CRISPR-DS: an efficient, low DNA input method for ultra-accurate sequencing

Daniela Nachmanson[1], Shenyi Lian[1,2], 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[3,4], Scott R. Kennedy*[1], Rosa Ana Risques*[1]

[1]Department of Pathology, University of Washington, Seattle, WA 98195, USA. [2]Department of Pathology, Key Laboratory of Carcinogenesis and Translational Research, Peking University Cancer Hospital and Institute, Beijing, PR, China. [3]Department of Medicine, Division of Hematology and Oncology, University of Washington, Seattle, WA 98195, USA. [4]TwinStrand Biosciences, Seattle, WA 98121, USA.

*****These authors contributed equally

Corresponding authors:

Scott Kennedy, PhD
Dept. of Pathology
University of Washington
Box 357470
1959 NE Pacific Ave.
Seattle, WA 98195-7705
(206) 543-5452
scottrk@uw.edu

Rosa Ana Risques, PhD
Dept. of Pathology
University of Washington
Box 357470
1959 NE Pacific Ave.
Seattle, WA 98195-7705
(206) 616 4976
rrisques@uw.edu

Key words: Target enrichment, Duplex Sequencing, Next-Generation Sequencing, NGS, CRISPR/Cas9

ABSTRACT

Conventional hybrid-capture based Next-Generation Sequencing suffers from frequent errors, PCR bias due to variable fragment length, and inefficient target enrichment. To address these shortcomings, we combined ultra-accurate Duplex Sequencing with CRISPR/Cas9 excision of target regions, which allows enrichment by size selection prior to library preparation. This high efficiency method, which we term CRISPR-DS, improves targeted sequence recovery, reduces PCR bias, and maximizes read usability from minimal DNA inputs.

Next generation sequencing (NGS) has revolutionized the fields of biology and medicine. However, conventional hybrid capture-based NGS suffers from abundant errors generated at sonication and end-repair [1], PCR [2], and sequencing [3]. These errors preclude the detection of low frequency mutations and, thus, multiple efforts have been directed at their reduction. The most successful approaches rely on the use of molecular barcodes to tag individual DNA molecules, which allows sequencing reads to be traced to their original molecule and their redundant information is used to generate an error-corrected consensus sequence. Several methods have implemented single-strand molecular barcodes [4-6], successfully decreasing the error rate of standard NGS from ~10^{-3} to ~10^{-5} and improving mutation detection (Supplementary Fig. 1). However, single-strand consensus making does not completely overcome early PCR errors caused by DNA damage, which typically affects one strand of DNA, but not the other, and are a pervasive problems in NGS [7]. These errors can be eliminated using double-strand molecular barcodes with independent consensus making for each strand of DNA. This approach, called Duplex Sequencing (DS) [8,9], compares the sequences of both strands to produce a highly accurate duplex consensus sequence (DCS). DS achieves an unprecedented error rate <10^{-7} [8,9], which allows for extremely high sensitivity mutation detection (Supplementary Fig. 1). We have previously demonstrated that DS can accurately detect a single mutant molecule amongst <24,000 normal genomes [10].

While Duplex Sequencing robustly eliminates NGS errors, the double-stranded nature of the barcode requires targeted sequencing to be accomplished by hybrid capture (as opposed to amplicon sequencing), which is well-known to be inefficient. When searching for low frequency mutations, the region of interest is usually small as a trade-off for the high sequencing depth required. Hybrid capture applied to small targets (<50 kb), only produces 5-10% on-target reads [11,12]. In this situation, a successful approach for target enrichment is to perform two consecutive rounds of capture [11]. This strategy enables high accuracy ultra-deep sequencing in small target regions [10,13,14], but is relatively inefficient and requires large amounts of DNA.

Here, we solve the problem of inefficient capture by performing an enrichment of the regions of interest using the programmable endonuclease CRISPR/Cas9. *In vitro* digestion with CRISPR/Cas9 has already been proven to be a useful tool for multiplexed excision of target regions for PCR-free NGS [15]. However, the innovation of our approach is that excised target regions are designed to be of predetermined, homogenous length (Fig. 1a), thus enabling size selection prior to library preparation (Fig. 1b). This enrichment is then coupled with double-stranded barcoding (Fig. 1c) to perform error removal identically to the standard DS method [9] (Fig. 1d). In contrast to the standard method, however, the new method, termed CRISPR-DS, achieves very high on-target enrichment with only one round of capture. This significantly decreases time and cost (Fig. 1e and Supplementary Fig. 2) and increases efficiency. In addition, CRISPR fragmentation can be designed to produce fragments of optimal length to cover the full sequencing read (Fig. 1f). Fragmentation for hybridization capture is usually performed with sonication, which often generates fragments that are too long and the sequencing reads do not overlap the region of interest, or too short and the sequencing reads overlap with each other and re-read the same sequence (Fig. 1f). Both scenarios waste sequencing resources and are eliminated with CRISPR-DS. Sonication is also a well-known source of errors [1] and generates ssDNA at the end of molecules that can introduce false mutations when converted into dsDNA by end-repair. These false mutations are prevented by CRISPR fragmentation because Cas9 produces blunt ends, which do not require end-repair. Thus, CRISPR-DS solves multiple common problems of NGS, including inefficient target enrichment, which is optimized by CRISPR-based size selection; sequencing errors, which are removed with double-strand molecular barcodes; and uneven fragment size, which is solved by predesigned CRISPR/Cas9 fragmentation.

To demonstrate CRISPR-DS, we designed guide RNAs (gRNAs) to excise the coding region of *TP53* and flanking intronic areas (Fig. 1a). Fragment size was set at ~500bp to maximize the sequencing length of a MiSeq v3 600 cycle kit and allow for sufficient clipping of low-quality bases at the ends of reads.

gRNAs were selected based on specificity score and fragment length (Supplementary Table 1, Supplementary Data 1, and Online Methods). Test samples with variable amounts of input DNA (10-250ng) were CRISPR/Cas9 digested followed by size selection with solid-phase reversible immobilization (SPRI) beads to remove undigested high molecular weight DNA and enrich for the excised fragments containing the targeted regions (Fig. 1b). Subsequent library preparation was performed according to the standard protocol for DS [8], using only one round of capture and minor modifications (See Online Methods). Briefly, DNA was A-tailed, ligated with DS adapters, amplified, purified by bead wash, and captured by hybridization with biotinylated 120bp DNA probes targeting *TP53* exons (Supplementary Table 2). Captured samples were amplified with index primers and sequenced in an Illumina MiSeq v3 600 cycle kit. The analysis pipeline was modified from the original version [8] to include consensus making prior to alignment, which significantly reduces data processing time (Supplementary Figure 3, Online Methods).

First, we sequenced three input amounts of the same DNA extracted from normal human bladder tissue with standard-DS and CRISPR-DS. CRISPR-DS with one round of capture achieved >90% raw reads on-target (e.g. covering *TP53)* (Table 1a), a significant improvement over standard-DS, which only achieved ~5% raw reads on-target with one round of capture (Table 1a), consistent with prior work [11]. For CRISPR-DS a second round of capture is unnecessary since it only minimally increases raw reads on-target (Supplementary Fig. 4). Standard-DS produced a recovery rate (percentage of input genomes recovered as sequenced genomes; also known as fractional genome-equivalent recovery) of ~1% across the different inputs while CRISPR-DS recovery rate ranged between 6 and 12%. Notably, the higher recovery allows 25ng of DNA prepared with CRISPR-DS, to produce a DCS depth (depth generated by DCS reads) comparable to 250ng with standard-DS. Side-by-side comparison of the two methods illustrated three additional technical advantages of CRISPR-DS (Fig. 2). First, homogeneous fragment size solves the problem of overrepresentation of short fragments due to PCR amplification bias (Fig 2a). These fragments consume an excess of sequencing reads that decreases efficiency and produces uneven coverage of the regions of interest. Second, distinct bands/peaks provide confirmation of correct library preparation prior to sequencing (Fig. 2b-d). Third, well-defined fragments created by targeted fragmentation fully span the desired target regions with homogeneous coverage (Fig. 2e).

Next, to validate the ability of CRISPR-DS to detect low-frequency mutations, we analyzed four peritoneal fluid samples collected during debulking surgery from women with ovarian cancer. The presence of a *TP53* tumor mutation in these samples was previously demonstrated by standard-DS [10]. With only 100ng of DNA (30-100 fold less than what was used for standard-DS) we obtained comparable DCS depth to standard-DS and identified the *TP53* tumor mutation in all cases (Table 1b). Recovery rates again ranged between 6 and 12%, representing an increase of 15x-200x compared to standard-DS with the same DNA.

We further confirmed the performance of CRISPR-DS in an independent set of 13 DNAs extracted from bladder tissue (Supplementary Table 3). We used 250ng and obtained a median DCS depth of 6,143x, corresponding to a median recovery rate of 7.4% in agreement with the two prior experiments. Reproducible performance was demonstrated with technical replicates for samples B2 and B4. All samples had >98% DCS reads on-target, but the percentage of raw reads on-target ranged from 43% to 98%. We noticed that the low target enrichment corresponded to samples with DNA Integrity Number (DIN) <7. DIN is a measure of genomic DNA quality ranging from 1 (very degraded) to 10 (not degraded) [16]. We reasoned that degraded DNA compromises enrichment by size selection and this 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 from two samples with degraded DNA (DINs 6 and 4) and demonstrated that this pre-enrichment increased raw reads on-target by 2-fold and DCS depth by 5-fold (Supplementary Fig. 5).

Finally, to directly quantify the degree of enrichment conferred simply by CRISPR/Cas9 digestion followed by size selection, 3 samples were sequenced without capture. 10-250ng of DNA were digested, size-selected, ligated, amplified, and sequenced. The percentage of raw reads on-target ranged from 0.2% to 5%, corresponding to ~2,000x to 50,000x fold enrichment (Supplementary Fig. 6). Notably, lower DNA inputs showed the highest enrichment, probably reflecting optimal removal of off-target, high molecular weight DNA fragments when they are in lower abundance. Future work will benefit from optimization of the size selection protocol to maximize enrichment with larger DNA amounts.

In summary, we have demonstrated that CRISPR/Cas9 fragmentation followed by size selection enables efficient target enrichment and eliminates 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 for low input DNA. 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.

METHODS

Methods, including statements of data availability and any associated accession codes and references are available in the online version of the paper.

ACKNOWLEDGEMENTS

We thank Shilpa Kumar for assistance with computational analysis, Emily Kohlbrenner for technical support and helpful discussions, 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 NIH under award numbers R01CA160674 and R01CA181308 to RAR; Mary Kay Foundation grant 045-15 to RAR.; and Cooperative Agreement Number W911NF-15-2-0127 to S.K. from the U.S. Army Research Office and the Defense Forensic Science Center (DFSC). 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 U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

AUTHOR'S 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.S., M.H., K.B., K.L.S. and B.F.K. carried out experiments and data analysis; M.T. provided samples and scientific input; Y.Z., and J.S. contributed to assay development and provided invaluable critical discussion; D.N. and R.A.R. wrote the paper.

COMPETING FINANCIAL INTERESTS

SK is a consultant and equity holder for TwinStrand Biosciences Inc. JJS is a founder and equity holder in TwinStrand Biosciences Inc.

ONLINE METHODS

Samples. The samples analyzed included de-identified human genomic DNA from peripheral blood, bladder with and without cancer, and peritoneal fluid DNA from a prior study [10]. Only peritoneal fluid samples had patient information available, which was necessary to confirm the tumor mutation. These 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. 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. De-identified frozen bladder samples were obtained from the University of Washington Genitourinary Cancer Specimen Biorepository and from not previously fixed 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. DNA had been previously extracted with the QIAamp DNA Mini kit (Qiagen, Inc., Valencia, CA, USA) and it had never been denatured, which is essential to preserve the double-strandedness of each molecule prior to ligation of DS adapters. DNA was quantified with a Qubit HS dsDNA kit (ThermoFisher Scientific). DNA quality was assessed with Genomic TapeStation (Agilent, Santa Clara, CA) and DNA integrity numbers (DIN) were recorded. Peripheral blood DNA and peritoneal fluid DNA had DIN>7 reflecting good quality DNA with no degradation. Bladder samples, however, were purposely selected to include different levels of DNA degradation. Samples B1 to B13 had DINs between 6.8 and 8.9 and were successfully analyzed by CRISPR-DS (Supplementary Table 3). 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 Supplementary Figure 5).

CRISPR guide design. CRISPR/Cas9 is a programmable endonuclease that uses a gRNA to identify the site of cleavage. gRNAs are composed of a complex of CRISPR RNA (crRNA), which contains the ~20bp unique sequence responsible for target recognition, and a trans-activating crRNA (tracrRNA), which has a universal sequence [17].To select the most optimal gRNAs to excise *TP53* exons we used the CRISPR MIT design website (http://CRISPR.mit.edu:8079/). The selection criteria were: (1) production of fragments of ~500bp covering the *TP53* coding region and (2) highest MIT website score (Supplementary Table 1 and Supplementary Data 1). For exon 7, a smaller size fragment was required in order to avoid a proximal poly-A tract. We designed a total of 12 gRNA, which excised *TP53* into 7 different fragments (Fig. 1a). All gRNA had scores >60. 10 gRNAs were successful with the first chosen sequence and 2 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 [18]. Successful guides produced a typical coverage pattern with sharp edges in region boundaries and proper DCS depth (Fig. 2e). Unsuccessful guides led to a drop in DCS depth and the presence of long reads that spanned beyond the expected cutting point. In order to simplify and speed up the assessment of guides, we designed a synthetic GeneBlock DNA fragment (IDT, Coralville, IA) that included all gRNA sequences interspaced with random DNA sequences (Supplementary Data 2). 3ng of GeneBlock 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, Santa Clara, CA, USA) (Supplementary Fig. 7). The presence of predefined fragment lengths confirms: (1) Proper gRNA assembly (2) The ability of the gRNA to cleave the designed site.

CRISPR/Cas9 in vitro digestion of genomic DNA. The *in vitro* digestion of genomic DNA with Cas9 Nuclease, *S. pyogenes*, requires the formation of a ribonucleoprotein complex, which both recognizes and cleaves a pre-determined 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 preferred because it eliminates competition between crRNAs. gRNAs are at risk of quick degradation and repeated cycles of freeze-thawing should be avoided. crRNAs and tracrRNAs (IDT, Coralville, IA) were complexed into gRNAs and then 30nM of gRNAs were incubated with Cas9 nuclease (NEB, Ipswich, MA) at ~30nM, 1x NEB Cas9 reaction buffer, and water in a volume of 23-27 µL at 25°C for 10 min. Then, 10-250ng of DNA was added for a final volume of 30 µL. The reaction was incubated overnight at 37°C and then heat shocked at 70°C for 10 min 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, Brea, CA, USA) were used to remove off-target, un-digested high molecular weight DNA. After heat inactivation, the reaction was combined with a 0.5x ratio of beads, briefly mixed and then incubated for 3 min to allow the high MW 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 1.8x ratio bead purification eluted into 50 µL of TE Low to exchange the buffer and remove small DNA contaminants.

Library preparation

A-tailing, and ligation. The fragmented DNA was A-tailed and ligated using the NEBNext Ultra II DNA Library Prep Kit (NEB, Ipswich, MA) according to manufacturer's protocol. The NEB end-repair and A-tailing (ERAT) reaction was incubated at 20°C for 30 min and 65°C for 30 min. 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. Then the NEB ligation mastermix and 2.5µl of DS adapters at 15 µM were added and incubated at 20°C for 15 min. Instead of relying on in-house manufactured adapters using previously published protocols [8,9], which tend to exhibit substantial batch-to-batch variability, we used a commercial adapter prototype of the structure shown in Fig 1d that were synthesized externally through arrangement with TwinStrand Biosciences. The two differences from the previous adapters are: (1) 10bp random, double stranded molecular tag instead of 12bp, which is more cost-effective and still produces a large excess of tag diversity; and (2) substitution of the previous 3' 5bp conserved sequence by a simple 3'-dT overhang to ligate onto the 5'-dA-tailed DNA molecules. The elimination of the conserved sequence removes potential phasing issues in the Illumina sequencers. Upon ligation, the DNA was cleaned by a 0.8X 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, Woburn, MA, USA). 50µl reactions were prepared including KAPA HiFi HotStart Real-time PCR Master Mix, 23µl of previously ligated and purified DNA and DS primers MWS13 and MWS20 [8,9] at a final concentration of 2 µM. The reactions were denatured at 98°C for 45 sec and amplified with 6-8 cycles of 98°C for 15 sec, 65°C for 30 sec, and 72°C for 30 sec, followed by final extension at 72°C for 1 min. Samples were amplified until they reached Fluorescent Standard 3, which typically takes 6-8 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, it prevents over-amplification, and indicates successful Cas9 cutting and ligation. A 0.8X ratio AMPure Bead wash was performed to purify the amplified fragments, which were eluted into 40µL of nuclease free water. Compared to standard-DS, CRISPR-DS offers two important advantages at the PCR step: (1) fragments have similar sizes, which reduces amplification bias towards small fragments (Fig. 2a) and produces a

more homogeneous coverage of the regions of interest (Fig. 2e), and (2) the predetermined fragment size allows for accurate assessment by TapeStation 4200 (Agilent Technologies, Santa Clara, CA, USA) of successful library preparation up to this step. In standard-DS, PCR products are a wide range of sizes due to sonication and present as a wide smear which is difficult to compare between samples. CRISPR-DS, however, produces discrete peaks that are clearly indicative of successful cutting and ligation and are amenable of comparison for quality control across samples (Fig. 2b-d).

Capture and post-capture PCR. *TP53* xGen Lockdown Probes (IDT, Coralville, IA) were used to perform hybridization capture for *TP53* exons as previously reported, with minor modifications. From the pre-designed 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) (Supplementary Table 2). Each CRISPR/Cas9 excised fragment was covered by at least 2 probes and a maximum of 5 probes (Supplementary data 1). To produce the capture probe pool, each of the probes for a given fragment was pooled in equimolar amounts, producing 7 different pools, one for each fragment. Then the 7 fragment 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 2-3 uses. Excessive rounds of freeze-thaw severely impact the efficiency of the protocol. Hybridization capture was performed according to the IDT protocol, except for 3 modifications. First, we used blockers MWS60 and MSW61, which are specific to DS adapters, as described [8,9]. Second, we used 75µl of Dynabeads M-270 Streptavidin beads instead of 100µl. Third, the post-capture PCR was performed with the KAPA Hi-Fi HotStart PCR kit (KAPA Biosystems, Woburn, MA, USA) using MWS13 and indexed primer MWS21 at a final concentration of 0.8 µM. The reaction was denatured at 98°C for 45 sec and then amplified for 20 cycles at 98°C for 30 sec, 60°C for 45 sec, and 72°C for 45 sec, followed by extension at 72°C for 60 sec. The PCR product was purified with a 0.8X 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 (Figure 2b-d). 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, Woburn, MA, USA). The library was sequenced on the MiSeq Illumina platform using a v3 600 cycle kit (Illumina, San Diego, CA, USA) as specified by the manufacturer. For each sample, we allocated ~7-10% of a lane corresponding to ~2 million reads. Each sequencing run was spiked with approximately 1% PhiX control DNA.

Data processing. A custom bioinformatics pipeline was created to automate analysis from raw FASTQ files to text files (Supplementary Fig 3). This pipeline includes two major modifications compared to the previously described method for DS analysis [8,9]: (1) the retention of paired read information and (2) consensus-making performed prior to alignment. Paired-end reads are essential to the analysis of CRISPR-DS data, but are also an important improvement for the analysis of DS in general, as they allow critical quality control of fragment size and removal of potential technical artifacts related to short fragments. Consensus making prior to alignment is also a convenient feature for DS in general, not just CRISPR-DS. The originally described DS analysis pipeline performed consensus after all reads were mapped to the reference genome, whereas this pipeline performs consensus as the initial step, solely reliant on the bases read by the sequencer. This is expected to improve consensus making and also reduces the time required for data processing as alignment is the most computationally intensive step in

the protocol. In this pipeline, consensus is executed by a custom python script called UnifiedConsensusMaker.py. This script takes all reads that are derived from the same tag, compares the base called at each position, and produces a single-stranded consensus (SSCS) read. The SSCS reads for each complementary pair of tags are then compared position by position to create a double-stranded consensus (DCS) read (Fig. 1e). Two FASTQ files are made containing the resulting SSCS reads and DCS reads. Note that DCS reads correspond to original DNA molecules so the average DCS depth is an estimation of the number of genomes sequenced. Recovery rate (also called fractional genome-equivalent recovery) is calculated as average DCS depth (sequenced genomes) divided by number of input genomes (one ng of DNA corresponds to ~330 haploid genomes). Raw reads on-target were calculated by counting the number of reads whose genomic coordinates fell within the upstream and downstream CRISPR/Cas9 cut sites with a 100bp added window on either side.

Paired-end, DCS FASTQ files are then aligned to the reference genome of interest, in this case human reference genome v38, using bwa-mem v.0.7.4[19] with default parameters. Mapped reads are re-aligned with GATK Indel-Realigner, and low quality bases are clipped from the ends with GATK Clip-Reads[18]. 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 7 bases from 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 ~80bp, are trimmed back using fgbio ClipOverlappingReads. This algorithm performs even clipping from the two ends of the paired reads until they meet, which maximizes the use of sequencing bases with high PHRED quality scores. A pileup file is created from the resulting file using SAMtools mpileup[19]. The pileup file is then filtered using a custom python script with a BED file for targeted genomic positions. The BED file can be easily created using the coordinates of the CRISPR/Cas9 gRNAs. Then the filtered pileup file is processed by a custom-made script, mut-position.1.33.py, which creates a tab delimited text file with mutation information called 'mutpos'. The mutpos includes a summary of the DCS depth and the mutations at each position sequenced. Software for CRISPR-DS is available at https://github.com/risqueslab/CRISPR-DS.

Standard-DS experiments

Three amounts of DNA (25ng, 100ng, and 250ng) 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 [8], with the exception that the KAPA Hyperprep kit (KAPA Biosystems, Woburn, MA, USA) was used for end-repair and ligation and the KAPA Hi-Fi HotStart PCR kit (KAPA Biosystems, Woburn, MA, USA) 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.

CRISPR-DS target enrichment experiments

Two different experiments were performed to characterize CRISPR-DS target enrichment. The first experiment consisted on the comparison of one vs. 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 subject 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 vs. 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. Different DNA amounts (from 10ng to 250ng) 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 3,280bp.

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, Beverly, MA). Two bladder DNAs with DINs of 6 and 4 were run using a 0.75% gel cassette and high-pass setting to obtain >8kb fragments. Size selection was confirmed by TapeStation (Supplementary Fig. 5a). Then 250ng 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 (Supplementary Fig. 5b). Alternative methods for size selection such as AMPure beads might be suitable to perform this enrichment.

Statement of data availability. Sequencing data that supports the findings of this study have been deposited in the Sequence Read Archive (BioProject ID: PRJNA412416).

Statistics and reproducibility. No sample sizes were predetermined in this study. For each experiment, the number of samples analyzed is indicated and the data is shown for all samples. No P values were calculated during this study. Median and mean values are used as descriptive statistics in the test. Two samples were analyzed as technical replicates within the same experiment to demonstrate reproducibility of percentage of raw reads on-target and recovery rate.

SUPPLEMENTARY TABLES

Supplementary Table 1. crRNA sequences for *TP53* CRISPR/Cas9 digestion
Supplementary Table 2. *TP53* hybridization capture probes
Supplementary Table 3. CRISPR-DS sequencing results for 15 samples processed with 250ng input DNA

SUPPLEMENTARY FIGURES

Supplementary Figure 1. Comparison of mutation limit detection by sequencing accuracy
Supplementary Figure 2. Timeline of library preparation for CRISPR-DS and standard-DS
Supplementary Figure 3. Overview of CRISPR-DS data processing
Supplementary Figure 4. Target enrichment for CRISPR-DS with one vs. two captures
Supplementary Figure 5. Pre-enrichment for high molecular weight DNA with BluePippin
Supplementary Figure 6. Target enrichment by size selection
Supplementary Figure 7. GeneBlock as a control to test *TP53* guide RNAs

SUPPLEMENTARY DATA

Supplementary Data 1. *TP53* sequence with crRNA and capture probes
Supplementary Data 2. GeneBlock sequence

REFERENCES

1. G. Park, J. K. Park, S. H. Shin et al., *Genome biology* 18 (1), 136 (2017).
2. J. M. Kebschull and A. M. Zador, *Nucleic acids research* 43 (21), e143 (2015).
3. I. Akogwu, N. Wang, C. Zhang et al., *Human genomics* 10 Suppl 2, 20 (2016).
4. J. B. Hiatt, C. C. Pritchard, S. J. Salipante et al., *Genome research* 23 (5), 843 (2013).
5. I. Kinde, J. Wu, N. Papadopoulos et al., *Proceedings of the National Academy of Sciences of the United States of America* 108 (23), 9530 (2011).
6. D. I. Lou, J. A. Hussmann, R. M. McBee et al., *Proceedings of the National Academy of Sciences of the United States of America* 110 (49), 19872 (2013).
7. L. Chen, P. Liu, T. C. Evans, Jr. et al., *Science* 355 (6326), 752 (2017).
8. S. R. Kennedy, M. W. Schmitt, E. J. Fox et al., *Nature protocols* 9 (11), 2586 (2014).
9. M. W. Schmitt, S. R. Kennedy, J. J. Salk et al., *Proceedings of the National Academy of Sciences of the United States of America* 109 (36), 14508 (2012).
10. J. D. Krimmel, M. W. Schmitt, M. I. Harrell et al., *Proceedings of the National Academy of Sciences of the United States of America* 113 (21), 6005 (2016).
11. M. W. Schmitt, E. J. Fox, M. J. Prindle et al., *Nature methods* 12 (5), 423 (2015).
12. M. Winters, C. Monroe, J. L. Barta et al., *Forensic Sci Int-Gen* 29, 61 (2017).
13. E. H. Ahn, S. H. Lee, J. Y. Kim et al., *Cancer research* 76 (15), 4569 (2016).
14. S. R. Kennedy, J. J. Salk, M. W. Schmitt et al., *PLoS genetics* 9 (9), e1003794 (2013).
15. G. Shin, S. M. Grimes, H. Lee et al., *Nature communications* 8, 14291 (2017).
16. H. Jung, S. Ji, S. Song et al., *Application Note Nucleic Acid Analysis* (2014).
17. F. A. Ran, P. D. Hsu, J. Wright et al., *Nature protocols* 8 (11), 2281 (2013).
18. J. T. Robinson, H. Thorvaldsdottir, W. Winckler et al., *Nature biotechnology* 29 (1), 24 (2011).
19. H. Li and R. Durbin, *Bioinformatics* 26 (5), 589 (2010).

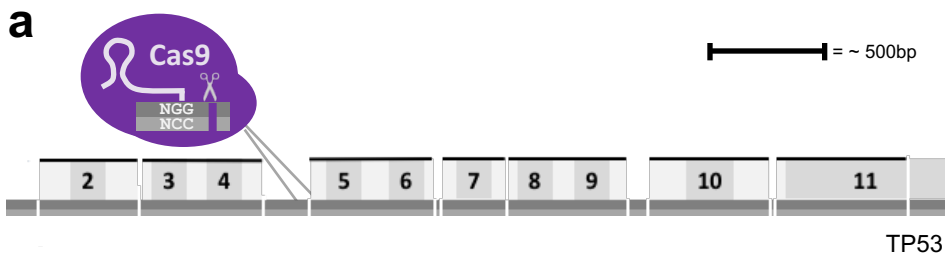
Table 1a. Comparison of Standard-DS vs. CRISPR-DS for the same sample with 3 different DNA amounts

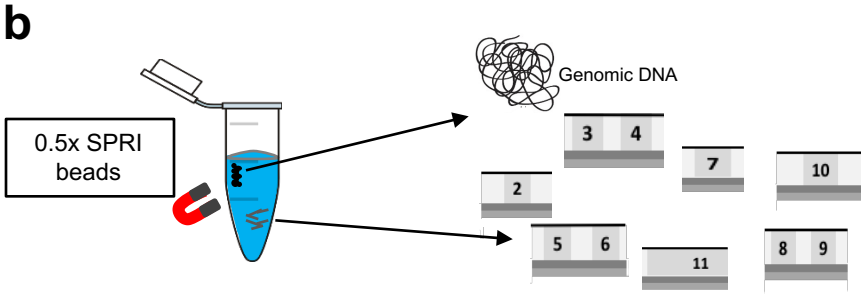
Method	Sample	Input DNA (ng)	Rounds of Hybridization Capture	Raw Reads On Target (%)	Median DCS depth (TP53 exons 2-11)	Recovery Rate
STANDARD-DS	B9	250	2	99.1%	946	1.1%
		100	2	99.3%	306	0.9%
		25	2	99.4%	100	1.2%
		250	1	1.3%	215	0.3%
		100	1	5.6%	296	0.9%
		25	1	5.1%	94	1.1%
CRISPR-DS	B9	250	1	98.7%	5167	6.3%
		100	1	98.2%	3219	9.8%
		25	1	99.0%	967	11.7%

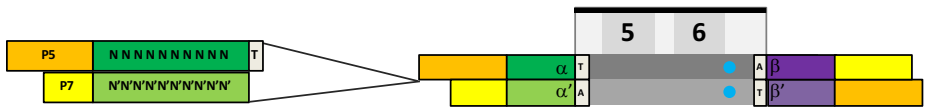
Table 1b. Comparison of Standard-DS vs. CRISPR-DS for 4 different samples with *TP53* mutations

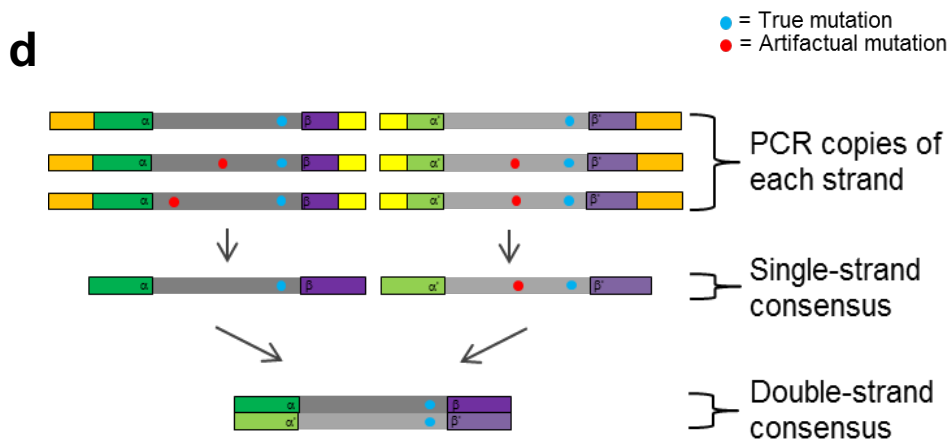
Method	Sample	Input DNA (ng)	Raw Reads On Target (%)	Median DCS depth (TP53 exons 4-10)	Recovery Rate	Tumor Mutation	Tumor Mutation Frequency
STANDARD DS	PF1	9,196	92.4%	2742	0.09%	chr17:g.7578275G>A	68.5%
	PF2	3,000	92.8%	5381	0.54%	chr17:g.7577548C>T	1.2%
	PF3	10,186	95.9%	1866	0.06%	chr17:g.7578403C>T	1.6%
	PF4	7,436	95.4%	2029	0.08%	chr17:g.7578526C>T	0.6%
CRISPR-DS	PF1	100	76.6%	2039	6.18%	chr17:g.7578275G>A	68.4%
	PF2	100	94.3%	2831	8.58%	chr17:g.7577548C>T	1.0%
	PF3	100	87.6%	3801	11.52%	chr17:g.7578403C>T	0.4%
	PF4	100	96.5%	2194	6.65%	chr17:g.7578526C>T	0.1%

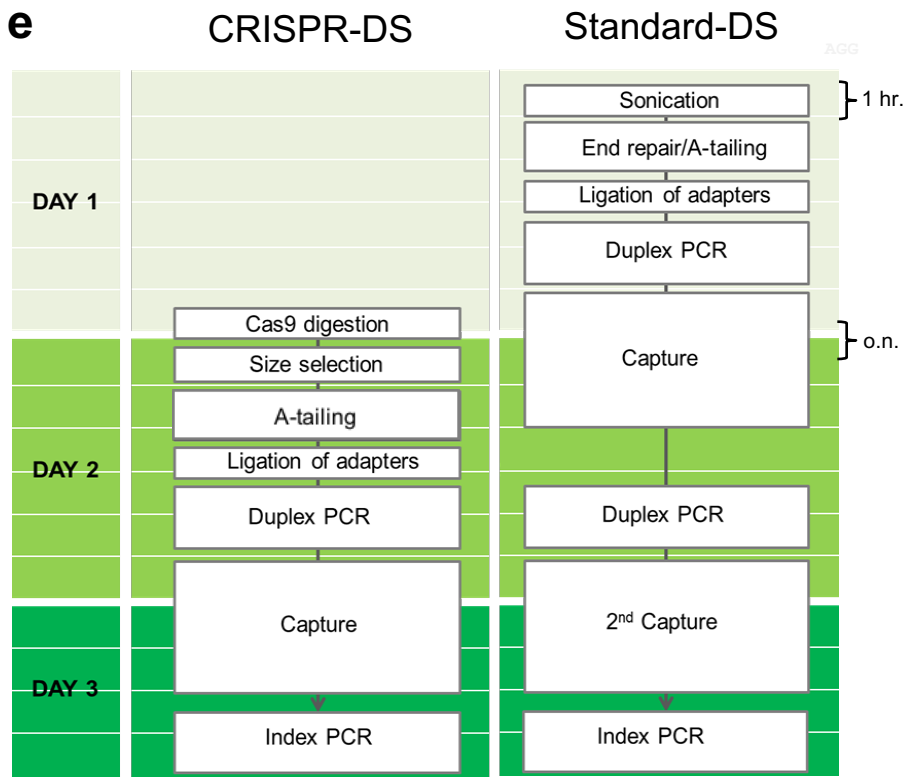
Figure 1











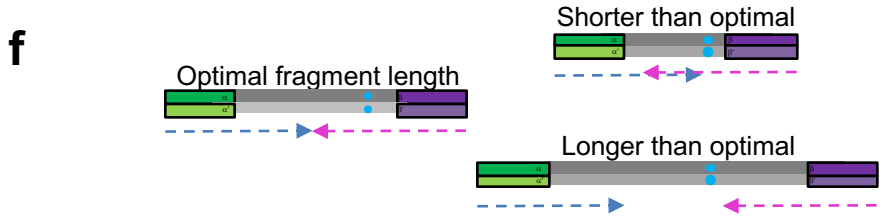


Figure 1. Schematic representation of key aspects of CRISPR-DS. (a) CRISPR/Cas9 digestion of *TP53*. Seven fragments containing all *TP53* coding exons were excised via targeted cutting using gRNAs. Dark grey represents reference strand and light grey represents the anti-reference strand. (b) Size selection using 0.5x SPRI beads. Uncut, genomic DNA binds to the beads and allows the recovery of the excised fragments in solution. (c) Double-stranded DNA molecule fragmented and ligated with DS-adapters. Adapters are double-stranded and contain 10-bp of random, complementary nucleotides and a 3'-dT overhang. (d) Error correction by DS. Reads derived from the same strand of DNA are compared to form a Single-Strand Consensus Sequence (SSCS). Then both strands of 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. (e) Comparison of library preparation steps for CRISPR-DS and standard-DS. CRISPR-DS reduces the workflow by nearly a day. Colored boxes represent 1h of time. (f) Sonication produces fragments that are either too short or too long, corresponding to redundant or lost information, respectively. CRISPR-DS produces optimally sized fragments which are perfectly covered by the sequencing reads.

Figure 2

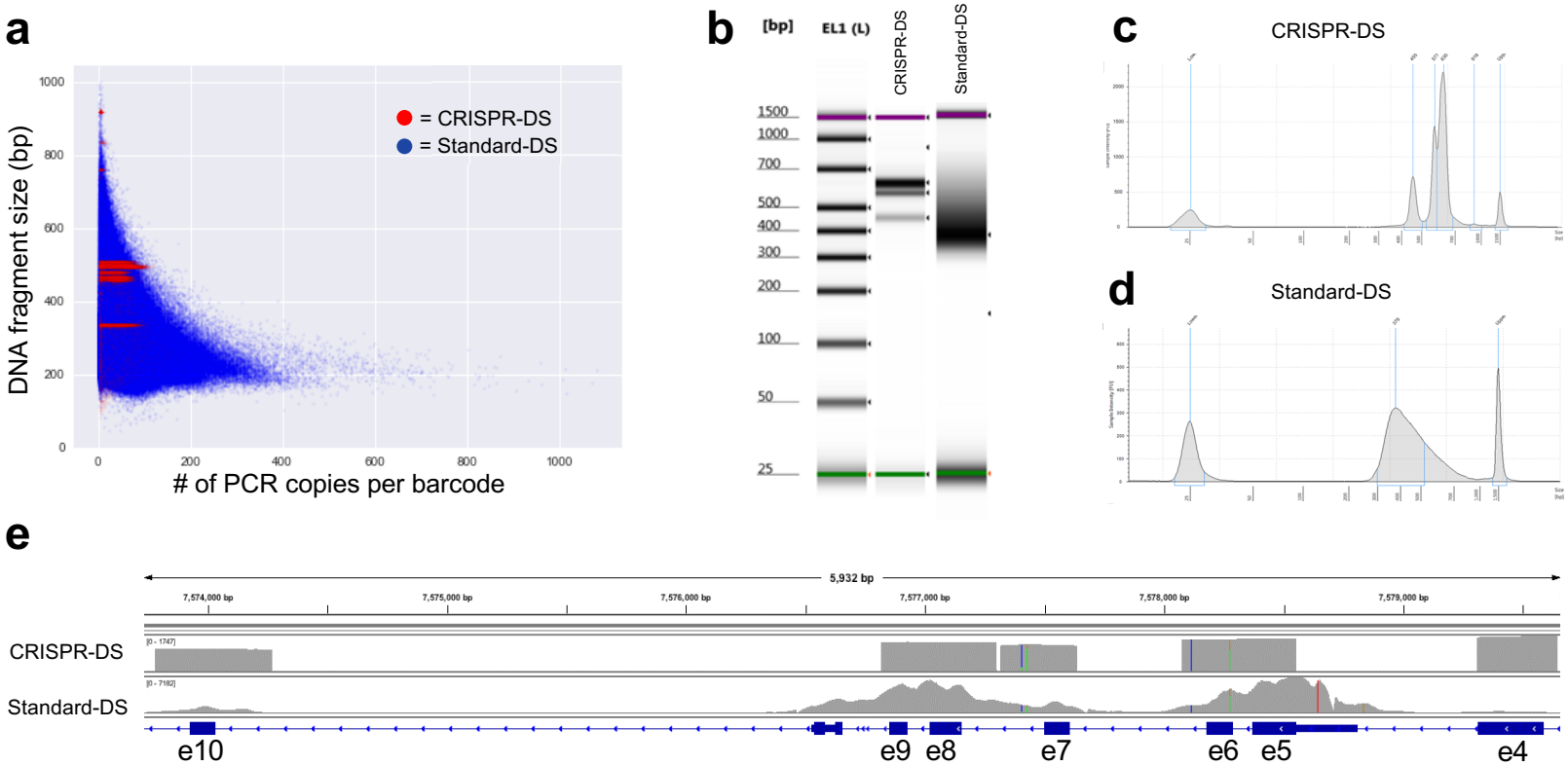


Figure 2. Side by side comparison of a test DNA sample processed with CRISPR-DS and standard-DS. (a) Dots represent original barcoded DNA molecules. Each DNA molecule has multiple copies generated at PCR (x-axis). In CRISPR-DS, all DNA molecules (red dots) have preset sizes (Y-axis) and generate similar number of PCR copies. In standard-DS, sonication shares DNA in variable fragment lengths (blue dots). Smaller fragments amplify better and generate an excess of copies that waste sequencing resources. (b-d) Visualization of library prepared with CRISPR-DS and standard-DS. (b) 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. CRISPR-DS electropherograms (c) allow visualization and quantification of peaks for quality control of the library prior to sequencing. Standard-DS electropherograms (d) show a diffuse peak that harbors no information about the specificity of the library. (e) Integrative Genomics Viewer of *TP53* coverage with DCS reads generated by CRISPS-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.