REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
31-03-2019	Final Report	21-Sep-2015 - 20-Sep-2018

4. TITLE AND SUBTITLE
Final Report: High Accuracy Genotyping of Complex Mixtures and Damaged DNA

5a. CONTRACT NUMBER
W911NF-15-2-0127

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Washington
Office of Sponsored Programs
4333 Brooklyn Ave NE Box 359472
Seattle, WA 98195 -9472

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
68102-LS-RIF.4

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Scott Kennedy
UU	UU	UU	UU		19b. TELEPHONE NUMBER 206-543-5452

RPPR Final Report
as of 03-Apr-2019

Agency Code:

Proposal Number: 68102LSRIF **Agreement Number: W911NF-15-2-0127**
INVESTIGATOR(S):

Name: Scott Kennedy
Email: scottrk@uw.edu
Phone Number: 2065435452
Principal: Y

Organization: **University of Washington**
Address: Office of Sponsored Programs, Seattle, WA 981959472
Country: USA
DUNS Number: 605799469 EIN: 916001537
Report Date: 20-Oct-2018 Date Received: 31-Mar-2019
Final Report for Period Beginning 21-Sep-2015 and Ending 20-Sep-2018
Title: High Accuracy Genotyping of Complex Mixtures and Damaged DNA
Begin Performance Period: 21-Sep-2015 **End Performance Period:** 20-Sep-2018
Report Term: 0-Other
Submitted By: Scott Kennedy Email: scottrk@uw.edu
Phone: (206) 543-5452
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals: Objective 1: Define and validate a standardized set of equipment and reagents and provide a step-by-step protocol that can be performed by a laboratory technician. This objective will entail developing quality control (QC) methods to monitor the success of the enzymatic steps required for sample processing in order to ensure consistent performance and maximize the success rate of genotyping difficult samples. Accuracy, precision, sensitivity, and specificity will be evaluated.
Objective 2: Provide a basic data analysis workflow for genotype calling. This objective will create a computational workflow for genotype calling and forensic identification from DS analyses of heterogeneous mixtures of low quality DNA. The workflow will have two operational modes: 1) If a target genotype is provided, then the workflow will determine if a particular individual or individual's relative DNA is present. 2) If no specific target genotype is provided, then the workflow will produce a list of possible genotypes present.
Objective 3: Validate protocol and equipment for use in genotyping complex DNA mixtures and highly damaged/degraded DNA. This objective will use DS to evaluate DNA samples subjected to a number of environmental conditions, such as increasing mixture complexity, extended exposure to UV/sunlight, elevated temperatures, and damaging chemicals. In this way, the limits of DS to genotype highly complex and damaged DNA will be defined.
Objective 4: Develop and provide a training program and material. Once DS has been optimized and validated for forensic analysis, we will develop and provide training and hands on experience for technical staff on how the system functions and how sample preparation is performed.

Objective 1: Protocol Standardization and Validation
Milestone 1 Tasks: 1) Demo equipment for compatibility and purchase 2) Initial trial run
Milestone 2 Tasks: Optimize 1)library preparation; 2) PCR conditions; 3) genomic capture
Milestone 3 Tasks: Validate 1) accuracy; 2) precision, 3) sensitivity, and 4) specificity of DS
Objective 2: Creation of Genotyping Workflow
Milestone 1 Tasks: 1) Basic improvements to data analysis pipeline
Milestone 2 Tasks: 2) Develop genotyping software; 2) Implement relevant statistical tests
Milestone 3 Tasks: 1) Simulate test data sets 2) test effectiveness of genotyping software
Objective 3: Validation of DS on damaged DNA and complex mixtures
Milestone 1: Tasks: 1) Validate on purified damaged DNA
Milestone 2: Tasks: 1) Determine dynamic range; 2) increase complexity
Milestone 3: Tasks: 1) Validate complex mixtures of damaged DNA.
Objective 4: Development of training material and transition to DFSC
Milestone 1: Tasks: 1) Create software documentation; 2) design course and training material

RPPR Final Report
as of 03-Apr-2019

Accomplishments: Objective 1: Protocol Standardization and Validation
Milestone 1 Tasks: 1) Demo equipment for compatibility and purchase 2) Initial trial run
-Demoed equipment was purchased and used for an initial trial of the Duplex Sequencing protocol

Milestone 2 Tasks: Optimize 1)library preparation; 2) PCR conditions; 3) genomic capture
-Due to the unexpectedly high inefficiency of the original published protocol, we developed two new approaches to Duplex Sequencing, termed CRISPR-DS and LS-AMP, that reduce targeted capture to either a single round or zero rounds, respectively. We were able to optimize the the CRISPR/Cas9 sites to allow read traversal in the direction that prevents read failure, which we noticed was an issue.
-Genomic capture continued to be an issue, likely due to the presence of STR sequence that flanked the probe target regions immediately adjacent to the STR. We were unable to devise a solution to this problem that involved targeted hybridization.

Milestone 3 Tasks: Validate 1) accuracy; 2) precision, 3) sensitivity, and 4) specificity of DS
-We were only able to partially validate our method due to the extremely high levels of off target sequence arising from poor targeted capture. We were able to show that we could reduce stutter rates by >10-fold, such that, on average, <0.5% of reads contained an erroneous length polymorphism. Specificity was poor

Objective 2: Creation of Genotyping Workflow
Milestone 1 Tasks: 1) Basic improvements to data analysis pipeline
-We tested our original data on STR sequences and found it inadequate for STR analysis due to the high rate of PCR stutter. This meant we had to redevelop our pipeline from scratch.

Milestone 2 Tasks: 2) Develop genotyping software; 2) Implement relevant statistical tests
-We redeveloped the pipeline from scratch by employing a third party software package, HipSTR, along with the fgbio software suite that is specifically designed to manipulate Duplex Sequencing data. We found that this approach, while successful, erroneously removed a very large fraction of reads from consideration and processing. We are unsure of the issue.

Milestone 3 Tasks: 1) Simulate test data sets 2) test effectiveness of genotyping software
-Unfinished at the time of project termination.

Training Opportunities: Nothing to Report

Results Dissemination: Dissemination took the form of one research paper and four talks.

Paper: Nachmanson D, Lian S, Schmidt EK, Hipp MJ, Baker KT, Zhang Y, Tretiakova M, Loubet-Senear K, Kohrn BF, Salk JJ, Kennedy SR*, Risques R-A* Targeted genome fragmentation with CRISPR/Cas9 improves hybridization capture, reduces PCR bias, and enables efficient high-accuracy sequencing of small targets. Genome Res 28: 1589-1599. *Co-senior author

Talks:
2016 “Removing PCR artifacts using massively parallel sequencing” National Institute of Justice Forensic Technology Center of Excellence

2017 “Removing PCR artifacts using massively parallel sequencing” 28th International Symposium on Human Identification

2018 “Removing sequencer and PCR artifacts using massively parallel sequencing platforms” Sequencing, Finishing, and Analysis in the Future Symposium

2018 “Removing sequencer and PCR artifacts using massively parallel sequencing platforms” Promega PowerTech Forensics Workshop

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: One patent application would produced during the project's period of performance.

Patent information: "Methods for Targeted Nucleic Acid Sequence Enrichment with Applications to Error Corrected Nucleic Acid Sequencing" Filed March 23, 2018

The application claims priorty to US provisional patent application No. 62/475,682, filed March 23, 2017 and US provisional patent application No 62/575,958, filed October 23, 2017.

PARTICIPANTS:

Participant Type: PD/PI
Participant: Scott R Kennedy
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Non-Student Research Assistant
Participant: Michael J Hipp
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Non-Student Research Assistant
Participant: Elizabeth K Schmidt
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: International Symposium on Human Identification
Date Received: 19-Oct-2018 Conference Date: 02-Oct-2017 Date Published:
Conference Location: Seattle WA
Paper Title: Removing Sequencer and PCR Artifacts for Forensic DNA Analysis on Massively Parallel Sequencing Platforms
Authors: Scott R Kennedy, Michael J Hipp
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 03-Apr-2019

PATENTS:

Intellectual Property Type: Patent Date Received: **18-Oct-2018**
Patent Title: Methods for targeted nucleic acid sequence enrichment with applications to error corrected nucleic acid sequencing
Patent Abstract: The advent of next-generation sequencing (NGS) in genomic research has enable the character
Patent Number: 72227-8137.WO00
Patent Country: USA
Application Date: 23-Mar-2018 Application Status: 1
Date Issued:

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
19-10-2018	Final Report	21-Sep-2015 - 20-Sep-2018

4. TITLE AND SUBTITLE

Final Report: High Accuracy Genotyping of Complex Mixtures and Damaged DNA

5a. CONTRACT NUMBER

W911NF-15-2-0127

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Washington
Office of Sponsored Programs
4333 Brooklyn Ave NE
Seattle WA 98195-9472

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)

ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

68102-LS-RIF.4

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Scott R Kennedy
UU	UU	UU	UU		19b. TELEPHONE NUMBER *(Include area code)* 206-543-5452

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

REPORT OF INVENTIONS AND SUBCONTRACTS

(Pursuant to "Patent Rights" Contract Clause) (See Instructions on back)

Form Approved
OMB No. 9000-0095
Expires Jan 31, 2008

PLEASE DO NOT RETURN YOUR COMPLETED FORM TO THE ABOVE ORGANIZATION. RETURN COMPLETED FORM TO THE CONTRACTING OFFICER.

1.a. NAME OF CONTRACTOR/SUBCONTRACTOR	c. CONTRACT NUMBER	2.a. NAME OF GOVERNMENT PRIME CONTRACTOR	c. CONTRACT NUMBER	3. TYPE OF REPORT (X one)
University of Washington	W911NF-15-2-0127	University of Washington	Same	a. INTERIM [X] b. FINAL

b. ADDRESS (Include ZIP Code)	d. AWARD DATE (YYYYMMDD)	b. ADDRESS (Include ZIP Code)	d. AWARD DATE (YYYYMMDD)	4. REPORTING PERIOD (YYYYMMDD)
4333 Brooklyn Ave NE Seattle WA 98195-0001	20150921	4333 Brooklyn Ave NE Seattle WA 98195-0001	20150921	a. FROM 20150921 b. TO 20180920

SECTION I - SUBJECT INVENTIONS

5. "SUBJECT INVENTIONS" REQUIRED TO BE REPORTED BY CONTRACTOR/SUBCONTRACTOR *(If "None," so state)*

NAME(S) OF INVENTOR(S) (Last, First, Middle Initial) a.	TITLE OF INVENTION(S) b.	DISCLOSURE NUMBER, PATENT APPLICATION SERIAL NUMBER OR PATENT NUMBER c.	ELECTION TO FILE PATENT APPLICATIONS (X) d.				CONFIRMATORY INSTRUMENT OR ASSIGNMENT FORWARDED TO CONTRACTING OFFICER (X) e.	
			(1) UNITED STATES		(2) FOREIGN			
			(a) YES	(b) NO	(a) YES	(b) NO	(a) YES	(b) NO
Kennedy, Scott, R; Seelig, Georg	Massively parallel single-cell transcriptomics and genomic variant detection	48361		X		X		X
Kennedy, Scott R; Hipp Michael J; Salk Jesse J; Schmidt, Elizabeth K	Methods for enrichment of genomic loci for duplex consensus sequencing	47836	X		X			
Kennedy, Scott R; Salk Jesse J; Risques Rosa Ana; Nachmanson, Daniela	CRISPR/Cas9 based methods for targeted genome enrichment	48192	X		X			

f. EMPLOYER OF INVENTOR(S) NOT EMPLOYED BY CONTRACTOR/SUBCONTRACTOR		g. ELECTED FOREIGN COUNTRIES IN WHICH A PATENT APPLICATION WILL BE FILED	
(1) (a) NAME OF INVENTOR (Last, First, Middle Initial)	(2) (a) NAME OF INVENTOR (Last, First, Middle Initial)	(1) TITLE OF INVENTION	(2) FOREIGN COUNTRIES OF PATENT APPLICATION
		Methods for enrichment of genomic loci for duplex consensus CRISPR/Cas9 based methods for targeted genome enrichment	European Union
(b) NAME OF EMPLOYER	(b) NAME OF EMPLOYER		
(c) ADDRESS OF EMPLOYER (Include ZIP Code)	(c) ADDRESS OF EMPLOYER (Include ZIP Code)		

SECTION II - SUBCONTRACTS *(Containing a "Patent Rights" clause)*

6. SUBCONTRACTS AWARDED BY CONTRACTOR/SUBCONTRACTOR *(If "None," so state)*

NAME OF SUBCONTRACTOR(S) a.	ADDRESS (Include ZIP Code) b.	SUBCONTRACT NUMBER(S) c.	FAR "PATENT RIGHTS" d.		DESCRIPTION OF WORK TO BE PERFORMED UNDER SUBCONTRACT(S) e.	SUBCONTRACT DATES (YYYYMMDD) f.	
			(1) CLAUSE NUMBER	(2) DATE (YYYYMM)		(1) AWARD	(2) ESTIMATED COMPLETION

SECTION III - CERTIFICATION

7. CERTIFICATION OF REPORT BY CONTRACTOR/SUBCONTRACTOR *(Not required if: (X as appropriate))* [ ] SMALL BUSINESS or [ ] NONPROFIT ORGANIZATION

I certify that the reporting party has procedures for prompt identification and timely disclosure of "Subject Inventions," that such procedures have been followed and that all "Subject Inventions" have been reported.

a. NAME OF AUTHORIZED CONTRACTOR/SUBCONTRACTOR OFFICIAL (Last, First, Middle Initial)	b. TITLE	c. SIGNATURE	d. DATE SIGNED
Becker, Dale	Program Coordinator, OSP	Dale Becker	10222018

DD FORM 882 INSTRUCTIONS

GENERAL

This form is for use in submitting INTERIM and FINAL invention reports to the Contracting Officer and for use in reporting the award of subcontracts containing a "Patent Rights" clause. If the form does not afford sufficient space, multiple forms may be used or plain sheets of paper with proper identification of information by item number may be attached.

An INTERIM report is due at least every 12 months from the date of contract award and shall include (a) a listing of "Subject Inventions" during the reporting period, (b) a certification of compliance with required invention identification and disclosure procedures together with a certification of reporting of all "Subject Inventions," and (c) any required information not previously reported on subcontracts containing a "Patent Rights" clause.

A FINAL report is due within 6 months if contractor is a small business firm or domestic nonprofit organization and within 3 months for all others after completion of the contract work and shall include (a) a listing of all "Subject Inventions" required by the contract to be reported, and (b) any required information not previously reported on subcontracts awarded during the course of or under the contract and containing a "Patent Rights" clause.

While the form may be used for simultaneously reporting inventions and subcontracts, it may also be used for reporting, promptly after award, subcontracts containing a "Patent Rights" clause.

Dates shall be entered where indicated in certain items on this form and shall be entered in six or eight digit numbers in the order of year and month (YYYYMM) or year, month and day (YYYYMMDD). Example: April 2005 should be entered as 200504 and April 15, 2005 should be entered as 20050415.

1.a. Self-explanatory.

1.b. Self-explanatory.

1.c. If "same" as Item 2.c., so state.

1.d. Self-explanatory.

2.a. If "same" as Item 1.a., so state.

2.b. Self-explanatory.

2.c. Procurement Instrument Identification (PII) number of contract (DFARS 204.7003).

2.d. through 5.e. Self-explanatory.

5.f. The name and address of the employer of each inventor not employed by the contractor or subcontractor is needed because the Government's rights in a reported invention may not be determined solely by the terms of the "Patent Rights" clause in the contract.

Example 1: If an invention is made by a Government employee assigned to work with a contractor, the Government rights in such an invention will be determined under Executive Order 10096.

Example 2: If an invention is made under a contract by joint inventors and one of the inventors is a Government employee, the Government's rights in such an inventor's interest in the invention will also be determined under Executive Order 10096, except where the contractor is a small business or nonprofit organization, in which case the provisions of 35 U.S.C. 202(e) will apply.

5.g.(1) Self-explanatory.

5.g.(2) Self-explanatory with the exception that the contractor or subcontractor shall indicate, if known at the time of this report, whether applications will be filed under either the Patent Cooperation Treaty (PCT) or the European Patent Convention (EPC). If such is known, the letters PCT or EPC shall be entered after each listed country.

6.a. Self-explanatory.

6.b. Self-explanatory.

6.c. Self-explanatory.

6.d. Patent Rights Clauses are located in FAR 52.227.

6.e. Self-explanatory.

6.f. Self-explanatory.

7. Certification not required by small business firms and domestic nonprofit organizations.

7.a. through 7.d. Self-explanatory.

1. List of Appendixes, Illustrations, and Tables

2. Statement of problem studied

Due to its high success rate and cost effectiveness, STR analysis for forensic genotyping has become the primary method in human identification casework. Since STR analysis depends on length variations in short poly-nucleotide repeats that are amplified using PCR, there are inherent limitations in this technology. For example, sample degradation due to environmental exposure leads to the breakdown of DNA molecules, which can result in significantly biased and artifactual peak heights, leading to inconclusive or erroneous genotyping calls[1]. Even in pristine samples, there are complicating factors that are intrinsic to current STR methods, such as spurious background peaks resulting from PCR stutter, co-migration, signal oversaturation, and machine noise[1,2]. Of particular concern are PCR stuttering artifacts, which arise from slippage of the DNA polymerase on the DNA template. Damaged or degraded DNA is particularly prone to this form of error due to the prevalence of DNA adducts that cause erroneous base pairings and enzyme stalling. While there are techniques that allow for the statistical exclusion of stutter peaks, DNA mixture samples of three or more contributors, especially combined with DNA damage and template degradation, present a significant challenge[3]. The recent advent of NGS offers the hope of being able to resolve complex DNA mixture samples. Unlike conventional STR analysis, which simply reports the average genotype of an aggregate population of molecules, NGS technology digitally tabulates the sequence of many individual DNA fragments, thus offering the unique ability to detect MAFs within a heterogeneous DNA mixture[4,5]. The use of NGS in forensic DNA analysis offers numerous advantages over conventional STR analysis; however the technology is not without its disadvantages. The most notable is that NGS is based on PCR during library construction, which has an associated stutter and base misincorporation rate[6]. These initial misincorporation and stutter events can be propagated to all of the reads, thus giving the appearance of a MAF in a putative DNA mixture. Similar to STR analysis, the chance of a false signal significantly increases on damaged DNA templates[7]. Furthermore, the ability to practically detect MAFs is limited to about 1-2% due to sequencing errors associated with various sequence contexts and base miscalls. Damaged DNA is known to worsen this background[7,8]. While better than current methods, the field of forensic DNA analysis has profound legal consequences, therefore, it is imperative that the detection of false MAFs be eliminated.

Summary of Methodology for Duplex Sequencing

The focus of the funded work was to develop Duplex Sequencing (DS) for use in forensic applications. DS relies on the concept of molecular tagging and the fact that the two strands of DNA contain complementary information[9]. Randomly sheared duplex DNA is tagged with a random, yet complementary, double-stranded nucleotide sequence (Fig. 1A). Following ligation, the individually labeled strands are PCR amplified such that there will be many duplicate “families” that share a common

tag sequence and are derived from a single parental strand of DNA (Fig. 1B). After sequencing, reads sharing the same tag sequence (i.e. tag family) are grouped together, and a consensus sequence is calculated for each family to create a single strand consensus sequence (SSCS), with each SSCS being derived from an individual molecule of ssDNA (Fig. 1C). This step filters out random sequencing or PCR errors. Importantly, the SSCS does not filter out artifactual mutations, such as base misincorporations and stutters, that occur during the first round of PCR. To remove these errors, the complementary tags derived from the same duplex DNA among the SSCS reads are compared to each other (Fig. 1C). The base identity at each position in a read is kept in the final consensus if the two strands match perfectly at that position. Apparent mutations occurring in only one of the SSCS reads will be filtered out. Upon remapping of the duplex consensus sequence (DCS) reads back to the reference genome, any deviations from the reference genome are considered true mutations.

3. Summary of the most important results.

Objective 1: ***Define and validate a standardized set of equipment and reagents and provide a step-by-step protocol that can be performed by a laboratory technician.***

Milestone 1: ***Standardization of Equipment***

We purchased an Agilent Tapestation 4200 for DNA quantification, an Agilent AriaMX qPCR thermocycler for qPCR quantification, and an Illumina MiSeq FGx sequencing platform. This last piece of equipment was bought using funds from the Defense University Research Instrumentation Program (DURIP).

Milestone 2: ***Optimization of Library Preparation*****:**

We focused on three areas related to library preparation. 1) Optimization of preparatory steps of sample DNA; 2) Optimization steps of PCR amplification; and 3) Optimization of target genome capture. Work on Milestone 2 focused on developing approaches to optimize each of these steps. The enzymatic preparatory steps can be further broken down into four main steps: DNA fragmentation, end-repair, 3´dA-tailing, and adapter ligation. In our originally published protocol, each step is separate and requires multiple expensive and time-consuming purification steps[10].

Single-tube library synthesis kits are made by several vendors and offer an easy and low cost way of performing the end-repair, 3'-dA-tailing, and ligation steps with minimal DNA loss from multiple reaction clean-up steps. New England Biolab and KAPA Biosystems are two vendors that make the most popular kits currently in use. As part of the reaction optimization portion of the cooperative agreement, we tested the efficiency of converting sheared DNA into adapter ligated sequencing library for the NEBNext Ultra kit, the NEBNext Ultra II, and KAPA Hyperprep kit (the Hyperprep-plus kit is incompatible with Duplex Sequencing and was not tested).

We tested the conversion efficiency of the three kits for three different DNA input amounts, 1µg, 500ng, 250ng. DNA was sonically sheared using a Covaris AFA system in 50µL of 10mM Tris-HCl pH8.0 (conditions outlined in Kennedy *et al.* [10]). Samples were end-repaired, dA-tailed, and ligated following the manufacturer's instructions for each kit. We used a qPCR based strategy that compares the change of the C(t) for a primer set that spans the adapter and a genomic DNA target (Fig. 2A) to a primer set that targets a small 90bp genomic DNA target to estimate the relative amount of ligated target DNA. In this particular assay, the genomic DNA target amplicons is only 90bp in size, whereas the ligated target amplicons is 300-900bp in size (i.e. shear fragment size), which results in a higher SYBRgreen signal for the ligated target amplicons and, consequently, a relatively lower C(t) when compared to the genomic DNA target amplicon. Consequently, the C(t) for the ligated target amplicons will get smaller when more DNA is ligated, whereas the C(t) for the genomic target DNA will remain constant. Therefore, an increase in the ΔC(t) indicates an increased ligation efficiency. The data show that for the 1µg and 500ng DNA inputs, no significant changes are seen between the three kits (Fig 2B,C). The 250ng DNA input shows a significant 2- fold (i.e. 1 C(t)) reduction in efficiency (i.e. lower ΔC(t)) for the NEB Ultra kit, relative to the NEB Ultra

II kit (Fig 1B,C). The NEB Ultra II kit trends to higher efficiency compared to both the NEB Ultra and KAPA Hyperprep kit for all DNA input amounts tested. Based on our results, the fact that the NEB Ultra kit is planned to be discontinued in the mid-term, and no significant cost differences between the KAPA Hyperprep and NEB Ultra II kits, we used the NEB Ultra II kits for the remainder of the project.

We next pursued optimization of PCR input. In order to observe true variants by DS, sequence changes must be present and complementary in both strands of DNA. Prior to sequencing, we will monitor the above steps in library preparation using qPCR with primers directed against targeted and non-targeted genes. The ligated product needs to be amplified prior to sequencing with the optimal amplification being an average of 3 to 10 sequenced PCR copies per starting DNA molecule. The conditions needed to consistently obtain these values must be standardized, since too many PCR copies of the same tag will reduce the number of unique families per DNA sample sequenced. Importantly, the number of DNA fragments used in the PCR reaction is the primary adjustable variable that dictates the number of sequencing reads that share the same tag sequence. A critical confounder to achieving the 3-10 PCR copies per tag has been the ligation step. Incompletely ligated molecules are unable to amplify during PCR, however, the presence of non-amplifiable molecules are also measured and will result in an overestimation of the amount amplifiable DNA being used in the PCR. This situation can lead to too many PCR copies per molecule and a substantial reduction in data yield.

Quantifying the amount of ligated target DNA is essential for the success of Duplex Sequencing. Technologies, such as Life Technologies Qubit and/or Agilent's TapeStation or Bioanalyzer can only quantify the total amount of DNA present in a library and not the quantity of a specific species of DNA molecules (i.e. amount of ligated DNA). To overcome this problem, we have designed a synthetic DNA using IDT's gBlocks product that is able to mimic the PCR product of a ligated fragment of DNA that is present in sequencing library preparation (Fig 2A). By making serial dilutions of the synthetic DNA and verifying the concentration each dilution using the qPCR and the P5 and P7 primers, we can then quantify the amount of *ligated* target DNA in the library preparation by using the P5 and a target specific primer.

To experimentally test this approach, we tested several primer pairs that target the sequencing adapter and different nuclear DNA sequences. After extensive primer optimization and testing (Data not shown), we settled on an Adapter/Target Ligation Standard (ATLiS) presented in Fig. 3A. The IDT gBlock was ordered and diluted in 50µL of 10mM Tris-HCl, pH8.0, as recommended by the manufacturer. A 1000-fold dilution of the master stock was made and quantified using the KAPA Library Quantification Kit (KAPA Biosystems) according to the manufacturer's instructions. The concentration of the 1000-fold dilution was found to be 9.1pM. A 10-fold dilution series (1mL/dilution), with the 1000-fold dilution being the first in the series, was made with 10mM Tris-HCl, pH8.0. The dilution series was then check for linearity and accuracy by qPCR using the P5 (*cyan*) and Human nuclear reverse (*orange*) primers (Fig. 3A). The resulting standard curve exhibited high linearity with a slope of -3.40 (expected is -3.32) and an efficiency of 97% (Fig. 3B).

In our initial trial run, <10% of the raw reads (out of ~$40x10^6$ reads) mapped to the CODIS20 loci, but with an optimal family size of ~20 (Fig. 4A). However, performing a second round of targeted capture can significantly improve the efficiency of capture for small genomic targets [11]. Implementing this second round of capture resulted in >90% of reads being localized to the genomic targets of interest (out of ~$40x10^6$ reads). However, as a consequence of adding the second round of targeted hybridization, we observed a higher than desired family size for the CODIS capture panels (Fig. 4B).

We hypothesized that a combination of stochastic PCR amplification, limited PCR cycles, and/or low targeted capture efficiency adversely affect the amount of tag diversity in the final sequencing library. The basic logic behind this hypothesis is as follows. Assuming the number of reads doesn't change between samples, the more unique tags present in the sequencing library, the less likely that any given tag will be sequenced, which, by definition, reduces the family size. Therefore, if we lost tag diversity (*i.e.* fewer unique tags), the higher the probability of any given tag will be sequenced, which increases the family size.

Prior to the 1st PCR amplification, each tag is present only once. Therefore, stochastic amplification may have a significant impact on tag representation. For example, consider a tag that undergoes replication on the 1st cycle, so that on PCR cycle 2 there are two copies. If both copies undergo replication, then during

PCR cycle 3 there will four starting copies. However, PCR is not 100% efficient. As such, a tag that amplified during the 1st cycle could fail to copy one or both initial copies during PCR cycle 2. Additionally, a small fraction of tags could fail to copy in both cycle 1 and 2 (for an amplification efficiency of 90% (P_{amp} = 0.9), this happens $(1 - p)^2 = 0.1^2 = 0.01$ or 1% of the time). If their luck evens out and both amplicons get amplified equally during subsequent cycles, the tag will appear at a copy number of about four times more than the tag that failed to amplify. This suggests that the distribution of copy numbers for P_{amp} = 0.9 will range over at least a factor of four.

This stochasticity is especially important when the number of PCR cycles is low and combined with targeted capture. Due to the amount of input DNA needed for the CODIS and SNP panels (250ng), only ~5-6 cycles can be performed before reagents in the PCR reaction are exhausted. Therefore, at most, there can only be 32-64 copies of each tag. Given that PCR is not 100% efficient, most tags will be represented by fewer than the expected number of copies. If we were to directly sequence from this pool of tags, then, provided enough reads were obtained, we would obtain reasonable family sizes. However, technical advice from IDT indicates that 95% of target DNA typically fails to capture. Therefore, on average, 95% the molecules containing the target loci fail to capture. If there are, at most, 32 starting copies (resulting from 5 PCR cycles) of each tag, then, on average, only 2 copies of each tag will make it through the 1st capture step, with some variance in the actual number. Because targeted capture can be thought of as a Bernoulli trial (*i.e.* a piece of target DNA can either be captured or not), then it can be modeled using a Binomial distribution with the number of trials, *n*, being the number of PCR copies of a particular tag (*i.e.* 32) and the probability of capture, *p*, being the percentage of copies that make it through capture (1-0.95=0.05). From these parameters we can estimate that the probability of a tag being completely lost due to not being sampled is $Pr(X=0) = Binom(32,0.05) = 0.193$. In other words, 19.3% of unique tag sequences will be completely lost and will not go on to be sequenced. This bottle neck could be even more severe if the failure to capture is worse than what IDT has indicated or PCR is less efficient than assumed. The loss of tag diversity would then cause the remaining tag copies to be "sampled" more frequently (assuming the number of reads is unchanged) during sequencing, which would increase the family size.

We tested his hypothesis by measuring the *per cycle* efficiency of PCR by quantifying the number of ligated genome equivalents used in the PCR reaction using the ATLiS construct for the standard curve. After cycling until saturation (as determined by qPCR) and purifying the PCR reaction using AMPure XP beads, we measure the resulting number of genome equivalents using the ATLiS standard curve. The ratio of the actual yield to the expected yield can then be used to calculate the *per* round PCR efficiency (Table 1). We found that the efficiency was very close to 90% for the two samples we tested. These values are similar to many previous reports that PCR efficiency ranges between 90-100%, suggesting that, at least for PCR input amounts ≤250ng, decreased PCR efficiency is not a significant confounder.

We next tested the efficiency of targeted DNA capture for the IDT xGen probe sets against the CODIS panel. Technical advice from IDT, the vendor supplying the targeted probes, indicates that only 5% of target DNA is typically captured. However, the stated efficiency is based on large, megabase sized, capture panels. We directly tested the capture efficiency by quantifying the amount of genome equivalents going into the capture and the amount of genome equivalents coming out of capture using the previously described ATLiS standard curve. Instead of observing a 5% capture efficiency, we observed a capture efficiency of only 0.06% (Table 1, Fig. 5), suggesting that the targeted capture step is the single biggest cause for our low efficiency. In addition, our approach of using sonication to randomly shear our sample DNA into fragments compatible with MPS library preparation resulted in sequencing reads that did not span the entire STR locus. Reads that fail to do so are not informative for accurate genotyping.

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 MPS [12] and has even been used for STR loci. Therefore, to simultaneously address these issue of limited efficiency of target selection and the failure to fully traverse the STR loci, we sought to use targeted genome fragmentation approach based on CRISPR/Cas9 digestion that produces DNA fragments of similar length. We reasoned that targeted *in vitro* CRISPR/Cas9 digestion could be used to excise similar length fragments covering the areas of interest, which could then be enriched by size selection prior to library preparation, thereby

eliminating one or both targeted capture steps. 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 Duplex Sequencing, we produced 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. The approach, termed CRISPR-DS, enables efficient target enrichment of small genomic regions, even coverage, ultra-accurate sequencing, and reduced DNA input. As a proof of principle, we developed the method for sequencing the exons of *TP53*.

The basic steps of the method is illustrated in Figure 6. First, target regions are excised from genomic DNA by multiplexed *in vitro* CRISPR/Cas9 digestion (Fig. 6A), followed by enrichment of the excised fragments by size selection using SPRI beads (Fig. 6B). The selected fragments are then coupled with the double-strand molecular barcodes used in DS (Fig. 6C). These fragments are then amplified and captured with biotinylated hybridization probes as previously described for DS[10]. We designed gRNAs to specifically excised the coding regions and their flanking intronic sequence of *TP53*. 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. We also designed guides for the CODIS20 STR loci.

We performed a side-by-side comparison of library performance and sequencing coverage of a sample DNA processed with CRISPR-DS versus standard-DS. Standard-DS for *TP53* had been previously performed using sonication and published protocols[10,13]. Visualization of the resulting sequencing library by gel electrophoresis showed that CRISPR restriction produced distinct bands/peaks (Fig. 7A,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 sizes. These fragments exhibited extremely uniform sequencing depth. In contrast, sonicated DNA fragments resulted in significant variability in depth across target regions (Fig 7D). 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. 7C). 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. 7C, *blue*). In contrast, targeted fragmentation produced a consistent number of PCR copies for all fragments, (Fig. 7C, *red*).

Although performing two rounds of capture substantially increases the number of on-target reads for standard-DS, 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 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. 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. Because the *TP53* target region only amounts to 0.0001% of the human genome, this corresponds to approximately 2000X to 50,000X enrichment, which matches or exceeds what is typically achieved with solution-based hybridization for small target size [11]. 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) [10] 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. A side-by-side comparison of CRISPR-DS versus standard-DS 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[9,13], standard-DS produced a recovery rate of ~1% across the different inputs, whereas CRISPR-DS recovery rate ranged between 6% and 12% (Fig. 8B). Notably, 25 ng of DNA prepared with CRISPR-DS produced a post-processing depth comparable to 250 ng with standard-DS (Fig. 8C), 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. A manuscript that describe our CRISPR-DS method were prepared and submitted to the *Genome Research* and accepted for publication on August 31st, 2018 [14].

Another issue that was highlighted during our work was the reliance on targeted hybridization, which is both costly and slow. Together, these limitations would likely preclude the deployment of DS (either standard DS or CRISPR-DS) in a forensic laboratory setting. It was suggested to us to potentially invent a way of performing DS that did not require these steps. To that end, we developed a purely PCR based enrichment approach compatible with Duplex Sequencing, which we termed Linked Strand Anchored Multiplex PCR (LS-AMP) (Fig. 9). Briefly, the LS-AMP approach begins with the fragmentation of the DNA sample, similar to the conventional DS library construction protocol. After end-repair and 3'-dA-tailing, every DNA fragment is ligated with DS adapters containing the random double- stranded barcodes (Fig 9, *Step 1*). All DNA molecules are PCR amplified using primers specific to the universal adapter sequences, resulting in multiple copies of DNA derived from each strand, along with the associated barcode (Fig. 9, *Step 2*). After removing reaction byproducts, the sample is evenly split into two separate tubes (Fig. 9, *Step 3*). This step results in an average of half of the copies of any given strand/barcode being found in each tube. It should be noted that the random nature in which PCR copies are split results in a variance about this mean. To take this variance into account, the hypergeometric distribution (*i.e.* probability of picking *k* barcode copies *without* replacement) can be used as a model to determine the minimum number of PCR copies of a barcode that are needed to maximize the chance that each tube contains at least one copy derived from both strands. Our model indicates that ≥4 PCR cycles in (*i.e.* 2^4=16 copies/barcode) during *Step 1* ensures a >99% probability that each barcode copy derived from each strand will be represented at least once in each tube. After splitting the sample into two tubes, target loci are enriched with multiplex PCR using primers specific for the adapter sequence and to the genetic loci of interest (Fig. 9, *Step 4*). However, the multiplexed loci-specific PCRs are performed so that the PCR products are derived from only one of the two strands. This is achieved as follows: In one tube, PCR is performed using a primer specific for the Read 1 (*i.e.* Illumina P5) adapter sequence (Fig. 9, *Step 4;* blue arrow), as well as primers specific to the genetic loci of interest containing Read 2 (*i.e.* Illumina P7) adapter sequences (Fig. 9, *Step 4;* black arrow w/orange tail). This ensures that amplification only occurs from DNA derived from one strand of the original parental DNA molecule. The same reaction is repeated in the second tube, but amplifying from the opposite strand of the same loci, but with the Read 1 and Read 2 primers roles reversed. Data are analyzed in an approach similar to DS, whereby reads sharing the same molecular barcode derived from the 'Top' or 'Bottom' strand (Which are found in Read 1 and Read 2, respectively) are separately grouped to form a consensus. These single-strand consensuses are then compared to the consensuses computed for the opposite strand and mutations kept if they appear in reads derived from both strands (*i.e.* form a DCS). After some optimization to increase specificity that made use of locus specific linear amplification and the use of nested PCR primers, we tested a draft protocol on a sample of 10ng human nuclear DNA, which corresponds to ~3200 haploid nuclear genomes. The result was a depth of ~1600X, which corresponds to ~50% efficiency, which is at least an order of magnitude higher than our current protocol. However, during the midpoint review, we were directed to cease development of LS-AMP, as it was determined that it was beginning to stray away from the original statement of work. However, we filed a patent application entitled "Methods for Targeted Nucleic Acid Sequence Enrichment with Applications to Error Corrected Nucleic Acid Sequencing" on 3/23/18. In addition, a manuscript describing LS-AMP is

currently under preparation and is expected to be submitted for publication sometime within the next few months. We will keep the ARO informed if and when additional manuscripts are to be sent for publication.

Milestone 3: ***Accuracy, Precision, Sensitivity, and Specificity*****:**

In order to begin testing accuracy, precision, sensitivity, and specificity, we needed to create a "gold standard" reference data set. To that end, we identified 50 samples from the 1000 Genomes Project that had all CODIS20 loci sequenced to a depth of at least 10X, as determined by the Erlich group and provided online (http://lobstr.teamerlich.org/download.html). After ordering the identified samples, we successfully genotyped by both PCR-CE (Promega PowerPlex Fusion 6C) and the Illumina ForenSeq platform, using the manufacturer protocols. All samples were in complete concordance with no differences in genotype calls between the two methods. Due to the large size of the data table that encompasses all 50 samples, we have not included the data in this final report, but can be provided upon request.

Having successfully genotyped the samples using two independent methods, we sought to determine the concordance between Duplex Sequencing and the ForenSeq platform, as well as quantified the stutter rates of the CODIS loci on the ForenSeq platform for each sample and compared it to the same samples sequenced with DS. The data show a markedly higher level of stutter compared to DS (Fig. 10A). Furthermore, we observed a linear increase in stutter amounts with repeat length for conventional PCR-CE and ForenSeq platform, but no increase in DS data. Representative plots are shown in Fig. 10B. As expected, we observed no false positive calls between Duplex Sequencing when compared to ForensSeq platform or PCR-CE. We did, however, notice frequent of instances of allelic dropout that seemed to repeatedly occur at specific loci. This prevented us from continuing on with our validation studies until we were able to determine the cause for the high level of false negatives. This issue caused a significant negative impact in our ability to complete the proposed work. We address the specifics of this issue in the section dedicated to Objective 2.

Objective 2: ***Provide a data analysis workflow for genotype calling.***

Completion of Objective 2 was severely hampered by delayed execution of the subcontractor agreement and then the subsequent under-performance of this subcontractor during the first year of the award. Starting in Year 2, we switch subcontractors to Fulcrum Genomics to develop a data processing pipeline.

Fulcrum Genomics explored three different options for creating the genotype caller: 1) Take previously created genotyping software (*i.e.* lobSTR, hipSTR, STRait Razor, etc) and modify/adapt it for use with Duplex Sequencing data, 2) Create a new genotyping software from scratch, or 3) A hybrid approach. After evaluation of options, it was decided to undertake a hybrid approach where each molecular barcode family would be grouped and submitted to the hipSTR program(ref), program to perform genotyping.

The basic pipeline works as follows:

A) Filter duplex source molecules that do not have enough observations (reads) on each strand respectively (ex. require at least some number of reads on both the top and bottom strand).

B) Learn the "stutter" error model using hipSTR for each duplex source molecule *(i.e. the raw data sharing the same molecular barcode)* for each locus as an independent "haploid" sample.

C) Apply the "stutter" model from (B) and genotype each barcode family, treating ***each strand of a duplex molecule*** as an independent haploid sample.

D) Filter genotypes for duplex source molecules where the two strand-specific genotypes do not agree.

E) Choose the most frequent two genotypes across the duplex source molecule haploid genotypes, requiring a minimum allele frequency for each.

To test this approach, we performed DS on the 50 samples from our validation samples and then performed our prototype genotyping approach. We were able to determine the amount of stutter observed across all loci and compared it the observed values obtained by PCR-CE and the Illumina Forenseq platform (Fig. 10A). This approach clearly demonstrates that DS is able to effectively remove PCR stutter.

However, in the course of our experiments, we noticed that some loci exhibited dramatically reduce depth. To further investigate this issue, we plotted the relative depth (defined as the ratio of total number of DCS reads before genotyping vs being filtered) against the STR length for several of the loci that show the genotyping issues. Our results showed that for loci that have a difficult time correctly genotyping, there is a negative correlation between repeat length and genotyping performance (Fig. 11). This finding indicates the reason behind why longer STRs are not being called correctly. Specifically, the longer the STR, the less final genotyping depth. If there are no reads, then the locus cannot be correctly genotyped. These data, in conjunction with our observation that HipSTR was filtering reads due to low quality, suggested that there is an issue with the actual sequencing of these loci by the sequencer itself. To further investigated this possibility, we observed the raw sequencing data by plotting the raw FASTQ files against the hg19 human genome without performing Duplex Sequencing error correction or genotyping. We noticed several aspects of the raw data that were concerning. 1) The base quality scores are dramatically reduced after reading through the STR; 2) A high percentage reads were soft-clipped almost immediately adjacent to the STR, which would have the effect of reducing the chance of having sufficient flanking sequence to be used by HipSTR for genotyping. However, these observations were based on all the raw reads that are not necessarily related to one another. Reads sharing the same molecular barcode may maintain a similar sequence (since they are PCR copies of one another), which would allow them to successfully form a Duplex Consensus Sequence and, thus, not account for our problem. We isolated individual molecular barcode families and visualized them separately using the Integrated Genome Browser (Fig. 12). The visualization confirmed that there is a significant problem with cluster quality once a read exits from the STR locus (each read is derived from a single cluster). This issue is not likely due to stutter arising during library preparation because all reads sharing a molecular barcode appear to vastly different from one another. We hypothesized that sequencing through a STR in one direction (where the read sequence gives the anti-reference sequence) may not perform well, whereas sequencing through a STR in the opposite direction (where the read sequence give the reference sequence) may perform better. Consistent with this possibility, the read orientation of the Illumina (Verogen) ForenSeq platform are in the opposite orientation from our data for every locus that performs poorly in our studies. To test this hypothesis, we PCR amplified both the PentaD and D12S391 loci using PCR primers that produced PCR amplicons that were the same as what is obtained in an Illumina/Verogen ForenSeq sequencing run. In one sample, the amplicons were designed such that read traversed the STR in the same sense as the reference genome, whereas in the second sample, the read traversed the STR in the anti-sense orientation as the reference genome. We hypothesized that in one orientation, the reads would fail after traversal through the STR, whereas cluster would show significantly reduced failure after traversing the STR in the opposite orientation. As can be seen in Figure 13, both PentaD (Fig 13A) and D12S391 (Fig 13B) show significant read failure when read traverse the STR in the anti-reference consistent with our previous observation, but show no sign of read failure when in the reference (forward) orientation. This confirms our hypothesis that the STR sequence itself leads to some state in the cluster that results in high error rates, which ultimately results in poor Duplex Sequencing performance. We redesigned the CRISPR/Cas9 gRNA sites and hybridization probes so that the sequencing reads will traverse the STR in the same orientation that we observe in the Illumina/Verogen ForenSeq kit. After solving this issue, the genotyping pipeline was able to successfully genotype our samples, but low depth remained an issue due to the reliance of the protocol on targeted DNA hybridization.

Objective 3: ***Validate SOP and equipment set for use in genotyping complex DNA mixtures and highly damaged/degraded DNA.***

Because of the inefficient nature of the standard protocol arising from targeted DNA capture and the recommendation that continued development of non-capture based approaches to Duplex Sequencing not be performed, we did not perform validation experiments related to DNA mixtures or damaged/degraded DNA.

Objective 4: ***Develop and provide a training program and material.***

A detailed step-by-step protocol was written and transferred to DFSC. A hands on training program was not provided.

4. Bibliography

1. Alaeddini R, Walsh SJ, Abbas A. Forensic implications of genetic analyses from degraded DNA-A review. Forensic Science International: Genetics. 2010;4: 148–157.
2. Pompanon F, Bonin A, Bellemain E, Taberlet P. Genotyping errors: Causes, consequences and solutions. Nat Genet. 2005;6: 847–859.
3. Butler JM, Buel E, Crivellente F, McCord BR. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. Electrophoresis. WILEY-VCH Verlag; 2004;25: 1397–1412.
4. Budowle B, Onorato AJ, Callaghan TF, Manna Della A, Gross AM, Guerrieri RA, et al. Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. Journal of Forensic Sciences. Blackwell Publishing Ltd; 2009;54: 810–821.
5. Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, et al. Quantification of rare allelic variants from pooled genomic DNA. Nat Meth. 2009;6: 263–265.
6. Metzker ML. Sequencing technologies - The next generation. Nat Genet. 2010;11: 31–46.
7. Shinde D, Lai Y, Sun F, Arnheim N. *Taq* DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$ microsatellites. 2003;31: 974–980.
8. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J Mol Diagn. 2013;15: 623–633.
9. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci USA. National Acad Sciences; 2012;109: 14508–14513.
10. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc. 2014;9: 2586–2606.
11. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, et al. Sequencing small genomic targets with high efficiency and extreme accuracy. Nat Meth. Nature Publishing Group; 2015;12: 423–425.
12. Shin G, Grimes SM, Lee H, Lau BT, Xia LC, Ji HP. CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. Nature Communications. Nature Publishing Group; 2017;8: 1–13.
13. Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ, et al. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic *TP53* mutations in noncancerous tissues. Proc Natl Acad Sci USA. National Acad Sciences; 2016;113: 6005–6010.
14. Nachmanson D, Lian S, Schmidt EK, Hipp MJ, Baker KT, Zhang Y, et al. 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). Genome Res. 2018;28: 1589-1599.

5. Figures

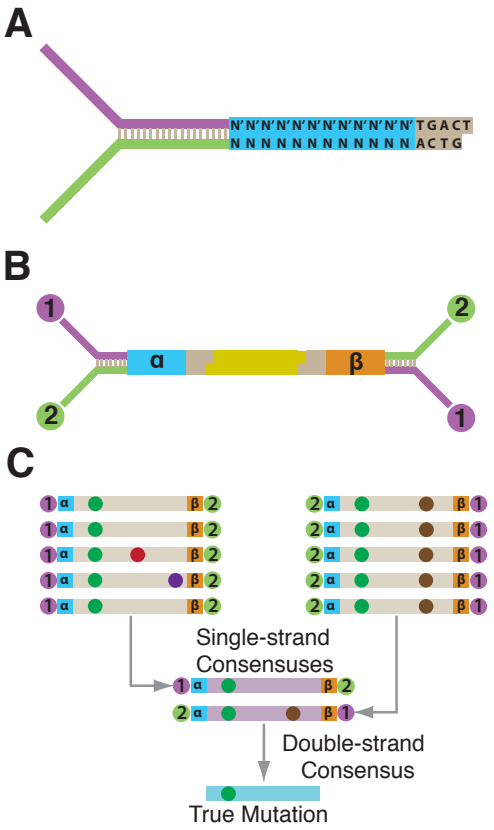

Fig. 1. Duplex Sequencing. (A) Adapter design containing the degenerate double- stranded barcode. **(B)** Ligation of adapters to sheared DNA (*yellow*) generates unique tags on each end (α and β). **(C)**PCR of the two strands produces two related but distinct products. Reads sharing a unique α and β are grouped into families. Mutations are of three types: sequencing mistakes or late arising PCR error (*blue or purple spots*); first round PCR errors (*brown spots*); true mutations (*green spots*).

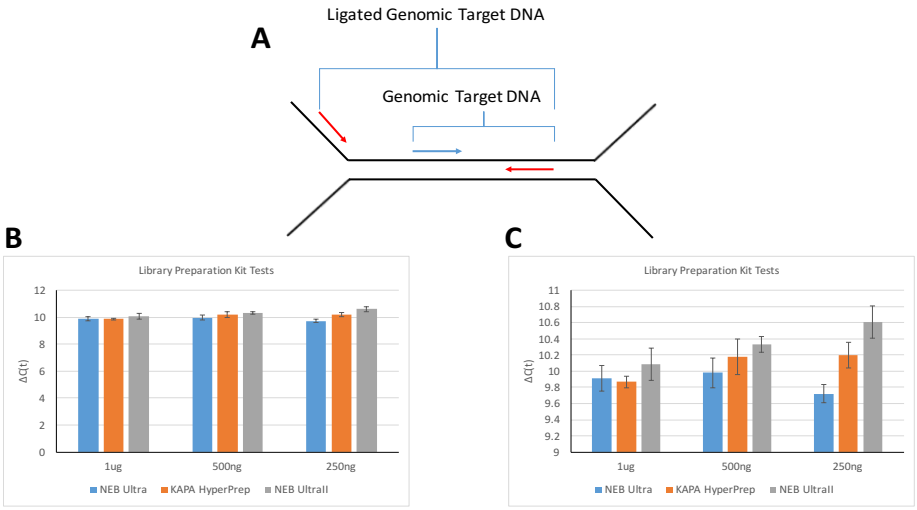

Fig. 2 Library preparation kit comparisons. (A) Schematic of the qPCR assay used to measure ligation efficiency. Only adapter ligated sample DNA can amplify with the red primer pair. However, all genomic target DNA can amplify with the blue and reverse red primer pair. The resulting ΔC(t) can be used to measure ligation efficiency. (B) Plot of the ΔC(t) between the Ligated Target DNA and the total Genomic DNA for three different DNA input amounts and three different vendors' kits (n=3 per reaction). Due to a significant disparity in the amplicons sizes between the two primer sets, a higher ΔC(t) indicates increased ligation efficiency. (C) Zoom in of the data presented in (B).

A

Human Nuclear ATLiS:

AATGATACGGCGACCACCGAcagcacgcgccgccagcacgtccgccggacaggctgcatcgtcagctcag
gaagcgtcctccggcgcagaagcggcatcagcaaaggccactgaagcggaaaaaagtgccgcagccgcagagtcctc
aaaaaacgcggcggccaccagtgccggtgcggcgaaaacgtcagaaacgaatgctgcagcgtcacaacaatcagcc
gccacgtctgcctccaccgcggccacgaaagcgtcagaggccgccacttcagcacgagatgcggtggcctcaaaagag
gcagcaaaatcatcagaaacgaacgcatcatcaagtgccggtcgtgcagcttcctcggcaacggcggcagaaaattctg
ccagggcggcaaaaacgtccgagacgaatgccaggtcatctgaaacagcagcggaacggaGAACAAGTTGGC
CTGCACTGTCGTATGCCGTCTTCTGCTTG

B

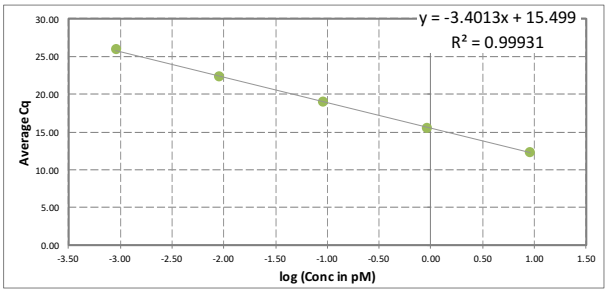

Fig. 3. Design of Standard Curve. (A) Design of the qPCR standard. Standard DNA is a 481bp synthesized fragment of lambda-phage genome. Colored bases denote a qPCR primer binding site and correspond to the corresponding oligonucleotides: P5-primer: 5'-AATGATACGGCGACCACCGA (*blue*); P7-primer: 5'-CAAGCAGAAGACGGCATACGA (*bold black*); Human Nuclear Reverse (p53): 5'-CAGTGCAGGCCAACTTGTTC (*orange*). (B) Plot of standard curve from data presented in Table 1. Standard 6 is not included in the plot and the that dilution is not expected to be used in future experiments.

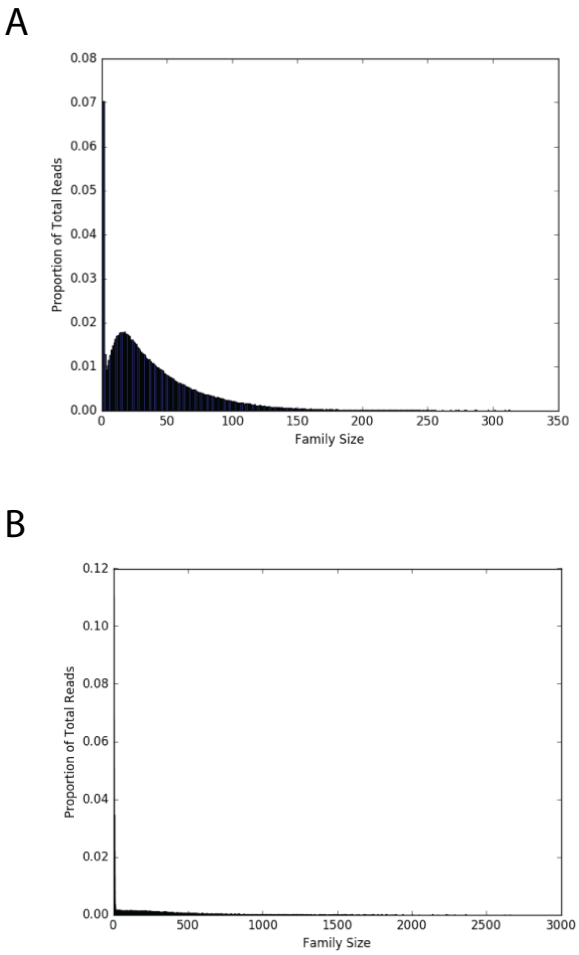

Fig. 4. Family size changes arising from additional targeted DNA capture. (A) Family size distribution with 250ng of genomic DNA with a single round of targeted capture of the CODIS20 loci and sequenced with $5x10^6$ reads. The peak family size is ~20, which is considered optimal. (B) Family size distribution with the same conditions as (A), except a second round of targeted capture was performed.

Table 1. Efficiency metric for PCR and two rounds of targeted capture.

Sample ID	PCR input (amoles)	# cycles(n)	Expected yield (amoles)*	Actual yield (amoles)	*per* round efficiency^	Post-capture yield (amole)	Capture efficiency
SNP	39.2	8	10048	3522	0.88	1.54	0.06%
CODIS	38.8	8	19875	4060	0.89	1.22	0.06%

*Calculated by Expected yield = $X*2^n$, where X is PCR input and n is the number of cycles

^Calculated by $eff = \sqrt[n]{\frac{actual}{expected}}$, where n is the number of cycles

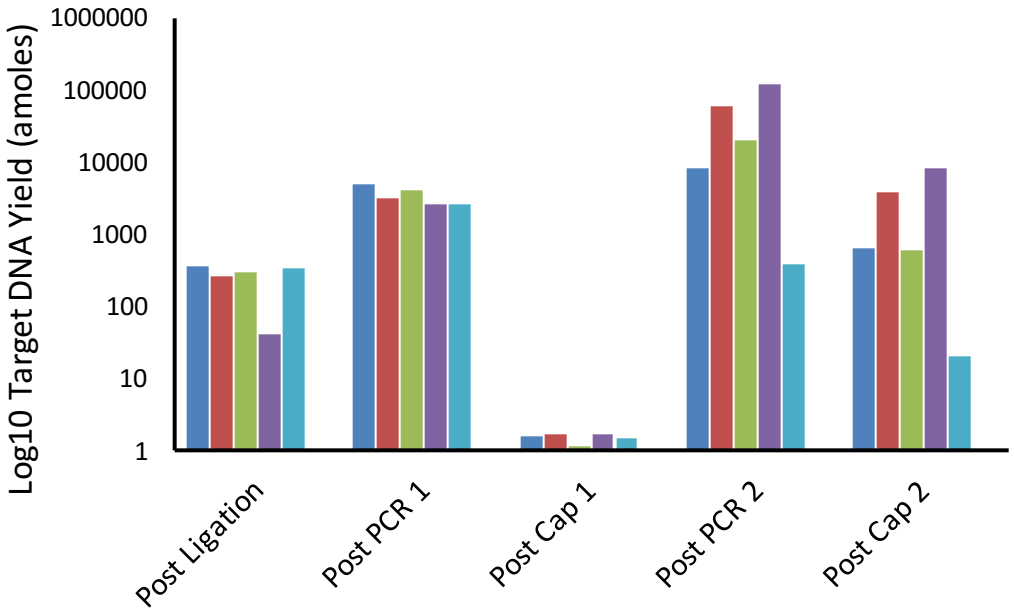

Fig. 5. First round of targeted capture is highly inefficient. The qPCR standard presented in Fig. 3B was used to quantify the absolute number of target molecules in each step of the DS protocol for five independent 250ng sample. The first targeted capture step (Post Cap 1) showed an extreme drop in the number of target loci, indicating that the vast majority of the sample diversity is lost at this step. The subsequent PCR step (Post PCR 2) is likely the cause of the extreme family size bias seen in Fig. 4B.

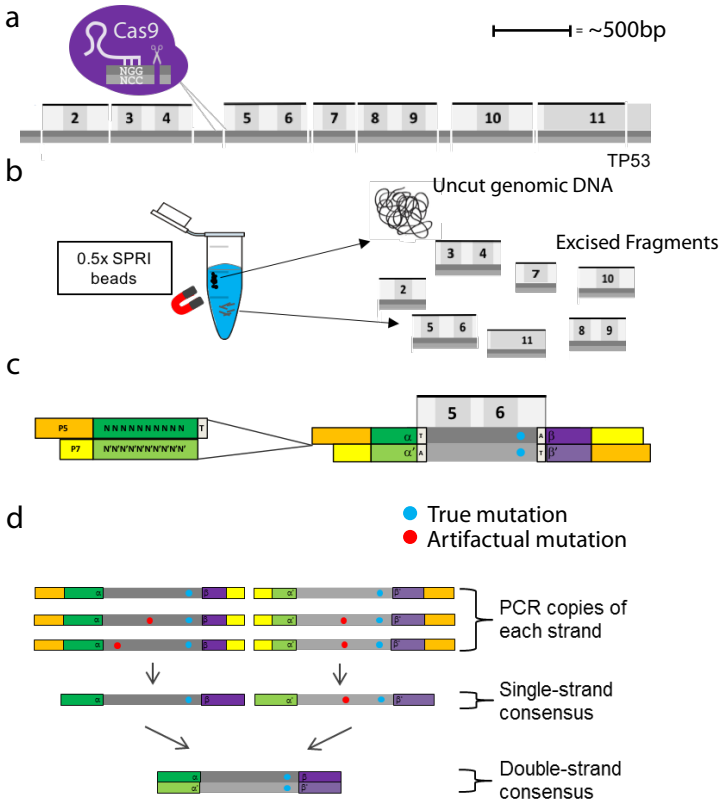

Fig. 6. Schematic representation of key aspects of CRISPR-DS. (a) CRISPR/Cas9 digestion of TP53. Seven fragments containing all *TP53* coding exons were excised via targeted cutting using gRNAs. Dark grey represents reference strand and light grey represents the anti-reference strand. (b) Size selection using 0.5x SPRI beads. Uncut, genomic DNA binds to the beads and allows the recovery of the homogenously sized 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.

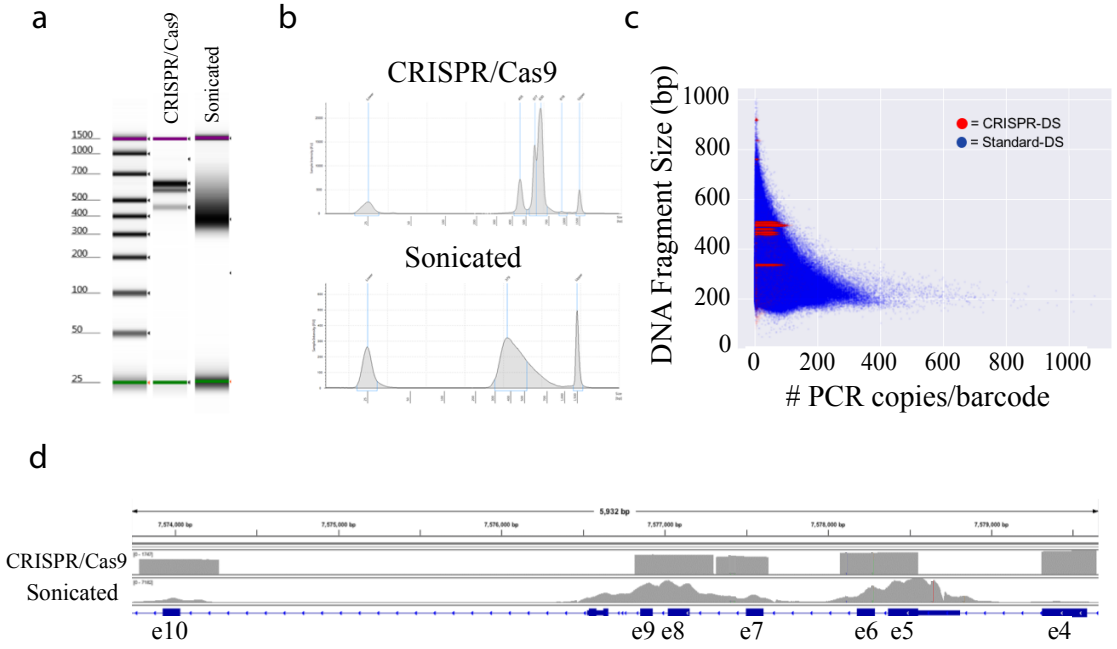

Fig. 7. Visualization of sequencing libraries and data prepared with CRISPR-DS and standard-DS. (A) TapeStation gels show distinct bands for CRISPR-DS as opposed to a smear for standard-DS. The size of bands corresponds to the CRISPR/Cas9 cut fragments with adapters. (B) CRISPR-DS electropherograms allow 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 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.

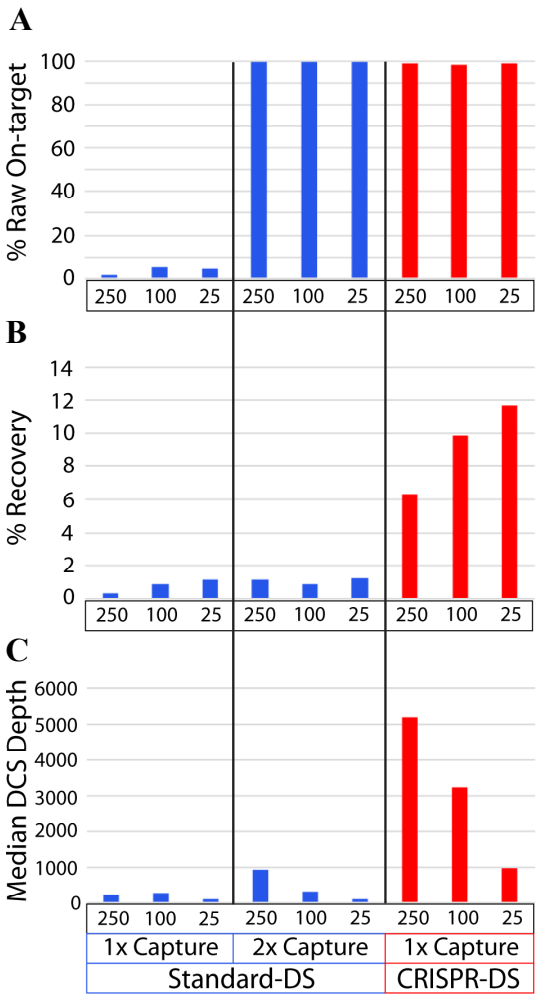

Fig. 8. Technical comparison of 250ng, 100ng and 25ng of DNA sequenced with both standard-DS and CRISPR-DS. Measurements were obtained by sequencing samples prepared with standard-DS (*blue*) using one and two rounds of hybridization capture and CRISPR-DS (*red*) with only one round of hybridization 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 5-10 times more sequencing depth for the same input DNA.

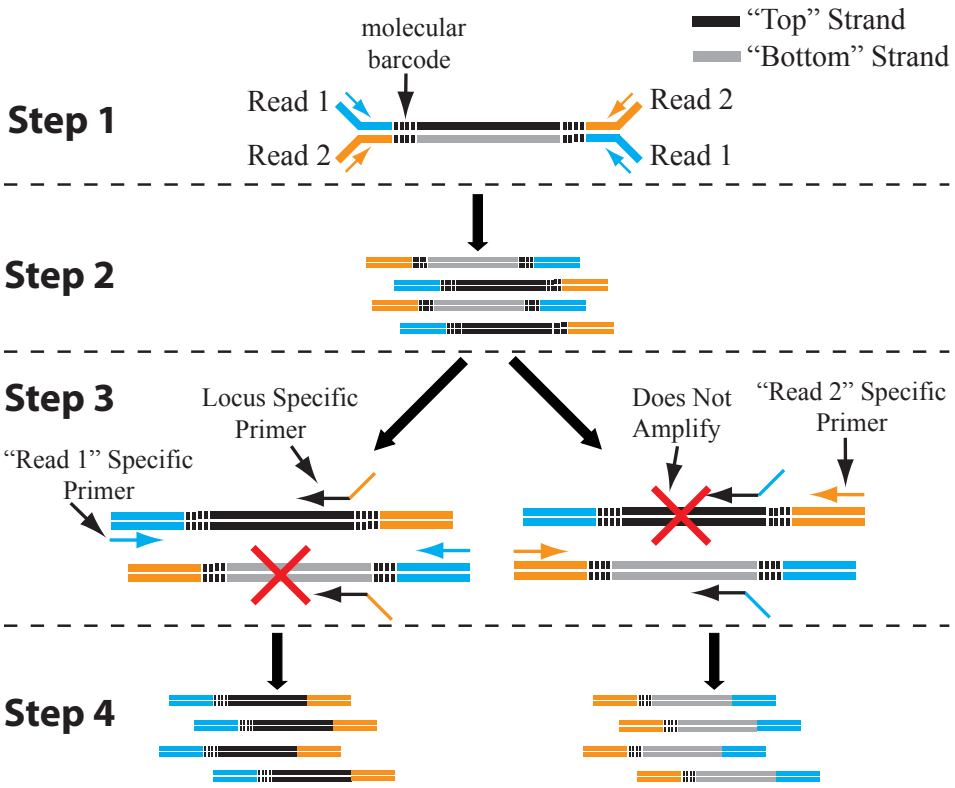

Fig. 9. The LS-AMP Approach. LS-AMP is broken up into four major steps. Double-stranded molecular barcodes are ligated to fragmented DNA and all molecules are PCR amplified to make two related but distinct products derived from the two strands (*black and grey*). For clarity, the orientation of molecules is maintained between steps. After splitting, target loci are PCR amplified from DNA derived from only one strand ("Top" strand on the left, "Bottom" strand on the right) by using a primer against the appropriate adapter sequence (*blue or orange arrow*) and a target specific primer that introduces the adapter sequence specific (black arrow w/orange or blue tail) for the Illumina platform. For clarity, only targeted genomic sequence are shown. After a final SPRI bead cleanup, the target amplicon library is ready for quantitation and sequencing. Error correction is performed in a similar manner as DS (Fig. 1C).

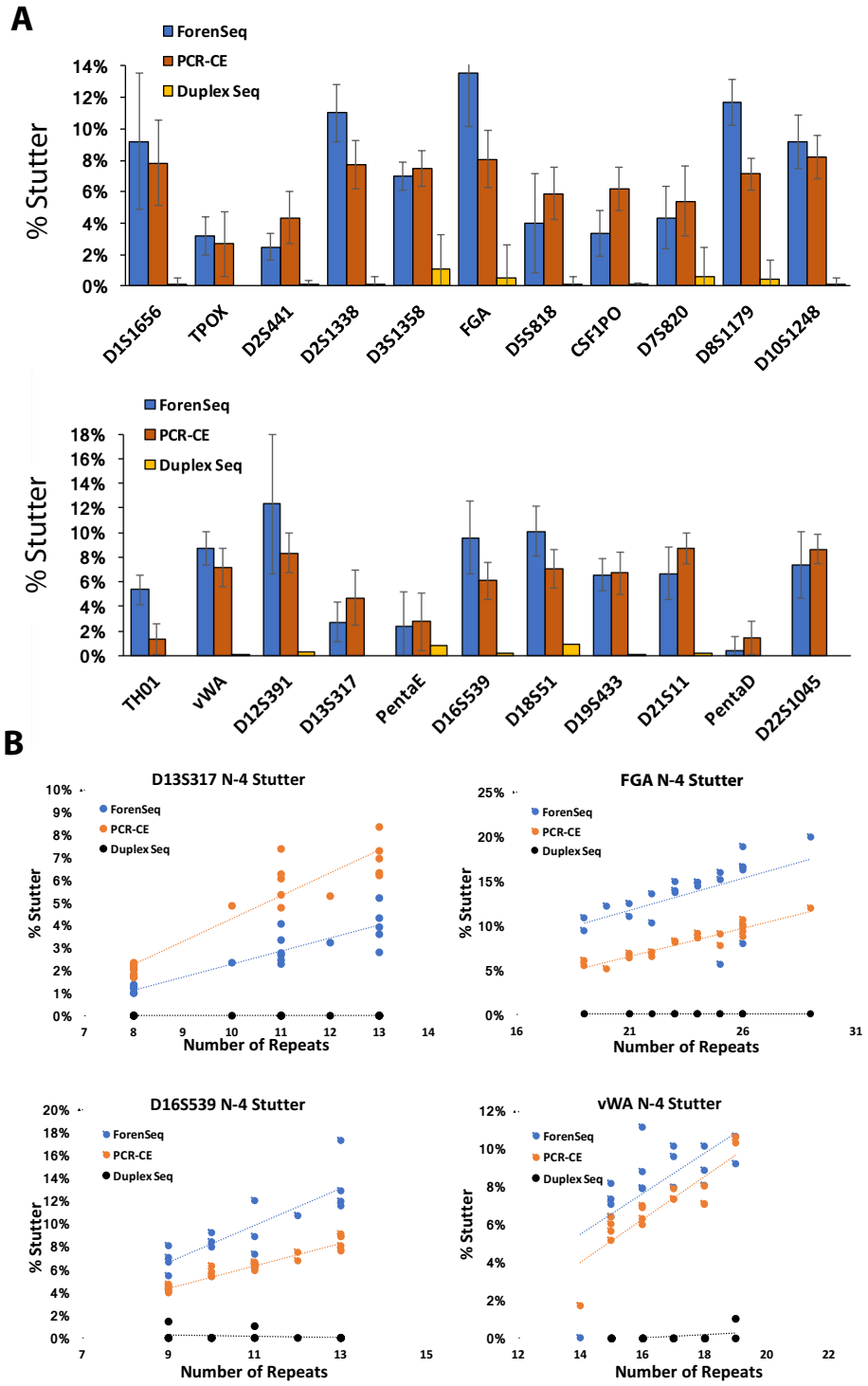

Fig. 10. Duplex Sequencing exhibits less PCR stutter compared to conventional genotyping methods. (A) Comparison of stutter levels for conventional PCR-CE (*orange*), Illumina ForenSeq system (*blue*), and DS (*yellow*) for each of the CODIS20 loci. (B) Correlation between STR repeat length and stutter rate for four representative CODIS loci. The other remaining CODIS loci show similar correlations.

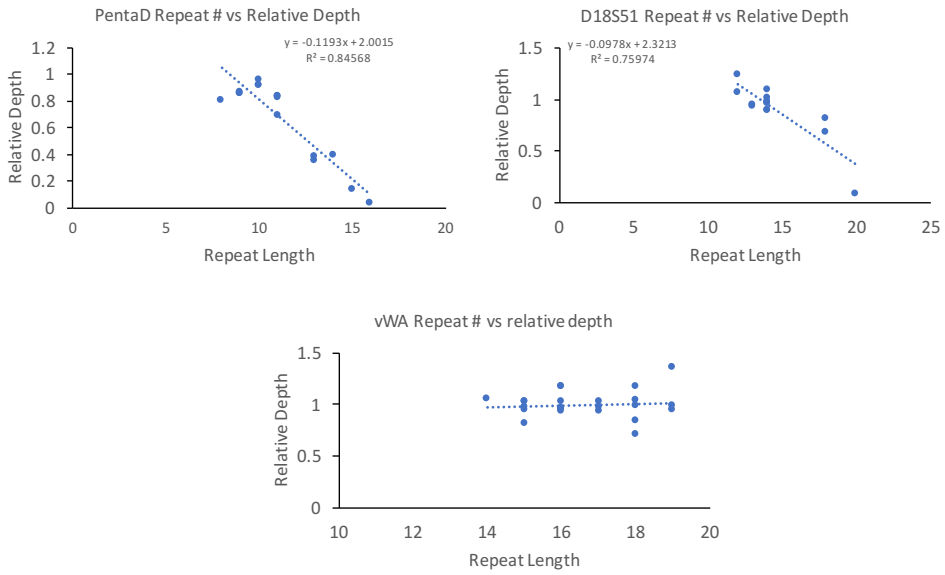

Fig. 11. Negative correlation between post-genotyping depth performance and STR length. Plots are representative plots from two loci that performed poorly (PentaD and D18S51) and one that consistently genotyped correctly (vWA). Data explain why longer length STRs do not report a genotype.

Fig. 12. Read failures at PentaD locus. Integrated Genome Viewer screenshot of a single molecular barcode family that becomes highly error prone after exiting the PentaD STR. Each read is derived from a single cluster on the sequencer. Data suggest that the sequencer itself is having difficulty properly sequencing the STR.

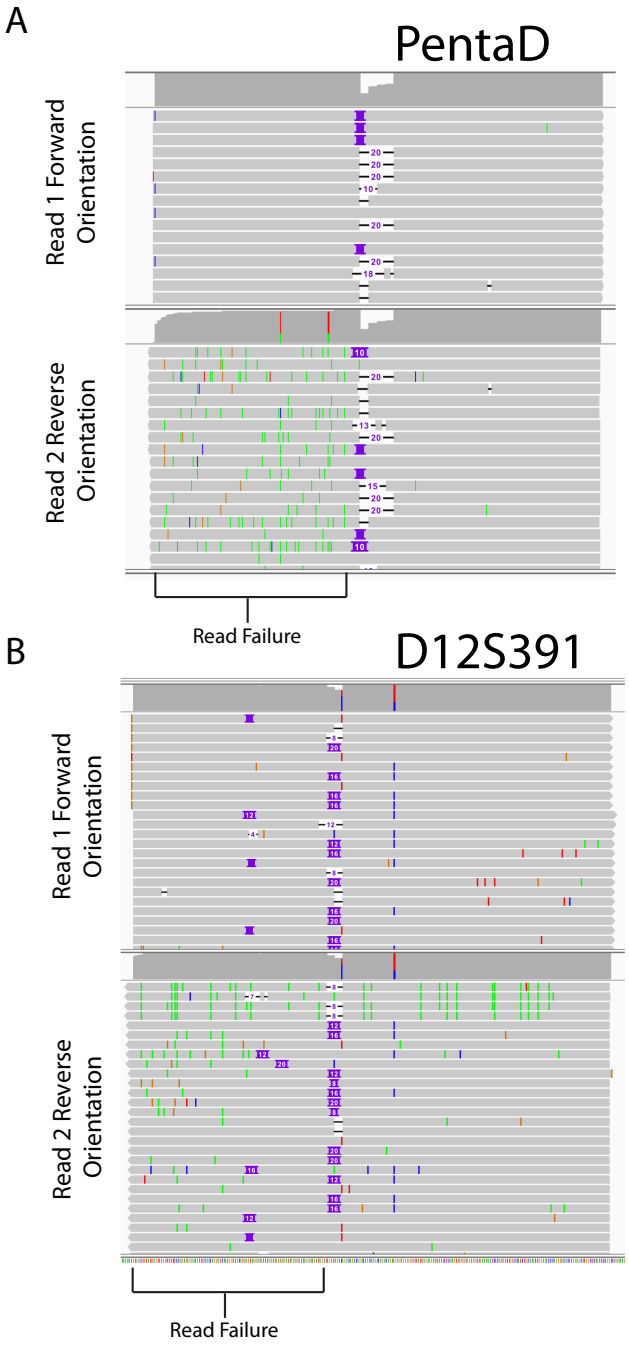

Fig. 13. Read orientation through STR locus affect read quality. (A) PentaD locus and (B) D12S391 both show a difference in read quality depending if the read traverses the STRs in the forward (i.e. reference) orientation (*top*) or in the reverse (i.e. anti-reference) orientation (*bottom*). In these representative cases, anti-reference mapping reads exhibited a loss of read quality after traversing the STR loci.