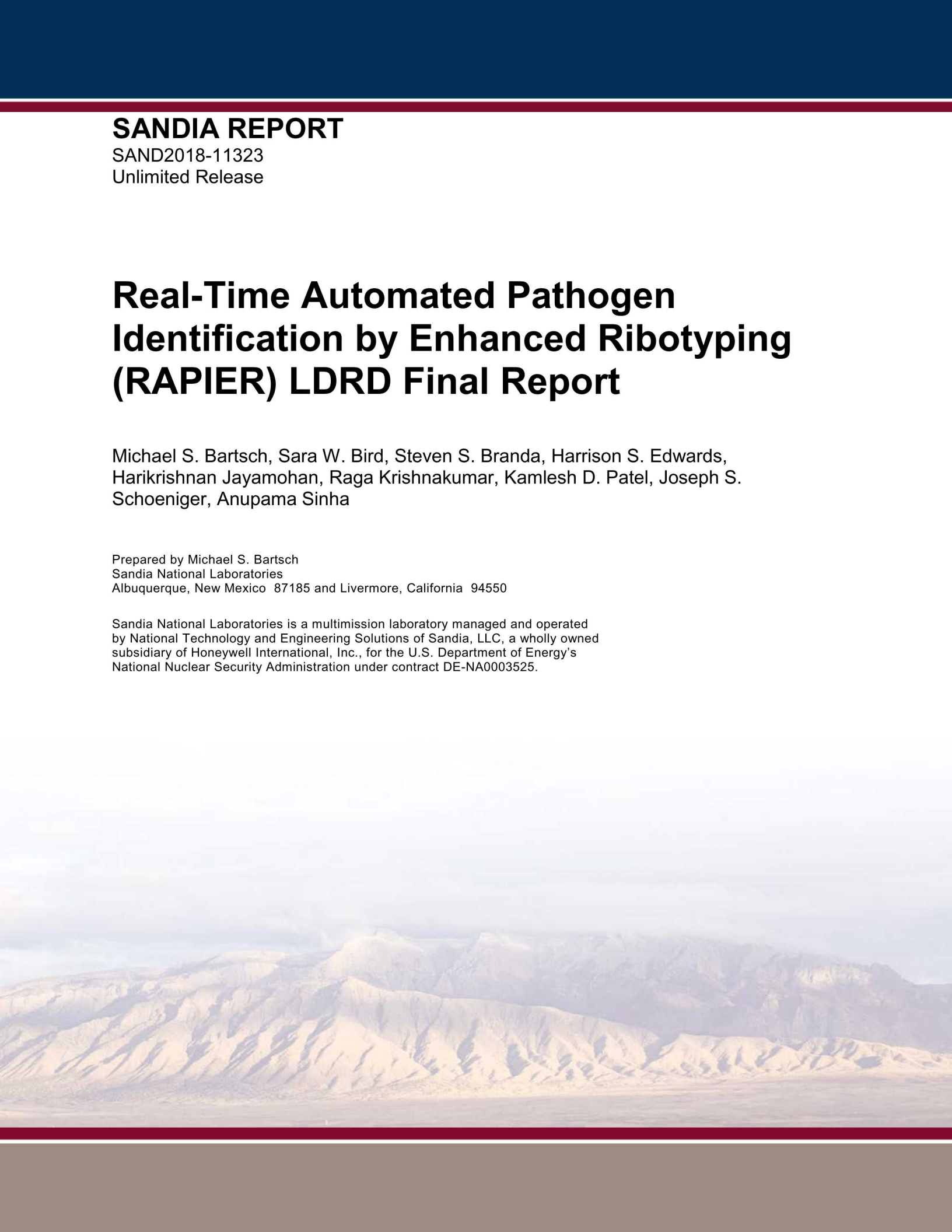
SANDIA REPORT
SAND2018-11323
Unlimited Release

Real-Time Automated Pathogen Identification by Enhanced Ribotyping (RAPIER) LDRD Final Report

Michael S. Bartsch, Sara W. Bird, Steven S. Branda, Harrison S. Edwards, Harikrishnan Jayamohan, Raga Krishnakumar, Kamlesh D. Patel, Joseph S. Schoeniger, Anupama Sinha

Prepared by Michael S. Bartsch
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.



Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods/

SAND2018-11323
Unlimited Release

Real-Time Automated Pathogen Identification by Enhanced Ribotyping (RAPIER) LDRD Final Report

Michael S. Bartsch[1]
Sara W. Bird[2,3]
Steven S. Branda[4]
Harrison S. Edwards[1,5]
Harikrishnan Jayamohan[1,6]
Raga Krishnakumar[7]
Kamlesh D. Patel1[1,8]
Joseph S. Schoeniger[7]
Anupama Sinha[7]

[1] Exploratory Systems Department
[2] Biotechnology and Bioengineering Department
[3] now uBiome, San Francisco, CA
[4] Bioenergy and Defense Technology Department
[5] now Inst. of Biomaterials and Biomedical Engineering, U. Toronto, Toronto, Canada
[6] now Roche Molecular Systems, Pleasanton, CA
[7] Systems Biology Department
[8] Purdue Partnerships Department

Department Name(s)
Sandia National Laboratories
P. O. Box 969
Livermore, California 94551-MS9292

Abstract

Funded through the IHNS/E&HS investment area for FY16-18, the RAPIER LDRD sought to evaluate the potential benefits and applicability of the new Oxford MinION nanopore sequencer to pathogen diagnostic applications in biodefense, biosurveillance, and global/public health. The project had four primary objectives: 1) to investigate the performance of the MinION sequencer while building facility with its operation, 2) to develop microfluidic library prep automation facilitating the use of the MinION in field-forward or point-of-care applications, 3) to leverage CRISPR/Cas9 technology to enable targeted identification of bacterial pathogens, and 4) to capitalize on the real-time data output capabilities of the MinION to enable rapid sequence-based

diagnostics. While the rapid evolution of the MinION sequencing technology during the course of the project posed a number of challenges and required a reassessment of initial project priorities, it also provided unique opportunities, notably culminating in our development of the RUBRIC real-time selective sequencing software.

ACKNOWLEDGMENTS

We would like to thank our project manager, James Carney and the LDRD investment area leads who supported this project, particularly Jennifer Gaudioso and Benjamin Wu. Thanks also to Matt Loose of the University of Nottingham for his early guidance in scaling the Read Until learning curve and to Julian Atienza, George Pimm, Richard Ronan, and Chris Wright of Oxford Nanopore Technologies for their advice and support in providing access to pre-release versions of the Read Until API. This work was supported by the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

TABLE OF CONTENTS

FIGURES

1. INTRODUCTION

The Oxford Nanopore Technologies (ONT) MinION sequencer is the first commercial DNA sequencing platform based on nanopore technology. In addition to its small size (Figure 1), the USB-powered MinION can be operated using a garden variety laptop PC, making it ideal for use in portable, field-forward, and point-of-care applications. Beyond its unique form factor, the MinION is noteworthy for its long read sequencing capability (up to 2.3 Mb demonstrated so far[1]), direct RNA sequencing capability[2], ability to perform methylation-aware basecalling, and real-time data output. Combined with its small operational and logistical footprint, the real-time data output of the MinION in particular suggests its potential in enabling sequence-based diagnostics. In contrast to conventional short read sequencers which operate in a batch mode where results are unavailable until sequencing has run to completion (e.g., 12-72 hours or more later), the real-time data output of the MinION means that diagnostically relevant information begins accumulating and can potentially be accessed and acted upon within minutes of starting a sequencing run.

Figure 1: Size comparison of the Illumina MiSeq next generation sequencer and the Oxford MinION nanopore sequencer (circled).

2. MINION SEQUENCER CHARACTERIZATION

At the start of the RAPIER project, one of the primary considerations in assessing whether the MinION nanopore sequencer could be effectively adapted to rapid, fieldable diagnostics was the accuracy of its basecalling algorithms. While current basecalling accuracy is in the 90-95% range (still significantly lower than the 99.9%+ accuracy typical of Illumina sequencing), MinION accuracy at the start of the RAPIER project was in the 60-65% range. As we familiarized ourselves with the operation of the sequencer and the requirements of its different library preparation protocols, we endeavored to perform a series of sequencing experiments aimed at understanding the nature, origin, and dependencies of the errors present in MinION basecalling. Of particular relevance to bacterial pathogen diagnostics was the question

whether the performance of the MinION was consistent across variations in bacterial genome composition, i.e., for organisms with high GC-content, low GC-content, and AT/GC-balanced genomes. Accordingly, we performed sequencing experiments using the MinION's (then gold-standard) 2D sequencing method, an approach that used a ligated hairpin adapter to enable back-to-back sequencing of both template and complement DNA strands, proving a means to reduce error rates by consensus basecalling. We tested DNA from three organisms, corresponding to high GC (*B. thailandensis*), mid GC (*E. coli*), and low GC (*C. difficile*) genomic content and assessed error mechanisms of the contemporary Metrichor basecaller at nucleotide, k-mer, and read scales. This work was published[3] in the open access journal Nature Scientific Reports in February, 2018.

3. MICROFLUIDIC LIBRARY PREP AUTOMATION

When the RAPIER project was conceived, one of the primary perceived benefits of the MinION nanopore sequencer was its small size and the ability to run it on USB power using only a laptop PC, making it ideally suited to field-forward, low-resource, or point-of-care applications. Despite these factors, its library preparation protocols—and particularly the higher fidelity 2D prep—nevertheless required significant laboratory infrastructure and conventional bench-prep instrumentation, significantly hampering the overall portability of a MinION-based workflow. Accordingly, we set out to develop a microfluidic library prep system suitable for automating this process for non-expert users operating outside the context of a well-provisioned laboratory environment.

Having previously developed integrated laboratory automation systems for library preparation[4,5], DNA forensics, and radionuclide purification centered around droplet-based electrowetting-actuated digital microfluidic platforms, we noted that this approach would likely be a poor match for fieldable MinION sequencing due to its complexity, expense, and lack of reliability. Instead, we sought to develop a simpler approach that nevertheless capitalized on some of the unique benefits of the digital droplet based approaches, notably the ability to operate across a wide range of volume scales and accomplish rapid mixing by aspirating/dispensing sample and reagents between a capillary tube and a droplet resting on a hydrophobic surface. The result was the ASPIRE system: Automated Sample Preparation by Indexed Rotary Exchange. The ASPIRE method used a very simple fluidic design consisting of a single transfer capillary coupled to a syringe pump as shown in Figure 1. Upstream of the open capillary end were a magnetic trap for capturing the beads used in the MinION library prep process and an inline heater providing sample incubation at elevated temperature for key protocol steps. Reagent droplets were pre-positioned on the free surface of a hydrophobically coated disc substrate resting on a turntable. During library preparation, the turntable would rotate, bringing the droplets into contact with the capillary tip in the correct order to implement the various steps of the library preparation process. At the conclusion of the semi-automated prep, the prepared library was dispensed from the ASPIRE transfer capillary for introduction to the MinION flowcell. This work is detailed in an upcoming peer-reviewed publication to be submitted to Lab on a Chip[6].

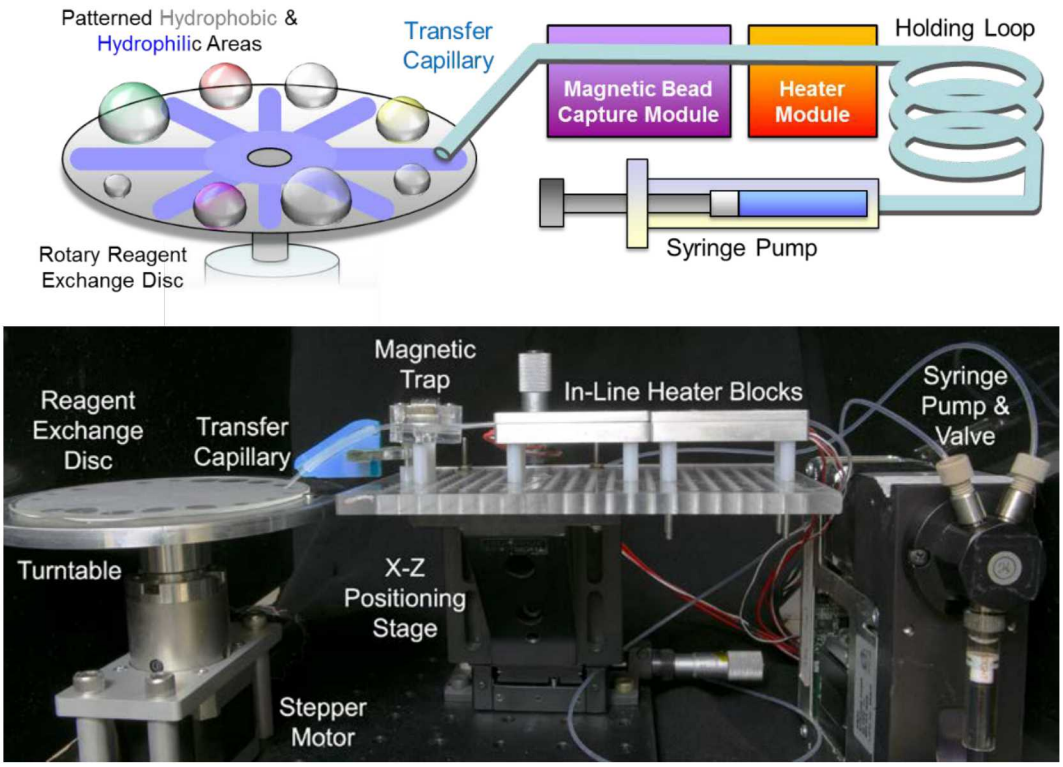


Figure 2: Conceptual schematic and photo of the prototype ASPIRE library prep automation system.

Unfortunately, the microfluidic automation element of the RAPIER project was both hampered by personnel problems and substantially overtaken by events in the rapidly evolving MinION sequencing space. Significantly, ONT's release of the rapid library prep kit for the MinION, consisting of mixing DNA sample and prep reagents together and incubating for 10 minutes at elevated temperature, removed much of the impetus for developing a highly engineered automation approach, as this prep would almost certainly be preferred for any field-forward applications. The more accurate but time-consuming (2+ hour) 2D library prep protocol, which was the original target of ASPIRE automation, was also discontinued. Lastly, while ASPIRE was still under development, ONT announced its own digital microfluidic library prep automation system, VolTRAX, which was designed to integrate directly with the MinION flowcell, further undercutting the value of any RAPIER/ASPIRE contribution on the library prep automation front. Accordingly, we shifted the focus of our efforts from library prep automation to more promising efforts like real-time selective sequencing, as described below.

4. TARGETING PATHOGENS WITH MOLECULAR BIOLOGY

A significant challenge in applying sequencing to pathogen diagnostics is the "needle in a haystack" problem of obtaining sufficient coverage of rare pathogen sequence against a background of predominantly non-informative host DNA. While a variety of amplification and hybridization-capture-based techniques have traditionally been used for target enrichment, the recent advent of CRISPR/Cas9 technology presented a novel

option for targeting bacterial DNA. While much attention has been given to the gene-editing capabilities of the Cas9 system, it also provides a means to make highly specific cuts in double stranded DNA at a location defined by the sequence of custom ~20 nt long guide RNA. As part of the RAPIER project, we sought to determine whether using CRISPR/Cas9-targeted cuts in conjunction with nanopore sequencing was possible, and if so, whether it could further enable sequence-based diagnostics. Toward this end, we developed a strategy to use Cas9 to target and excise bacterial ribosomal DNA (rDNA). rDNA represents a promising target for use in nanopore pathogen diagnostics because 1) all bacteria have rDNA and 2) the 16S-23S rDNA locus is ~5 kb in length and can readily be sequenced in its entirety by the MinION, and 3) the rDNA locus contains both highly conserved regions (suitable for Cas9 targeting) and highly variable regions (enabling bacterial discrimination & typing).

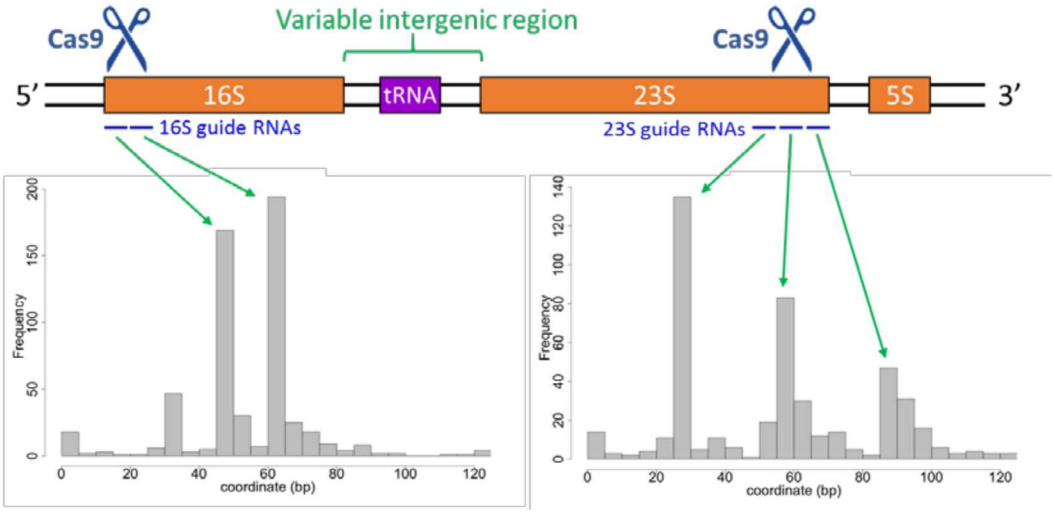


Figure 3: Validation of CRISPR/Cas9 guide RNA designs targeting the outer portions of 16S and 23S rDNA loci.

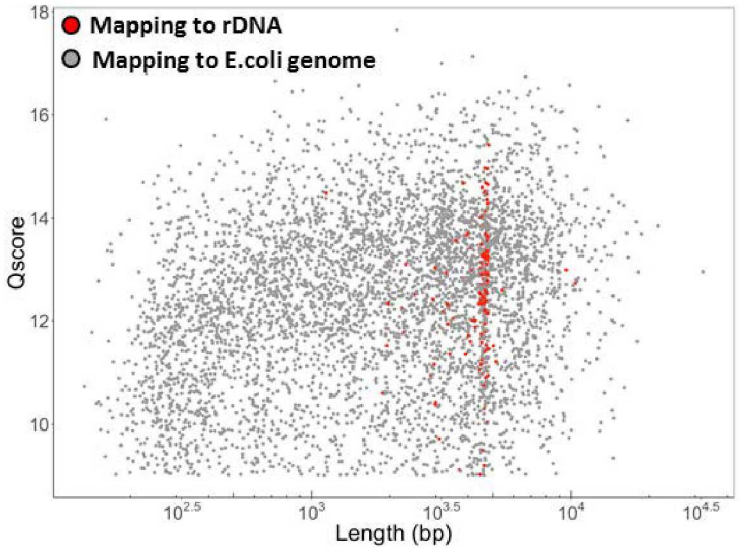


Figure 4: Quality Score vs. Length scatterplot showing the effect of CRISPR/Cas9 excision of the rDNA locus on the size distribution of rDNA-mapping reads produced by the MinION nanopore sequencer.

Initial RAPIER Cas9 experiments focused on guide RNA design and optimization to effectively target the outer portions of the 16S and 23S rDNA loci as shown in Figure 1 with a goal of producing double-stranded cuts on both sides to fully excise the rDNA sequence. Figure 2 shows the result of Cas9-cutting with optimized guide RNA and the distribution of resulting DNA fragments sequenced with the MinION.

While this RAPIER task showed that Cas9-cut DNA could in fact be sequenced by the MinION, the application of the CRISPR/Cas9 system itself proved to be significantly less straightforward than anticipated. In particular, the readily available *S. pyogenes* Cas9 variant exhibited much lower cutting activity than anticipated, apparently as a consequence of the enzyme cutting DNA and then remaining bound to one of the resulting fragments and unavailable to make subsequent cuts. Due to this unexpectedly low activity, effective Cas9 digests required very high concentrations of the enzyme. Moreover, experiments performed with mixed human/*E. coli* samples revealed a tendency of Cas9 cuts to be significantly less specific than expected, a result independently observed by others as well. In toto, these complications combined with project team member transitions, a lack of funds to support an in-depth Cas9 optimization study well beyond the original scope of the task, and the more promising prospects of real-time selective sequencing caused us to de-emphasize the Cas9 targeting task after the first half of the project. Additional Cas9 experiments were performed in conjunction with real-time selective sequencing, however, where the ability to select fragments on the basis of known cut locations is highly beneficial. In all such experiments, however, the limited efficacy realized for our Cas9 digests was fundamentally limiting.

5. REAL-TIME SELECTIVE SEQUENCING

At the start of the RAPIER project, there was an expectation that the poor basecalling accuracy of the sequencer represented a significant impediment to its use in pathogen detection (and many other applications). Nevertheless, it was anticipated that the ability to access and analyze MinION data in real-time, combined with the long read capabilities of the sequencer could potentially compensate for this deficiency. In particular, the original RAPIER project plan identified a task specifically aimed at circumventing the problematic and error-prone MinION basecalling process by applying pattern-matching algorithms to the raw MinION pore current data as a basis for identifying pathogen targets of interest.

In September, 2016, Matthew Loose and his team at the University of Nottingham published[7] an article in Nature Methods that simultaneously validated this aspect of the RAPIER project plan while, in combination with ongoing increases in MinION basecalling accuracy, rendering it entirely obsolete. The Loose publication revealed a new MinION functionality dubbed “Read Until” that took advantage of the real-time data output and individually addressable pore architecture of the MinION to enable *real-time selective sequencing*. Read Until provided the means to preview the data associated with an individual strand of DNA as it transited an individual nanopore, apply some analysis to that data to determine whether that particular molecule was of

interest, and then reverse the pore polarity to physically eject the DNA from the pore if it was deemed to be uninformative. Target DNA determined to be informative would be allowed to sequence to completion as usual. Performed across hundreds of pores and thousands to millions of DNA fragments during the course of a typical sequencing run, the net effect of this kind of selection is to enrich for the desired target sequence in the final pool of sequence data while depleting non-target sequence, providing a purely software-based target enrichment method requiring none of the target-specific front-end molecular biology sample/library preparation typically required for capture or amplification-based methods.

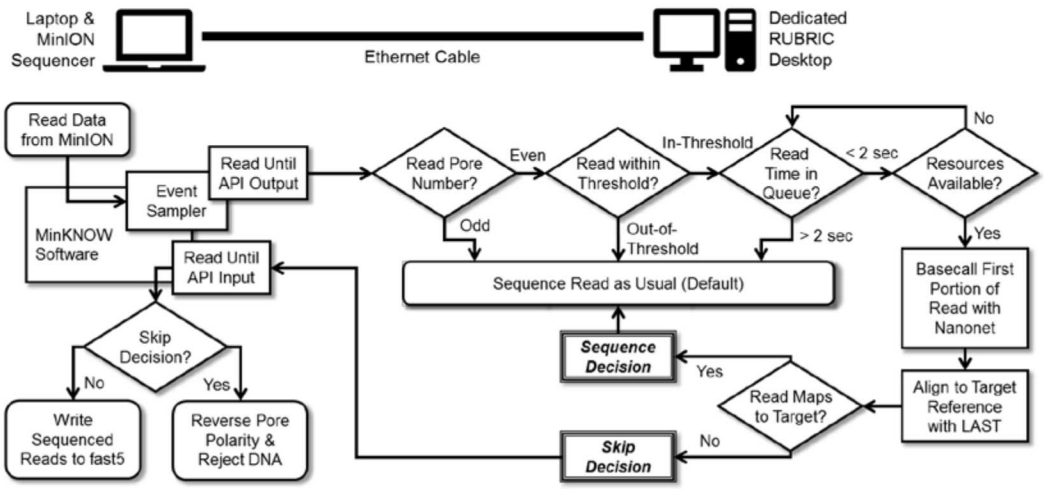


Figure 5: RUBRIC selective sequencing architecture and workflow illustrating the division of computational labor between ethernet linked PCs and the read pre-screening steps used to maximize the timeliness and relevance of reads admitted to the RUBRIC decision process.

Much like our original concept for avoiding inaccurate MinION basecalling, the Loose implementation of Read Until utilized a Dynamic Time Warping-based (DTW) pattern matching algorithm to enable real-time selection decisions. This method proved effective but required significant computing power (i.e., a 22-core cluster system) and lacked the scalability to be used with larger target references. More significantly, the DTW-based method required target patterns for matching to be synthesized in advance from genomic reference sequences, significantly constraining the flexibility and adaptability of the approach.

Recognizing the potential benefits and limitations of the Read Until method for rapid diagnostics, we recast our own real-time analysis task and set about adapting the open source code of the original method to a more generally applicable and less computationally intensive framework. The result of this effort was RUBRIC, Read Until with Basecall and Reference-Informed Critera, a version of Read Until that, as the name implies, utilized real-time basecalling and sequence alignment in ACGT-space to enable selection (Figure 3). Despite frequent setbacks due to rapidly-changing ONT software, hardware, and library preparation kits, the RUBRIC effort was ultimately successful, garnering significant attention for the RAPIER team within the nanopore sequencing community including invited talks at the annual London

Calling and Nanopore Community Meetings as well as presentations at Nanopore Day at UCSF.

Our RUBRIC work is detailed in a forthcoming manuscript to be published online via the BioRxiv preprint server ahead of submission for peer reviewed publication through Nature Methods[8]. In addition to demonstrating the operation of the method, characterizing its performance in detail, and showing its application to sample use cases including selecting 1% *E. coli* DNA from a background of 99% human DNA and selecting *E. coli* ribosomal DNA from genomic DNA, the publication also proposes a model framework for predicting the potential limits of real-time selective sequencing performance. At the time of this writing, we expect that our article will be only the third peer-reviewed publication on this topic—and the first describing a non-DTW implementation—and we believe that the insights offered by our modeling work in particular will provide significant benefit to the nanopore community at large, potentially making this a highly cited, high-impact publication.

6. PROGRAMMATIC

The RAPIER LDRD was funded at a nominal allocation of $530k/year but received plus-up funding of $135k, $40k, and $87k in FY16, FY17, and FY18, respectively.

7. TALKS & MEETINGS

- M.S. Bartsch, “Real-time selective sequencing with RUBRIC (Read Until with Basecall- and Reference-Informed Criteria)”, invited talk, London Calling 2018, London, 25 May, 2018.
- M.S. Bartsch, “Read Until with Basecall- and Reference-Informed Criteria (RUBRIC),” invited talk, Nanopore Day at UCSF, 13 March, 2018.
- R. Krishnakumar, “Characterizing the performance of the MinION for real-time detection,” invited talk, Nanopore Day at UCSF, 13 March, 2018.
- M.S. Bartsch, “The evolution of MinION selective sequencing: Read Until with Basecall- and Reference-Informed Criteria (RUBRIC),” Nanopore Community Mtg., NYC, 30 Nov., 2017. (poster + lightning talk)
- R. Krishnakumar, “Selective Long-Read Nanopore Sequencing for Real-time Point-of-need Pathogen Identification”, DTRA CBD S&T Conference, Long Beach, 28 Nov., 2017.
- R. Krishnakumar, “Real-time diagnostics using nanopore sequencing,” Molecular Tricon, San Francisco, CA, 24 Feb., 2017.
- H. Jayamohan, “Nanopore sequencing for real-time pathogen identification,” Molecular Tricon, San Francisco, CA, 24 Feb., 2017.
- R. Krishnakumar, “Real-Time Auttomated Pathogen Identification by Enhanced Ribiotyping (RAPIER)”, Oxford Nanopore Community Mtg., NYC, 1-2 Dec., 2016. (poster)

REFERENCES

1. Payne, A., Holmes, N., Rakyan, V. & Loose, M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*, doi:10.1101/312256 (2018).
2. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201-206, doi:10.1038/nmeth.4577 (2018).
3. Krishnakumar, R. *et al.* Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Sci Rep* **8**, 13, doi:10.1038/s41598-018-21484-w (2018).
4. Kim, H. *et al.* Automated Digital Microfluidic Sample Preparation for Next-Generation DNA Sequencing. *Jala* **16**, 405-414, doi:10.1016/j.jala.2011.07.001 (2011).
5. Kim, H. *et al.* A Microfluidic DNA Library Preparation Platform for Next-Generation Sequencing. *Plos One* **8**, doi:10.1371/journal.pone.0068988 (2013).
6. Jayamohan, H. *et al.* Preparing nanopore sequencing libraries with ASPIRE: Automated sample preparation by indexed rotary exchage. *Lab on a Chip* **in preparation** (2018).
7. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13**, 751-754, doi:10.1038/nmeth.3930 (2016).
8. Edwards, H. S., Krishnakumar, R., Sinha, A., Patel, K. D. & Bartsch, M. S. Real-time selective sequencing with RUBRIC: Read Until with Basecall- and Reference-Informed Criteria. *Nat. Methods* **submission pending** (2018).

DISTRIBUTION

1	MS0899	Technical Library	9536 (electronic copy)

For LDRD reports, add:

1	MS0359	D. Chavez, LDRD Office	1911

For CRADA reports add:

1	MS0115	OFA/NFE Agreements	10012

For Patent Caution reports, add:

1	MS0161	Legal Technology Transfer Center	11500


Sandia National Laboratories