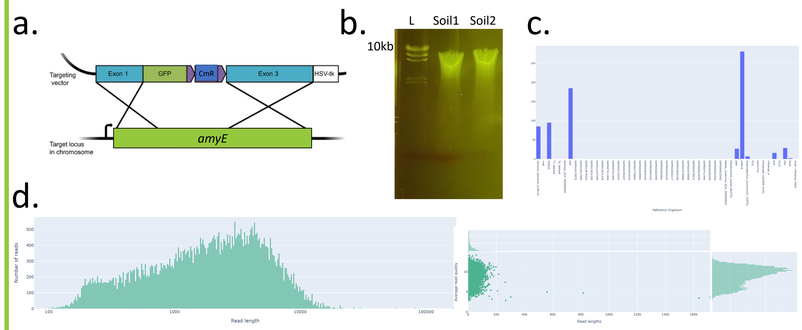
Field-Deployable Pan-genome Analysis Pipeline for Characterization of Genetic Variation and Identification of Novel Sequences

Bradley W. Abramson[1], Jared Haas[1], Lauren Brinkac[1], Granger Sutton[2], Katharine Jennings[1]
[1]Noblis Reston, VA USA. [2]JCVI La Jolla, CA USA

Abstract

Next Generation Sequencing (NGS) devices provide a means of detecting pathogenic or novel organisms via genetic screening in the field with a laptop but are often underutilized because the bioinformatic analysis required to interpret the sequencing data is complex. To make sequencing and analysis more applicable to users that lack necessary access to high performance computing systems, or simply lack the bioinformatics expertise, we have developed a simplified *de novo* genome assembly and pan genomic pipeline to characterize newly sequenced organisms that can run on a laptop. The Field-Deployable Pangenome Analysis Pipeline is extensible to a range of biological organisms (bacteria, plants, fungi, and viruses), and can rapidly identify and separate organisms at the contig level in a *de novo* metagenomic sample. Its innovative k-mer based pangenome graph (PGG) algorithm allows for rapid querying of newly sequenced genomes and accurate alignment to the PGG to determine diverse types of genome modifications (insertions, deletions, single nucleotide polymorphisms [SNPs], duplications, transversions, and rearrangements). Here we inserted a gene cassette into *Bacillus subtilis*, spiked a soil sample with the bioengineered strain, performed metagenomic sequencing with our field deployable lab and sequencer, and used our bioinformatic analysis pipeline on a laptop to determine the location of the insertion. This represents a streamlined method for determining genetic anomalies compared to all wildtype diversity of the species and may prove instrumental in determining novel genetic elements arising in evolving biothreats such as pathogenic clinical samples (i.e. MRSA strains) or bioengineered organisms. The deployment of this analytical capability to the field represents a step forward in early warning.

Generating a Known Bioengineered Validation Dataset

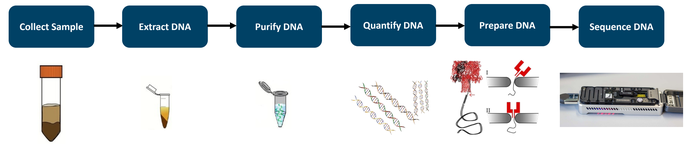


In the lab, we bioengineered a *Bacillus subtilis* 168 genome by stably inserting an antimicrobial resistance gene (Chloramphenicol resistance; CmR) and GFP into the alpha-amylase (*amyE*) gene via double-crossover homologous recombination (Figure a). This genome was sequenced previously to ensure the insert was stably integrated. Additionally, the bioengineered *B. subtilis* was mixed with soil and DNA was extracted using the portable sequencing kit (Figure 2b). 70ng/ul isolated DNA was prepared for sequencing using Oxford Nanopore Technologies Field Sequencing Kit (SQK-LRK001) and sequenced on a Flongle flow cell connected to a laptop. Fast base calling was performed in real-time via Minknow (v22.03.6) and guppy (v6.0.7). Total DNA extraction time was 25 minutes, 10 minutes library preparation, and ~8 hours of sequencing. Resultant fastq files were used for metagenomic analysis (Figure c). Read length and quality are shown in Figure d.

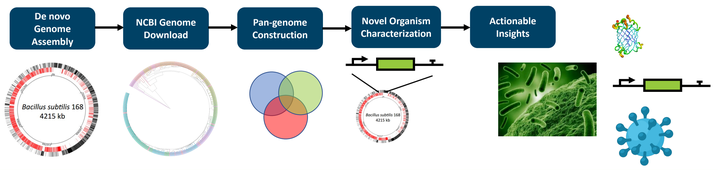
All-in-One Sequencing and Analysis

Figure 1: A field-deployable kit with necessary reagents to perform sequencing in austere environments was created. Testing was performed on a soil samples spiked with a known genetically engineered *Bacillus subtilis* sample. Sequencing and bioinformatic analysis was performed.

Portable Sampling and Processing Equipment Package



Pan-genome Genetic Anomaly Detection Pipeline



Automated Pan-genome Construction

A simple Linux command kicks off PGG construction of a species of interest as ./PGGrun_v1.sh construct “Bacillus subtilis”. The single command downloads complete genomes for that species and uses MASH to determine average nucleotide identity (ANI) for all genomes and discards potentially mislabeled genomes. The resultant genomes are then used for tree building and pan-genome graph construction by a k-mer anchored PanOCT algorithm.

Genomes:	194
Genomes <0.01%:	110
PGG cluster:	76530
Identical clusters:	42.94%
Identical edges:	48.13%
Single copy cluster:	7.55%
Unique cluster:	6.72%
Unique alleles:	1.22%
Frameshift Clusters:	0.78%

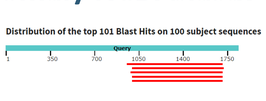
Genetic Anomaly Detection and Insights

A single 1841 bp region is detected as an insertion

Output includes tabular data of location of genetic anomaly and information describing type of insertion, deletion, tandem-duplication or rearrangement. A resulting fasta file of the anomaly sequence is generated for downstream analysis

Anomaly NCBI blastn

Anomaly NCBI blastx



Actionable Insights with Rapid Database Analysis

A curated database of protein sequences is easily customizable and provides rapid descriptions of the anomalous sequence(s) if there are general targets of interest. Automatically produced table of functional significance. Future work will include more robust gene predictions and visual tools.

Target Hits	Subjects	Database	Anomaly Detection
1	33173	Biosynthetic_mibig_prot_seqs_2.0	0.0%
0	8	Control_Elements_for_Engineering.pro	0.0%
481	748	NobleCombi_protein_FPbase_protein_seqs	47.7%
43	6320	NCBI_AMR_reference_database	0.7%
0	6093	Plasmid_UniVec	0.0%
0	3137	Plasmid_UniVec_Core	0.0%
0	2108	T3DB_toxin_targets	0.0%
22	2979	CARD_ARO_database	0.7%
0	19	Toxin_CARD_protein_fasta_protein_knockout_model	0.0%
0	13	Toxin_CARD_protein_fasta_protein_overexpression_model	0.0%
0	163	Toxin_CARD_protein_fasta_protein_variant_model	0.0%
0	48	Toxin_TADB_type_I_pro_T s	0.0%
0	47	Toxin_TADB_type_I_pro_T_exp s	0.0%
0	6192	Toxin_TADB_type_II_pro_AT s	0.0%
0	105	Toxin_TADB_type_II_pro_AT_exp s	0.0%
0	5	Toxin_TADB_type_II_pro_RE s	0.0%
0	6192	Toxin_TADB_type_II_pro_T s	0.0%
0	105	Toxin_TADB_type_II_pro_T_exp s	0.0%
0	7	Toxin_TADB_type_III_pro_T s	0.0%
0	4	Toxin_TADB_type_IV_pro_AT s	0.0%
0	4	Toxin_TADB_type_IV_pro_T s	0.0%
0	1	Toxin_TADB_type_V_pro_AT	0.0%
0	1	Toxin_TADB_type_V_pro_T s	0.0%
0	1	Toxin_TADB_type_VI_pro_AT s	0.0%
0	1	Toxin_TADB_type_VI_pro_T s	0.0%
0	7216	Toxin_uniprot_KW-0800	0.0%
0	28924	Virulence_VFDB_Full	0.0%

Acknowledgements/References

This research is based upon work supported [in part] by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under Finding Engineering Linked Indicators (FELIX) program contract #N6600118C-4506.

- Sutton G, Fogel GB, Abramson B et al. A pan-genome method to determine core regions of the Bacillus subtilis and Escherichia coli genomes [version 2; peer review: 2 approved]. F1000Research 2021, 10:286 (https://doi.org/10.12688/f1000research.51873.2)
- Derrick E. Fouts, Lauren Brinkac, Erin Beck, Jason Inman, Granger Sutton, *PanOCT*: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species, *Nucleic Acids Research*, Volume 40, Issue 22, 1 December 2012, Page e172, https://doi.org/10.1093/nar/gks757
- GGRaSP: a R-package for selecting representative genomes using Gaussian mixture models. Thomas H Clarke, Lauren M Brinkac, Granger Sutton, and Derrick E Fouts. Bioinformatics, bty300, https://doi.org/10.1093/bioinformatics/bty300