Detecting genetic engineering via portable nanopore sequencing and reachback data analysis tools

Maria Arévalo[1], Pierce Roth[1,2], Samir Deshpande[1], Jessica Hill[1,2], Alvin Liem[1,2], Mark Karavis[1], Adina Doyle[1,3], Jackie Harris[1], Sarah Katoski[1], and R. Cory Bernhards[1]

[1]U.S. Army Combat Capabilities Development Command Chemical Biological Center, Aberdeen Proving Ground, MD; [2]DCS Corp., Belcamp, MD; [3]Excet, Inc., Springfield, VA

Abstract

Recent advances in nanopore sequencing technology have enabled rapid, low-cost, and portable DNA/RNA sequencing analysis in any environment. Untargeted nanopore sequencing allows for the identification of any biological threat, including those that are genetically modified to evade existing detection capabilities. We have developed untargeted nanopore sequencing systems and rapid protocols that can be used by a variety of operators in the field, including those without laboratory training. The goal is to quickly identify if a biological threat is present, then the data can be sent to reachback facilities for further analysis. This reachback analysis will include genome assembly and the utilization of software tools to detect the presence of genetic modifications. The software tools developed under the IARPA Finding Engineering-Linked Indicators (FELIX) program are currently being evaluated against nanopore sequencing datasets. Once downselected, further development of these tools will be needed. These tools utilize artificial intelligence and machine learning (AI/ML) so re-training of the algorithms will be important as code is modified and more data becomes available. Upgrades will also be important as new genetic engineering methods emerge. It is important to note that these software currently require substantial computational infrastructure to run and will so for the foreseeable future. Some also require trained bioinformaticians with assistance from subject matter expert biologists in order to make proper assessments. To support the biodefense community, DEVCOM CBC is currently establishing a reachback capability for this type of analysis, which will be critical as genetic engineering becomes more and more prevalent. Being able to quickly identify genetically modified threats will serve to protect the warfighter as well as the general population.

Figure 1. Nanopore Sequencing in the Field.

Current Capabilities: A Case Study

Our untargeted nanopore sequencing systems and protocols are developed for used by a variety of operators in the field (**Fig. 1**). Upon identification of a biological threat, the data can be forwarded to reachback facilities for further analysis. Developing this reachback capability will be key in propelling us toward the future to detect, deter, and defeat engineered threats developed by an adversary. The current capabilities are highlighted in the context of a case study.

Case Study Background: In 2021, DEVCOM CBC was asked to sequence and analyze an unexpected contaminant found in a biomanufacturing facility.

Methods: The sample was received and cultured. DNA from a single colony liquid culture was sequenced for 20 h. The organism was identified and sequences were assembled for further analysis. Our pipeline is summarized in **Fig. 2**.

Results*: Bacillus oceanisediminis* was identified as the contaminating organism (**Table 1**). Assemblies yielded expected chromosomal and plasmid DNA, plus one smaller and unexpected plasmid (**Table 1**). When blasted (blastx), the smaller plasmid contained regions with homology to: 1) a hypothetical protein containing a helix-turn-helix DNA binding domain 2) CamS family sex pheromone, and 3) tyrosine recombinase/integrase. The hits came from the Bacillales order and top hits were to *Cytobacillus firmus*, a close relative to *B. oceanisediminis.* It was ultimately determined that the strain was not genetically engineered.

Testing New Systems and Pipelines

Revolutionary advances in genome editing and synthetic biology approaches have made bioengineering more accessible. Thus, it is important to develop tools that will detect misuse of these capabilities. The IARPA FELIX program has developed software to detect and characterize signatures of engineering in any organism, including in operationally challenging samples. Four computational platforms have been transitioned to DEVCOM CBC and are being evaluated against nanopore sequencing datasets. The tools and their attributes are summarized in **Table 2**. Genetically engineered bacterial, bacteriophage, and fungal T&E strains have also been transitioned to us for testing. While genetic engineering was ultimately not suspected in the case study above, these tools may help streamline our screening process and provide further insights.

Developing State of the Art Infrastructure & Reachback Capability

As shown in **Table 2**, current, state-of-the-art tools to detect genetic engineering require an investment in computational power. DEVCOM CBC has acquired a server with the following specifications: 2 TB RAM, 48 cores (2 x 24 core CPUs), 2x RTX2080Ti GPUs, 1x RTX2080 GPU. It has a 1 TB SSD, but is also connected to a 40 TB main storage system. We are in the process of installing and testing ENDAR software from Gingko Bioworks, as this tool was developed to handle both short reads and long nanopore reads from the outset. The goal is to evaluate all four platforms for integration into our analyses and reachback capabilities.

DEVCOM CBC @ DTRA CBD S&T Conference
Scan the QR Code to view all of CBC's
2022 DTRA CBD S&T Conference materials
https://cbc.devcom.army.mil/cbdst-conference/

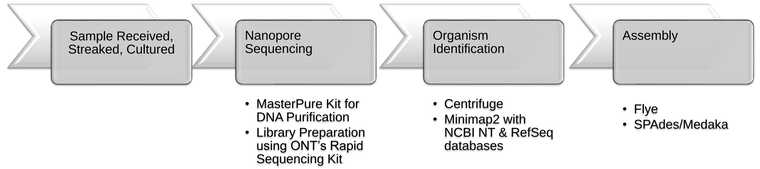

Figure 2. Analysis Pipeline for Unknown Bacterial Contaminant.

Table 1. Aggregated Analyses of Cultured Contaminant.

Tool	Liquid Culture
Identification - Centrifuge	*B. oceanisediminis* 2691
Identification - minimap2 with NT database	• *B. oceanisediminis* 2691, complete genome (expected size 5,461,741 b) • *B. oceanisediminis* strain 2691 plasmid pBO1, complete sequence (expected size 386,866 b)
Identification - minimap2 with RefSeq database	Same as minimap2 with NT database
Assembly - Flye	Contig_2: 5,289,748 b, circular Contig_1: 342,048 b, circular Contig_3: 13,250 b, circular
Assembly - SPAdes/Medaka	Chromosome: 5,289,748 b Plasmid: 342,048 b

Table 2. Software Tools Developed & Transitioned Under IARPA FELIX Program.

Performer (Tool)	Description	Computational Requirements	Input	Characteristics
Broad Institute	• Automated software for assembly, annotation, & analysis to detect non-natural or synthetic features	• Runs containerized on Linux systems, & supports the SGE (Sun Grid Engine) for scheduling • Development on standalone systems: 2.6 GHz Core i9-7980 (128 GB RAM) or 2.2 GHz Xeon E5-2650 (220 GB RAM) utilizing up to 36 concurrent processes • Installation requires Singularity 3 & Java 11	• Raw NGS sequences • Short & long reads, but preferably both for hybrid assembly	• Automated genome assembly and annotation; determination of taxonomic representation • Comprehensive heuristics for signature detection based on homology to sequences found in NCBI databases • Deep-learning based analysis to classify uncertain sequences • Findings scored and summarized into annotated list of sequences of interest
Ginkgo Bioworks (ENDAR)	• Modular pipeline • Detects indicators of engineering. • Extensive in-house organism engineering dataset & large collection of simulated and public data sources used to train models	• Meant to run on a single server with a large amount of RAM and a fast disk • ENDAR on Amazon's Elastic Compute Cloud (EC2) r5.8xlarge: Intel Xeon Platinum 8000 series processor, up to 3.1 GHz, with 256 GB RAM	• Output files from NGS • Short read: Illumina paired-end sequencing • Long read: Oxford Nanopore and PacBio • >200 b reads	• DNA assembly into contiguous sequences, identification of genes and genomic parts, matches to reference genomes • Expansive reference database containing data from in-house engineering, simulated *in silico* sequences, and public sources • Focus on synbio parts and ontologies • Synthesizability module - potential to generate sequences using modern methods • Identifies areas that should be examined by an analyst
Noblis (IMAGED)	• Modular pipeline • Bioinformatics tools for genetic and proteomic analysis and pan-genome mapping • Identification of structural changes to the genome	• The majority of the IMAGED platform runs on a standard Linux server with 3 TB RAM and 224 processors	• NGS data; short reads	• Proteomic analysis and pan-genome mapping for identifying structural changes of the genome, ensembled through a neural network for detecting engineering • Incorporation of algorithms based on evolutionary biology • *In silico* proteomic analysis of sequence data • Training database built from bacteria and plants engineered in-house, simulated sequence data, and multiple public sources • Generation of whole genome map for structural analysis and comparison
Raytheon BBN (GUARDIAN)	• Pipeline leveraging anomaly analysis software developed for malware detection • Data ensembled through a neural network for detecting engineering	• Complete System: AMD Opteron(tm) Processor 6380 (64 Cores) Intel(R) Xeon(R) Gold 6136 CPU @ 3.00 GHz (48 Cores) Intel(R) Xeon(R) Gold 5217 CPU @ 3.00 GHz (32 Cores), 17.9 TB Disk Space	• Paired-end short-read Illumina sequencing data (a pair of FASTQ files) and optionally long read nanopore sequencing data (also FASTQ).	• Established algorithm for rare anomaly detection • Six modules: five detect sequence inserts and one detects sequence deletions • Modeling engineering through natural language processing • Expansive set of fungal engineering targets generated from in-house engineering & multiple public data sources

Acknowledgements : This research was funded by the Defense Threat Reduction Agency Joint Science and Technology Office (DTRA JSTO) and the Joint Program Executive Office for Chemical, Biological, Radiological and Nuclear Defense (JPEO-CBRND). The views expressed in this poster are those of the authors and do not necessarily reflect the official policy or position of the Department of Defense or the U.S. Government.