinucleotide repeat motif (Fig. 1d) with a median repeat length of 51 bp (Extended Data Fig. 2h). Nearly all breaks associated with WRN-deficiency in both KM12 and HCT116 cells occurred at TA dinucleotide repeats (Fig. 1e). However, only about 8% of all TA repeats in the reference genome (5,400 out of 66,644) were associated with DSB (Extended Data Fig. 2h). We conclude that DSBs flank a fraction of TA repeats. Hereafter, we refer to these sites as "broken (TA)n repeats," and to the DSBs themselves as "TA breaks".

Cruciform structures form at (TA)n repeats in plasmids in *E. coli* and yeast when their length exceeds roughly 20-22 repeat units[13-15]. Long (TA)n tracts also cause replication fork stalling and chromosome fragility at late replicating common fragile sites (CFS)[16-18]. Exogenous replication stress further enhances replication fork collapse at simple repeats[12,19], including CFSs. Accordingly, we found that WRN depletion induces DNA breakage precisely at (TA)n repeats within multiple CFSs (Extended Data Fig. 3a) and at palindromic TA-rich repeats (Extended Data Fig. 3b), which have been proposed to form cruciform

structures[20]. These data suggest that (TA)n repeats at these sites might fold into secondary structures that are targeted by WRN.

MUS81-EME1 shatters chromosomes in WRN-depleted MSI cells

MUS81-EME1 is a structure-specific endonuclease that processes late recombination intermediates at CFSs[21]. MUS81-EME1 forms a complex with the scaffolding protein SLX4 that hyperactivates it at the G2/M boundary[22]. Recent studies indicate that the yeast MUS81-EME1 homolog (Mus81-Mms4) causes DSBs at (TA)n repeats when the tract exceeds a threshold length for forming cruciform structures[17]. To determine whether the DSBs that accumulate in WRN-depleted MSI cells are MUS81- and SLX4-dependent, we depleted these factors prior to WRN depletion (Extended Data Fig. 4a). MUS81 or SLX4 depletion dramatically reduced chromosome shattering (Fig. 2a, Extended Data Fig. 4b). Consistent with this result, depleting MUS81 and SLX4 strongly reduced pKAP1 signaling (Extended Data Fig. 4c) and DSB formation at (TA)n repeats (Fig. 2b, Extended Data Fig. 4d). Thus, MUS81 and SLX4 nucleases induce toxic chromosome breakage when WRN is depleted from MSI cells.

MSI correlates with susceptibility to recombinant MUS81-EME1 cleavage

To test whether MUS81-EME1 acts directly on secondary structures in MSI cells, we treated agarose-embedded KM12 DNA with recombinant MUS81-EME1 *in situ* prior to performing END-seq. MUS81-EME1 promotes the resolution of cruciform structures by a nick and counter-nick mechanism (Extended Data Fig. 4e), and we found that MUS81-EME1 generated recurrent and reproducible DSBs (Fig. 2c-d , Extended Data Fig. 4f). These overlapped strikingly with DSBs generated from WRN depletion in these cells (Extended Data Fig. 4g-h). These data show that secondary structures accumulate and can be cleaved by

MUS81-EME1 even at baseline conditions in MSI cells. In contrast to DSBs that were highly resected after WRN depletion, *in situ* cleavage by MUS81-EME1 led to accumulation of reads precisely adjacent to the border of (TA)n repeats (Fig. 2d, Extended Data 4i). Thus, the broad distribution of END-seq reads in the WRN-depleted samples are indeed caused by 5' end-processing *in vivo* post MUS81 cleavage.

We hypothesized that WRN is recruited to unwind DNA secondary structures before they can be cleaved by physiologically active MUS81. To test this, we incubated agarose-embedded KM12 DNA with recombinant human WRN prior to MUS81-EME1 treatment *in situ,* followed by END-seq detection of DSBs. While incubation with WRN alone did not result in DSB formation, WRN pre-treatment substantially decreased MUS81-EME1-cleavage (Fig. 2e, Extended Data Fig. 4j). These results suggest that WRN melts secondary DNA structures at (TA)n repeats in MSI cells.

If structure-forming (TA)n repeats are responsible for massive breakage in MSI lines, there should be fewer such structures in MSS cells. To test this, we performed *in situ* MUS81-EME1 cleavage assay in two MSS cell lines (SW837 and RPE-1) and compared these to MSI cell lines KM12 and HCT116. While the MSI cells displayed an overlapping set of strong MUS81-EME1 cleavage sites (Extended Data Fig. 5a, b), the substrates for MUS81-EME1 cleavage were significantly reduced in the MSS cells. Thus, structure-forming (TA)n repeats accumulate in much higher abundance in MSI compared to MSS cells.

Structure forming (TA)n repeats are susceptible to replication fork stalling and collapse

DNA polymerase stalling generates RPA-bound ssDNA, which activates checkpoint kinase ataxia telangiectasia and Rad3 related (ATR) to protect the replication fork. RPA ChIP-seq in WRN-proficient KM12 cells revealed an enrichment of RPA in the vicinity of broken (TA)n sites, suggesting that stalled forks spontaneously accumulate at these sites (Fig. 3a). WRN is

recruited to stalled replication forks in a manner that requires ATR phosphorylation[23]. Consistent with this, HCT116 cells expressing WRN mutants WRN^{3A} or WRN^{6A} containing alanine substitutions at identified ATR phosphorylation sites[23] showed an increase in KAP1 phosphorylation when endogenous WRN was ablated by an siRNA targeting the 5' untranslated region (Extended Data Fig. 5c). We therefore hypothesized that stalled forks at structure-forming (TA)n repeats would be susceptible to fork collapse upon ATR kinase inhibition (ATRi). To test this, we treated MSI and MSS cells with ATRi as well as low dose aphidicolin (APH) to partially inhibit DNA polymerase elongation. As determined by END-seq, replication forks collapsed into DSBs preferentially at (TA)n repeats, including those within CFSs (Fig. 3b, c, Extended Data Fig. 5d). The frequency fork collapse at (TA)n repeats in MSI cells was at least 30-fold higher than in MSS cells (Fig. 3b, c), and these sites largely overlapped with DSBs generated in the absence of WRN (Extended Data Fig. 5e). Thus, secondary structure-forming (TA)n repeats are associated with replication fork stalling which is overcome by WRN, possibly through activation by ATR.

(TA)n repeats inhibit DNA synthesis *in vitro*

In MSI cancers, insertion/deletions of a few nucleotides are commonly found at mononucleotide repeats[5]. Among dinucleotides, (TA)n repeats are reported to be the least frequently altered (Extended Data Fig. 6a)[24]. To determine why (TA)n repeats are susceptible to breakage, we performed a PCR-based size analysis of several (TA)n repeats in a panel of MSI and MSS cell lines (Fig. 4a, Extended Data Fig. 6b, Supplementary Table 1). We examined eight (TA)n repeats that were broken and three that were not broken upon WRN depletion in MSI cells. All 11 sites that we queried harbored uninterrupted stretches of more than 20 (TA)n units (in the hg19 reference genome), and were amplified with primers flanking the (TA)n repeat.

While MSS cell lines showed the predicted PCR products at broken and non-broken sites (Fig. 4a, Extended Data Fig. 6b), many recurrently broken sites failed to be amplified in KM12 cells (Fig. 4a). These same sites frequently could not be amplified with genomic DNA from other MSI cell lines, highlighting that identical sites are affected across distinct cancers (Extended Data Fig. 6b). A notable exception was HCT116, which showed the expected banding pattern. However, as shown below, this likely reflects that only one of the two alleles is of normal size and can therefore be amplified.

The absence of PCR products at multiple broken (TA)n repeats in MSI cell genomic DNA was unlikely to be caused by large deletions that included primer binding sites, as these sites were highly covered in END-seq and whole genome sequencing data (below). We therefore speculated that these sites contain large, structure-forming expansions that might inhibit polymerase extension during PCR amplification. To test this hypothesis, we performed PCR-free whole genome sequencing of KM12 and HCT116 cells, enabling us to reach an average sequencing depth greater than 100x. By inspection, it was immediately clear that broken (TA)n repeats displayed markedly lower read depth compared to flanking regions and non-broken (TA)n repeats (Extended Data Fig. 6c). Indeed, broken (TA)n repeats (mapped in WRN-depleted MSI cells) displayed the most significant drop in coverage relative to other classes of mono- and di-nucleotide repeats (Fig. 4b, Extended Data Fig. 6d). Finally, as assessed by CRISPR-Cas9 and RNA interference fitness screens[2], the degree of dependence on WRN for human cancer cell survival was inversely correlated with sequencing coverage at broken (TA)n sites (Fig. 4c). Thus, structure forming (TA)n repeats are highly recalcitrant to sequencing. While not previously documented as mutated in MSI cancers, (TA)n repeats are nevertheless predictive of WRN dependency.

MSI cells accumulate large-scale (TA)n repeat expansions

Since the identification of large-scale changes in short tandem repeats is challenging with short read sequencing, we utilized the ExpansionHunter algorithm[25], which has been used to detect various large pathogenic repeat expansions. This analysis revealed that broken (TA)n repeats exhibit large-scale expansions compared to non-broken (TA)n repeats (Extended Data Fig. 7a, b). Consistent results were obtained with the expanded short tandem repeat algorithm (exSTRa)[26] (Extended Data Fig. 7c).

Since large expansions should increase the size of (TA)n repeats relative to the reference genome, we performed Southern blotting with probes spanning two broken (B1 and B2) and two non-broken sites (NB1 and NB3) previously analyzed by PCR (Fig. 4d, Extended data Fig. 7d-e). For this analysis, we compared two MSI (KM12 and HCT116) and two MSS (RPE-1 and eHAP) cell lines. At the non-broken (TA)n repeats, we detected bands corresponding to the expected sizes in all cell lines (Extended Data Fig. 7d). Additionally, neither broken (TA)n region was altered in control RPE-1 or eHAP cell lines (Extended Data Fig. 7d). For broken site B1, we detected one allele of the expected size, but also one expanded allele in both KM12 and HCT116 (Fig. 4d, Extended Data Fig. 6b). The shorter, non-expanded allele likely corresponds to the PCR product detected at B1 in both cell lines. At broken site B2, we detected one expanded allele in HCT116, while the other allele was the expected size (Fig. 4d), corresponding to the product detected by PCR (Extended Data Fig. 6b). In KM12 cells, broken site B2 exhibited distinct expansions on the two alleles (Fig. 4d, Extended Data Fig. 7e), and no PCR product was detectable at this site (Fig. 4a). Taken together, our data suggest that (TA)n repeats undergo large-scale expansion, and that expanded alleles inhibit DNA synthesis during PCR and short read whole genome sequencing.

Long-read sequencing technology has successfully characterized expansions of simple tandem repeats responsible for various diseases[27]. To determine whether a similar approach

could uncover variations in secondary structure-forming (TA)n repeats in MSI cells, we utilized the Pacific Biosciences (PacBio) continuous long-read (CLR) sequencing platform[27]. With this platform, MSI (HCT116 and KM12) and MSS (SW620 and SW837) colon cancer cell lines displayed similar coverage at broken (TA)n and other simple repeats (Extended Data Fig. 7f). We detected a large range of expansions at different broken (TA)n sites in MSI cells, with median repeat length expanding from 54bp in the hg19 reference genome to 91bp and 125bp in HCT116 and KM12, respectively (Fig. 4e). In contrast, the median lengths of non-broken (TA)n repeats in HCT116 (75bp) and KM12 (74bp) were much more similar to the reference genome (72bp). MSS cell lines did not show substantial expansions at either broken or non-broken (TA)n sites (Fig. 4e). Motif analysis within the broken sites revealed that the expanded alleles consisted of almost pure (TA)n (Extended Data Fig. 7g). Thus, MSI predisposes (TA)n repeat tracts to undergo large-scale expansions.

Length of uninterrupted TA repeats and replication timing contribute to breakage

Of all annotated (TA)n repeats in the reference genome, only 8% display DSBs upon WRN depletion in MSI cells (Extended Data Fig. 2h). Thus, it remains unclear what characterizes the underlying expansion and breakage specifically at these sites and why susceptible sites are recurrent across multiple distinct MSI cell lines. Changes in the length of microsatellite DNA are thought to arise from replication slippage caused by the transient dissociation of the polymerase from replicating DNA strands followed by self-annealing and misaligned reassociation[28]. Since the probability of self-annealing increases with the purity and length of the repeat sequence and decreases with repeat interruptions[28], we speculated that predisposition to large-scale expansions of (TA)n repeats, and therefore breakage, would be influenced by the exact sequence composition of the repeat.

Indeed, we found broken sites have a higher TA content, fewer interruptions in the (TA)n repeat, and longer uninterrupted (TA)n sequences relative to non-broken sites as assessed both in the reference genome (Extended Data Fig. 8a-c) and in long-read sequencing reads from MSI cells (Extended Data Fig. 8d-f). Moreover, longer pure (TA)n repeats were much more likely to expand (Extended Data Fig. 8g). Quantitative modeling showed that three features of (TA)n repeats were predictive of breakage in WRN-deficient cells: the probability that they form secondary structures as measured by MUS81-EME1 cleavage *in situ*, the sizes of expansions determined by long-read sequencing, and the likelihood that they occur in late replicating regions (Extended Data Fig. 8h, j). While these features cannot fully predict the propensity of (TA)n repeats to break upon WRN deficiency, the analysis demonstrates that longer, uninterrupted (TA)n repeats are more likely to undergo expansion in MSI cell lines, where they form secondary structures and disrupt DNA replication. In the absence of WRN, (TA)n repeats that are replicated in late S phase are less likely to be resolved prior to mitosis, when they are recognized and cleaved by MUS81 to generate DSBs.

Relationship between expandable repeats and genome stability

Secondary structure forming (TA)n tracts within CFSs cause replication fork stalling and chromosome breakage. CFSs are also associated with deletions in multiple tumor types[29], likely due to their susceptibility to replication fork collapse. Since broken (TA)n sites in MSI cells are also susceptible to replication fork stalling and collapse (Fig. 3), we hypothesized they could be hotspots for deletions in MSI cancers. Although genomic instability in MSI cancers is most frequently associated with small insertions and deletions, large Kb-Mb scale deletions of unknown etiology have been detected[30]. We therefore analyzed data from pan-cancer whole genome sequencing cohorts in uterine corpus endometrial carcinoma (UCEC), colon adenocarcinoma (COAD) and gastric adenocarcinoma (STAD)[31]. Breakpoints

300 associated with deletions (ranging in size from 459bp to 176Mb) in MSI (24) and MSS (93)
301 cancers were identified[5]. In several cases, one or both deletion breakpoints mapped precisely
302 to broken (TA)n repeats (Extended Data Fig. 9a, b). We then calculated the enrichment of
303 different annotated repeats at the tumor breakpoints compared to their enrichment at random
304 breakpoints with comparable chromosome- and size- distributions, and found that MSI tumor
305 breakpoints showed the greatest enrichment at broken (TA)n repeats (Extended Data Fig. 9c).
306 In contrast, these sites were not enriched for deletion breakpoints in MSS tumors (Extended
307 Data Fig. 9c). Thus, in MSI cancers, large scale deletions are frequently associated with
308 broken (TA)n sites, suggesting that they are inherently fragile.
309 Based on our results, we propose the following model (Extended Data Fig. 8k): MMR
310 deficient cells undergo microsatellite instability, which gradually manifests as large-scale
311 expansions at (TA)n repeats. The most susceptible sites are those that contain pure,
312 uninterrupted TA-dinucleotide repeats. Over the course of months to years, (TA)n repeats
313 reach a threshold length above which they extrude into cruciform-like structures, perhaps as a
314 result of negative supercoiling during nucleosome removal ahead of the replication fork.
315 These structures would stall replication forks and trigger ATR-dependent WRN
316 phosphorylation, which promotes unwinding of the secondary structure to complete
317 replication. In the absence of WRN, the structure-specific MUS81-EME1 endonuclease, and
318 its scaffold SLX4, cleave these structures in an attempt to salvage the replication fork.
319 However, thousands of concerted MUS81 cleavage events lead to extensive DNA end-
320 resection, RPA exhaustion (Fig. 1b-c)[8], chromosomal fragmentation, and cell death.
321 Thirty years ago, Vogelstein and colleagues discovered the *DCC* (Deleted in Colon
322 Cancer) gene, whose expression is greatly reduced in colon cancer[32]. By Southern blot
323 analysis, they observed that several MSI tumor cell lines harbored "insertions" up to 300 bp at
324 a locus containing an uninterrupted $(TA)_{22}$ repeat just downstream of *DCC* exon 7. However,

numerous attempts to clone or amplify alleles with insertions failed, leading them to conclude that the inserted sequence might form an unusual DNA structure. Strikingly, we detected a MUS81-EME1 sensitive DNA structure precisely at the same (TA)n repeat (Extended Data Fig. 10a), and long-read sequencing confirmed a (TA)n expansion at that locus (Extended Data Fig. 10b, right highlighted sequence). Thus, our data demonstrate that rather than a non-templated insertion, DCC contains a structure-forming, MSI-expanded (TA)n repeat.

Anti-PD-1 antibodies have been approved for use in patients with cancers with mismatch repair deficiency or MSI, independent of the cancer lineage. The therapeutic response is correlated with the number indels in coding regions, which can generate immunogenic neoantigens[33,34]. (TA)n expansions are mostly localized outside of coding regions, and therefore might not generate immunogenic neoantigens. However, DSBs at (TA)n motifs are capable of triggering innate cytosolic DNA- and RNA-dependent sensing and signaling pathways[35,36], indicating that WRN inhibition has the potential for further stimulating immune responses. Although indel mutations of a few nucleotides and large-scale (TA)n expansions are both features of MSI, further studies will be necessary to determine whether these defects arise through different mechanisms, in distinct MSI tumors, or within different subclones. In summary, our findings provide a mechanistic explanation for WRN dependence with MSI and identify a novel biomarker to guide selection of patients where WRN inhibition may be effective in combination with immune checkpoint blockade or as an independent line of therapy.

References

1 Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511-516, doi:10.1038/s41586-019-1103-9 (2019).

2 Chan, E. M. *et al.* WRN helicase is a synthetic lethal target in microsatellite unstable cancers. *Nature* **568**, 551-556, doi:10.1038/s41586-019-1102-x (2019).

3 Kategaya, L., Perumal, S. K., Hager, J. H. & Belmont, L. D. Werner Syndrome Helicase Is Required for the Survival of Cancer Cells with Microsatellite Instability. *iScience* **13**, 488-497, doi:10.1016/j.isci.2019.02.006 (2019).

4 Lieb, S. *et al.* Werner syndrome helicase is a selective vulnerability of microsatellite instability-high tumor cells. *Elife* **8**, doi:10.7554/eLife.43333 (2019).

5 Fujimoto, A. *et al.* Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res*, doi:10.1101/gr.255026.119 (2020).

6 Dudley, J. C., Lin, M. T., Le, D. T. & Eshleman, J. R. Microsatellite Instability as a Biomarker for PD-1 Blockade. *Clin Cancer Res* **22**, 813-820, doi:10.1158/1078-0432.CCR-15-1678 (2016).

7 Chu, W. K. & Hickson, I. D. RecQ helicases: multifunctional genome caretakers. *Nat Rev Cancer* **9**, 644-654, doi:10.1038/nrc2682 (2009).

8 Toledo, L. I. *et al.* ATR prohibits replication catastrophe by preventing global exhaustion of RPA. *Cell* **155**, 1088-1103, doi:10.1016/j.cell.2013.10.043 (2013).

9 Canela, A. *et al.* DNA Breaks and End Resection Measured Genome-wide by End Sequencing. *Mol Cell* **63**, 898-911, doi:10.1016/j.molcel.2016.06.034 (2016).

10 Paiano, J. *et al.* ATM and PRDM9 regulate SPO11-bound recombination intermediates during meiosis. *Nat Commun* **11**, 857, doi:10.1038/s41467-020-14654-w (2020).

11 Khil, P. P., Smagulova, F., Brick, K. M., Camerini-Otero, R. D. & Petukhova, G. V. Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. *Genome Res* **22**, 957-965, doi:10.1101/gr.130583.111 (2012).

12 Tubbs, A. *et al.* Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell* **174**, 1127-1142 e1119, doi:10.1016/j.cell.2018.07.011 (2018).

13 Bowater, R., Aboul-ela, F. & Lilley, D. M. Large-scale stable opening of supercoiled DNA in response to temperature and supercoiling in (A + T)-rich regions that promote low-salt cruciform extrusion. *Biochemistry* **30**, 11495-11506, doi:10.1021/bi00113a003 (1991).

14 Dayn, A. *et al.* Formation of (dA-dT)n cruciforms in Escherichia coli cells under different environmental conditions. *J Bacteriol* **173**, 2658-2664, doi:10.1128/jb.173.8.2658-2664.1991 (1991).

15 McClellan, J. A., Boublikova, P., Palecek, E. & Lilley, D. M. Superhelical torsion in cellular DNA responds directly to environmental and genetic factors. *Proc Natl Acad Sci U S A* **87**, 8373-8377, doi:10.1073/pnas.87.21.8373 (1990).

16 Zlotorynski, E. *et al.* Molecular basis for expression of common and rare fragile sites. *Mol Cell Biol* **23**, 7143-7151, doi:10.1128/mcb.23.20.7143-7151.2003 (2003).

17 Kaushal, S. *et al.* Sequence and Nuclease Requirements for Breakage and Healing of a Structure-Forming (AT)n Sequence within Fragile Site FRA16D. *Cell Rep* **27**, 1151-1164 e1155, doi:10.1016/j.celrep.2019.03.103 (2019).

18 Wang, H. *et al.* CtIP maintains stability at common fragile sites and inverted repeats by end resection-independent endonuclease activity. *Mol Cell* **54**, 1012-1021, doi:10.1016/j.molcel.2014.04.012 (2014).

19 Shastri, N. *et al.* Genome-wide Identification of Structure-Forming Repeats as Principal Sites of Fork Collapse upon ATR Inhibition. *Mol Cell* **72**, 222-238 e211, doi:10.1016/j.molcel.2018.08.047 (2018).
20 Inagaki, H. *et al.* Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Res* **19**, 191-198, doi:10.1101/gr.079244.108 (2009).
21 Minocherhomji, S. & Hickson, I. D. Structure-specific endonucleases: guardians of fragile site stability. *Trends Cell Biol* **24**, 321-327, doi:10.1016/j.tcb.2013.11.007 (2014).
22 Wyatt, H. D., Laister, R. C., Martin, S. R., Arrowsmith, C. H. & West, S. C. The SMX DNA Repair Tri-nuclease. *Mol Cell* **65**, 848-860 e811, doi:10.1016/j.molcel.2017.01.031 (2017).
23 Ammazzalorso, F., Pirzio, L. M., Bignami, M., Franchitto, A. & Pichierri, P. ATR and ATM differently regulate WRN to prevent DSBs at stalled replication forks and promote replication fork recovery. *EMBO J* **29**, 3156-3169, doi:10.1038/emboj.2010.205 (2010).
24 Cortes-Ciriano, I., Lee, S., Park, W. Y., Kim, T. M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* **8**, 15180, doi:10.1038/ncomms15180 (2017).
25 Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res* **27**, 1895-1903, doi:10.1101/gr.225672.117 (2017).
26 Tankard, R. M. *et al.* Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *Am J Hum Genet* **103**, 858-873, doi:10.1016/j.ajhg.2018.10.015 (2018).
27 Mitsuhashi, S. & Matsumoto, N. Long-read sequencing for rare human genetic diseases. *J Hum Genet* **65**, 11-19, doi:10.1038/s10038-019-0671-8 (2020).
28 Khristich, A. N. & Mirkin, S. M. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J Biol Chem* **295**, 4134-4170, doi:10.1074/jbc.REV119.007678 (2020).
29 Glover, T. W., Wilson, T. E. & Arlt, M. F. Fragile sites in cancer: more than meets the eye. *Nat Rev Cancer* **17**, 489-501, doi:10.1038/nrc.2017.52 (2017).
30 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).
31 Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).
32 Fearon, E. R. *et al.* Identification of a chromosome 18q gene that is altered in colorectal cancers. *Science* **247**, 49-56, doi:10.1126/science.2294591 (1990).
33 Ding, L. & Chen, F. Predicting Tumor Response to PD-1 Blockade. *N Engl J Med* **381**, 477-479, doi:10.1056/NEJMcibr1906340 (2019).
34 Mandal, R. *et al.* Genetic diversity of tumors with mismatch repair deficiency influences anti-PD-1 immunotherapy response. *Science* **364**, 485-491, doi:10.1126/science.aau0447 (2019).
35 Feng, X. *et al.* ATR inhibition potentiates ionizing radiation-induced interferon response via cytosolic nucleic acid-sensing pathways. *EMBO J*, e104036, doi:10.15252/embj.2019104036 (2020).
36 Harding, S. M. *et al.* Mitotic progression following DNA damage enables pattern recognition within micronuclei. *Nature* **548**, 466-470, doi:10.1038/nature23470 (2017).

Figure Legends

Fig. 1: WRN depletion in MSI cells induces recurrent DSBs at (TA)n dinucleotide repeats.

(a) Representative metaphase spreads from KM12 NT, KM12-shWRN and SW837-shWRN cells. Cells were treated with DMSO (NT) or doxycycline (shWRN) for 48 hrs.

(b) Genome browser screenshots displaying END-seq and RPA-ssDNA ChIP-seq profiles as normalized read density (reads per million, RPM) for KM12-shWRN cells. Positive and negative strand END-seq reads are displayed in black and grey, and positive and negative strand RPA-ssDNA ChIP-seq reads in blue and red, respectively.

(c) Genome browser screenshot zoomed in on the highlighted region in panel b (orange box). The light blue highlight indicates the gap region between the left and right ends of the DSB.

(d) Motif analysis for sequence enrichment in the gap between positive and negative END-seq peaks in KM12-shWRN cells.

(e) Fraction of END-seq peaks occurring at sites of (TA)n repeats for KM12-shWRN and HCT116-shWRN cells treated with either doxycycline (shWRN) or WRN siRNAs (siWRN) for 72 hrs.

Fig. 2: TA breaks are dependent on structure-specific endonucleases MUS81-EME1 and SLX4.

(a) Quantification of metaphases displaying chromosome shattering (defined as the absence of intact chromosomes) in KM12-shWRN cells treated with doxycycline plus non-targeting (siCTRL), MUS81 (siMUS81), or SLX4 (siSLX4) siRNAs for 72hrs. Student's t-test values are shown.

(b) Quantification of END-seq signal intensity at broken TA repeats in KM12-shWRN cells with same treatment as in panel (a). *** indicates a p-value < 10^{-16}.

(c) Genome browser screenshot for KM12-shWRN cells treated with doxycycline (shWRN, top), and DMSO-treated cells processed with purified recombinant MUS81-EME1 enzyme *in situ* (bottom).

(d) Zoom-in of TA break at highlighted region in Fig. 2c (orange box).

(e) Quantification of END-seq peak intensity for DMSO-treated KM12-shWRN cells processed *in situ* with either purified recombinant MUS81-EME1, WRN, or WRN followed by MUS81-EME1. For the latter, Proteinase K digestion was performed between the two enzymatic treatments. *** indicates a p-value < 10^{-16}.

Fig. 3: Replication stalling and collapse at (TA)n repeats in MSI cell lines.

(a) Composite plot of RPA-ssDNA ChIP-seq signal(blue: RPA-ChIP, red: input DNA) around broken (TA)n repeats in KM12 cells.

(b) Genome browser screenshot displaying END-seq profiles for KM12-shWRN cells, HCT116-shWRN, RPE-1-shWRN, and eHAP-shWRN cells treated with doxycycline (shWRN) for 72hrs or APH+ATRi for 8 hrs. MSI cells are marked in red, MSS cell lines in blue.

(c) Barplots indicating percentage of total DSBs located at (TA)n repeats after APH+ATRi treatment, as shown in panel (b). MSI cell are marked in red, MSS cell lines in blue.

Fig. 4: (TA)n repeats undergo large-scale expansion in MSI cell lines.

(a) Agarose gels showing PCR fragments (or lack thereof) at (TA)n repeats in KM12 (MSI) and SW837 (MSS) cells. B1-B8 were chosen based on the presence of END-

seq peaks upon WRN depletion in KM12 cells. NB1-NB3 were chosen for similar (TA)n repeat lengths as broken sites without breakage upon WRN depletion in KM12 cells. Ladder fragment sizes (in bp) are displayed. For gel source data, see Supplementary Figure 1.

(b) Boxplots displaying coverage at different classes of repeats in WGS data from KM12 cells. Dotted red lines indicate the average coverage over the genome.

(c) Cell lines plotted by their average WRN dependency score and sequencing coverage of broken (TA)n loci. fpbm: fragments per base per million.

(d) Southern blots for two different genomic regions containing broken (TA)n repeats corresponding to the same sites in panel (a). Red markers and dotted lines represent expected fragment sizes. For gel source data, see Supplementary Figure 1.

(e) Boxplot of long-read sequencing data demonstrating total length of broken and non-broken (TA)n in indicated cell lines compared to hg19 reference genome. MSI (red), MSS (blue). *** p-value $<10^{-5}$. Wilcox testing for the alternative hypothesis that the broken (TA)n has a greater group mean than the non-broken (TA)n was used. p values were corrected for multiple testing using the Benjamini-Hochberg method.

Extended Data Figure Legends

Extended Data Fig. 1: WRN depletion induces DNA damage in different MSI cell lines.

(a) Western Blot analysis of MLH1, MSH2, and GAPDH protein levels in human stomach epithelial cells (HSEC) following CRISPR/Cas9 knockout. sg*Luc*: control sgRNA targeting *luciferase*. sg*MLH1* and sg*MSH2*: sgRNAs targeting *MLH1* and *MSH2*, respectively. For gel source data, see Supplementary Figure 1.

(b) Relative viability 7 days after sgRNA transduction in HSEC. Negative controls targeting chromosome 2 intergenic sites: sgCh2.2 and sgCh2.4. Pan-essential control: sg*PolR2D*. Experimental sgRNA targeting *WRN*: sg*WRN2*, sg*WRN3*. Two-tailed Student's t-test values are shown.

(c) Example of flow cytometry gating strategy used in Extended Data Fig. 1d, 4c.

(d) Flow cytometry profiles for exponentially growing KM12-shWRN cells treated with DMSO (NT) or doxycycline (shWRN) for 72 hrs. EdU was added during the last 30 minutes before harvesting cells. Percentage of cells in the gates are indicated.

(e) Western blot analysis of WRN protein levels in KM12-shWRN and SW837-shWRN treated with DMSO or doxycycline for 72 hrs. For gel source data, see Supplementary Figure 1.

(f) Western blot analysis of WRN and pKAP1 protein levels in KM12-shWRN and KM12-shWRN.C911 (non-targeting shRNA) treated with DMSO or doxycycline for 72 hrs. For gel source data, see Supplementary Figure 1.

Extended Data Fig. 2: WRN depletion induces recurrent and overlapping DSBs in MSI cells.

(a) Genome browser screenshot displaying END-seq profiles as normalized read density (reads per million, RPM) for HCT116-shWRN cells and KM12-shWRN treated with DMSO (NT) or doxycycline (shWRN), or transfected with non-targeting siRNAs (siCTRL) or WRN siRNAs (siWRN) for 72 hrs.

(b) Scatterplots of END-seq peak intensity between replicate experiments of KM12-shWRN and HCT116-shWRN cells treated with doxycycline for 72 hrs. Pearson correlation coefficients are indicated.

(c) Venn diagrams showing overlap between peaks detected in HCT116-shWRN and KM12-shWRN cells treated with either doxycycline (shWRN) or WRN siRNAs (siWRN) for 72 hrs (p-value $< 10^{-16}$ for both comparisons).

(d) Quantification of END-seq peak intensity for KM12-shWRN and SW837-shWRN cells treated with doxycycline for 72 hrs. *** indicates a p-value $< 10^{-16}$.

(e) Venn diagram showing overlap between peaks identified from END-seq and RPA-ssDNA ChIP-seq for KM12-shWRN cells treated with doxycycline for 72 hrs (p-value $< 10^{-16}$).

(f) Composite plot of END-seq (black: positive strand reads, grey: negative strand reads) and RPA-ssDNA ChIP-seq (blue: positive strand reads, red: negative strand reads) signal around DSB sites in KM12-shWRN cells treated with doxycycline for 72 hrs.

(g) Heatmap displaying intensity of END-seq signal in KM12-shWRN cells treated with doxycycline for 72 hrs, relative to the center of the gap between positive and negative strand peaks. Sites are ordered by the size of the gap, from smallest to largest.

(h) Calculated size distribution from the reference genome of (TA)n repeats either located in gaps between positive and negative END-seq peaks (black, broken sites) or located elsewhere in the genome (grey, non-broken sites), determined from KM12-shWRN cells treated with doxycycline for 72 hrs.

Extended Data Fig. 3: WRN depletion induces DNA breakage in common fragile sites and palindromic TA-rich repeats in MSI cells.

(a) Genome browser screenshot displaying END-seq profiles of common fragile sites FRA16D, FRA3B, FRA10B, and FRA7I as normalized read density (reads per million, RPM) for KM12-shWRN cells treated with DMSO (NT) or doxycycline (shWRN) for 72 hrs. Indicated is the number of uninterrupted (TA)n repeat units in the hg19 reference genome at DSB sites.

(b) Genome browser screenshot displaying END-seq profiles of PATRRs on chromosomes 11 and 22 as normalized read density (reads per million, RPM) for KM12-shWRN cells treated with DMSO (NT) or doxycycline (shWRN) for 72 hrs.

Extended Data Fig. 4: (TA)n repeat-forming repeats in MSI cell lines are substrates for MUS81-EME1.

(a) qRT-PCR quantification of MUS81 and SLX4 mRNA levels in KM12-shWRN cells transfected with non-targeting siRNAs (siCTRL), MUS81 siRNAs (siMUS81), or SLX4 siRNAs (siSLX4).

(b) Representative images of metaphase spreads from KM12-shWRN cells treated with doxycycline (shWRN) and either non-targeting siRNAs (siCTRL), MUS81 siRNAs (siMUS81), or SLX4 siRNAs (siSLX4) for 48 hrs.

(c) Flow cytometric profiles for KAP1 phosphorylation in exponentially growing KM12-shWRN cells treated with doxycycline (shWRN), plus non-targeting siRNAs (siCTRL), MUS81 siRNAs (siMUS81), or SLX4 siRNAs (siSLX4) for 72 hrs.

(d) Genome browser screenshot displaying END-seq profiles as normalized read density (reads per million, RPM) for KM12-shWRN cells treated with DMSO (NT), plus non-targeting siRNAs (siCTRL), MUS81 siRNAs (siMUS81), or SLX4 siRNAs (siSLX4) for 72 hrs.

(e) Schematic representation of DNA cruciform cleavage by MUS81-EME1 structure-specific endonuclease.

(f) Venn diagram displaying overlap of END-seq peaks between two replicate experiments of DMSO-treated KM12-shWRN cells processed with purified recombinant MUS81-EME1 enzyme *in situ* (MUS81-EME1)(p-value < 10^{-16}).

(g) Venn diagram showing overlap in TA breaks between KM12-shWRN treated with doxycycline (shWRN) for 72 hrs, and DMSO-treated cells processed with MUS81-EME1 enzyme *in situ* (MUS81-EME1) (p-value < 10^{-16}).

(h) Venn diagram displaying overlap between END-seq peaks from DMSO-treated KM12-shWRN and HCT116-shWRN cells processed *in situ* with MUS81-EME1 (p-value < 10^{-16}).

(i) Genome-wide aggregate analysis of END-seq signal around TA breaks from KM12-shWRN cells treated with doxycycline for 72 hrs (shWRN - black: positive strand reads, grey: negative strand reads), or DMSO-treated KM12-shWRN cells processed with purified recombinant MUS81-EME1 enzyme *in situ* (blue: positive strand reads, red: negative strand reads).

(j) Genome browser screenshot displaying END-seq profiles for DMSO-treated KM12-shWRN cells (WRN proficient) processed *in situ* with either purified recombinant WRN, MUS81-EME1, or WRN followed by MUS81-EME1. For the latter, Proteinase K digestion was performed between the two enzymatic treatments.

Extended Data Fig. 5: Structure-forming repeats in MSI cells activate ATR.

(a) Genome browser screenshot displaying END-seq profiles for DMSO-treated KM12-shWRN, HCT116-shWRN, SW837-shWRN and RPE-1-shWRN cells processed *in situ* with purified recombinant MUS81-EME1. Cells are indicated as MSI (red label) or MSS (blue label).

(b) Quantification of END-seq peak intensity for libraries displayed in panel a.

(c) Western blot analysis of WRN and pKAP1 levels in HCT116 cells expressing WT WRN, or ATR phosphorylation mutants WRN^{3A} or WRN^{6A}. Endogenous WRN was depleted using an siRNA targeting WRN 5'UTR. For gel source data, see Supplementary Figure 1.

(d) Genome browser screenshot displaying END-seq profiles within FRA3B on chromosome 3 as normalized read density (reads per million, RPM) for KM12-shWRN cells, HCT116-shWRN, RPE-1-shWRN, and eHAP-shWRN cells treated with doxycycline (shWRN) for 72hrs or APH+ATRi for 8 hrs.

(e) Venn diagrams displaying overlap of DSBs detected after WRN depletion or APH+ATRi treatment in KM12 and HCT116 cells.

Extended Data Fig. 6: (TA)n repeat sequences are underrepresented in whole-genome sequencing data from MSI cells.

(a) Barplots indicating the percentage of recurrent mutations in different classes of repeats (left panel - mono, di, tri and tetra) and a barplot (right panel) showing the number of various dinucleotide repeats in the 1000 altered loci. The plots were based on sequencing analysis from[24], which considered microsatellites smaller than 40bp.

(b) Agarose gels showing PCR fragments (or lack thereof) of sites of different (TA)n repeats in one MSS and four MSI cell lines. Broken sites B1-B8 were chosen based on the presence of END-seq peaks upon WRN depletion in KM12 cells. Sites NB1-NB3

were chosen with similar (TA)n repeat lengths as broken sites, but were not broken upon WRN depletion in KM12 cells. Fragment sizes (in bp) are displayed. For gel source data, see Supplementary Figure 1.

(c) Genome browser screenshots of short read PCR-free whole genome sequencing reads, indicating coverage, in KM12 and HCT116 cell lines. Shown are two regions containing (TA)n repeats, one that displays END-seq peaks upon WRN depletion in KM12 (left panels), and one that does not (right panels). Regions correspond to equivalent PCR sites in Fig. 4a and Extended Data Fig. 5b.

(d) Boxplots displaying coverage at different classes of mono- and di-nucleotide repeats in PCR-free whole-genome sequencing libraries made from HCT116 cells. (TA)n repeats are split into those that overlap END-seq peaks upon shWRN induction, and those that do not harbor DSBs. Dotted red lines indicate the average coverage over the genome.

Extended Data Fig. 7: (TA)n repeats undergo large-scale expansions in MSI cells.

(a) Cumulative fraction of expanded (TA)n repeats in KM12 and HCT116, based on ExpansionHunter analysis of PCR-free whole genome sequencing data. TA repeats were split into broken (red) and non-broken (black) based on presence or absence of END-seq peaks after WRN depletion in KM12 cells.

(b) Graphical representation of a (TA)n repeat expansion in HCT116. This site has 33 (TA)n repeat units in the reference genome; ExpansionHunter identified an expansion to 86-87 repeat units based on PCR-free whole genome sequencing of HCT116.

(c) Empirical cumulative distribution function based on the length by which each read overlaps the (TA)n repeat shown in panel b as identified by exSTRa.

(d) Southern blots for two different genomic regions containing non-broken (TA)n repeats corresponding to the same sites in Fig. 4a and Extended Data Fig. 6b. Red markers and dotted lines represent expected fragment sizes. For gel source data, see Supplementary Figure 1.

(e) Southern blots for broken (TA)n repeat B2 (top) and non-broken (TA)n repeat NB3 (bottom) in MSS (blue) and MSI (red) cell lines, confirming expansion of broken (TA)n repeats in MSI cell lines. Red markers and dotted lines represent expected fragment sizes based on the reference genome. For gel source data, see Supplementary Figure 1.

(f) Boxplots displaying coverage at different classes of repeats in long-read sequencing libraries made from MSI (red) and MSS (blue) cells.

(g) Motif analysis for sequence enrichment at broken (TA)n in the KM12 cell line from long-read sequencing data.

Extended Data Fig. 8: Large-scale expansions occur at long, uninterrupted (TA)n repeat sequences.

(a) Boxplot showing, in the hg19 reference genome, the proportion of (TA)n repeat units found within the full annotated sequence at broken or non-broken (TA)n repeats in KM12 cells. *** indicates a p-value $< 10^{-16}$.

(b) Boxplot showing, in the hg19 reference genome, the proportion of the longest run of uninterrupted (TA)n within the full annotated sequence at broken or non-broken (TA)n repeats in KM12 cells. *** indicates a p-value $<10^{-16}$.

(c) Boxplot showing, in the hg19 reference genome, the length (bp) of the longest uninterrupted (TA)n dinucleotide repeats within the full annotated sequence at broken or non-broken (TA)n repeats in KM12 cells. *** indicates a p-value < 10^{-16}.

(d) Boxplot showing, in long read sequencing data, the proportion of (TA)n repeat units found within the full sequence at broken or non-broken (TA)n repeats in KM12 cells. *** indicates a p-value < 10^{-16}.

(e) Boxplot showing, in long read sequencing data, the proportion of the longest run of uninterrupted (TA)n within the full sequence at broken or non-broken (TA)n repeats in KM12 cells. *** indicates a p-value <10^{-16}.

(f) Boxplot showing, in long read sequencing data, the length (bp) of the longest uninterrupted (TA)n dinucleotide repeat within the full sequence at broken or non-broken (TA)n repeats in KM12 cells. *** indicates a p-value < 10^{-16}.

(g) Boxplot showing, in the hg19 reference genome, lengths (bp) of the longest uninterrupted (TA)n repeats for expanded or non-expanded (TA)n repeats as determined through long read sequencing data from KM12 cells.

(h) Multiple linear regression model predicting END-seq peak intensity of KM12 shWRN cells treated with doxycycline (shWRN) for 72 hrs derived from END-seq intensity of MUS81-EME1 cleavage *in situ*, replication timing, and expanded length of broken (TA)n. The Pearson correlation coefficient is indicated (see panel j).

(i) END-seq intensity of broken (TA)n repeats in KM12 shWRN cells treated with doxycycline (shWRN) for 72 hrs grouped by replication timing values from late replicating to early replicating.

(j) Multiple linear regression was performed to predict END-seq peak intensity of KM12 shWRN cells treated with doxycycline (shWRN) for 72 hrs based on following parameters: END-seq intensity of MUS81-EME1 cleavage *in situ*, replication timing,

and expanded length of broken (TA)n. END-seq intensity of shWRN END-seq and MUS81-EME1 cleavage were calculated using RPKM in ±1kb window around broken (TA)n. Mean value was used for replication timing quantification. Expanded lengths were identified from long read sequencing data. Estimates of the standardized regression coefficients (β) are shown, along with t statistics and P–values based on the standardized coefficients.

(k) Model for MSI cell dependence on WRN. Large scale expansions of (TA)n repeats are associated with microsatellite instabilty in MMR deficient cells. When (TA)n reach above a critical length, they extrude into cruciform-like structures, which stall replication forks and activate ATR kinase, which in turn phosphorylates WRN and other substrates to complete DNA replication. In the absence of WRN, MUS81-EME1/SLX4 cleaves secondary structures at (TA)n repeats, thereby shattering the chromosomes.

Extended Data Fig. 9: Deletion breakpoints in MSI cancers are enriched at (TA)n repeats.

a) Genome browser screenshot of a broken (TA)n (defined from KM12), MSI deletion (derived from a patient sample), and END-seq profile (in WRN-depleted KM12 cells). The sequences around the breakpoints are shown in the inset.

b) Junctions associated with 6 different MSI deletions from patients. Seq1 represents the sequence from -50bp to left breakpoint and Seq2 represents the sequence from right breakpoint to +50bp.

c) Enrichment of simple repeats, broken and non-broken (TA)n, and LINE, SINE and LTR at patient deletion breakpoints relative to their overlap with random deletion breakpoints of the same size (enrichment value=1). B(TA)n-B(TA)n represents cases

in which both breakpoints overlap with broken (TA)n repeats; B(TA)n- represents cases in which only one breakpoint overlaps with a broken (TA)n repeat. B(TA)n, broken TA repeat; NB(TA)n, nonbroken TA repeat; LTR, long terminal repeat; SINE, short interspersed nuclear element; LINE, long interspersed nuclear element.

Extended Data Fig. 10: DNA breaks within *DCC* gene body

(a) Genome browser screenshots within *DCC* gene displaying END-seq profiles as normalized read density (reads per million, RPM) for KM12-shWRN cells treated DMSO (NT), doxycycline (shWRN) for 72 hrs, or MUS81-EME1 *in situ*.

(b) Zoom-in view of region including exons 6 and 7 of *DCC* gene, containing two (TA)n repeats displaying END-seq peaks. The highlighted sequences below were extracted from long read sequencing reads in KM12 cells. The (TA)n repeat in intron 7 is where Vogelstein and colleagues previously detected an insertion.

Methods

Cell lines and cell culture

Cell lines containing doxycycline-inducible shWRN cassette (KM12, HCT116, SW837, RPE-1) were generated as previously described [2]. Cell lines were grown in medium supplemented with 10% fetal bovine serum (FBS), penicillin (100 µg/mL) / streptomycin (100 µg/mL), and L- glutamine (Gibco, 292 µg/mL) unless stated otherwise. KM12, SW837, and SW48 were grown in RPMI1640 (Gibco), HCT116 in McCoy's 5A (Gibco), RPE-1 in DMEM/F12 (Gibco), SW620 in Leibovitz's L-15 (Gibco), OVK18 and LS180 in MEMα (Gibco) supplemented with 15% FBS. Human primary stomach epithelial cells (HSEC) were obtained from Cell Biologics (H-6039) and cultured with complete human epithelial cell medium (CellBiologics H6621). Independent HCT116 clones were used for whole genome sequencing and continuous long-read sequencing. Cell lines were tested for mycoplama contamination.

Generation of MMR-deficient HSEC cell lines.

Transduction of HSEC with Cas9 was performed with pLX311-Cas9 (Addgene #118018) followed by blasticidin (4µg/mL) selection. Subsequently, we transduced pXPR_BRD003 carrying sgRNAs targeting *MLH1* (target sequence TTTGGCCAGCATAAGCCATG), *MSH2* (CCGGTCGAAAAGGCGCACTG), or *luciferase* (ACAACTTTACCGACCGCGCC) to generate stable cell lines following puromycin selection (2ug/mL).

Cell viability assay

Cas9-expressing HSEC with sgRNAs targeting *MLH1*, *MSH2*, or *luciferase* were transduced with pXPR_BRD003 harboring the following sgRNAs. sgRNAs targeting *WRN* include sg*WRN*2 (ATCCTGTGGAACATACCATG) and sg*WRN*3 (GTAGCAGTAAGTGCAACGAT). Two negative controls targeting intergenic sites on chromosome 2 were used: Chr.2-2 sgRNA (GGTGTGCGTATGAAGCAGTG) and Chr.2-4

774 sgRNA (GCAGTGCTAACCTTGCATTG). sgRNA targeting *PolR2D*
775 (AGAGACTGCTGAGGAGTCCA) was used as a pan-essential control. All sgRNAs were
776 inserted in the pXPR_BRD003 lentiviral vector and inserts were verified by Sanger
777 sequencing. Viability was determined using CellTiter-Glo (Promega G7572) 7 days following
778 lentiviral transduction.

779 **Protein depletion and exogenous replication stress**

780 Expression of shWRN was induced by adding 1μg/mL doxycycline to cell culture medium for
781 indicated time. The following siRNAs were used for experiments: siGenome Human siRNA
782 Smartpool targeting WRN, MUS81, and SLX4, as well as non-targeting control pool (Horizon
783 Discovery). WRN 5'UTR (AAACCCGAGAAGAUCCAGUCCAACA) was described in [3]
784 and ordered from ThermoFisher. Cells were transfected using RNAiMax Transfection
785 Reagent (ThermoFisher) according to manufacturer's instructions. To induce exogenous
786 replication fork collapse, cells were treated with aphidicolin (0.2μg/mL, Sigma Aldrich) for
787 24 hrs, and ATR inhibitor AZ20 (10μM, SelleckChem) was added for the final 8 hrs.

788 **RNA isolation and qRT-PCR**

789 Total RNA was extracted from cells using Trizol Reagent (Invitrogen), and cDNA was made
790 using SuperScript II Reverse Transcriptase (ThermoFisher), according to manufacturer's
791 instructions. RT-PCR was performed using iTaq Universal SyBR Green (BioRad), samples
792 were run and analyzed on a BioRad CFX96 Real-Time PCR detection system. Primer
793 sequences were as follows.

794 MUS81 Fw: 5' GCTGCTCCGAGAGCTACAG 3'

795 MUS81 Rv: 5' CAGGGTTTGCTGGGTCTCTA 3'

796 SLX4 Fw: 5'AGTGTGCTGTGAAGATGGAG 3'

797 SLX4 Rv: 5' CCGTTTCAGACCTCTACTGTG 3'

798 ßACTIN Fw: 5' CGTCACCAACTGGGACGACA 3'

799 ßACTIN Rv: 5’ CTTCTCGCGGTTGGCCTTGG 3’

800

801 **Metaphase analysis**

802 Cells were arrested at mitosis with 0.04 μg/ml colcemid (Roche) for 16h and metaphase
803 chromosome spreads were prepared as previously described[37].

804 **Western blotting**

805 Cells were lysed in a buffer containing 50 mM Tris-HCl (pH 7.5), 200 mM NaCl, 5% Tween-
806 20, 2% Igepal CA-630, 2 mM PMSF, 50 mM β-glycerophosphate (Merck) and protease
807 inhibitor cocktail tablet (cOmplete Mini, Roche Diagnostics). Equal amounts of lysates were
808 loaded into precast mini-gels (Invitrogen) and resolved by SDS-PAGE. Transfer of proteins
809 onto nitrocellulose membranes and incubation with primary/secondary antibodies were
810 performed according to standard procedures. Visualization of protein bands was achieved by
811 fluorescence imaging on the Odyssey Clx system (LI-COR Biosciences). Antibodies and
812 dilutions used were as follows: Anti-WRN (1:5000, Novus), anti-pKAP1 (Bethyl
813 Laboratories, 1:100), anti-tubulin (1:5,000, sigma), IRDye 680RD Goat anti-Mouse IgG
814 (1:2000, LI-COR Biosciences), IRDye 800CW Goat anti-Rabbit IgG (1:2000, LI-COR
815 Biosciences).

816 **Flow cytometry**

817 For cell cycle analysis, exponentially growing cells were incubated with 10mM (5-ethynyl-
818 2’ -deoxyuridine) for 30 min at 37°C and stained using the Click-IT EdU Alexa Fluor 488
819 Flow Cytometry Assay Kit (ThermoFisher) according to the manufacturer’s instructions.
820 DNA content was measured by DAPI (4′,6-diamidino-2-phenylindole, 0.5μg/mL). For
821 pKAP1 staining, fixed and permeabilized cells were washed in PBS+2%FSB, incubated for 1
822 hour with pKAP1 antibody (Bethyl Laboratories, 1:100), washed, stained with anti-rabbit
823 Alexa Fluor 647 antibody (ThermoFisher, 1:2000).

824

825 **Protein purification**

826 Expression and purification of MUS81-$_{FLAG}$EME1[22] and WRN[38] were performed as
827 previously described. Human MUS81 (residues 246–551) and double 8xHIS-EME1 (residues
828 246-570) were expressed in a bicistronic expression vector and purified as previously
829 described[39], minus the cation exchange chromatography step.

830 **Enzymatic reactions**

831 For enzymatic reactions, END-seq plugs were incubated in the presence of 50 nM of the
832 indicated enzyme for 1.5 hours at 37°C. MUS81/EME1 enzyme reactions were done in buffer
833 consisting of 25 mM Tris-HCl pH 8.0, 30 mM NaCl, 3 mM $MgCl_2$, 100 ng/µL BSA, 5%
834 glycerol, and 1 mM DTT. WRN enzyme reactions were done in buffer consisting of 30 mM
835 HEPES pH 7.4, 40 mM KCl, 8 mM $MgCl_2$, 100 ng/µL BSA, 5% glycerol and 2 mM ATP.
836 For *in situ* experiments in which plugs were treated with both WRN and MUS81/EME1, after
837 WRN incubation, agarose plugs were treated with proteinase K in lysis buffer at 50°C for 1
838 hour and then washed in 10 mM Tris-HCl pH 8.0 solution, before incubation with MUS81-
839 EME1.

840 **END-seq**

841 For END-seq, 7-8 million cells were haRvested, embedded in 1% agarose plugs, and
842 processed as previously reported[9,12].

843 **RPA ChIP-seq**

844 20 million cells were harvested and ChIP-seq was performed as previously reported[12].

845 **PCR-free whole genome sequencing**

846 Genomic DNA was isolated from cultured cells by phenol-chloroform extraction. To provide
847 sufficient mass and maximize library diversity, two identical library preparations were
848 performed in parallel and combined for each sample. DNA (1000 ng) was fragmented on a

Covaris S2 sonicator, using 50 µl microTUBE AFA Snap-Caps (Intensity=5, Duty Cycle=5%, Cycles per Burst=200, Time=35 sec). Sequencing libraries were prepared using PCR-free KAPA HyperPrep Kit (Roche) according to the manufacturer's instructions. During ligation, standard adaptors were substituted for xGen Dual Index UMI Adapters (IDT). Following ligation, libraries were size selected using SPRIselect Beads (Beckman Coulter) twice, at concentrations of 0.7x and 0.5x by volume.

Sequencing of END-seq and ChIP seq libraries

END-seq and ChIP-seq libraries were sequenced on the Nextseq 500 or Nextseq 550 platform (Illumina), using 75bp single-end read kits. WGS libraries were sequenced on the Novaseq 6000 platform (Illumina), using 250bp paired-end kits.

Continuous long-read sequencing (CLR)

High molecular weight genomic DNA was purified using the MagAttract HMW DNA kit (Qiagen). For CLR library preparation, ≥5 ug of high molecular weight genomic DNA (more than 50% of fragments ≥40 kb) was sheared to ~40 kb using the Megaruptor 3 (Diagenode B06010003), followed by DNA repair and ligation of PacBio adapters using the PacBio SMRTbell Express Template Prep Kit 2.0 (100-938-900). Libraries were then size-selected for >30 kb using a BluePippin instrument with 0.75% agarose cassettes (Sage Science). Following quantification with the Qubit dsDNA High Sensitivity assay (Thermo Q32854), libraries were diluted to 50 pM per SMRT cell, hybridized with PacBio V2 sequencing primer, and bound with SMRT seq polymerase using Sequel II Binding Kit 2.0 (PacBio 101-842-900). CLR sequencing was performed on the Sequel II instrument using 8M SMRT Cells (101-389-001) and Sequel II Sequencing 2.0 Kit (101-820-200), with a 15 hour movie time per SMRT cell. Initial quality filtering, basecalling, and adapter marking was done automatically on board the Sequel II.

Polymerase chain reaction

PCR reactions consisted of 1x KAPA LongRange Buffer, 0.5 U KAPA LongRange HotStart Polymerase (Roche), 1.75 mM $MgCl_2$, 300 µM dNTPs, 0.4 µM of each primer, and 75 ng of genomic DNA used as template. Samples were initially denatured at 95°C for 3 minutes, underwent 32 cycles of: 1) denaturation at 95°C for 30 seconds and 2) annealing/extension at 60°C for 3 minutes, followed by a final extension at 60°C for 10 minutes. PCR products were separated on a 2% agarose gel and visualized by 0.5 µg/mL ethidium bromide staining.

PCR primer sequences are as follows, FW/RV indicate forward and reverse primers, respectively:

B1 Fw: 5'-GCAACCAGCTGTTTTTGTGA-3'

B1 Rv: 5'-GCAATAGTATGCAGCTTGCCC-3'

B2 Fw: 5'-TTTGCATCCTGCTTTTCTCATCT-3'

B2 Rv: 5'-GAAGAGGTGCCTGGTAGCTG-3'

B3 Fw: 5'-TTTGGCTTAGGGGAAGTGTGG-3'

B3 Rv: 5'-GTTTTGAGCATGCTGACCTGA-3

B4 Fw: 5'-GCAAGAACCAAATGCTGCAC-3'

B4 Rv: 5'-ACTCCTGTTGCTCAGGCAAT-3'

B5 Fw: 5'-TGTCCGTGTCTCGAGGAGT-3'

B5 Rv: 5'-GTGTCTCCATCCATTGTTCTGC-3'

B6 Fw: 5'-TGCTTTCAACCTGCCCAAAC-3'

B6 Rv: 5'-GCACTTGAGCCTTGCTGGTA-3'

B7 Fw: 5'-TGTGGTTGTCTTCTCCACCC-3'

B7 Rv: 5'-AGCTGGGTGTTAAGGGATGAA-3'

B8 Fw: 5'-ATGGGATGGCCACACTGAAG-3';

B8 Rv: 5'-AACTTGCCTTTCACCTGCCT-3';

NB1 Fw: 5'-ATAGTCTGTCTCCCGCAGTCT-3';

NB1 Rv: 5’-GAGACCGCCGATTAGCATTC-3’

NB2 Fw: 5’-TACCTGACAGAACCACTGGC-3’

NB2 Rv: 5’-GACAAGGATTCCCCTCCTGC-3’

NB3 Fw: 5’-GTGGTGTGGTAAAGGGACCA-3’

NB3 Rv: 5’-CCTCTCCCTGTTAAGTCATTACC-3’

Genomic location and reference genome size for TA loci amplified by PCR

Genomic location and reference genome size of sites selected for PCR amplification are shown below . Each site amplified contains an uninterrupted stretch of (TA)n repeats and was amplified with primers localized in flanking areas outside of the (TA)n repeat. Broken sites B1-B8 were chosen based on the presence of END-seq peaks upon WRN depletion in KM12 cells. Sites NB1-NB3 were chosen with similar TA repeat lengths as broken sites but were not broken upon WRN depletion in KM12 cells.

Southern blotting

Native (1 % agarose) Southern blot analyses of broken and non broken (TA)n dinucleotide lengths were carried out using standard Southern blot techniques. Genomic DNA (10 ug) from indicated cell lines were digested with the indicated restriction enzymes: B1: EcoRV; B2: HindIII; NB1: ApaLI; NB3: HindIII. Probes were generated by PCR amplification using genomic DNA from the RPE-1 cell line using the following oligonucleotides:

B1 Fw: 5′-GTCAAGGAAAGCCAAGAATTGGAA-3′

B1 Rv: 5′-AGAGTTGGAATTTGAACCAAAGC-3′

B2 Fw: 5’-GAAACTGCCAAACACAGGGC-3’

B2 Rv: 5’-TCTTCAGCACAGGAGCAAGG-3’

NB1 Fw: 5’-GTAAGGAGTGTGTGTGGGGG-3’

NB1 Rv: 5’-ACTCATGGAGAATAGACTACGACT-3’

NB3 Fw: 5'-AGCCATGGTGTGTTTCTGGT-3'

NB3 Rv: 5'-GCCGTCTTCGAACCTGTAGA-3'

Quantification and Statistical Analysis

1) Genome Alignment: For Single-end sequencing, NextSeq 500/550 was utilized for single end sequencing using 75bp reads. The sequenced tags were aligned using Bowtie2 [40] with the parameters -N 0 -k 1 -q --local --fast-local. For paired end sequencing: NovoSeq 6000 was utilized for paired-end sequencing using 2x250bp reads. The sequenced tags were aligned using Bowtie2 with the parameters -N 0 – local.
2) Peak calling: Peaks were called for single end sequencing experiments using MACS 1.4.3[41] using the parameters -p 1e-5 --nolambda --nomodel --keep-dup = all (keep all redundant reads). For each case the experimental sample was doxycycline (shWRN) treated sample compared with DMSO (NT) sample as control. There were some additional filtering criteria as explained below:
    a) shWRN- ENDseq, RPA ChIPseq

The output of the peak calling was filtered by a 20-fold enrichment over background and a minimum size of 1kb. The resulting regions were merged using bedtools [42] merge -d 1000 function to define the list of regions.

    b) siWRN- ENDseq, MUS81-EME1- ENDseq

The output of the peak calling was filtered by a 20-fold enrichment over background. The resulting regions were merged using bedtools merge -d 1000 function to define the list of regions. For each region defined above the summit in positive and negative strand reads was evaluated using bedtools coverage -d function. The bedtools intersect function was used to filter the regions for which the positive summit was downstream

of the negative summit and the distance between the two summits was set to be between 30 to 2,500bp. This defined the final list of peaks for each experiment.

3) Motif Finding: For each peak the taller of the positive or negative summit was identified. From this taller summit a 1kb window was generated downstream (upstream) if the summit was on negative (positive) strand. MEME suite [43] was used to identify the common sequence motif in these genomic windows.
4) For each annotated (TA)n repeat sequence extracted from the hg19 genome sequence, three parameters were calculated: (a) TA proportion - the cumulative TA length divided by the length of the full annotated (TA)n repeat; (b) uninterrupted (TA)n proportion - the length of the longest continuous TA sequence divided by the length of the full annotated (TA)n repeat; (c) uninterrupted (TA)n length - the length of the longest continuous TA sequence within the annotated (TA)n repeat. These three parameters were plotted separately to compare between (TA)n repeats at broken and non-broken sites of shWRN-KM12 cells. This data was plotted in Fig. 6a-c, respectively. The same was done to calculate the proportion and length of (TA)n repeats at annonated (TA)n sequences in the KM12 cell line using long read sequencing data (Extended Data Fig. 7a-c).
5) Visualization of Sequencing Data:
    a) The alignment output sam files were converted and sorted into bam files using samtools [44] view function. For WGS paired-end data the bam files are directly visualized using IGV [45] from the Broad Institute. The following steps were carried out for single end sequencing experiments.
    b) The bamtobed function in bedtools package was used to convert the bam files to bed files

c) The genomecov function in bedtools package was used to convert the bedfiles into bedGraph. This file was converted to bigwig file using the UCSC tools package and visualized on the UCSC Genome Browser[46].

6) Expansion detection by ExpansionHunter and exSTRa: Expansions are identified for broken and non-broken (TA)n from the HCT116 and KM12 whole genome sequencing data by ExpansionHunter v3.2.2 [25,47]. The supporting reads for expansions are visualized by GraphAlignmentViewer (https://github.com/Illumina/GraphAlignmentViewer). Empirical cumulative distribution function for the TA repeat located at hg19 coordinates chr8:106950919-106950985 was generated using exSTRa v0.89.0 [26] and Bio-STR-exSTRa v1.1.0 using the default parameter settings.

7) Long read sequencing data processing: We aligned CLR reads to the hg19 reference sequence using minimap2[48] (with parameters -ayYL --MD --eqx -x map-pb) and called indel/structural variants with the pbsv tool using the --tandem-repeats argument for increased sensitivity to repeat expansion/contraction variants in accordance with the software's documentation (Pacific Biosciences. pbsv https://github.com/PacificBiosciences/pbsv). We extracted indel variants passing built-in filters and overlapping with repetitive loci of interest (6,915 "broken" loci, 59,729 "non-broken" loci) using the bedtools closest utility[42]. We computed coverage over each interval using the mosdepth[49] package and determined the repetitive motifs present using the Tandem Repeat Finder (trf) tool[50], lowering the score threshold in order to see all possible motifs (i.e. with parameters 2 5 7 80 10 0 2000 -l 6 -ngs -h). We quantified the copy number of each motif in the reference haplotype and the alternate haplotype (computed by permuting the reference haplotype with the variants

997 detected in each locus). We report the copy number change for the highest copy
998 number motif per locus.
999 8) Cancer cell line encyclopedia (CCLE) whole genome sequencing analysis: For 321
1000 cancer cell lines from the CCLE2 project[51] that had WGS data we determined the
1001 number of reads at each AT-rich interval using 'samtools bedcov'[44]. For each interval,
1002 we then defined: fragments per base (fpb) = number of reads coveraing the interval /
1003 (length of the interval in bases * 2). We then normalized this value across all intervals
1004 within the genome as: fragments per base per million (fpbm) = (fpb / total fpb) * 1
1005 million. This latter quantity (fpbm) was used to determine the read coverage at
1006 broken/non-broken loci. We determined overall read counts at loci for with high AT
1007 repeats
1008 9) MSI deletion analysis: MSI deletions in uterine corpus endometrial carcinoma
1009 (UCEC), colon adenocarcinoma (COAD), and gastric adenocarcinoma (STAD) [5] were
1010 identified from ICGC (https://dcc.icgc.org/releases/PCAWG/msi and
1011 https://dcc.icgc.org/releases/PCAWG/consensus_sv). In total, 396 deletions from 24
1012 MSI samples and 4501 deletions from 93 MSS samples were analyzed. 1000 random
1013 sets were generated using the bedtools[42] shuffle command in order to estimate
1014 enrichments for different repeat annotations relative to random.
1015 10) Statistical Analysis:
1016 a) Venn Diagrams
1017 For Venn diagrams 10,000 random sites were generated for one of the peak
1018 lists (while keeping others the same) using bedtools shuffle command. The
1019 significance was generated using the pnorm function in R.
1020 b) Boxplots

1021 The statistical significance of the data represented in boxplots are calculated
1022 using the wilcox.test function in R.
1023

37 Callen, E. *et al.* 53BP1 Enforces Distinct Pre- and Post-resection Blocks on Homologous Recombination. *Mol Cell* **77**, 26-38 e27, doi:10.1016/j.molcel.2019.09.024 (2020).
38 Palermo, V. *et al.* CDK1 phosphorylates WRN at collapsed replication forks. *Nat Commun* **7**, 12880, doi:10.1038/ncomms12880 (2016).
39 Chang, J. H., Kim, J. J., Choi, J. M., Lee, J. H. & Cho, Y. Crystal structure of the Mus81-Eme1 complex. *Genes Dev* **22**, 1093-1106, doi:10.1101/gad.1618708 (2008).
40 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
41 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
42 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
43 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208, doi:10.1093/nar/gkp335 (2009).
44 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
45 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
46 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102 (2002).
47 Dolzhenko, E. *et al.* ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754-4756, doi:10.1093/bioinformatics/btz431 (2019).
48 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
49 Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867-868, doi:10.1093/bioinformatics/btx699 (2018).
50 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).
51 Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508, doi:10.1038/s41586-019-1186-3 (2019).

Acknowledgements

We thank Roackie Awasthi for assistance with Southern Blotting; David Goldstein, Bao Tran and the CCR Genomics core for sequencing support; Michael Lawrence for computational assistance. Work in S.C.W.'s laboratory is supported by the Francis Crick Institute (FC10212) and the European Research Council (ERC-ADG-666400). The Francis Crick Institute receives

core funding from Cancer Research UK, the Medical Research Council, and the Wellcome Trust. K.F. is the recipient of fellowships from the Benzon Foundation and the Lundbeck Foundation. The P.J.M laboratory is funded by the MRC MR/R009368/1; A.C-M. is the recipient of a fellowship from AstraZeneca; E.C.M is supported by the Damon Runyon Cancer Research Foundation, and E.C.M and A.J.B are supported by a pilot grant from the Dana-Farber Department of Medical Oncology. The A.N. laboratory is supported by the Intramural Research Program of the NIH, an Ellison Medical Foundation Senior Scholar in Aging Award (AG-SS- 2633-11), the Department of Defense Idea Expansion (W81XWH-15-2-006) and Breakthrough(W81XWH-16-1-599) Awards, the Alex's Lemonade Stand Foundation Award, and an NIH Intramural FLEX Award.

Author Contribution

N.V.W. set up the project, performed END-seq and flow cytometry experiments upon WRN/MUS81/SLX4 depletion, and performed preliminary analysis of END-seq data; W.J.N. performed MUS81-EME1 *in situ* END-seq and PCR; A.T. performed END-seq, Southern Blotting and designed ATR mutant WRN cDNA; E.M.C. generated the inducible shWRN in KM12 and HCT116, performed and analyzed the HSEC western blot and viability experiments, long-read sequencing, and analyzed the CCLE/WRN dependency data;E.C. performed ATRi END-seq experiments, Western blotting, and metaphase analysis. V.T. performed RPA ChIP-seq; K.F. performed the HSEC and long-read sequencing experiments; N.W. performed Western blotting and helped generate WRN^{3A} and WRN^{6A} cells; J.N. and J.K. analyzed the CCLE/WRN dependency data; S.S. analyzed END-seq, RPA ChIP-seq experiments; W.W. analyzed WGS, PacBio coverage across repeats, deletion breakpoints in MSI cancers, and performed quantitative modeling; F.B. analyzed nucleotide composition of

broken vs. non broken repeats and replication timing; E.D. performed ExpansionHunter and exSTRa bioinformatic analysis; M.E. supervised computational work; K.G., Y.H., A.A.B., J.T.S., and N.K. analyzed the data and designed bioinformatic pipelines; R.W. prepared WGS libraries; A.C.M. and K.F. provided recombinant MUS81-EME1; J.S. provided recombinant WRN; B.E.H. provided advice about PCR across repeats; K.U. provided advice about repeat expansion biology; C.H.F. provided advice about secondary structure biology; R.M.B. provided advice about WRN helicase; S.W. provided advice about structure specific nucleases and recombination intermediates, P.J.M. helped design *in situ* experiments with recombinant proteins, P.S.M., provided advice on WGS experiments and analyses; A.J.B. and A.N. supervised the project; N.V.W., W.J.N, A.T., E.M.C., A.J.B. and A.N. wrote the manuscript with comments from the authors.

Author Information

These authors contributed equally: Niek van Wietmarschen, Sriram Sridharan, William J. Nathan, Anthony Tubbs, Edmond M. Chan

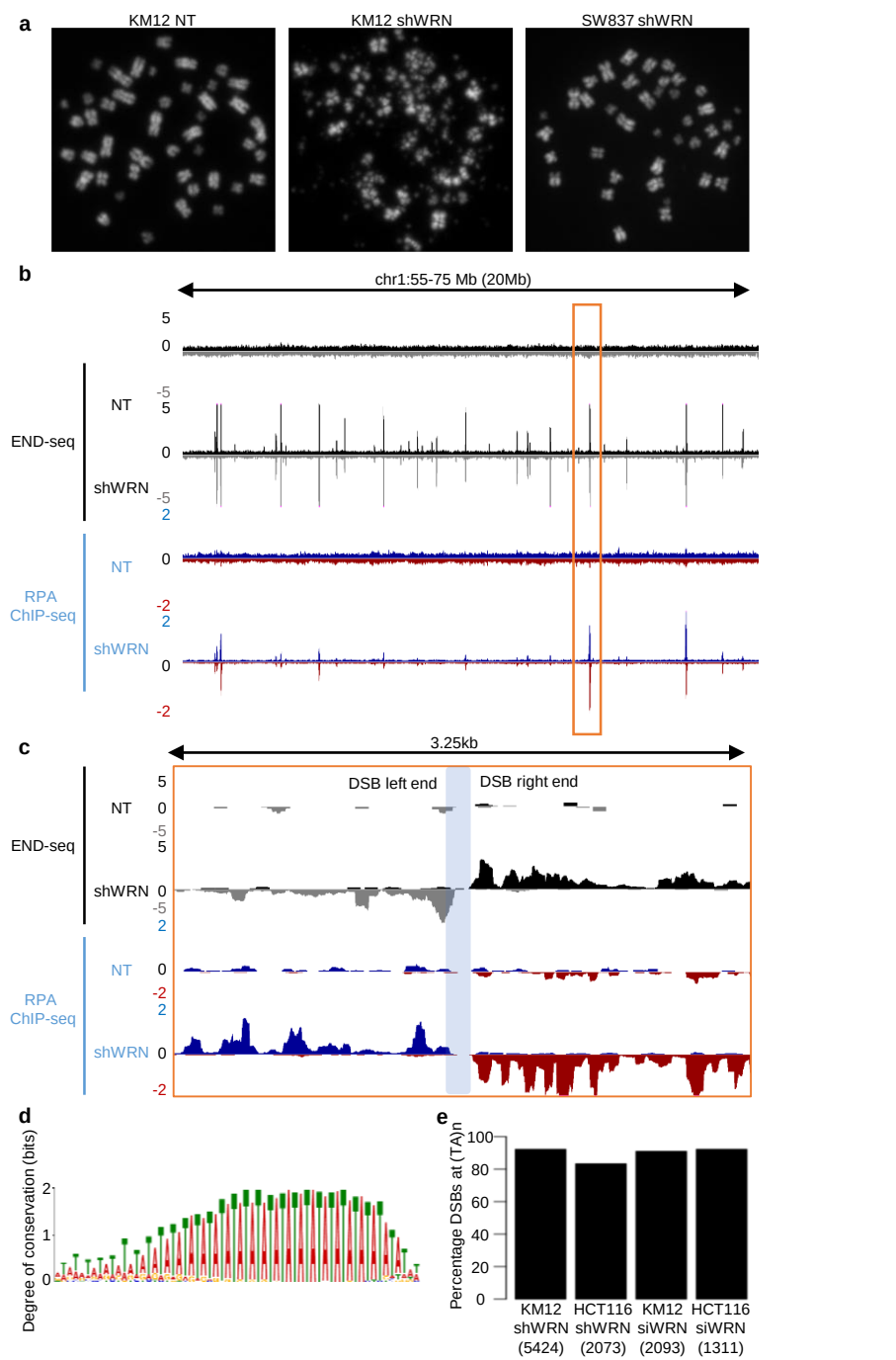
Correspondence and requests from materials: Andre Nussenzweig

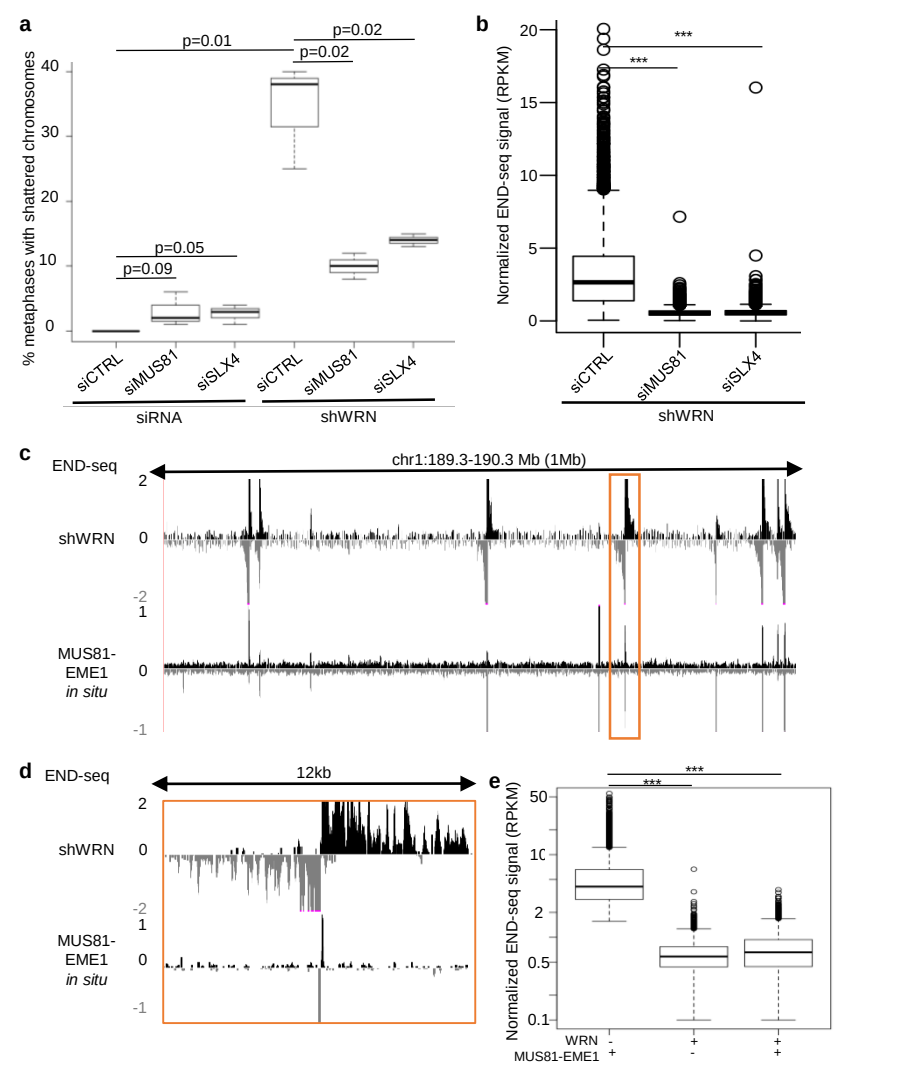
Competing Interest

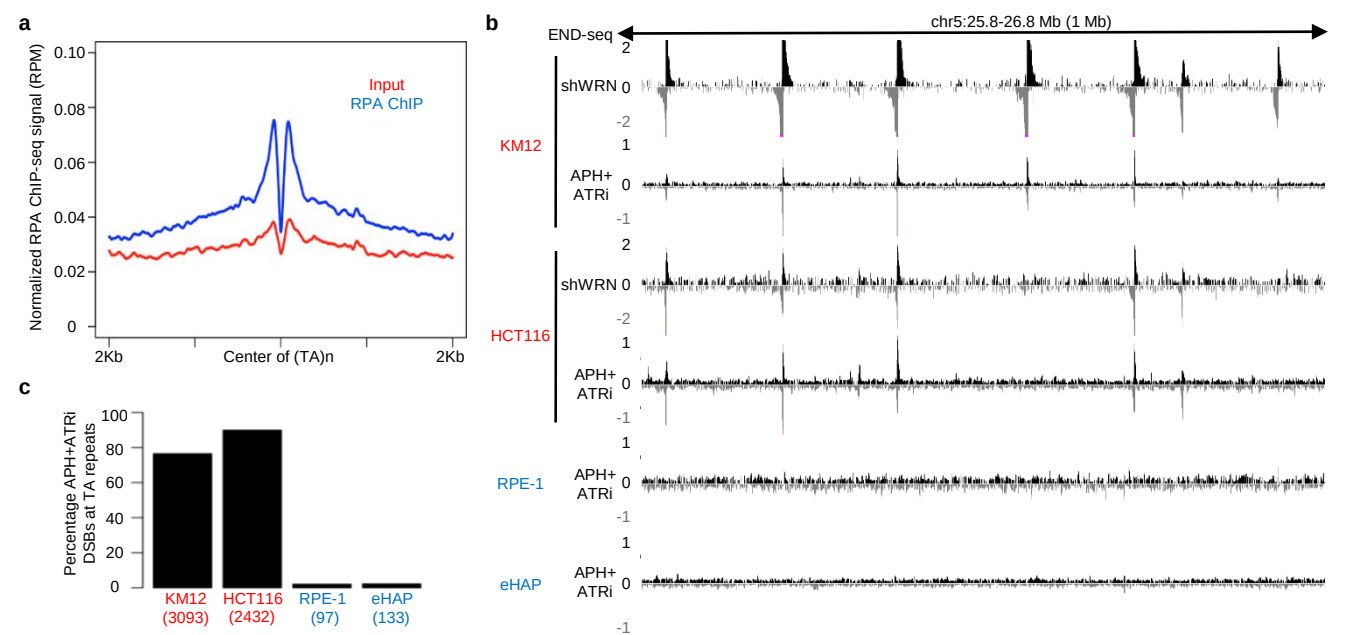
Adam J. Bass has research support from Bayer, Merck and Novartis.

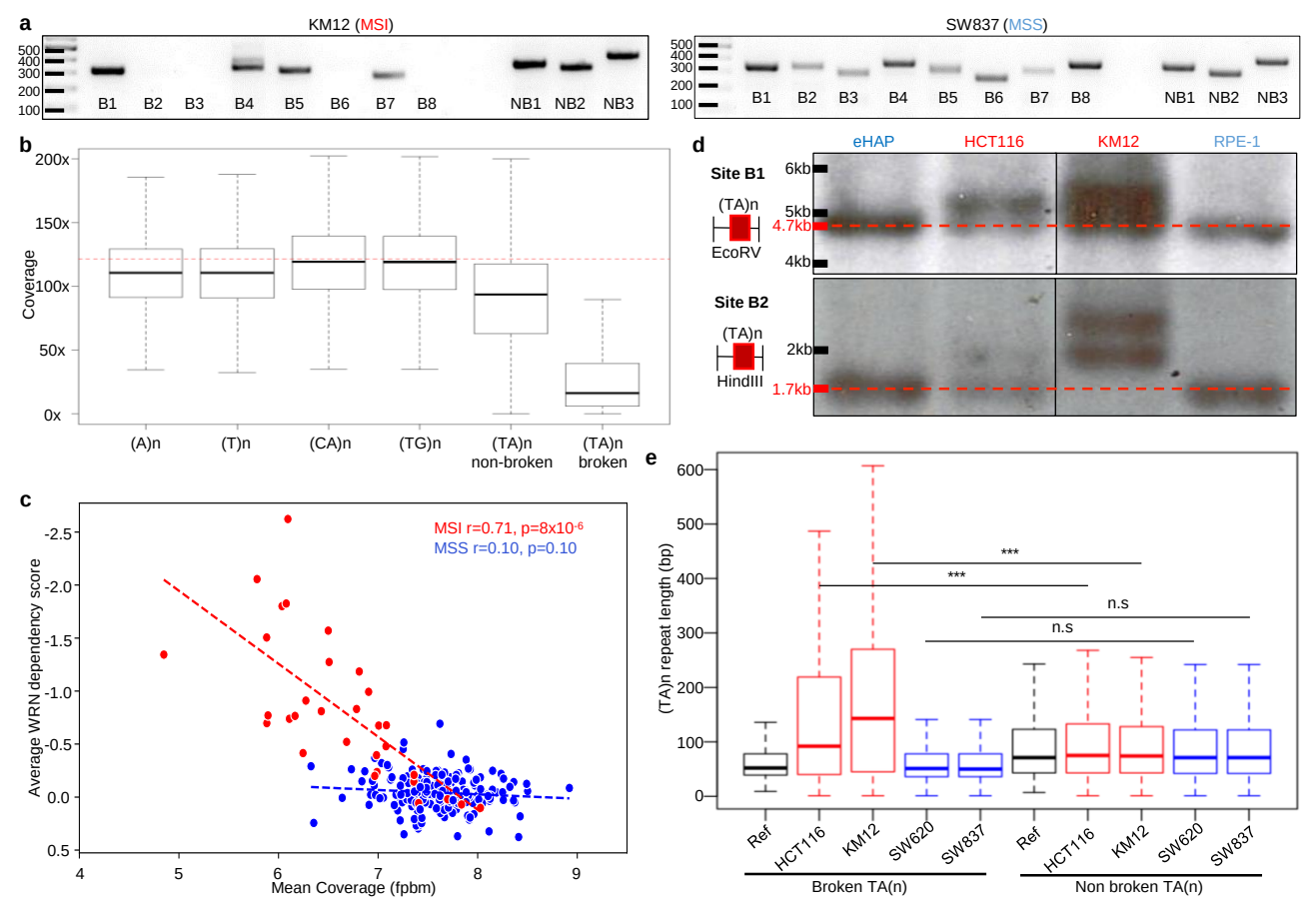
Data Availability

END-seq, ChIP-seq, WGS and Pacbio CLR data have been deposited in the Gene Expression Omnibus (GEO) database under the accession number GSE149709. Source data are provided with this paper.


a
KM12 NT
KM12 shWRN
SW837 shWRN
b
chr1:55-75 Mb (20Mb)
END-seq
NT
shWRN
RPA ChIP-seq
NT
shWRN
c
3.25kb
DSB left end
DSB right end
END-seq
NT
shWRN
RPA ChIP-seq
NT
shWRN
d
Degree of conservation (bits)
e
Percentage DSBs at (TA)n
KM12 shWRN (5424)
HCT116 shWRN (2073)
KM12 siWRN (2093)
HCT116 siWRN (1311)


a
% metaphases with shattered chromosomes
p=0.01
p=0.02
p=0.02
p=0.05
p=0.09
siCTRL
siMUS81
siSLX4
siCTRL
siMUS81
siSLX4
siRNA
shWRN
b
Normalized END-seq signal (RPKM)
***
***
siCTRL
siMUS81
siSLX4
shWRN
c
END-seq
chr1:189.3-190.3 Mb (1Mb)
shWRN
MUS81-EME1 in situ
d
END-seq
12kb
shWRN
MUS81-EME1 in situ
e
Normalized END-seq signal (RPKM)
***
***
WRN
MUS81-EME1


a
Normalized RPA ChIP-seq signal (RPM)
0.10
0.08
0.06
0.04
0.02
0
Input
RPA ChIP
2Kb
Center of (TA)n
2Kb
b
END-seq
chr5:25.8-26.8 Mb (1 Mb)
KM12
shWRN
APH+ ATRi
HCT116
shWRN
APH+ ATRi
RPE-1
APH+ ATRi
eHAP
APH+ ATRi
c
Percentage APH+ATRi DSBs at TA repeats
100
80
60
40
20
0
KM12 (3093)
HCT116 (2432)
RPE-1 (97)
eHAP (133)


a
KM12 (MSI)
500
400
300
200
100
B1
B2
B3
B4
B5
B6
B7
B8
NB1
NB2
NB3
SW837 (MSS)
b
200x
150x
100x
50x
0x
Coverage
(A)n
(T)n
(CA)n
(TG)n
(TA)n non-broken
(TA)n broken
d
eHAP
HCT116
KM12
RPE-1
Site B1
(TA)n
EcoRV
6kb
5kb
4.7kb
4kb
Site B2
(TA)n
HindIII
2kb
1.7kb
c
MSI r=0.71, p=8x10-6
MSS r=0.10, p=0.10
Average WRN dependency score
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
4
5
6
7
8
9
Mean Coverage (fpbm)
e
(TA)n repeat length (bp)
600
500
400
300
200
100
0
***
***
n.s
n.s
Ref
HCT116
KM12
SW620
SW837
Ref
HCT116
KM12
SW620
SW837
Broken TA(n)
Non broken TA(n)