Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS)

Daniela Nachmanson,[1] Shenyi Lian,[1,3] Elizabeth K. Schmidt,[1] Michael J. Hipp,[1] Kathryn T. Baker,[1] Yuezheng Zhang,[1] Maria Tretiakova,[1] Kaitlyn Loubet-Senear,[1] Brendan F. Kohrn,[1] Jesse J. Salk,[2,4] Scott R. Kennedy,[1,5] and Rosa Ana Risques[1,5]

[1]*Department of Pathology, University of Washington, Seattle, Washington 98195, USA;* [2]*Department of Medicine, Division of Medical Oncology, University of Washington, Seattle, Washington 98195, USA*

Next-generation sequencing methods suffer from low recovery, uneven coverage, and false mutations. DNA fragmentation by sonication is a major contributor to these problems because it produces randomly sized fragments, PCR amplification bias, and end artifacts. In addition, oligonucleotide-based hybridization capture, a common target enrichment method, has limited efficiency for small genomic regions, contributing to low recovery. This becomes a critical problem in clinical applications, which value cost-effective approaches focused on the sequencing of small gene panels. To address these issues, we developed a targeted genome fragmentation approach based on CRISPR/Cas9 digestion that produces DNA fragments of similar length. These fragments can be enriched by a simple size selection, resulting in targeted enrichment of up to approximately 49,000-fold. Additionally, homogenous length fragments significantly reduce PCR amplification bias and maximize read usability. We combined this novel target enrichment approach with Duplex Sequencing, which uses double-strand molecular tagging to correct for sequencing errors. The approach, termed CRISPR-DS, enables efficient target enrichment of small genomic regions, even coverage, ultra-accurate sequencing, and reduced DNA input. As proof of principle, we applied CRISPR-DS to the sequencing of the exonic regions of *TP53* and performed side-by-side comparisons with standard Duplex Sequencing. CRISPR-DS detected previously reported pathogenic *TP53* mutations present as low as 0.1% in peritoneal fluid of women with ovarian cancer, while using 10- to 100-fold less DNA than standard Duplex Sequencing. Whether used as standalone enrichment or coupled with high-accuracy sequencing methods, CRISPR-based fragmentation offers a simple solution for fast and efficient small target enrichment.

[Supplemental material is available for this article.]

In the last decade, next-generation sequencing (NGS) has revolutionized the fields of biology and medicine. However, standard NGS suffers from two major problems that negatively impact multiple applications: the limited efficiency of current target selection methods and the high error rate of the sequencing process. Targeted genome enrichment is essential to many applications that do not require whole-genome sequencing, and it is performed either by PCR or by hybridization capture. PCR is simple and efficient but does not scale well and suffers from biases that result in uneven coverage and false mutation calls (Kebschull and Zador 2015; Samorodnitsky et al. 2015). Hybridization capture improves coverage uniformity and mutation call accuracy but has low recovery, especially when the target region is small, which leads to the requirement of larger amounts of DNA (Samorodnitsky et al. 2015). An additional complication is that DNA is typically fragmented by sonication, which introduces DNA damage resulting in sequencing errors (Park et al. 2017). Moreover, the heterogeneous fragment sizes generated by sonication are subject to PCR bias and contribute to uneven coverage. An alternative option to sonication is enzymatic fragmentation. This method resolves some issues but introduces different artifacts that also result in sequencing errors (Knierim et al. 2011). Thus, at the library preparation step, both methods of target selection suffer important limitations that lead to nonoptimal sequencing outcomes, including uneven coverage, introduction of false mutations, and low recovery.

The second major problem of NGS is the high error rate inherent to the sequencing process. Illumina currently offers the most accurate sequencing platform with an estimated error rate of 10^{-3} (Goodwin et al. 2016). This error rate, however, translates

Present addresses: [3]Key laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Pathology, Peking University Cancer Hospital and Institute, 100142 Beijing, China; [4]TwinStrand Biosciences, Seattle, WA 98121, USA
[5]These authors contributed equally to this work.
Corresponding authors: scottrk@uw.edu, rrisques@uw.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.235291.118.

into thousands of false calls in each sequencing run and precludes the detection of low-frequency mutations, which is critical for applications such as forensics, metagenomics, and oncology (Salk et al. 2018). Sequencing errors can be significantly reduced by the use of molecular barcodes, which are random DNA sequences attached to the original DNA molecules before or during PCR. Single-stranded molecular barcodes, also known as "unique molecular identifiers" (UMIs), produce a consensus with the reads derived from one DNA strand (Kinde et al. 2011), whereas double-stranded molecular barcodes introduce an additional level of correction by allowing the comparison of independent consensus sequences derived from the two complementary strands of the original DNA molecule (Schmitt et al. 2012). This additional level of correction is essential for removing polymerase errors occurring in the first round of PCR and subsequently propagated to all reads derived from a given DNA strand (Arbeithuber et al. 2016). Duplex Sequencing (DS), the method that pioneered double-strand molecular barcodes (Schmitt et al. 2012; Kennedy et al. 2014), has an estimated error rate $<10^{-7}$, two orders of magnitude less than single-strand molecular barcode methods. This translates into the confident detection of mutations present at frequencies $<10^{-5}$, whereas single-strand molecular barcode methods show substantial decrease in accuracy at frequencies $\lesssim 10^{-3}$ (Salk et al. 2018). The extreme sensitivity of DS has been used in a variety of applications including the detection of very low-frequency somatic mutations in cancer and aging (Kennedy et al. 2013; Ahn et al. 2016; Hoekstra et al. 2016; Krimmel et al. 2016; Reid-Bayliss et al. 2016).

DS successfully addresses the problem of sequencing errors, but it suffers from the limitations of hybridization capture, which is required to perform target selection while preserving the strand recognition of molecular barcodes. As described above, hybridization capture is highly inefficient when selecting small targets (Winters et al. 2017), estimated at only 5%–10% of reads on-target for targets <50 kb (Schmitt et al. 2015). In DS, as well as in other panel-based sequencing approaches, the region of interest is usually designed to be small as a cost-effective trade-off for higher sequencing depth. In this situation, a successful approach for target enrichment is to perform two consecutive rounds of capture (Schmitt et al. 2015). However, this approach results in a time consuming, costly, and inefficient protocol that requires large amounts of DNA (Kennedy et al. 2014). For example, in DS at least 1 µg of DNA has historically been needed to produce depths greater than 3000× (Krimmel et al. 2016), which is prohibitive in many applications that rely on small samples.

Here, we present CRISPR-DS, a new method that addresses the two main problems of NGS: limited efficiency of target selection and high error rate. Target selection is facilitated by an enrichment of the regions of interest using the CRISPR/Cas9 system. In vitro digestion with CRISPR/Cas9 has been proven to be a useful tool for multiplexed excision of large megabase fragments and repetitive sequence regions for PCR-free NGS (Bennett-Baker and Mueller 2017; Shin et al. 2017). We reasoned that targeted in vitro CRISPR/Cas9 digestion could be used to excise similar length fragments covering the area of interest, which could then be enriched by size selection prior to library preparation. We designed this method to enable target enrichment while simultaneously eliminating sonication-related errors and biases arising from random genome fragmentation. In addition, by pairing this approach with double-strand molecular barcoding, we aimed to produce a method that preserves the sequencing accuracy of DS while increasing the recovery rate, thus enabling low DNA input and a simplified protocol for translational applications.

Results

Design of CRISPR-DS based on CRISPR/Cas9 target fragmentation and double-strand molecular barcodes

CRISPR-DS is based on in vitro CRISPR/Cas9 excision of target sequences to generate DNA molecules of uniform length that are then enriched by size selection. The versatility, specificity, and multiplexing capabilities of the CRISPR/Cas9 system enable its application for the excision of any target region of interest by simply designing guide RNAs (gRNA) to the desired cutting points. As a proof of principle, we developed the method for sequencing the exons of *TP53*. Further, in order to achieve high recovery and sequencing accuracy, we combined it with DS. The main steps of the protocol are illustrated in Figure 1. First, target regions are excised from genomic DNA by multiplexed in vitro CRISPR/Cas9 digestion (Fig. 1A), followed by enrichment of the excised fragments by size selection using SPRI beads (Fig. 1B). The selected fragments are then coupled with the double-strand molecular barcodes used in DS (Fig. 1C; Kennedy et al. 2014). These fragments are then amplified and captured with biotinylated hybridization probes as previously described for DS (Kennedy et al. 2014), with the exception that only one round of hybridization capture is required due to the prior enrichment of target fragments (see below). Finally, the library is sequenced, and the resulting reads are analyzed to perform error correction based on the consensus sequences of both strands of each DNA molecule (Fig. 1D; Kennedy et al. 2014). Due to the requirement of only one round of hybridization capture, the workflow of CRISPR-DS is almost one day shorter than standard-DS (Fig. 2; Supplemental Fig. S1), enabling a more cost-efficient and applicable method.

CRISPR/Cas9-cut fragments can be designed to be of homogenous length, reducing PCR bias and producing uniform coverage

Typically, genome fragmentation is performed with sonication, which generates randomly sized fragments that have different amplification efficiencies (Dabney and Meyer 2012). Short fragments are preferentially amplified, resulting in uneven coverage of the regions of interest and decreased recovery. In DS, amplification bias introduces an additional problem because short fragments produce an excess of PCR copies that do not further aid error reduction. To produce a consensus, only three PCR copies of the same molecule are required. Additional copies waste resources because they produce sequencing reads but do not generate additional data. By using CRISPR/Cas9, gRNA can be designed such that restriction with Cas9 produces fragments of predefined, homogeneous size. We reasoned that these fragments would eliminate PCR bias, leading to homogeneous sequencing coverage and minimizing wasted reads that are PCR copies of the same original molecule.

To test this approach, we designed gRNAs to specifically excise the coding regions and their flanking intronic sequence of *TP53* (Fig. 1A). Fragment length was designed to be ~500 bp in order to maximize read space of an Illumina MiSeq v3 600 cycle kit while allowing for sequencing of the molecular barcode (10 bp) and 3′-end clipping of 30 bp to remove low-quality bases produced in the later sequencing cycles. gRNAs were selected based on the highest specificity score that produced appropriate fragment length (Supplemental Table S1; Supplemental Data S1; Hsu et al. 2013). The fragment comprising exon 7 was designed to be shorter than the rest (336 bp) to avoid a homopolymeric run of T's in the

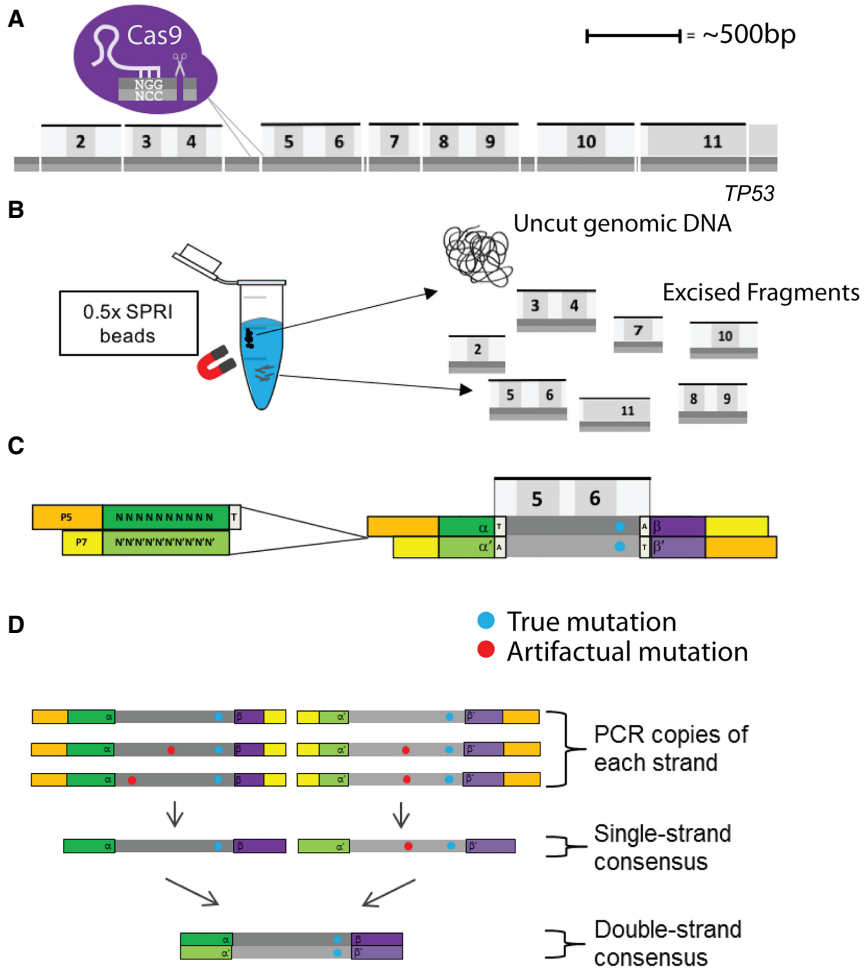

Figure 1. Schematic representation of key aspects of CRISPR-DS. (*A*) CRISPR/Cas9 digestion of *TP53*. Seven fragments containing all *TP53* coding exons were excised via targeted cutting using gRNAs. Dark gray represents reference strand, and light gray represents the anti-reference strand. (*B*) Size selection using 0.5× SPRI beads. Uncut, genomic DNA binds to the beads and allows the recovery of the homogenously sized excised fragments in solution. (*C*) Double-stranded DNA molecule fragmented and ligated with double-stranded DS adapters. Adapters contain 10 bp of random, complementary nucleotides and a 3′-dT overhang. (*D*) Error correction by DS. After creating single-strand consensus sequence (SSCS) reads, SSCS reads derived from the same original DNA molecule are compared with one another to create a double-strand consensus sequence (DCS). Only mutations found in both SSCS reads are counted as true mutations in DCS reads.

flanking intronic region which induced poor base quality in reads that span this region (Supplemental Fig. S2).

We performed a side-by-side comparison of library performance (Fig. 3A–C) and sequencing coverage (Fig. 3D) of a sample DNA processed with CRISPR-DS versus standard-DS (Methods). Standard-DS for *TP53* had been previously performed using sonication and published protocols (Kennedy et al. 2014; Krimmel et al. 2016). Visualization of the resulting sequencing library by gel electrophoresis showed that CRISPR restriction produced distinct bands/peaks (Fig. 3A,B) corresponding to the predesigned size of target fragments as opposed to the characteristic "smear" of libraries prepared by sonication. The discrete peaks allow confirmation of correct library preparation and target enrichment, preventing the sequencing of suboptimal libraries. Sequencing and mapping of the libraries demonstrated that targeted Cas9 restriction results in well-defined DNA fragments corresponding to the expected size (Fig. 3D). Importantly, these fragments exhibited extremely uniform sequencing depth. In contrast, sonicated DNA fragments resulted in significant variability in depth across target regions. Because DS reads correspond to individual DNA molecules, the uniform depth achieved by CRISPR-DS indicates a homogenous representation of the original genomic DNA in the final sequencing output, confirming the proper excision of all fragments.

The ability to uniformly control the DNA insert size should not only provide homogenous depth, but also a more uniform number of copies of each molecule, minimizing the waste of unnecessary reads to produce a consensus sequence. We examined this possibility by counting the number of PCR copies for each molecular barcode and plotting it as a function of the DNA fragment size (Fig. 3C). Sonicated DNA exhibited a strongly negative association between DNA fragment size and the number of PCR copies as expected because small DNA fragments are preferentially amplified (Fig. 3C, blue). In contrast, targeted fragmentation produced a consistent number of PCR copies for all fragments, including the smaller exon 7 fragment (Fig. 3C, red).

CRISPR/Cas9-cut fragments can be designed to be of optimal length to maximize read usage

An additional disadvantage of the variable fragment size produced by sonication is inefficient read usage: fragments that are too short generate overlapping reads that waste sequencing space, whereas fragments that are too long get sequenced on the ends, leaving captured but unsequenced DNA in the middle (Fig. 4A). The programmable nature of Cas9 can be leveraged to reduce the amount of data "lost" by generating optimal length fragments tailored to the preferred number of sequencing cycles. To illustrate the improvement in read usage, we quantified the amount of deviation from the optimal fragment size (defined as the total number of sequencing cycles minus the total length of the molecular barcodes and 3′-end clipping) of seven samples independently processed with sonication and targeted fragmentation. Sonication produced significant variability in the amount of deviation from the optimal fragment size with a large fraction of fragments being twice the optimal size for one of the samples (Fig. 4B,C; Supplemental Fig. S3). Indeed, only 9.1 ± 4.2% of reads had inserts that were within 10% deviation from the optimal fragment length. Even samples with more stringent size selection had only ~61% of reads within the 10%-deviation window (Fig. 4C; Supplemental Fig. S3). In contrast, the same samples fragmented with Cas9 had 71.0 ± 3.2% of reads within the same window range, with the vast majority of the reads outside the window being due to the

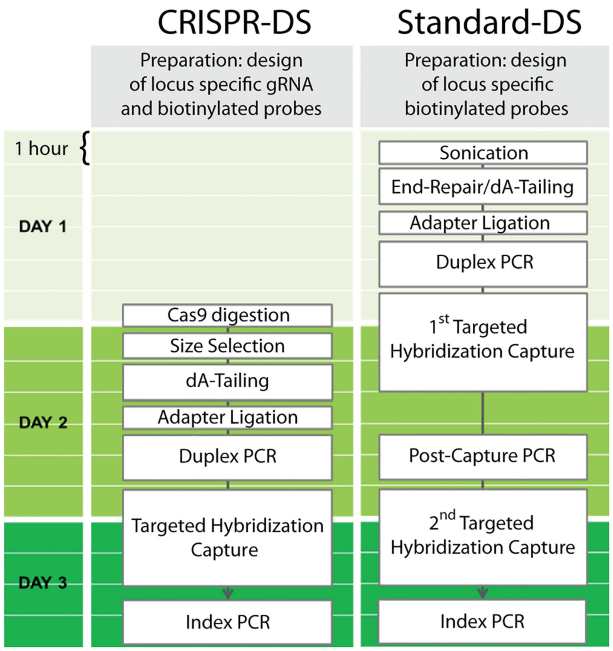

Figure 2. Comparison of library preparation protocols for standard-DS versus CRISPR-DS. The primary differences between the CRISPR-DS and standard-DS library preparation are the fragmentation methods and the number of hybridization capture steps. Instead of fragmentation by sonication as performed in standard-DS, CRISPR-DS relies on an in vitro excision of target regions by CRISPR/Cas9 followed by size selection for the excised fragments. Although this method requires additional preparation to design locus-specific gRNAs, this is a one-step process that then reduces the protocol by nearly a day. The reduction is achieved by the elimination of the second round of hybridization capture, which is required for sufficient target enrichment in the standard-DS protocol but not in CRISPR-DS. Colored boxes represent 1 h of time.

purposefully shorter exon 7 fragment (Fig. 4B,C; Supplemental Figs. S2, S3). Exclusion of exon 7 from this analysis improved the percent of reads within the 10%-deviation window to 94.3 ± 2.1%. These data indicate that targeted fragmentation can tightly control the fragment size to optimize read usage, thereby increasing the efficiency of sequencing.

CRISPR/Cas9 fragmentation enables target enrichment by size selection, eliminates one round of hybridization capture, and increases sequencing yield

Although performing two rounds of capture substantially increases the number of on-target reads for standard-DS and other small target applications, the process is time consuming, expensive, and requires additional PCR steps that introduce further bias (Schmitt et al. 2015). We hypothesized that target enrichment via size selection of CRISPR/Cas9-digested fragments would sufficiently enrich for on-target DNA fragments and eliminate the need for a second capture. To test this hypothesis, we performed CRISPR/Cas9 digestion of targeted *TP53* exons (Fig. 1A) on a range of DNA input amounts (10–250 ng) followed by SPRI size selection to remove undigested high molecular weight DNA fragments (>1 kb in size). The selected DNA fragments were ligated to DS adapters, PCR amplified, and sequenced (Methods). No hybridization capture or any other type of target enrichment was performed. Mapping of raw reads revealed between 0.2% and 5% reads on-target (i.e., covering *TP53*) (Table 1). Because the *TP53* target region only amounts to 0.0001% of the human genome, this corresponds to approximately 2000× to 50,000× enrichment, which matches or exceeds what is typically achieved with solution-based hybridization for small target size (Schmitt et al. 2015; Winters et al. 2017). Notably, lower DNA inputs showed the highest enrichment, potentially reflecting more efficient digestion or improved removal of off-target, high molecular weight DNA fragments when they are in lower abundance.

These results suggested that a simple size selection step can be used in lieu of a targeted hybridization enrichment step. To test this possibility, we performed a side-by-side comparison of standard-DS (both with one and two rounds of hybridization capture) (Kennedy et al. 2014) and CRISPR-DS with only one round of hybridization capture. Three input amounts of the same control DNA extracted from normal human bladder tissue were sequenced in parallel for each of the methods. CRISPR-DS with one round of capture achieved >90% raw reads on-target (Fig. 5A), a significant improvement over standard-DS, which only achieved ~5% raw reads on-target with a single capture, consistent with prior work (Schmitt et al. 2015). In an independent experiment, we tested the reproducibility of this result with three different DNA samples that were sequenced with CRISPR-DS using one and two rounds of capture (Supplemental Fig. S4). Confirming the prior result, the three samples produced >90% raw reads on-target using only one round of capture. The second round of capture only minimally increased raw reads on-target and is, therefore, unnecessary.

The side-by-side comparison of CRISPR-DS versus standard-DS also demonstrated a substantial increase in recovery using CRISPR-DS. Sequencing recovery, also referred to as yield, is typically measured as the fraction or percentage of sequenced genomes equivalents compared to input genomes. Consistent with prior studies (Schmitt et al. 2012; Krimmel et al. 2016), standard-DS produced a recovery rate of ~1% across the different inputs, whereas CRISPR-DS recovery rate ranged between 6% and 12% (Fig. 5B). Notably, 25 ng of DNA prepared with CRISPR-DS produced a post-processing depth comparable to 250 ng with standard-DS (Fig. 5C), indicating that size selection for excised fragments not only removes a step from the library preparation, but increases the recovery of input DNA, thereby enabling deep sequencing with greatly reduced DNA requirements.

Validation of CRISPR-DS recovery in an independent set of samples, including low-quality DNA

We further confirmed the performance of CRISPR-DS in an independent set of 13 DNA samples extracted from bladder tissue (Supplemental Table S3). We used 250 ng and obtained a median DCS depth of 6143×, corresponding to a median recovery rate of 7.4%, in agreement with the prior experiment. Reproducible performance was demonstrated with technical replicates for two samples (B2 and B4) (Supplemental Table S3). All samples had >98% reads on-target after consensus making, but the percentage of on-target raw reads ranged from 43% to 98%. We noticed that the low target enrichment corresponded to samples with DNA integrity number (DIN) less than 7. DIN is a measure of genomic DNA quality ranging from 1 (very degraded) to 10 (not degraded) (Jung et al. 2014). We reasoned that degraded DNA compromises enrichment by size selection, and the poor yield could be mitigated by removing low molecular weight DNA prior to CRISPR/Cas9 digestion. To test this hypothesis, we used the pulse-field feature of the BluePippin system to select high molecular weight DNA (>8 kb) from two samples with degraded DNA (DINs 6 and 4).

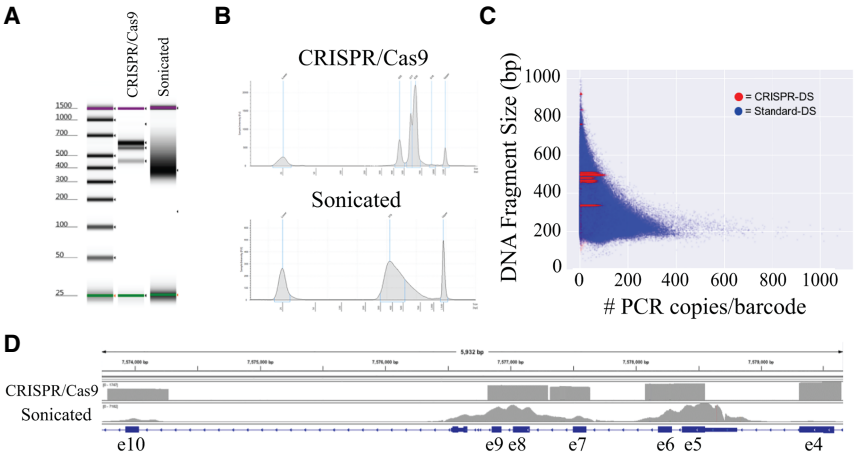

Figure 3. Visualization of sequencing libraries and data prepared with CRISPR-DS and standard-DS. (*A*) TapeStation gels show distinct bands for CRISPR-DS as opposed to a smear for standard-DS. The size of bands corresponds to the CRISPR/Cas9-cut fragments with adapters. (*B*) CRISPR-DS electropherograms allow visualization and quantification of peaks for quality control of the library prior to sequencing. Standard-DS electropherograms show a diffuse peak that harbors no information about the specificity of the library. (*C*) Dots represent original barcoded DNA molecules. Each DNA molecule has multiple copies generated at PCR (*x*-axis). In CRISPR-DS, all DNA molecules (red dots) have preset sizes (*y*-axis) and generate a similar number of PCR copies. In standard-DS, sonication shears DNA into variable fragment lengths (blue dots). Smaller fragments amplify better and generate an excess of copies that waste sequencing resources. (*D*) Integrative Genomics Viewer of *TP53* coverage with DCS reads generated by CRISPR-DS and standard-DS. CRISPR-DS shows distinct boundaries that correspond to the CRISPR/Cas9 cutting points and an even distribution of depth across positions, both within a fragment and between fragments. Standard-DS shows the typical "peak" pattern generated by random shearing of fragments and hybridization capture, which leads to variable coverage.

This pre-enrichment resulted in successful removal of low molecular weight products and increased on-target raw reads by twofold and DCS depth by fivefold (Supplemental Fig. S5). These results indicate that enrichment of high molecular weight DNA could be used as a solution for successful CRISPR-DS performance in partially degraded DNA.

Validation of CRISPR-DS for the detection of low-frequency mutations

Because CRISPR-DS uses a double-strand barcoding technique identical to standard-DS, it should theoretically have the same ability to identify low-frequency mutations. To prove this point, we sequenced a defined mixture of mutations with both CRISPR-DS and standard-DS. Two samples with known *TP53* variants were mixed at dilutions of 1:1, 1:10, 1:100, and 1:1000. Because the spiked-in sample was heterogenous, this experiment yielded a wide range of expected MAF to be compared with the MAF obtained by CRISPR-DS and standard-DS (Supplemental Fig. S6). The two methods showed very high correlations between expected and observed MAF (CRISPR-DS $R^2 = 0.980$, standard-DS $R^2 = 0.984$), as well as very high correlation between observed MAFs with each method ($R^2 = 0.996$). Most important, both methods could detect mutations at frequencies of roughly 0.001, and CRISPR-DS, but not standard-DS, detected an expected mutation at frequency of 0.0005. No additional, unexpected mutations were detected with any of the methods. Thus, these data demonstrate that the extremely high sensitivity and accuracy of double-strand molecular barcoding used by standard-DS is preserved with CRISPR-DS.

To validate the sensitivity of CRISPR-DS with clinical samples, we analyzed four peritoneal fluid samples collected during gynecological surgery from women with ovarian cancer and previously analyzed for *TP53* mutations using the standard-DS protocol (Krimmel et al. 2016). The tumor mutation was previously identified in the four samples: in one sample at a high frequency (68.5%) and at a very low frequency (around or below 1%) in the remaining three samples. CRISPR-DS detected the tumor mutation in all samples at frequencies comparable to what was reported in the original study (Table 2; Krimmel et al. 2016). In addition to the tumor mutation, standard-DS also revealed the presence of extremely low frequency (<0.1%) *TP53* mutations in these samples. These "biological background" mutations are not tumor-derived, but age-related (Krimmel et al. 2016). Standard-DS detected between one and five biological background mutations in each of the samples, representing an overall mutation frequency of about 1×10^{-6}. Similarly, CRISPR-DS identified biological background mutations in the four samples at a comparable overall mutation frequency (Supplemental Fig. S7). These results indicate that CRISPR-DS matches the performance of standard-DS in clinical samples (Krimmel et al. 2016).

Table 2 also illustrates a critical advantage of CRISPR-DS compared to standard-DS in terms of translational applicability: the reduced requirement of input DNA. Standard-DS of these peritoneal fluid samples required between 3 and 10 µg of DNA to compensate for the ~1% recovery rate of standard-DS and to achieve the high depth necessary to detect low-frequency tumor mutations. With CRISPR-DS, we only used 100 ng of DNA (30- to 100-fold less than what was used for standard-DS) and obtained comparable DCS depth to standard-DS (Table 2). Recovery rates ranged between 6% and 12%, as in prior experiments (Fig. 5; Supplemental Table S3). These results represent an efficiency increase of 15× to 200× compared to standard-DS with the same DNA. Notably, CRISPR-DS not only preserved sensitivity for mutation detection, increased sequencing recovery, and reduced DNA input, but also shortened the protocol by nearly one day (Supplemental Fig. S1), making it a more cost-effective option for accurate deep sequencing of samples with limited DNA amounts.

Discussion

Here we have developed a new approach for target enrichment based on CRISPR/Cas9 fragmentation followed by size selection, and we combined this approach with DS, producing a new method called CRISPR-DS. CRISPR-DS merges the increased efficiency provided by CRISPR-based targeted genome fragmentation with the high accuracy of sequencing provided by double-strand molecular barcodes, thus enabling ultra-accurate sequencing of small target regions using minimal DNA inputs. In addition to CRISPR-DS, the CRISPR-based target enrichment approach can be used in combination with other methods for targeted sequencing to improve recovery of small targets and to reduce PCR bias and uneven coverage arising from random fragment sizes.

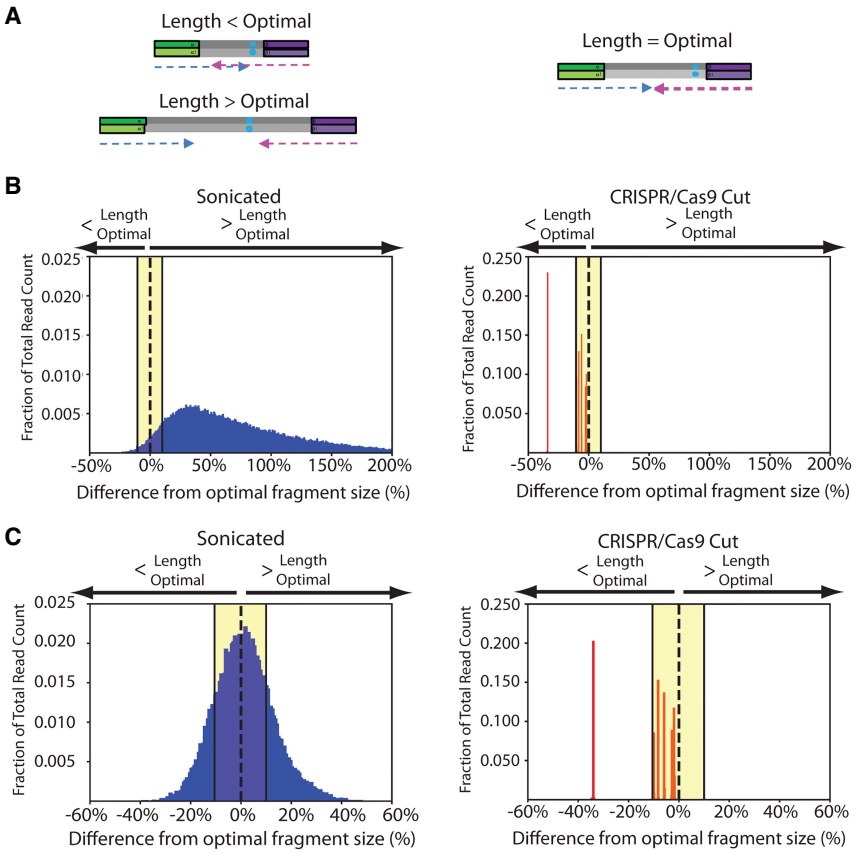

Figure 4. CRISPR/Cas9 fragmentation produces optimal fragment lengths. (*A*) Sonication produces fragments that are either too short or too long, corresponding to redundant or lost information, respectively. CRISPR-DS produces optimally sized fragments that are perfectly covered by the sequencing reads. (*B*,*C*) Comparison of histograms of the insert sizes of two samples prepared with standard-DS (blue, *left* panels), which uses sonication for fragmentation, and CRISPR-DS (red, *right* panels), which uses CRISPR/Cas9 digestion for fragmentation. The *x*-axes represent the percent difference from the optimally sized fragment, e.g., fragment size that matches the sequencing read length after adjustments for molecular barcodes and clipping. Yellow shading highlights the range of fragment sizes that are within 10% difference from optimal size.

Targeted sequencing remains a cost-effective alternative to whole-genome sequencing, especially when high depth is desired (Samorodnitsky et al. 2015). In multiple applications, such as oncology, the goal is to sequence a small panel of relevant genes with high accuracy in order to find low-frequency mutations. Although the selected target panel can be amplified by PCR, this method creates uneven coverage and false mutations, thus hybridization capture is typically preferred (Samorodnitsky et al. 2015). Hybridization capture improves coverage uniformity and removes certain artifactual mutations but does not resolve these issues completely. A major disadvantage in hybridization-based sequencing methods is the reliance on sonication for genome fragmentation, which generates DNA fragments of random size. We demonstrated that this size heterogeneity produces two problems that can be solved by replacing sonication with CRISPR-based genome fragmentation. The first problem is PCR bias, which results in the preferential amplification of short DNA fragments. PCR bias leads to wasted reads that contain an excess of PCR copies of the same molecule. Although these reads can be removed bioinformatically (Li 2011), the amplification advantage of certain molecules can lead to uneven coverage and reduced recovery (Kozarewa et al. 2015). In methods that use molecular barcodes, such as DS, three PCR copies are typically sufficient to generate a consensus sequence (Kennedy et al. 2014). Thus, additional sequencing of PCR copies does not produce additional data and only wastes resources. We demonstrated that with CRISPR-based fragmentation, all fragments amplify similarly. This homogeneous amplification translates into uniform coverage across all targeted regions, a critical feature when the goal is to detect low-frequency mutations in selected panel of genes.

The second problem associated with the heterogeneous fragment sizes relates to reduced data yield at the read level. Because sonication allows minimal control over fragment size, a large proportion of fragments are typically too short or too long compared to the optimal length size determined by the number of sequencing cycles. When reads are too short, paired-end reads overlap and the middle region is double-sequenced. Conversely, when reads are too long, the middle part of the DNA fragment, which may contain a variant or region of interest, remains unsequenced. This inefficient read usage is solved with CRISPR-based target selection because the fragments are tailored to the desired read length.

CRISPR-based target fragmentation also offers two additional advantages. First, homogeneously sized DNA fragments can be visualized to confirm library target enrichment prior to sequencing. In sonication-based hybridization capture, the gel electrophoresis for a target-enriched library looks identical to a library with no target enrichment. This issue can result in the costly waste of a sequencing run in which the majority of reads are in off-target regions. We show that the defined fragment lengths created by CRISPR-based digestion produce distinct peaks that are easily visualized and confirm that the sequencing library is target-enriched. A second advantage of Cas9 digestion over sonication is the elimination of sonication-

Table 1. Target enrichment due to size selection

Sample	DNA input (ng)	Reads on-target precapture (%)	Fold enrichment
B9	25	0.76	7527
	200	0.25	2452
	250	0.21	2037
PF1	10	2.85	28,139
	25	1.99	19,583
	100	0.68	6667
	250	0.70	6878
PF5	10	5.05	49,794
	25	0.96	9456
	100	0.34	3321
	250	0.22	2217

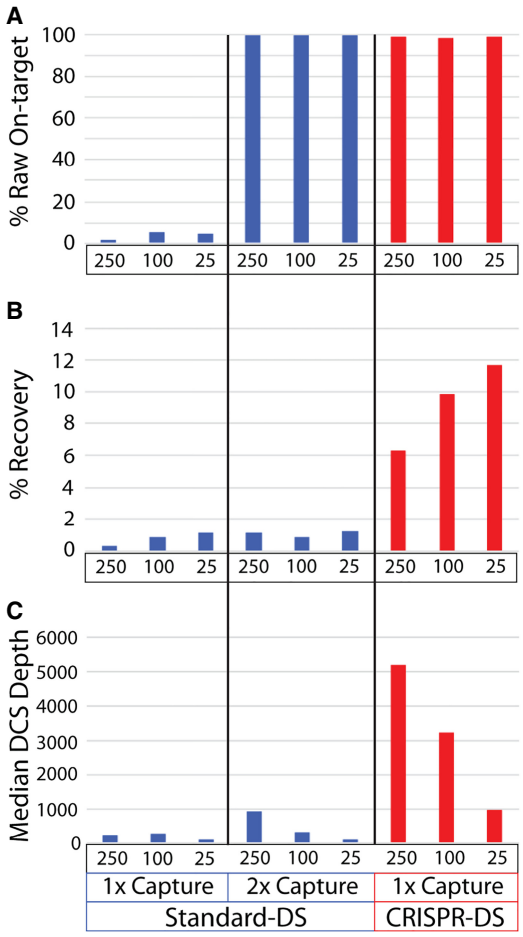

Figure 5. Technical comparison of 250, 100, and 25 ng of DNA sequenced with both standard-DS and CRISPR-DS. Measurements were obtained by sequencing samples prepared with standard-DS (blue) using one and two rounds of hybridization capture and CRISPR-DS (red) with only one round of hybridization capture. (*A*) The percentage of raw sequencing reads on-target (covering *TP53*) post-capture(s) was comparable between standard-DS with two rounds of capture and CRISPR-DS with one round of capture, demonstrating the target enrichment efficiency of the novel method. (*B*) Percentage recovery was calculated as the percentage of genomes in input DNA that produced DCS reads. CRISPR-DS increases recovery thanks to the initial CRISPR-based target enrichment, which eliminates one round of hybridization capture. (*C*) After creating DCS reads, the median DCS depth across all targeted regions was calculated for each input amount. The increased recovery enabled by CRISPR-DS translates into five to ten times more sequencing depth for the same input DNA.

induced sequencing errors (Park et al. 2017) and the preservation of double-stranded DNA at the ends of fragments. Sonication produces ssDNA at the end of molecules, which is susceptible to damage and converted into "pseudo-dsDNA" by end repair. This process has the potential to introduce false variant calls, but it is prevented by CRISPR-DS because Cas9 produces blunt ends that do not require end repair.

In the context of small target sequencing by hybridization capture, the major advantage introduced by CRISPR-based target enrichment is increased recovery, that is, percentage of input genomes that produce sequencing data. Hybridization capture is notably inefficient, especially for small target regions (Schmitt et al. 2015; Winters et al. 2017). As demonstrated with our experiments and in agreement with prior studies, the average recovery rate of DS is ~1%, which translates to at least 1 µg of DNA being needed to produce an average depth of approximately 3000×. This recovery is improved 10-fold by the addition of CRISPR-based target enrichment and the elimination of one round of capture. We demonstrated that by simply excising the genomic regions of interest and performing size selection, we can achieve a level of enrichment comparable to a single round of capture. By performing this step prior to library preparation, only one round of hybridization capture is needed, greatly minimizing DNA loss and increasing recovery. Therefore, using CRISPR-based target enrichment prior to DS achieves the same depth with 10 times less DNA.

Although CRISPR-DS addresses several needs in targeted NGS, it could still benefit from optimizations. First, improvements could be made to increase the recovery of degraded samples. Currently, in order to perform efficient target enrichment with CRISPR/Cas9 digestion and size selection, degraded samples must be preprocessed to remove low molecular weight fragments. We performed this preprocessing using electrophoretic size selection with the BluePippin system. However, to minimize loss of DNA, high molecular weight DNA could be selected with alternative methods such as micro-column filters. Second, although CRISPR-DS is highly efficient with a wide range of DNA inputs (10–500 ng), we noticed that the best recovery was achieved with smaller starting DNA amounts (10–25 ng). Because our goal was to achieve higher depth with low DNA input, this was not problematic. However, further efforts can be directed to improve recovery from larger DNA inputs, since this would be required if extremely high (>10,000×) target depths are desired. Last, although CRISPR-DS provides an effective solution for small target region deep sequencing, the method becomes costly for deep sequencing of large genomic regions, an inherent problem of deep sequencing. Nevertheless, fragmentation by CRISPR/Cas9 followed by size selection as a generic target enrichment technique can easily be scaled to many genomic regions because each region only requires the addition of the appropriate gRNAs for target excision. Sequencing of larger fragments can be achieved by tiling the gRNA along the desired fragment. Regarding the additional cost arising from the synthesis of gRNAs, it is important to note that a typical synthesis results in enough gRNA for thousands of cutting reactions. Over time this upfront cost becomes minimal compared with the substantial savings in time and reagents generated by the elimination of the second round of hybridization capture and by a more efficient use of sequencing space. Thus, CRISPR-DS becomes more economical than standard-DS in the long term, especially for small to moderate size panels (1–100 kb) that are deployed on large numbers of samples.

In conclusion, we have demonstrated that CRISPR/Cas9 fragmentation followed by size selection enables efficient target enrichment by increasing the recovery of hybridization capture and eliminating the need for a second round of capture for small target regions. In addition, it eliminates PCR bias, maximizes the use of sequencing resources, and produces homogeneous coverage. This fragmentation method can be applied to multiple sequencing modalities that suffer from these problems. Here we have applied it to DS in order to produce CRISPR-DS, an efficient, highly accurate sequencing method with significantly reduced input DNA requirements. CRISPR-DS has broad application for the sensitive identification of mutations in situations in which samples are DNA-limited, such as forensics and early cancer detection.

Table 2. Comparison of standard-DS versus CRISPR-DS for four different samples with *TP53* mutations

Method	Sample	Input DNA (ng)	Raw reads on-target (%)	Median final depth[a]	Recovery (%)	Tumor mutation	Mutant allele fraction (%)
Standard-DS	PF1	9196	92.4	2742	0.09	Chr 17: g.7578275G>A	68.5
	PF2	3000	92.8	5381	0.54	Chr 17: g.7577548C>T	1.2
	PF3	10,186	95.9	1866	0.06	Chr 17: g.7578403C>T	1.6
	PF4	7436	95.4	2029	0.08	Chr 17: g.7578526C>T	0.6
CRISPR-DS	PF1	100	76.6	2039	6.18	Chr 17: g.7578275G>A	68.4
	PF2	100	94.3	2831	8.58	Chr 17: g.7577548C>T	1.0
	PF3	100	87.6	3801	11.52	Chr 17: g.7578403C>T	0.4
	PF4	100	96.5	2194	6.65	Chr 17: g.7578526C>T	0.1

[a]After final Duplex Sequencing data processing is performed.

Methods

Samples

The samples analyzed included deidentified human genomic DNA from peripheral blood, bladder with and without cancer, and peritoneal fluid DNA from a prior study (Krimmel et al. 2016). Only peritoneal fluid samples had patient information available, which was necessary to confirm the tumor mutation. The peritoneal fluid samples were obtained from the University of Washington Gynecologic Oncology Tissue Bank, which collected specimens and clinical information after informed consent under protocol number 27077 approved by the University of Washington Human Subjects Division institutional review board. Frozen bladder samples were obtained from the University of Washington Genitourinary Cancer Specimen Biorepository and from unfixed or frozen autopsy tissue with waiver of consent under protocol number 52389 approved by the Fred Hutchinson Cancer Research Center Human Subjects Division institutional review board. The remainder of the study samples were used solely to illustrate technical aspects of the technology, no patient information was available, and interpretation of the mutational status of *TP53* is not reported. DNA was extracted with the QIAamp DNA Mini kit (Qiagen) with care being taken to not heat the sample above the recommended 56°C, which is essential to preserve the double-stranded nature of each DNA molecule prior to ligation of DS adapters. DNA was quantified with a Qubit HS dsDNA kit (Thermo Fisher Scientific). DNA quality was assessed with Genomic TapeStation tapes (Agilent) and DNA integrity numbers (DIN) were recorded. Peripheral blood DNA and peritoneal fluid DNA had DIN greater than 7, reflecting good-quality DNA with no degradation. Bladder samples, however, were purposely selected to include different levels of DNA degradation. Samples B1–B13 had DINs between 6.8 and 8.9 and were successfully analyzed by CRISPR-DS (Supplemental Table S3). Samples B14 and B16 had DINs of 6 and 4, respectively, and were used to demonstrate pre-enrichment of high molecular weight DNA with the BluePippin system (see below and Supplemental Fig. S5).

CRISPR guide design

CRISPR/Cas9 uses a gRNA to identify the site of cleavage. gRNAs are composed of a complex of CRISPR RNA (crRNA), which contains the ~20 bp unique sequence responsible for target recognition, and a *trans*-activating crRNA (tracrRNA), which has a universal sequence (Ran et al. 2013). To select the best gRNAs to excise *TP53* exons, we used the CRISPR MIT design website (http://CRISPR.mit.edu). The selection criteria were (1) production of fragments of ~500 bp covering exons 2–11 of *TP53*, and (2) highest MIT website score (Supplemental Table S1; Supplemental Data S1). Additionally, we recommend avoiding gRNAs that cover sites with known polymorphisms or mutational hotspots because this could potentially decrease the affinity of the gRNA and lead to reduced fragment depth. For exon 7, a smaller size fragment was required in order to avoid a proximal poly(T) repeat (Supplemental Fig. S2). We designed a total of 12 gRNA, which excised *TP53* into seven different fragments (Fig. 1A). All gRNA had scores above 60. Ten gRNAs were successful with the first chosen sequence, and two had to be redesigned due to poor cutting. Initially, the quality of the cut was assessed by reviewing the alignment of the final DCS reads with Integrative Genomics Viewer (Robinson et al. 2011). Successful guides produced a typical coverage pattern with sharp edges in region boundaries and proper DCS depth (Fig. 3D). Unsuccessful guides led to a drop in DCS depth and the presence of long reads that spanned beyond the expected cutting point. To simplify and speed up the assessment of guides, especially with scores below 80, as well as to assess the activity of the Cas9/gRNA complex over time, we designed a synthetic gBlock DNA fragment (IDT) that included all gRNA sequences interspaced with random DNA sequences (Supplemental Data S2). Three nanograms of gBlock DNA were digested with each of the gRNAs using the CRISPR/Cas9 in vitro digestion protocol described below. After digestion, the reactions were analyzed by TapeStation 4200 (Agilent Technologies) (Supplemental Fig. S9). The presence of predefined fragment lengths confirms (1) proper gRNA assembly, (2) the ability of the gRNA to cleave the designed site, and (3) proper nuclease activity of Cas9.

CRISPR/Cas9 in vitro digestion of genomic DNA

The in vitro digestion of genomic DNA with *S. pyogenes* Cas9 Nuclease requires the formation of a ribonucleoprotein complex, which both recognizes and cleaves a predetermined site. This complex is formed with gRNAs (crRNA + tracrRNA) and Cas9. For multiplex cutting, the gRNAs can be complexed by pooling all the crRNAs, then complexing with tracrRNA, or by complexing each crRNA and tracrRNA separately, then pooling. The second option is recommended by manufacturers because it eliminates competition between crRNAs; however, in the limited set of gRNAs tested here, both methods of complexing were comparable. Decreased efficiency over time has been observed due to degradation of Cas9 and gRNA. Thus, exposure to room temperature and repeated cycles of freeze-thawing should be avoided. The crRNAs and tracrRNAs (IDT) were complexed into gRNAs by incubating 5 min at 95°C, and then 30 nM of gRNAs were incubated with Cas9 nuclease (NEB) at ~30 nM, 1× NEB Cas9 reaction buffer, and water in a volume of 23–27 µL for 10 min at 25°C. Then, 10–250 ng of DNA was added for a final volume of 30 µL. The reaction was incubated overnight at 37°C and then heat shocked for 10 min at 70°C to inactivate the enzyme.

Size selection

Size selection for the predetermined fragment length is critical for target enrichment prior to library preparation. AMPure XP Beads (Beckman Coulter) were used to remove off-target, undigested high molecular weight DNA. After heat inactivation, the reaction was combined with a 0.5× ratio of beads, briefly mixed, and then incubated for 3 min to allow the high molecular weight DNA to bind. The beads were then separated from the solution with a magnet, and the solution containing the targeted DNA fragment length was transferred into a new tube. This was followed by a standard AMPure 1.8× ratio bead purification eluted into 50 µL of TE Low to exchange the buffer and remove small DNA contaminants.

A-tailing and ligation

The fragmented DNA was A-tailed and ligated using the NEBNext Ultra II DNA Library Prep Kit (NEB) according to the manufacturer's protocol. The NEB end repair and A-tailing (ERAT) reaction were incubated for 30 min at 20°C and for 30 min at 65°C. Note that end repair is not needed for CRISPR-DS because Cas9 produces blunt ends, but the ERAT reaction was used for convenient A-tailing. The NEB ligation master mix and 2.5 µL of DS adapters at 15 µM were added and incubated for 15 min at 20°C according to the manufacturer's instructions. Instead of relying on in-house manufactured adapters using previously published protocols (Schmitt et al. 2012; Kennedy et al. 2014), which tend to exhibit substantial batch-to-batch variability, we used a commercial adapter prototype of the structure shown in Figure 1C that was synthesized externally through arrangement with TwinStrand Biosciences. The two differences from the previous adapters are (1) 10 bp random double-stranded molecular tag instead of 12 bp, and (2) substitution of the previous 3′ 5 bp conserved sequence by a simple 3′-dT overhang to ligate onto the 5′-dA-tailed DNA molecules. Upon ligation, the DNA was cleaned by a 0.8× ratio AMPure Bead purification and eluted into 23 µL of nuclease free water.

PCR

The ligated DNA was amplified using KAPA Real-Time Amplification kit with fluorescent standards (KAPA Biosystems). Fifty-microliter reactions were prepared including KAPA HiFi HotStart Real-time PCR Master Mix, 23 µL of previously ligated and purified DNA, and DS primers MWS13, 5′-AATGATACGGCGACCACC GAG-3′, and MWS20, 5′-GTGACTGGAGTTCAGACGTGTGC-3′ (Schmitt et al. 2012; Kennedy et al. 2014) at a final concentration of 2 µM. The reactions were denatured for 45 sec at 98°C and amplified with six to eight cycles for 15 sec at 98°C, for 30 sec at 65°C, and for 30 sec at 72°C, followed by final extension for 1 min at 72°C. Samples were amplified until they reached Fluorescent Standard 3, which typically takes six to eight cycles depending on the amount of DNA input. Reaching Fluorescent Standard 3 produces a sufficient and standardized number of DNA copies into capture across samples and prevents overamplification. A 0.8× ratio AMPure Bead wash was performed to purify the amplified fragment and eluted into 40 µL of nuclease free water.

Capture and post-capture PCR

TP53 xGen Lockdown Probes (IDT) were used to perform hybridization capture for *TP53* exons as previously reported with minor modifications (Krimmel et al. 2016). From the predesigned IDT *TP53* Lockdown probes, we selected 21 probes that cover the entire *TP53* coding region (exon 1 and part of exon 11 are not coding) (Supplemental Table S2). Each CRISPR/Cas9 excised fragment was covered by at least two probes and a maximum of five probes (Supplemental Data S1). To produce the capture probe pool, each of the probes for a given fragment was pooled in equimolar amounts, producing seven different pools, one for each fragment. The pools were mixed again in equimolar amounts, except for the pools for exon 7 and exons 8–9, which were represented at 40% and 90%, respectively. The decrease of capture probes for those exons was implemented after observing consistent overrepresentation of these exons at sequencing. The final capture pool was diluted to 0.75 pmol/µL. Of note, it is essential to dilute the capture pool in low TE (0.1 mM EDTA) and to aliquot it in small volumes suitable for two to three uses. Excessive rounds of freeze-thaw severely impact the efficiency of the protocol. Hybridization capture was performed according to the IDT protocol, except for three modifications. First, we used blockers MWS60, 5′-AATGA TACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCT CTTCCGATCTIIIIIIIIIIIITGACT-3′ and MSW61, 5′-GTCAIIIIIIIIIII AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′, which are specific to DS adapters. Second, we used 75 µL of Dynabeads M-270 Streptavidin beads instead of 100 µL. Third, the post-capture PCR was performed with the KAPA HiFi HotStart PCR kit (KAPA Biosystems) using MWS13 and indexed primer MWS21 at a final concentration of 0.8 µM. The reaction was denatured for 45 sec at 98°C and then amplified for 20 cycles of 30 sec at 98°C, of 45 sec at 60°C, and of 45 sec at 72°C, followed by extension for 60 sec at 72°C. The PCR product was purified with a 0.8× AMPure Bead wash.

Sequencing

Samples were quantified using the Qubit dsDNA HS Assay Kit, diluted, and pooled for sequencing. The sample pool was visualized on the Agilent 4200 TapeStation to confirm library quality. The TapeStation electropherogram should show sharp, distinct peaks corresponding to the fragment length of the designed CRISPR/Cas9 cut fragments (Fig. 3B,C). This step can also be performed for each sample individually, prior to pooling, to verify the performance of each individual sample. The final pool was quantified using the KAPA Library Quantification kit (KAPA Biosystems). The library was sequenced on the MiSeq Illumina platform using a v3 600 cycle kit (Illumina), as specified by the manufacturer. For each sample, we allocated ~7%–10% of a lane corresponding to about 2 million reads. Each sequencing run was spiked with ~1% PhiX control DNA.

Standard-DS experiments

Three amounts of DNA (25, 100, and 250 ng) from normal human bladder sample B9 were sequenced with standard-DS with one round and two rounds of capture to provide direct comparison with CRISPR-DS. Standard-DS was performed as previously described (Kennedy et al. 2014), with the exception that the KAPA HyperPrep kit (KAPA Biosystems) was used for end repair and ligation, and the KAPA HiFi HotStart PCR kit (KAPA Biosystems) was used for PCR amplification. Hybridization capture was performed with xGen Lockdown probes that covered *TP53* exons 2–11, the same that were used for CRISPR-DS. Samples were sequenced on ~10% of a HiSeq 2500 Illumina platform to accommodate shorter fragment lengths. Data analysis was perform with the standard-DS analysis pipeline (https://github.com/risqueslab/DuplexSequencingScripts).

CRISPR-DS target enrichment experiments

Two different experiments were performed to characterize CRISPR-DS target enrichment. The first experiment consisted of comparing one versus two rounds of capture. Three DNA samples were

processed for CRISPR-DS and split in half after one hybridization capture. The first half was indexed and sequenced, and the second half was subjected to an additional round of capture, as required in the original DS protocol. Then the percentage of raw reads on-target (covering *TP53* exons) was compared for one versus two captures. The second experiment assessed the percentage of raw reads on-target without performing hybridization capture to determine the enrichment produced exclusively by size selecting CRISPR-excised fragments. Fold enrichment was calculated as the fraction of on-target raw reads divided over the expected fraction of on-target reads given the size of the target region (bases in the target region/total genome bases). Different DNA amounts (from 10 to 250 ng) of three different samples were processed with the protocol described above until first PCR, that is, prior to hybridization capture. Then the PCR product was indexed and sequenced. The percentage of raw reads on-target was calculated, and the fold enrichment was estimated considering the size of the targeted region, which is 3280 bp.

Pre-enrichment for high molecular weight DNA

Selection of high molecular weight DNA improves the performance of degraded DNA in CRISPR-DS. We performed this selection using a BluePippin system (Sage Science). Two bladder DNAs with DINs of 6 and 4 were run using a 0.75% gel cassette and high-pass setting to obtain >8 kb fragments. Size selection was confirmed by TapeStation (Supplemental Fig. S5A). Then, 250 ng of DNA before BluePippin and 250 ng of DNA after BluePippin were processed in parallel with CRISPR-DS. The percentage of raw reads on-target as well as average DCS depth was quantified and compared (Supplemental Fig. S5B). Alternative methods for size selection such as AMPure beads might be suitable to perform this enrichment.

Data processing

A custom bioinformatics pipeline was created to automate analysis from raw FASTQ files to text files (Supplemental Fig. S8). The primary modification of this pipeline is performing consensus making prior to alignment rather than after, as previously described for DS analysis (Schmitt et al. 2012; Kennedy et al. 2014). In this pipeline, consensus is executed by custom Python and Bash scripts. After consensus calling, the resulting processed FASTQ files are aligned to the reference genome of interest, in this case human reference genome v38, using BWA-MEM v.0.7.4 (Li and Durbin 2009) with default parameters. Mapped reads are realigned with GATK Indel-Realigner, and low-quality bases are clipped from the ends with GATK ClipReads (https://software.broadinstitute.org/gatk/). Because of the expected decrease in read quality in the latest cycles of sequencing, we performed a conservative clipping of 30 bases from the 3′ end and another seven bases from the 5′ end were clipped to avoid the occasional extra overhang left by incorrectly synthesized adapters. In addition, overlapping areas of read-pairs, which in our *TP53* design spanned ~80 bp, are trimmed back using fgbio ClipOverlappingReads (https://github.com/fulcrumgenomics/fgbio). Software for CRISPR-DS can be found in Supplemental Code S1 and is available at https://github.com/risqueslab/CRISPR-DS.

Data analysis

Recovery rate (also called fractional genome-equivalent recovery) was calculated as average DCS depth (sequenced genomes) divided by number of input genomes (1 ng of human genomic DNA corresponds to about 330 haploid genomes). The number of on-target raw reads was calculated by counting the number of reads within a 100-bp window on either side of the CRISPR/Cas9 cut sites. Optimal fragment size (Fig. 4B,C; Supplemental Fig. S3) was calculated as the sequencing read length minus the barcode sequence and minus clipped off bases for poor quality at the ends of reads. For peritoneal fluid samples sequenced with both CRISPR-DS and standard-DS, *TP53* biological background mutation frequency was calculated as the number of *TP53* mutations in *TP53* exons 4 to 10 (excluding the tumor mutation) divided by the total number of nucleotides sequenced in those exons. The 95% confidence intervals were calculated in R using the Clopper-Pearson "exact" method for binomial distribution (R Core Team 2017).

Software availability

Software for CRISPR-DS data analysis is available in Supplemental Code S1 as well as at https://github.com/risqueslab/CRISPR-DS.

Data access

Sequencing data from this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA412416.

Competing interest statement

S.R.K. is a consultant and equity holder for TwinStrand Biosciences, Inc. J.J.S. is a founder and equity holder in TwinStrand Biosciences, Inc. R.A.R. is the principal investigator on a NIH SBIR R44CA221426 subcontract research agreement with TwinStrand Biosciences, Inc.

Acknowledgments

We thank Shilpa Kumar for assistance with computational analysis, Emily Kohlbrenner for technical support and helpful discussions, Penny Faires for critical reading and copyediting of the manuscript, and the Genitourinary Cancer Specimen Biorepository for providing access to bladder cases (Director Dr. Colm Morrissey, PhD). We thank the University of Washington Gynecologic Oncology Tissue Bank for providing peritoneal fluid DNA and the Brigham and Women's Hospital/Harvard Cohorts Biorepository for sending archived samples from the Nurses' Health Study for pilot testing. Research reported in this publication was supported by grants from the National Institutes of Health under award numbers R01CA160674 and R01CA181308 to R.A.R.; Mary Kay Foundation grant 045-15 to R.A.R.; Rivkin Center for Ovarian Cancer grant 567612 to R.A.R.; and Cooperative Agreement Number W911NF-15-2-0127 from the Department of Defense Army Research Office/Defense Forensic Science Center (DFSC) as well as grant W81XWH-16-1-0579 from the Department of Defense Congressionally Directed Medical Research Program to S.R.K. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, DFSC, or the US Government. The US Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

Author contributions: S.R.K. conceived the idea; D.N., S.R.K., and R.A.R. developed the method; D.N. and R.A.R. designed the experiments; D.N., S.L., E.K.S., M.J.H., K.T.B., K.L.-S., B.F.K., R.A.R., and S.R.K. carried out experiments and/or performed data analysis; M.T. provided samples and scientific input; Y.Z. and J.J.S. contributed to assay development and provided invaluable critical discussion; D.N., S.R.K., and R.A.R. wrote the paper.

References

Ahn EH, Lee SH, Kim JY, Chang CC, Loeb LA. 2016. Decreased mitochondrial mutagenesis during transformation of human breast stem cells into tumorigenic cells. *Cancer Res* **76:** 4569–4578.

Arbeithuber B, Makova KD, Tiemann-Boege I. 2016. Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA Res* **23:** 547–559.

Bennett-Baker PE, Mueller JL. 2017. CRISPR-mediated isolation of specific megabase segments of genomic DNA. *Nucleic Acids Res* **45:** e165.

Dabney J, Meyer M. 2012. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52:** 87–94.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17:** 333–351.

Hoekstra JG, Hipp MJ, Montine TJ, Kennedy SR. 2016. Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. *Ann Neurol* **80:** 301–306.

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31:** 827–832.

Jung H, Ji S, Song S, Park Y, Yang JSE. 2014. The DNA Integrity Number (DIN) provided by the genomic DNA ScreenTape assay allows for streamlining of NGS on FFPE tissue samples. Agilent Technologies, Inc., publication number 5991-5360EN.

Kebschull JM, Zador AM. 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* **43:** e143.

Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. 2013. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet* **9:** e1003794.

Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC, Risques RA, et al. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9:** 2586–2606.

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci* **108:** 9530–9535.

Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One* **6:** e28240.

Kozarewa I, Armisen J, Gardner AF, Slatko BE, Hendrickson CL. 2015. Overview of target enrichment strategies. *Curr Protoc Mol Biol* **112:** 7.21.1–7.21.23.

Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ, Loeb LA, Swisher EM, Risques RA. 2016. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic *TP53* mutations in noncancerous tissues. *Proc Natl Acad Sci* **113:** 6005–6010.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27:** 2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Park G, Park JK, Shin SH, Jeon HJ, Kim NKD, Kim YJ, Shin HT, Lee E, Lee KH, Son DS, et al. 2017. Characterization of background noise in capture-based targeted sequencing data. *Genome Biol* **18:** 136.

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8:** 2281–2308.

Reid-Bayliss KS, Arron ST, Loeb LA, Bezrookove V, Cleaver JE. 2016. Why Cockayne syndrome patients do not get cancer despite their DNA repair deficiency. *Proc Natl Acad Sci* **113:** 10151–10156.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26.

Salk JJ, Schmitt MW, Loeb LA. 2018. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* **19:** 269–285.

Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, Damodaran S, Bhatt D, Reeser JW, Datta J, et al. 2015. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Hum Mutat* **36:** 903–914.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci* **109:** 14508–14513.

Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, Loeb LA. 2015. Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods* **12:** 423–425.

Shin G, Grimes SM, Lee H, Lau BT, Xia LC, Ji HP. 2017. CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nat Commun* **8:** 14291.

Winters M, Monroe C, Barta JL, Kemp BM. 2017. Are we fishing or catching? Evaluating the efficiency of bait capture of CODIS fragments. *Forensic Sci Int Genet* **29:** 61–70.

Received January 31, 2018; accepted in revised form August 31, 2018.