REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
25-02-2019	Final Report	15-Apr-2017 - 14-Apr-2018

4. TITLE AND SUBTITLE
Final Report: Massively Parallel Sequencing Platform for High Accuracy Variant Detection

5a. CONTRACT NUMBER
W911NF-17-1-0158

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611103

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)
70049-LS-RIP.2

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: 70049LSRIP **Agreement Number: W911NF-17-1-0158**

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: 14-May-2018 Date Received: 25-Feb-2019
Final Report for Period Beginning 15-Apr-2017 and Ending 14-Apr-2018
Title: Massively Parallel Sequencing Platform for High Accuracy Variant Detection
Begin Performance Period: 15-Apr-2017 **End Performance Period:** 14-Apr-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: Purchase an Illumina MiSeq FGx sequencing platform for two ongoing and future studies.

Accomplishments: C.1. Project 1: High Accuracy Genotyping of Complex Mixtures and Damaged DNA
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. Even with 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. Of particular concern are PCR stuttering artifacts, which arise from slippage of the DNA polymerase of 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. 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. 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. 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. 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]. 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.
The purpose of the effort and the associated development under the DURIP Grant No. W911NF- 17-1-0158 was to acquire a massively parallel sequencing platform to facilitate the development of a high accuracy sequencing method for forensic human identification conducted as part of ARO/RIF Cooperative Agreement W911NF-15-2-0127. The purchased equipment was used to more fully develop Duplex Sequencing for forensic applications. This work 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. Focus areas 2) and 3)

required the use of the purchased equipment.
Optimization steps of PCR amplification. For Duplex Sequencing to function efficiently, 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.
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. By making serial dilutions of the synthetic DNA and verifying the concentration each dilution using the qPCR, we can then quantify the amount of ligated target DNA in the library preparation by using the P5 and a target specific primer. We then used the Illumina MiSeq FGx sequencer to sequence the test samples and determine the family size for the given DNA input.
Optimization of target genome capture. During our development phase, we noticed that <10% of the raw reads mapped to the CODIS20 loci, but with an optimal family size of ~20. However, performing a second round of targeted capture resulted in >90% of reads being localized to the genomic targets of interest. However, as a consequence of adding the second round of targeted hybridization, we observed a higher than desired family size is observed for the CODIS capture panels. Therefore, we wanted to develop an approach that removed one or both rounds of targeted capture. Again, the purchased Illumina MiSeq FGx was used for these experiments.
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 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 and a subset of CODIS20 loci. The basic steps of the method is illustrated in Figure 2. First, target regions are excised from genomic DNA by multiplexed in vitro CRISPR/Cas9 digestion (Figure 2A), followed by enrichment of the excised fragments by size selection using SPRI beads (Fig. 2B). The selected fragments are then coupled with the double-strand molecular barcodes used in DS (Figure 2C). These fragments are then amplified and captured with biotinylated hybridization probes using standard solution-based hybridization methods. In our proof-of-principal, we designed gRNAs to specifically excised the coding regions and their flanking intronic sequence of TP53 to be ~500 bp in size 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.
We performed CRISPR/Cas9 dig

Training Opportunities: Nothing to Report

Results Dissemination: Nothing to Report

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

RPPR Final Report
as of 03-Apr-2019

PARTICIPANTS:

Participant Type: PD/PI
Participant: Scott R Kennedy
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

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

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Services Directorate (9000-0095). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

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-17-1-0158	University of Washington	Same	a. INTERIM [ ] b. FINAL [X]
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	20170415	4333 Brooklyn Ave NE Seattle, WA 98195-0001	20170415	a. FROM 20170415 b. TO 20180414

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.	d. ELECTION TO FILE PATENT APPLICATIONS (X): (1) UNITED STATES (a) YES	(1) UNITED STATES (b) NO	(2) FOREIGN (a) YES	(2) FOREIGN (b) NO	e. CONFIRMATORY INSTRUMENT OR ASSIGNMENT FORWARDED TO CONTRACTING OFFICER (X): (a) YES	(b) NO
Kennedy, Scott R; Sellig, Georg	Massively parallel single-cell transcriptomics and genomic variant detection	48361		X		X		X
Kennedy, Scott R; Salk Jesse S; 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

(1) (a) NAME OF INVENTOR *(Last, First, Middle Initial)*	(2) (a) NAME OF INVENTOR *(Last, First, Middle Initial)*
(b) NAME OF EMPLOYER	(b) NAME OF EMPLOYER
(c) ADDRESS OF EMPLOYER *(Include ZIP Code)*	(c) ADDRESS OF EMPLOYER *(Include ZIP Code)*

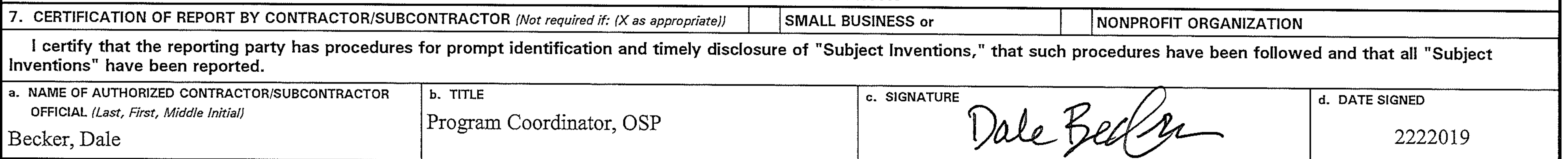
g. ELECTED FOREIGN COUNTRIES IN WHICH A PATENT APPLICATION WILL BE FILED

(1) TITLE OF INVENTION	(2) FOREIGN COUNTRIES OF PATENT APPLICATION
CRISPR/Cas9 based methods for targeted genome enrichment	European Union

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.	d. FAR "PATENT RIGHTS": (1) CLAUSE NUMBER	(2) DATE *(YYYYMM)*	DESCRIPTION OF WORK TO BE PERFORMED UNDER SUBCONTRACT(S) e.	f. SUBCONTRACT DATES *(YYYYMMDD)*: (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	2222019

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.

Final Report: DURIP Grant Number: W911NF-17-1-0158
Sponsor: Army research Office (ARO)
PI: Scott R Kennedy
Program Title: Massively Parallel Sequencing Platform for High Accuracy Variant Detection

A. Abstract

Sequencing technologies have rapidly progressed in the past 15 years from “classic” Sanger sequencing to the current generation of massively parallel sequencing (MPS) platforms, which are capable of simultaneously sequencing hundreds of millions of small DNA fragments. Due to how these technologies work, it is possible to digitally tabulate the sequence of individual DNA molecules, thus allowing for the detection of sequences occurring in the minority of DNA molecules. While a significant improvement over original Sanger methods, MPS technology is not able to detect a variant in <1% of sequencing reads. To overcome the high error rate of MPS, we developed a highly accurate MPS methodology termed Duplex Sequencing (DS). DS dramatically improves the ability to detect low level variants and delivers a >10,000-fold improvement in accuracy compared to conventional MPS. In this DURIP program an Illumina MiSeq FGx massively parallel sequencing platform was purchased to facilitate the development of DS for use in a variety of applications of interest to the DoD. The instrument is also available to other research groups at the University of Washington, where it has been used to attract significant external funding focused on early detection of ovarian cancer and forensics, both of significant interest to the DoD.

B. Budget

The following table details the actual spent budget and the list of purchased equipment that are used in this setup.

Part #	Description	Manufacturer	Quantity	Unit Price	Price
SY-411-1001	MiSeq FGx Sequencing System	Illumina	1	$124,999.60	$124,999.60*
SE-550-1001	ForenSeq Universal Analysis Software and Server	Illumina	1	$16,500.00	$16,500*
TG-143-1001	MiSeq FGx Reagent Kit	Illumina	1	$1,400.00	$1,400.00
TG-450-1001	ForenSeq DNA Signature Prep Kit	Illumina	1	$15,500.00	$15,500.00
				Sub-Total:	$158,399.60
				Tax (9.6%)	$1,622.40
				Grand Total	$160,022.00

*Tax exempt

C. Supporting Information

Figure 1 shows the final purchased Illumina FGx sequencing platform.

Figure 1. Photo of deployed Illumina MiSeq FGx.

C.1. Project 1: High Accuracy Genotyping of Complex Mixtures and Damaged DNA

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. Even with 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. Of particular concern are PCR stuttering artifacts, which arise from slippage of the DNA polymerase of 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. 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. 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. 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. 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]. 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.

The purpose of the effort and the associated development under the DURIP Grant No. W911NF-17-1-0158 was to acquire a massively parallel sequencing platform to facilitate the development of a high accuracy sequencing method for forensic human identification conducted as part of ARO/RIF Cooperative Agreement W911NF-15-2-0127. The purchased equipment was used to more fully develop Duplex Sequencing for forensic applications. This work 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. Focus areas 2) and 3) required the use of the purchased equipment.

Optimization steps of PCR amplification. For Duplex Sequencing to function efficiently, 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.

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. By making serial dilutions of the synthetic DNA and verifying the concentration each dilution using the qPCR, we can then quantify the amount of *ligated* target DNA in the library preparation by using the P5 and a target specific primer. We then used the **Illumina MiSeq FGx sequencer** to sequence the test samples and determine the family size for the given DNA input.

Optimization of target genome capture. During our development phase, we noticed that <10% of the raw reads mapped to the CODIS20 loci, but with an optimal family size of ~20. However, performing a second round of targeted capture resulted in >90% of reads being localized to the genomic targets of

interest. However, as a consequence of adding the second round of targeted hybridization, we observed a higher than desired family size is observed for the CODIS capture panels. Therefore, we wanted to develop an approach that removed one or both rounds of targeted capture. Again, the purchased Illumina MiSeq FGx was used for these experiments.

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 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* and a subset of CODIS20 loci. The basic steps of the method is illustrated in Figure 2. First, target regions are excised from genomic DNA by multiplexed *in vitro* CRISPR/Cas9 digestion (Figure 2A), followed by enrichment of the excised fragments by size selection using SPRI beads (Fig. 2B). The selected fragments are then coupled with the double-strand molecular barcodes used in DS (Figure 2C). These fragments are then amplified and captured with biotinylated hybridization probes using standard solution-based hybridization methods. In our proof-of-principal, we designed gRNAs to specifically excised the coding regions and their flanking intronic sequence of *TP53* to be ~500 bp in size 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.

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. 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. Notably, 25 ng of DNA prepared with CRISPR-DS produced a post-processing depth comparable to 250 ng with standard-DS (Figure 3), 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 accepted for publication in *Genome Research* (See Section C3).

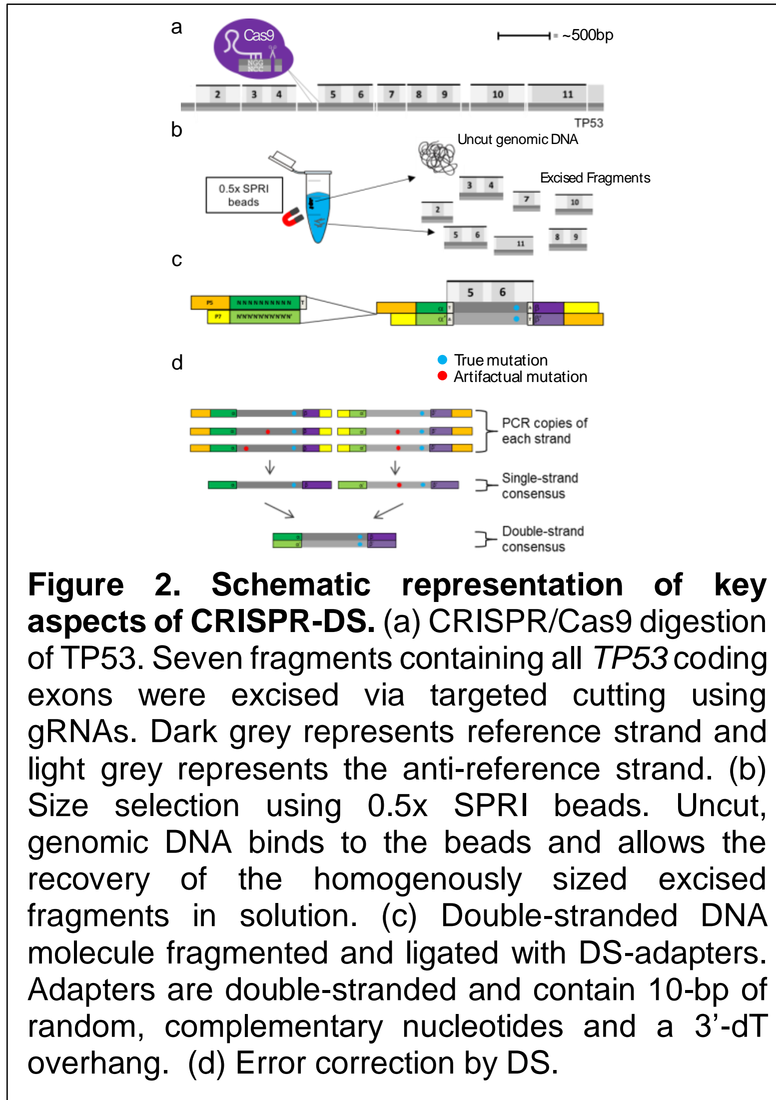


Figure 2. Schematic representation of key aspects of CRISPR-DS. (a) CRISPR/Cas9 digestion of TP53. Seven fragments containing all *TP53* coding exons were excised via targeted cutting using gRNAs. Dark grey represents reference strand and light grey represents the anti-reference strand. (b) Size selection 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.

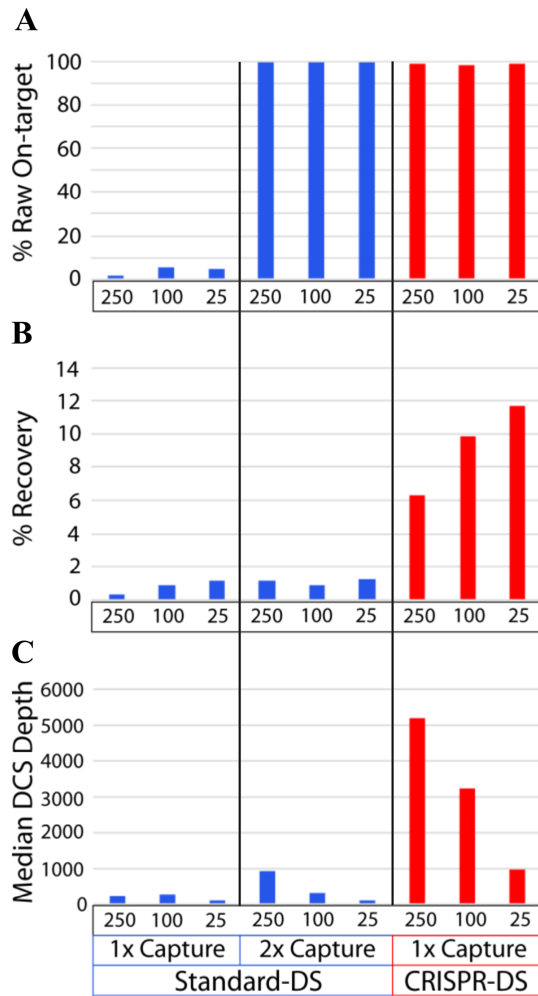


Figure 3. Technical comparison of 250ng, 100ng and 25ng of DNA sequenced with both standard-DS and CRISPR-DS. Measurements were obtained by sequencing samples prepared with standard-DS (*blue*) using one and two rounds of hybridization capture and CRISPR-DS (*red*) with only one round of hybridization 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.

C.2. Project 2: Genetic Determinants Influencing the Maternal Transmission and Somatic Accumulation of Mitochondrial Mutations

The focus of the grant is on mitochondrial diseases. Owing to their evolutionary history, mitochondria have retained a small circular genome that encodes essential components of the electron transport chain, which provides the majority of the energy required for an organism to live. Mutations in the mitochondrial genome can cause the electron transport chains to become dysfunctional. These mutations can be potentially passed to the offspring of the female carrier or accumulate with advancing

age. If these mutations reach a high enough level within a tissue, a number of devastating diseases can occur. Remarkably, there is significant evidence to suggest that mechanisms have evolved to prevent the transmission or accumulation of the most deleterious of these mutations. However, even in the presence of these selection mechanisms, diseases involving mtDNA mutations occur. The goal of the funding is to understand the mechanism(s) that are involved in preventing the accumulation of these mutations either through the germline or during development and aging of the somatic tissue by identifying candidate genes involved in this process. To begin to identify these mechanisms, we are using the common fruit fly, *Drosophila melanogaster*, as a model organism to pursue two aims. The first aim will focus on using a novel, high accuracy, next-generation sequencing methodology to perform a genome wide screen for genes involved in selecting against the accumulation and/or transmission of mtDNA mutations. The second aim will focus on characterizing the effects of putative modifiers on mitochondrial bioenergetics and morphology in hopes of providing insight into how these modifiers function.

We used the **Illumina MiSeq FGx sequencer** purchased by the W911NF-17-1-0158 funds to help develop a Drosophila model with germline specific expression of the cytidine deaminase APOBEC1. Briefly, Cyo/UAS-APOBEC1 flies were crossed with a strain of flies that expresses the transcription factor that stimulates the UAS sequence, Gal4, under the control of three different ovarian drivers (MTD): the Bloomington Drosophila Stock Center ID 31777 P{w[+mC]=otu-GAL4::VP16.R}1, w[*]; P{w[+mC]=GAL4-nos.NGT}40; P{w[+mC]=GAL4::VP16nos.UTR} CG6325[MVD1] (Fig.1). *otu* and *nos* drivers have previously been reported to be expressed in early oogenesis. Since both stocks (UAS-APOBEC- 1 and 31777) were found in homozygous dose, the first progeny produced from these crosses (F1) would be expected to be heterozygous for both constructs and therefore, expressing Gal4 in early stages of oogenesis of the F1 females and consequently stimulating the expression of APOBEC-1.

To test whether the heterozygous F1 females were expressing APOBEC and inducing mtDNA mutations in their germ-line as expected, the rate of low level mtDNA mutations was assessed by using the **Illumina MiSeq FGx sequencer** to perform mtDNA Duplex Sequencing of the F1 heterozygous females and their F2 progeny. Our data revealed a significantly increased frequency of point mutations in the F1 mothers and the F2 w daughters with G-A/C-T mutations as the predominant point mutation. Bioinformatic analysis of the sequence surrounding these G-A/C-T mutations were consistent with the APOBEC-1 mutation signature described in cancer. This difference was not seeing in F1 or F2 flies with the catalytically dead APOBEC-1. The increased frequency of mutations in the F1 progenitor suggests that both MTD-driven Gal4 and UAS_APOBEC-1 are being expressed very early in the embryogenesis of the F1 female progenitor and, because mutations are detected in the fly head (somatic tissue), these mutations are being expanded clonally as the fly developed and reached adulthood. This observation was confirmed by performing a replicate study with 3 additional independently obtained F1 females and this time including the original progenitor (PO) to exclude the possibility that the observed mutations were inherited.

In a replication study, heterozygous F1 female virgins were crossed to w1118 males (the same genetic background of both stocks) instead of Cyo/Gla males and the rate of mtDNA mutations in the F1 mother and F2 daughters was performed as the initial study. By using w1118 males, this replication study eliminated phenotypic confounders that were associated with the Cyo/Gla background. In both the initial and the replication studies, the F2 progeny was also found to have an increased frequency of point mutations although in the initial study the frequency of point mutations in the F2 progeny is lower than in the replication. Since the F1 female flies are expected to express both transgenes in their germ-line, the mutations detected in F2 flies could be the result of both inheritance of mtDNA mutations from F1 mothers and induction of new mtDNA mutations by APOBEC during oogenesis. Interestingly, the clonality of each individual point mutation was found to increase from F1 to F2 females in both the initial and the replication studies while mutations in the F1 females did not exceed clonality of 0.1 many mutations detected in F2 females were above or double this clonality. This observation suggests that mtDNA point mutations detected by DS in this study are the result of a combination of mutagenesis in the early embryogenesis of the F1 female and in early cells in their germarium, the increased clonality

of the mutations detected in F2 females may be the result of clonal expansion of mutations that escape the mitochondrial bottleneck

The sequencing data was further analyzed to identify specific mutations inherited from the F1 female and past down in each generation. More detailed examination of specific G-A/C-T mutations inherited through 4 generations revealed different patterns of frequency and clonality. First, mutations that were shared from F1 to F4 progeny were considered to have originated during F1 embryogenesis. These shared mutations were observed to have increasing clonality from F1 to F3 which decreases in F4 progeny in most of the lineages. This pattern was similar between synonymous and nonsynonymous mutations. We are now using the characterized experimental system to study germline transmission of mtDNA mutations.

C.3. Publications using DURIP purchased equipment.

At the time of this final report, the Illumina MiSeq FGx sequencer has been used in one published manuscript and one manuscript under revision at *Nature Communications*

1) 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, 2018.

2) Andreazza S, Samstag CL, Sanchez-Martinez A, Fernandez-Vizarra E, Gomez-Duran A, Lee JJ, Tufi R, Hipp MJ, Schmidt EK, Nicholls TJ, Chinnery PF, Minczuk M, Pallanck LJ, Kennedy SR, Whitworth AJ. A new mitochondrial DNA mutator model shows that quality not quantity of mutations affects organismal fitness in *Drosophila*. Nature Communications [Under revision]

C.4. Role of DURIP purchased equipment in other DoD-funded research

Below is a list of current or previous DoD-funded research supported by the equipment purchased as part of W911NF-17-1-0158.

W81XWH-18-1-0339
PI: Scott Kennedy
Period of Performance: 8/1/18-7/31/20
Funding Agency: DoD/CDMRP Lung Cancer Research Program
Direct Funds: $400,000
Title: *Developing High-Accuracy Sequencing Methods for Use in Early Cancer Detection, Disease Stratification, and Chemotherapy Resistance with Cell-Free DNA*
Description: The presence of ctDNA in blood has been recognized for 40 years, but the recent emergence of ultra-sensitive methods has enhanced its status to a potentially transformative biomarker for the diagnosis and surveillance of cancer. The greatest challenge has been the identification of very low amounts of ctDNA within blood samples with variable amounts of cfDNA. This proposal focuses on developing and validation the required technical capabilities (e.g. sensitivity, multiplexing, scalability, and low DNA input) of SPLiT-Seq to enable the use of liquid biopsies for non-small cell lung cancer diagnostics and treatment evaluation.
Specific Aims:
Specific Aim 1: Fully develop SPLiT-Seq to make it amenable for ctDNA detection in non-small cell lung cancer.
Specific Aim 2: Demonstrate the utility of SPLiT-Seq for non-small cell lung cancer detection with ctDNA.

C.5. Role of DURIP purchased equipment in other non-DoD-funded research

Below is a list of current or previous non-DoD-funded research supported by the equipment purchased as part of W911NF-17-1-0158.

2017-DN-BX-0160
PI: Scott Kennedy
Period of Performance: 1/1/18-12/31/20
Funding Agency: National Institute of Justice/ Research and Development in Forensic Science for Criminal Justice Purposes
Direct Funds: $509,728
Title: *Developing SPLiT-Seq: A Massively Parallel Sequencing Method to Remove PCR Artifacts for Forensic STR Genotyping*
Description: The utility of modern forensic DNA analysis is greatly limited by the presence of DNA damage or multiple contributors. The goal of this proposal is develop and extend the high accuracy Duplex Sequencing methodology for use in forensic DNA analysis. While the underlying method being developed as part of this proposal is the same as is proposed in the DoD/CDMRP Lung Cancer Research Program proposal, the genomic loci being targeted are very different. Knowledge gained from this proposal will be helpful for the Lung Cancer Research Program work.
<u>Specific Aims:</u>
Specific Aim 1: Fully develop and optimize a SPLiT-Seq protocol for genotyping the CODIS20 loci and evaluate the assay's accuracy, precision, sensitivity, and specificity.
Specific Aim 2: Validate SPLiT-Seq for genotyping DNA mixtures.

Minnesota Ovarian Cancer Alliance Early Detection Award
PI: Barbara Norquist
Period of Performance: 09/1/2018 – 08/31/2019
Funding Agency: Minnesota Ovarian Cancer Alliance
Direct Funds: $60,000
Title: *Detection of Ovarian Cancer using Uterine Lavage and Crispr-Duplex Sequencing*
Description: The goal of this project is to detect tumor-derived *TP53* mutations from minimally invasively collected uterine lavage samples to develop an early detection test for ovarian cancer.
<u>Specific Aims:</u>
Specific Aim 1: Collection of neoplastic DNA through uterine lavage, pap smears, and tumor biopsy in women with advanced ovarian cancer
Specific Aim 2: Detection of *TP53* mutations in uterine lavage and Pap smear specimens from women with advanced ovarian cancer using Crispr-DS sequencing.

R01 CA181308
PI: Risques
Period of Performance: 1/4/2014 – 3/31/2019
Funding Agency: NIH/NCI
Direct Funding: $1,037,500
Mitochondrial Biomarkers for Early Detection of Cancer in Ulcerative Colitis
Description: The goal of this project is to develop a clinically feasible mitochondrial biomarker for cancer detection and prediction in ulcerative colitis according to the guidelines of the Early Detection Research Network. This biomarker will be based on immunohistochemistry of mitochondrial proteins and NanoString analysis of mitochondrial transcripts.
<u>Specific Aims:</u>
Specific Aim 1: Development of a mitochondrial biomarker based on IHC
Specific Aim 2: Development of a mitochondrial biomarker based on NanoString
Specific Aim 3: Validation of the mitochondrial biomarkers

Rivkin Foundation Research Grant
PI: Risques
Period of Performance: 01/4/2017-3/31/2018

Funding Agency: Rivkin Center
Direct Funding: $74,943
Title: *Characterization of TP53 mutations in BRCA carcinogenesis*
Description: This project will uses Duplex Sequencing to characterize with extremely high resolution the *TP53* mutational landscape that precedes the development of ovarian cancer in BRCA mutation carriers.
<u>Specific Aims:</u>
Specific Aim 1: Comparison of *TP53* mutations in histologically normal fallopian tube epithelium (fimbria) and blood of women BRCA+ and age-matched BRCA- controls
Specific Aim 2: Comparison of *TP53* mutations in normal fallopian tube epithelium (fimbria) and blood of women BRCA+ that developed HGSOC and women BRCA+ cancer-free

Mary Kay Foundation Cancer Research Grant 045-15
PI: Rosana Risques
Period of Performance: 7/1/2015-6/30/2017
Funding Agency: Mary Kay Foundation
Direct funding: $86,956
Title: *Early detection of ovarian cancer by ultra-sensitive sequencing of Pap smear DNA*
Description: The goal of this project is to develop a method for early detection of ovarian cancer using Duplex Sequencing to analyze p53 mutations in DNA extracted from Pap smears.
<u>Specific Aims:</u>
Specific Aim 1: To determine the sensitivity for the detection of ovarian tumor DNA in Pap smears using Duplex Sequencing
Specific Aim 2: To determine the specificity for the detection of ovarian tumor DNA in Pap smears using Duplex Sequencing
Specific Aim 3: To extend the Duplex Sequencing assay to more common somatic passenger mutations

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*
02/04/2019

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
04/15/2017-4/14/18

4. TITLE AND SUBTITLE
Massively Parallel Sequencing Platform for High Accuracy Variant Detection

5a. CONTRACT NUMBER
W911NF-17-1-0158

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kennedy, Scott R

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Washington
4333 Brooklyn Ave NE
Seattle WA 98195-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

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

10. SPONSOR/MONITOR'S ACRONYM(S)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Sequencing technologies have rapidly progressed in the past 15 years from “classic” Sanger sequencing to the current generation of massively parallel sequencing (MPS) platforms, which are capable of simultaneously sequencing hundreds of millions of small DNA fragments. Due to how these technologies work, it is possible to digitally tabulate the sequence of individual DNA molecules, thus allowing for the detection of sequences occurring in the minority of DNA molecules. While a significant improvement over original Sanger methods, MPS technology is not able to detect a variant in <1-5% of sequencing reads. This lower bound arises from a multiple sources, including enzymatic manipulations used in preparation

15. SUBJECT TERMS
Duplex Sequencing; High Accuracy Sequencing

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Scott R Kennedy
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU		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. DATE 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. F33315-86-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.

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.