precisionFDA

CDRH ID-NGS Biothreat Challenge

August 3 - October 18

To encourage the development and improvement of Infectious disease next-generation sequencing (ID-NGS) analytical methods, precisionFDA recently launched the ID-NGS Biothreat Challenge! Be a part of the Challenge and test your algorithms on blinded mock-clinical and in silico metagenomics samples. Visit precision.fda.gov/challenges/3 to learn more and join today!

Diagnostic Genomes & The PrecisionFDA Biothreat Challenge

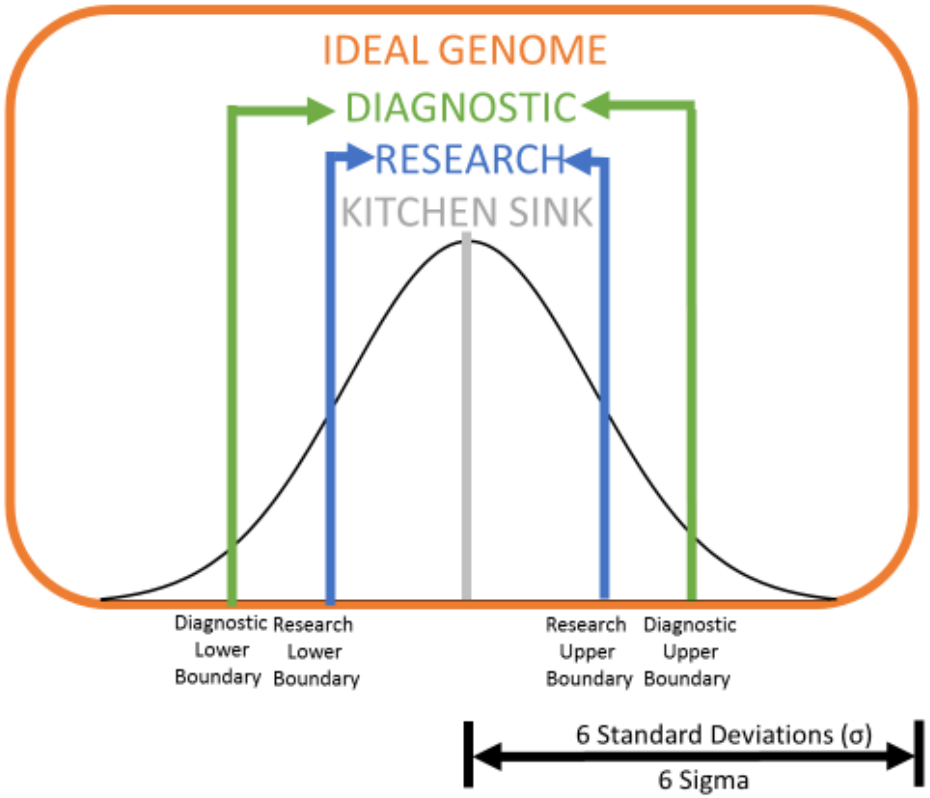
Heike Sichtig, PhD
SME/Principal Investigator
FDA Genomics Working Group Chair

Reference Genomes For Diagnostic Use

- ✓ Support for *in silico* validation

Proposed Quality Metrics

- ❑ Identified by orthogonal reference method
- ❑ Sequenced and de-novo assembled using 2 sequencing methodologies
- ❑ High depth of sequencing coverage
- ❑ Minimum of 20X over 95 percent of the assembled and polished core genome
- ❑ Taxonomy-specific ANI thresholds that are sufficient for identification
- ❑ Placed within a pre-established phylogenetic tree
- ❑ Sample specific metadata, raw reads, assemblies, annotation and details of the bioinformatics pipeline are available



FDA-ARGOS DIAGNOSTIC DATABASE EFFORT

Latest Accomplishments and Next Steps

U.S. FOOD & DRUG ADMINISTRATION

UNIVERSITY of MARYLAND SCHOOL OF MEDICINE INSTITUTE FOR GENOME SCIENCES

NCBI

✓ FDA established a government-academic-clinical partnership with 35+ collaborators

American Type Culture Collection/ BEI	Bernard Nocht Institute for Tropical Medicine, Germany	Biodefense and Emerging Infections Research Repository	British Columbia Centre for Disease Control (BCCDC)	Children's National Medical Center	Defense Threat Reduction Agency (DTRA)	George Washington University
IMMSA Consortium	Joint Program Executive Office for Chemical and Biological Defense (JPEO-CBD)	Lawrence Livermore National Lab (LLNL)	Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures	Los Alamos National Lab (LANL)	Mayo Clinic	National Biodefense Analysis and Countermeasures Center
National Center for Biotechnology Information (NCBI)	National Institute of Allergy and Infectious Diseases (NIH-NIAID)	New York State Wadsworth Laboratories	Public Health Agency Canada (PHAC)	Public Health England (PHE)	Rockefeller University	Rutgers University
Stanford University Medical Center	University of California, San Francisco (UCSF)	University of Colorado Denver	University of Ibadan, Nigeria	University of Louisville	University of Maryland School of Medicine (UMD)/ Institute for Genome Sciences (IGS)	University of Michigan
University of North Carolina at Chapel Hill	University of Texas Medical Branch (UTMB)	University of Washington School of Medicine	U.S. Army Edgewood Chemical Biological Center (ECBC)	U.S. Army Medical Research Institute for Infectious Diseases (USAMRIID)	U.S. Food and Drug Administration (CDRH, CBER, CFSAN, CVM)	Weill Cornell Medicine

✓ Optimized Collaborator and Microbe Specific Sample Collection Protocols

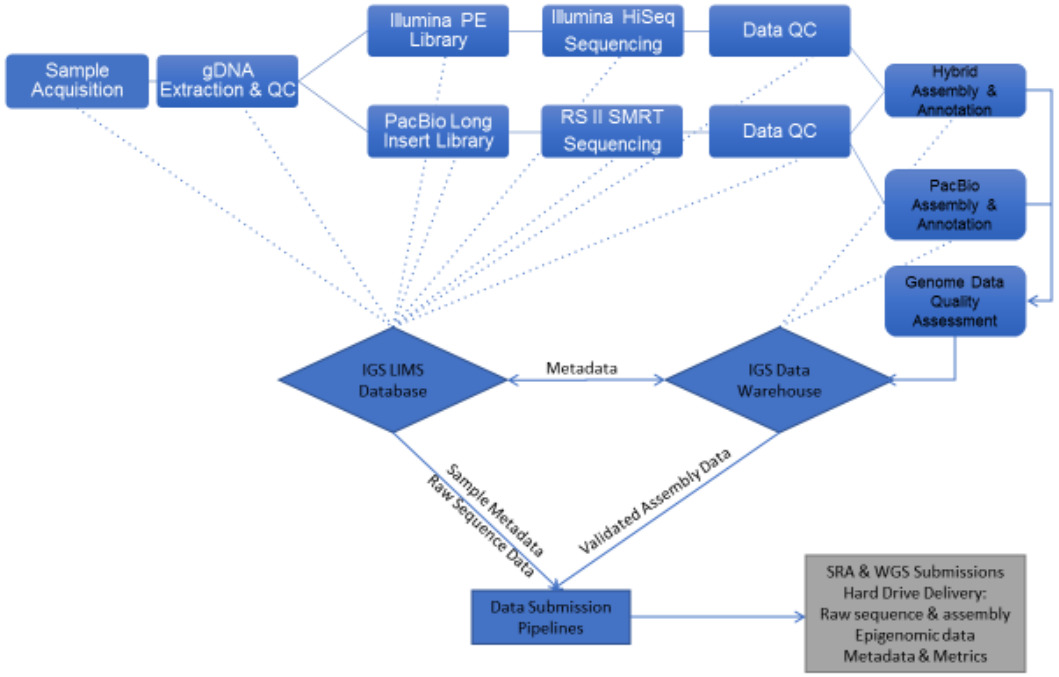
FDA-ARGOS reference genomes are generated in 3 phases:

Phase 1- collection of a previously identified microbe and nucleic acid extraction

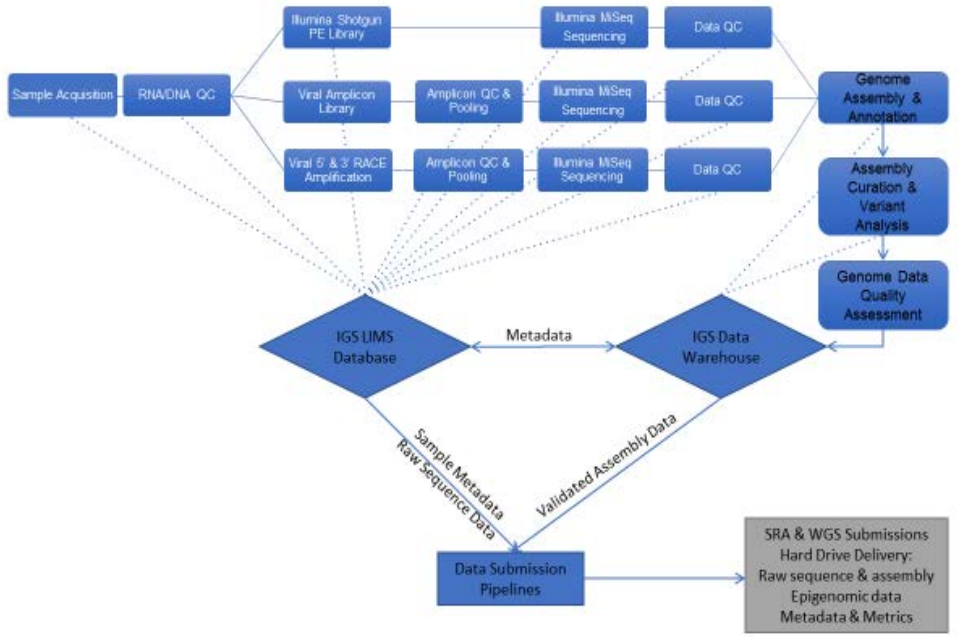
Phase 2- sequencing and de novo assembly at UMD (workflows below)

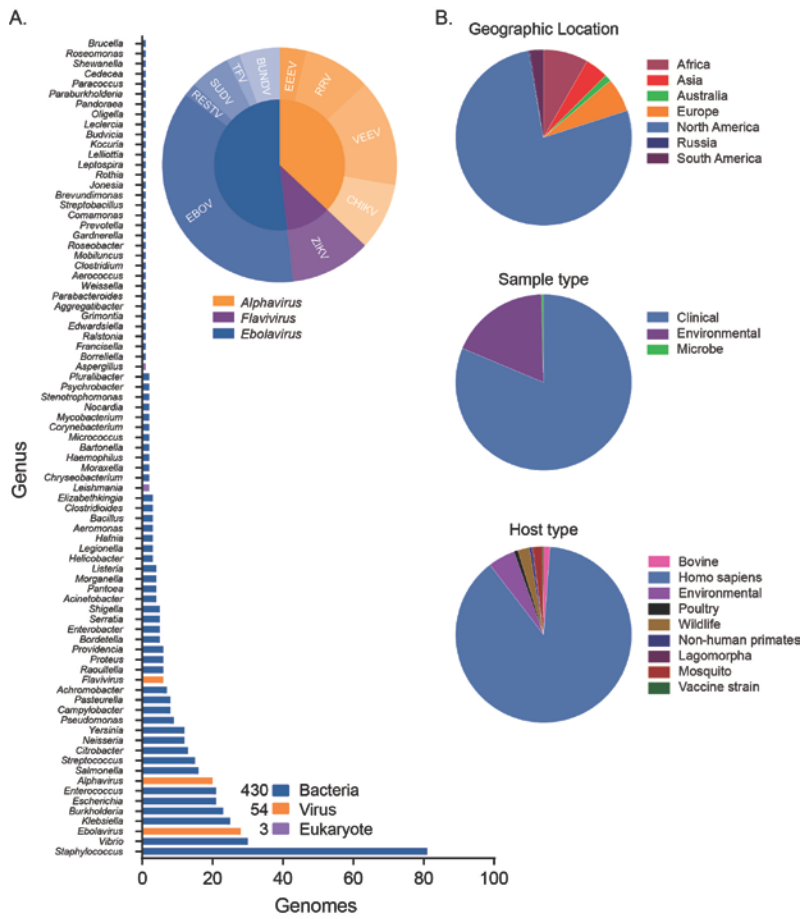
Phase 3- Recognition and data deposit in NCBI databases

Bacterial/Fungal/Parasite Workflow



Viral Workflow

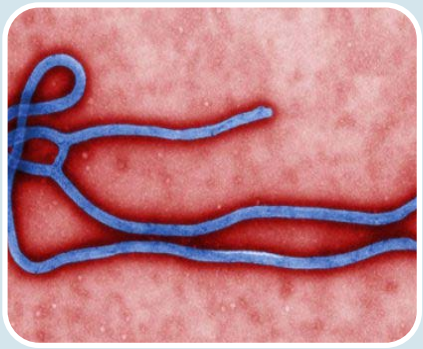


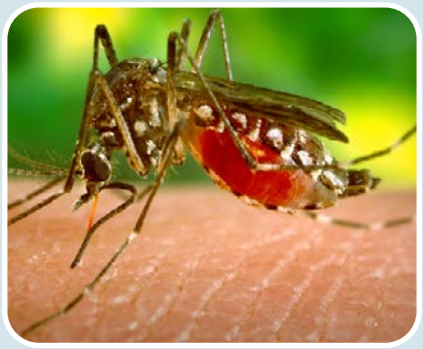


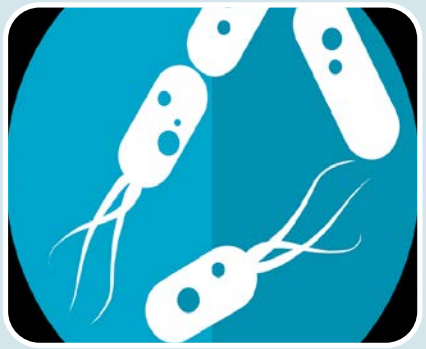
FDA-ARGOS Sample Status

- 1814 samples have been received to date
- 1180 samples have passed sample QC
  - 988 bacterial samples
    - 734 registered with NCBI
    - 53 assembled/in annotation
    - 131 sequencing pipeline
    - 70 abandoned (mixed samples, contamination)
  - 192 viral samples
    - 120 registered with NCBI
    - 29 assembled/waiting on NCBI for annotations
    - 43 sequencing pipeline

✓ Developed Regulatory-Grade Reference Genomes for Microbial Standards Efforts







Ebola

- National Institute of Allergy and Infectious Diseases (NIH-NIAID)
- Public Health Agency Canada (PHAC)
- Public Health England (PHE)
- U.S. Army Medical Research Institute for Infectious Diseases (USAMRIID)

Zika

- US Food and Drug Administration (FDA CBER)
- Public Health Agency Canada (PHAC)
- Biodefense and Emerging Infections Research Resources Repository

Biothreat

- U.S. Army Medical Research Institute for Infectious Diseases (USAMRIID)
- U.S. Army Edgewood Chemical Biological Center (ECBC)

Microbiome

- ZYMO RESEARCH

Mixed Microbial Reference Materials

- National Institute of Standards and Technology (NIST)

✓ Developed Database Access and Outreach Materials

Landing Page for FDA-ARGOS @NCBI BioProject 231221

https://www.ncbi.nlm.nih.gov/bioproject/?term=FDA-ARGOS

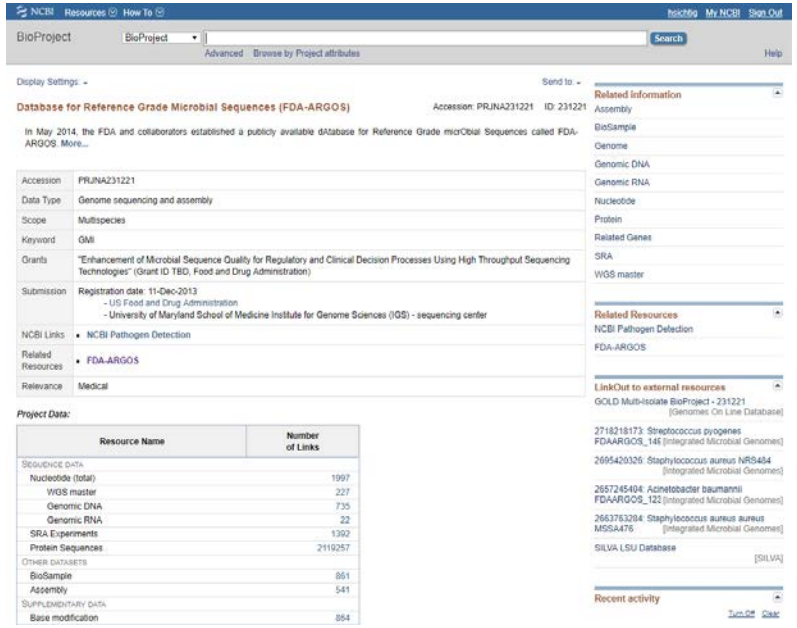
>> To get all associated genbank entries, select the Nucleotide database and enter this search term: '231221[BioProject]'

GenBank records (annotations, not RefSeq):
https://www.ncbi.nlm.nih.gov/nuccore?term=231221%5BBioProject%5D

BioSamples:
https://www.ncbi.nlm.nih.gov/biosample?Db=biosample&DbFrom=bioproject&Cmd=Link&LinkName=bioproject_biosample&LinkReadableName=BioSample&ordinalpos=1&IdsFromResult=231221

Assemblies:
https://www.ncbi.nlm.nih.gov/assembly?LinkName=bioproject_assembly_all&from_uid=231221

Raw reads:
https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=231221



- http://www.fda.gov/argos
- mailto:FDA-ARGOS@fda.hhs.gov
- FDA-ARGOS: A Public Quality-Controlled Genome Database Resource for Infectious Disease Sequencing Diagnostics and Regulatory Science Research Available on bioRxiv
- National Institute of Standards and Technology (NIST) Report "Standards for Pathogen Detection via Next-Generation Sequencing"
- Decoding Ebola: Next Generation Sequencing of the Ebola Genome for the FDA ARGOS Database
- American Society for Microbiology (ASM) Report "Applications of Clinical Microbial Next-Generation Sequencing"

Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Comment on this paper

FDA-ARGOS: A Public Quality-Controlled Genome Database Resource for Infectious Disease Sequencing Diagnostics and Regulatory Science Research

Heike Sichtig, Timothy Minogue, Yi Yan, Christopher Stefan, Adrienne Hall, Luke Tallon, Lisa Sadzewicz, Suvarna Nadendla, William Klimke, Eneida Hatcher, Martin Shumway, Dayanara Aldea, Jonathan Allen, Jeffrey Koehler, Tom Slezak, Stephen Lovell, Randal Schoepp, Uwe Scherf

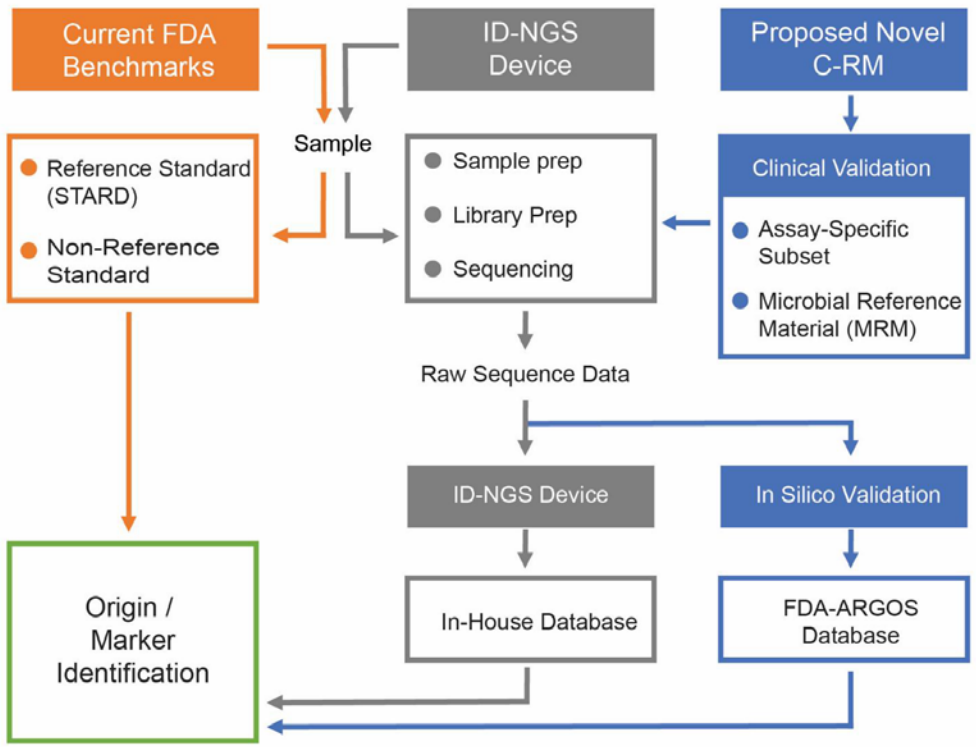
doi: https://doi.org/10.1101/482059

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract

Infectious disease next generation sequencing (ID-NGS) diagnostics are on the cusp of revolutionizing the clinical market. To facilitate this transition, FDA proactively invested in tools to support innovation of emerging technologies. FDA and collaborators established a publicly available database, FDA dAtabase for Regulatory-Grade micrObial Sequences (FDA-ARGOS), as a tool to fill reference database gaps with quality-controlled genomes. This manuscript discusses quality control metrics for the proposed FDA-ARGOS genomic resource and outlines the need for quality-controlled genome gap filling in the public domain. Here, we also present three case studies showcasing potential applications for FDA-ARGOS in infectious disease diagnostics, specifically: assay design, reference database and *in silico* sequence comparison in

A.



B.

FDA-ARGOS Quality Metrics

a. Organisms identified prior to sequencing by orthogonal reference method

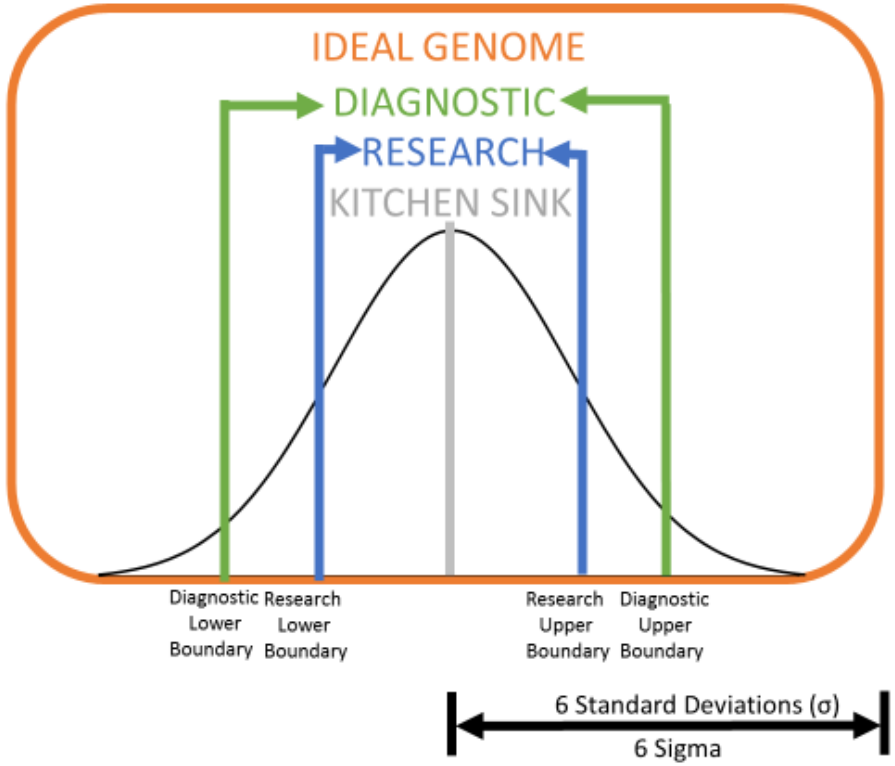
b. Two sequencing methodologies
- Bacterial - Long read and short read
- Viral - Shotgun, amplicon and RACE

c. De-novo assembly

C.

FDA-ARGOS Data Requirements

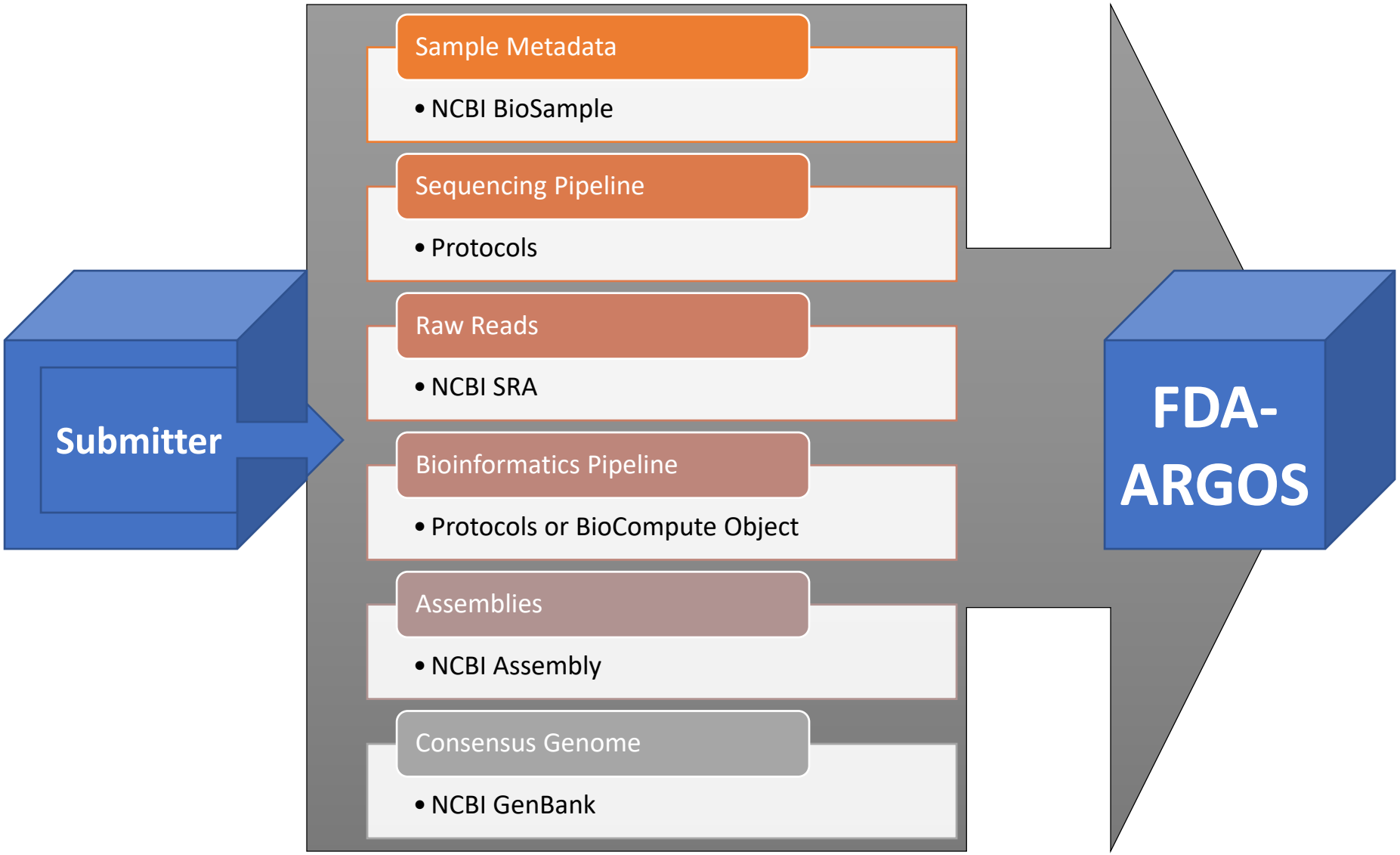
a. Sample Name (Sample ID)
b. Raw Reads (SRA Accession)
c. Assemblies (Chromosome, Plasmid, WGS Accession)
d. Annotations (GenBank Accession)
e. 10-meta data (Biosample Accession)



WORK IN PROGRESS
External Genome Qualification

- Based on FDA-ARGOS Reference Genome Characteristics
- Open-source tool
  - Genome quality (e.g. coverage, ANI, GC, assembly size)
  - Genome continuity (e.g. N50, L50, num contigs)
  - Metadata (e.g. species name, submitter, orthogonal identification method)
- Current work on boundary finding is challenging
- Looking at TCC and NCTC 3000 efforts

External Genome Submission (NCBI, BioSample)



Acknowledgements

FDA-ARGOS team members include representatives from the:

- U.S. Food and Drug Administration
- U.S. Department of Defense
- National Institutes of Health
- Institute for Genome Sciences at University of Maryland

UNIVERSITY of MARYLAND SCHOOL OF MEDICINE INSTITUTE FOR GENOME SCIENCES

NCBI

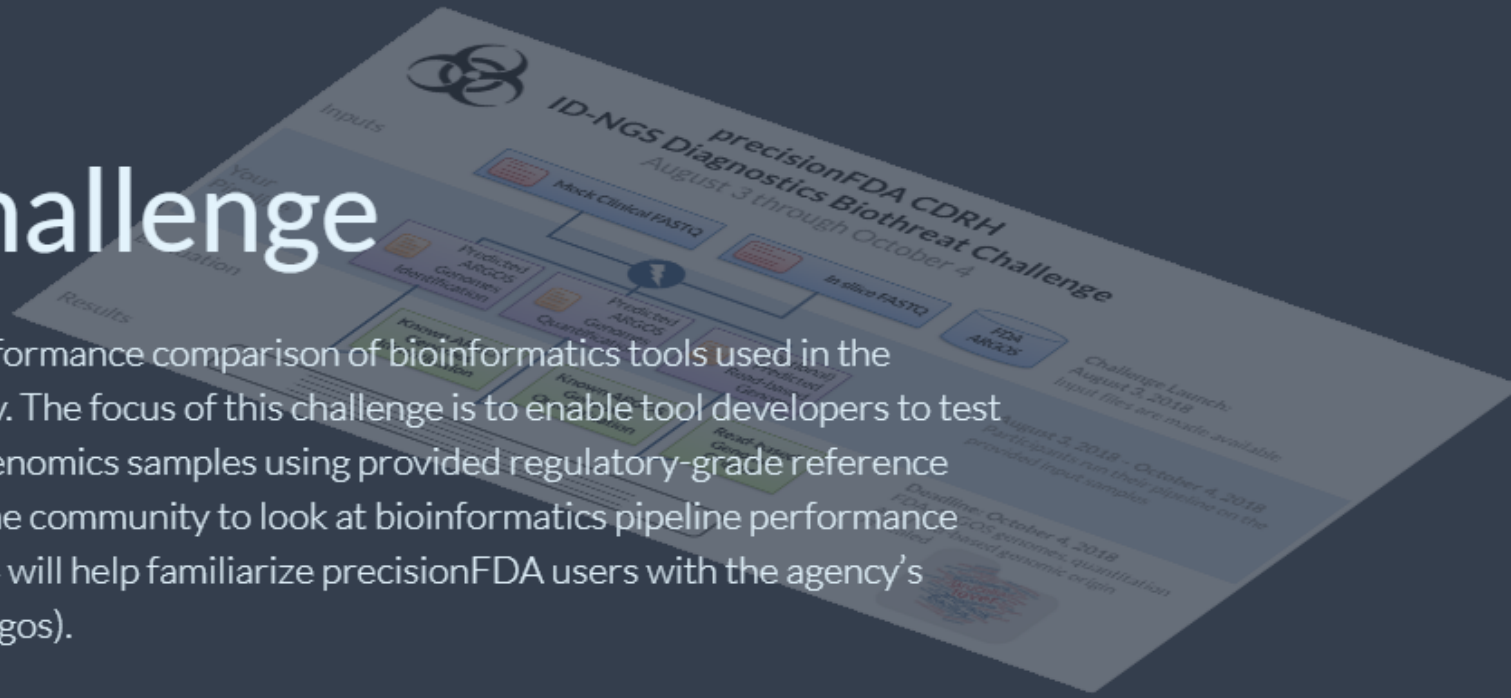
Funding Agencies

FDA's Office of Counterterrorism and Emerging Threats
Joint Program Executive Office for Chemical and Biological Defense (JPEO-CBD)

FDA U.S. FOOD & DRUG ADMINISTRATION

American Type Culture Collection/ BEI
Bernard Nocht Institute for Tropical Medicine, Germany
Biodefense and Emerging Infections Research Resources Repository
British Columbia Centre for Disease Control (BCCDC)
Children's National Medical Center
Defense Threat Reduction Agency (DTRA)
George Washington University
IMMSA Consortium
Joint Program Executive Office for Chemical and Biological Defense (JPEO-CBD)
Lawrence Livermore National Lab (LLNL)
Leibniz Institute (DSMZ)
Los Alamos National Lab (LANL)
Mayo Clinic
National Biodefense Analysis and Countermeasures Center
National Institute of Allergy and Infectious Diseases (NIH-NIAID)
National Institute of Standards and Technology (NIST)
New York State Wadsworth Laboratories
Public Health Agency Canada (PHAC)
Public Health England (PHE)
Rockefeller University
Rutgers University
Stanford University Medical Center
Tetracore
University of California, San Francisco (UCSF)
University of Colorado Denver
University of Ibadan, Nigeria
University of Louisville
University of Michigan
University of North Carolina at Chapel Hill
University of Texas Medical Branch (UTMB)
University of Washington School of Medicine
U.S. Army Edgewood Chemical Biological Center (ECBC)
U.S. Army Medical Research Institute for Infectious Diseases (USAMRIID)
U.S. Food and Drug Administration
Weill Cornell Medicine

PrecisionFDA
CDRH Biothreat Challenge



Provide challenge data sets and reference standards for performance comparison of bioinformatics tools used in the biothreat and infectious disease NGS diagnostics community. The focus of this challenge is to enable tool developers to test their algorithms on blinded mock-clinical and in silico metagenomics samples using provided regulatory-grade reference genomes from the FDA-ARGOS database. This will enable the community to look at bioinformatics pipeline performance using a fixed reference genome data standard. The challenge will help familiarize precisionFDA users with the agency's innovative FDA-ARGOS database resource (www.fda.gov/argos).

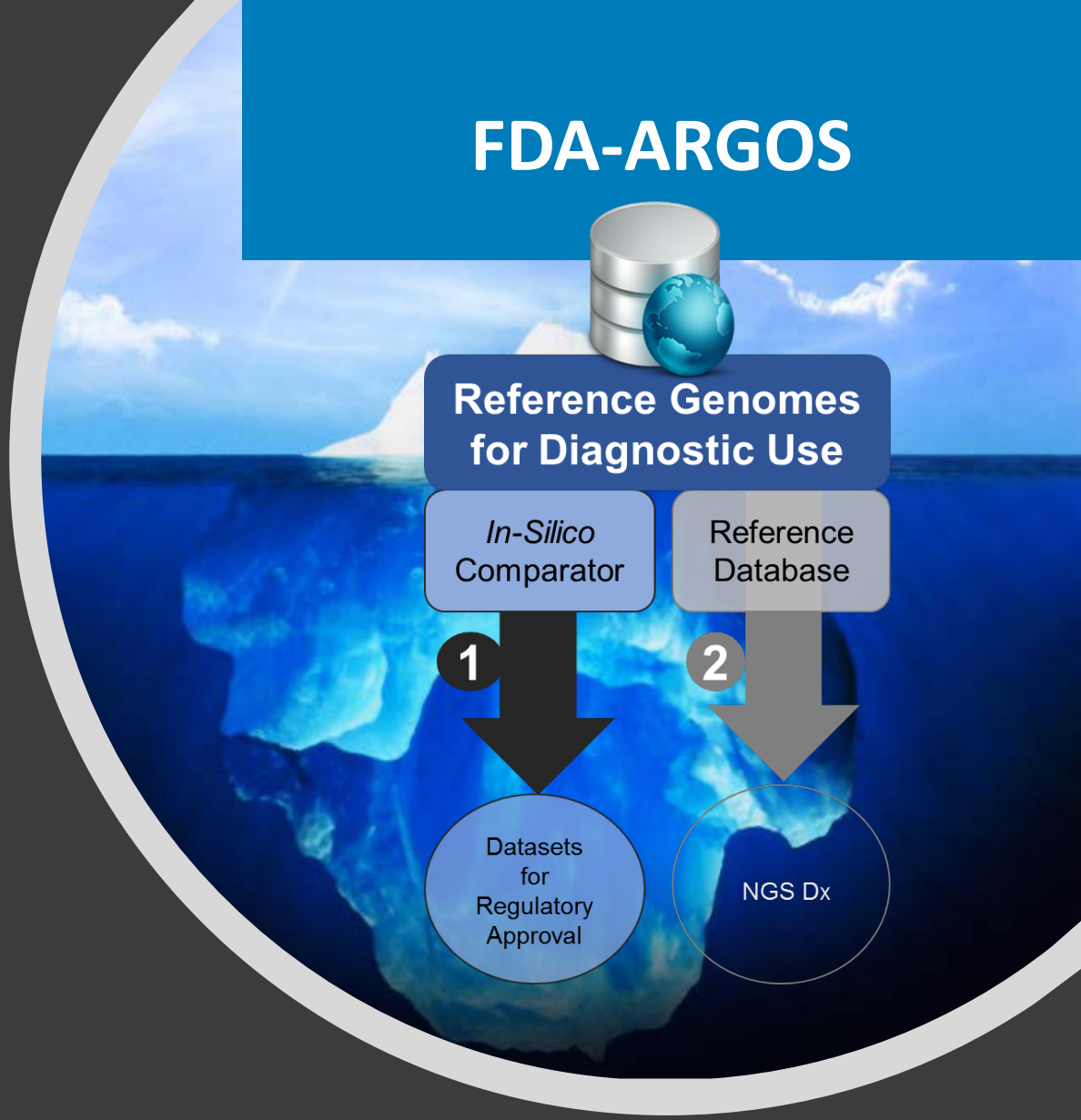
STARTS	ENDS
2018-08-04 00:00:00 UTC	2018-10-19 03:00:00 UTC

View Challenge

FDA ARGOS Team – DoD USAMRIID Collaboration on Biothreat Detection

- The FDA ARGOS database (www.fda.gov/argos) generates and publishes quality-controlled microbial reference genomes for diagnostic use, which enable ID-NGS developers to perform *in silico* validation of their workflows.

- FDA partnered with USAMRID to collect, sequence and publish quality-controlled biothreat reference genomes for diagnostic use.

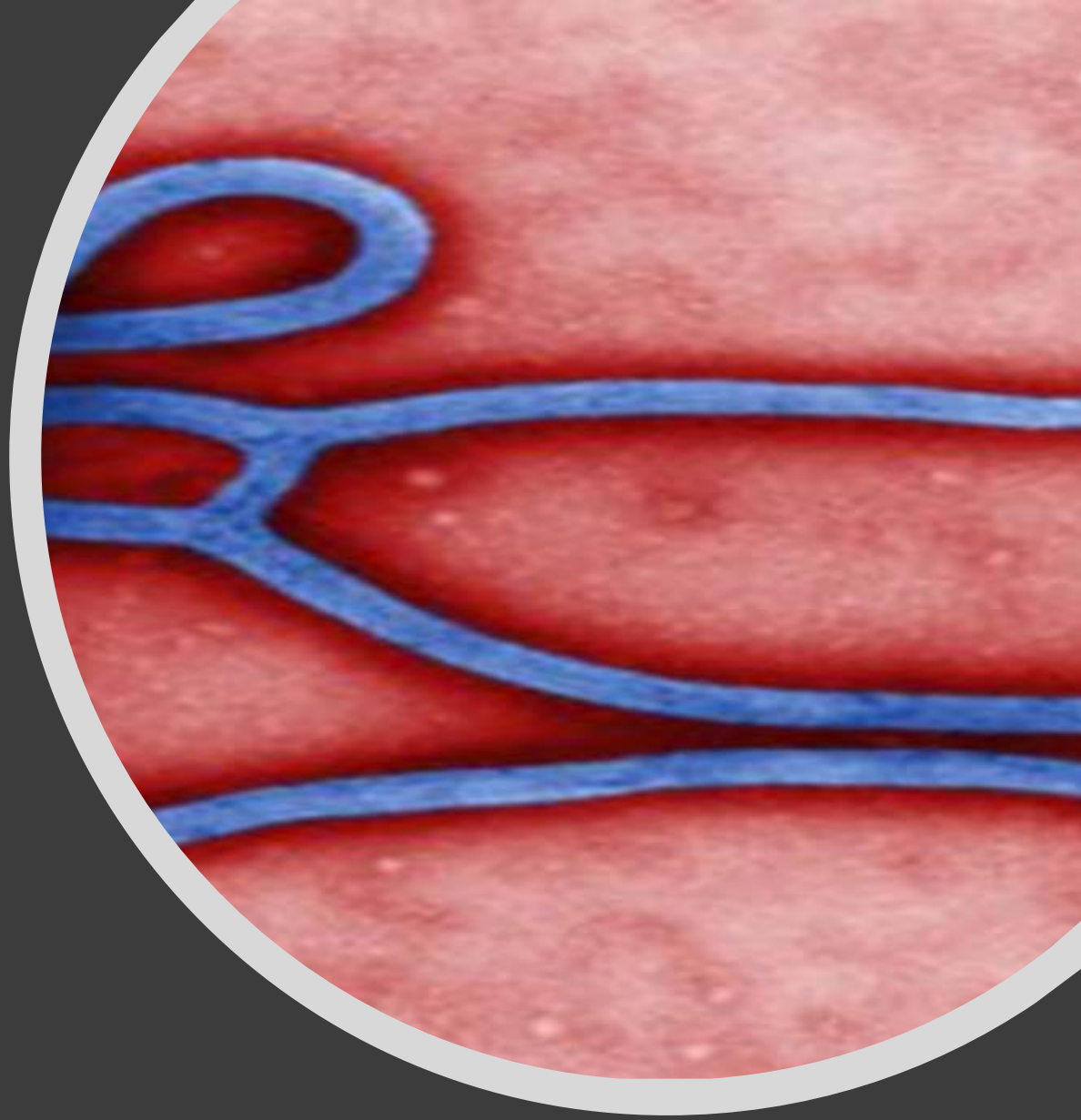


Biothreat Challenge Motivation

- The Ebola outbreak in West Africa in 2014 used advanced infectious disease (ID) detection technology based on next generation sequencing (NGS) to determine index case and potential novel microbes.
- Current ID-NGS technology is still evolving and typically involves complex laboratory and bioinformatics workflows.
- The use of NGS provides a biased-free, detailed view of infectious microorganisms that promises to enable faster detection, traceback, and selection of therapeutics without prior knowledge of disease cause.

➢ **To reach these objectives, ID NGS computational workflows must be independently evaluated and validated.**

https://precision.fda.gov/experts/6/blog



precisionFDA

A community platform for NGS assay evaluation and regulatory science exploration.

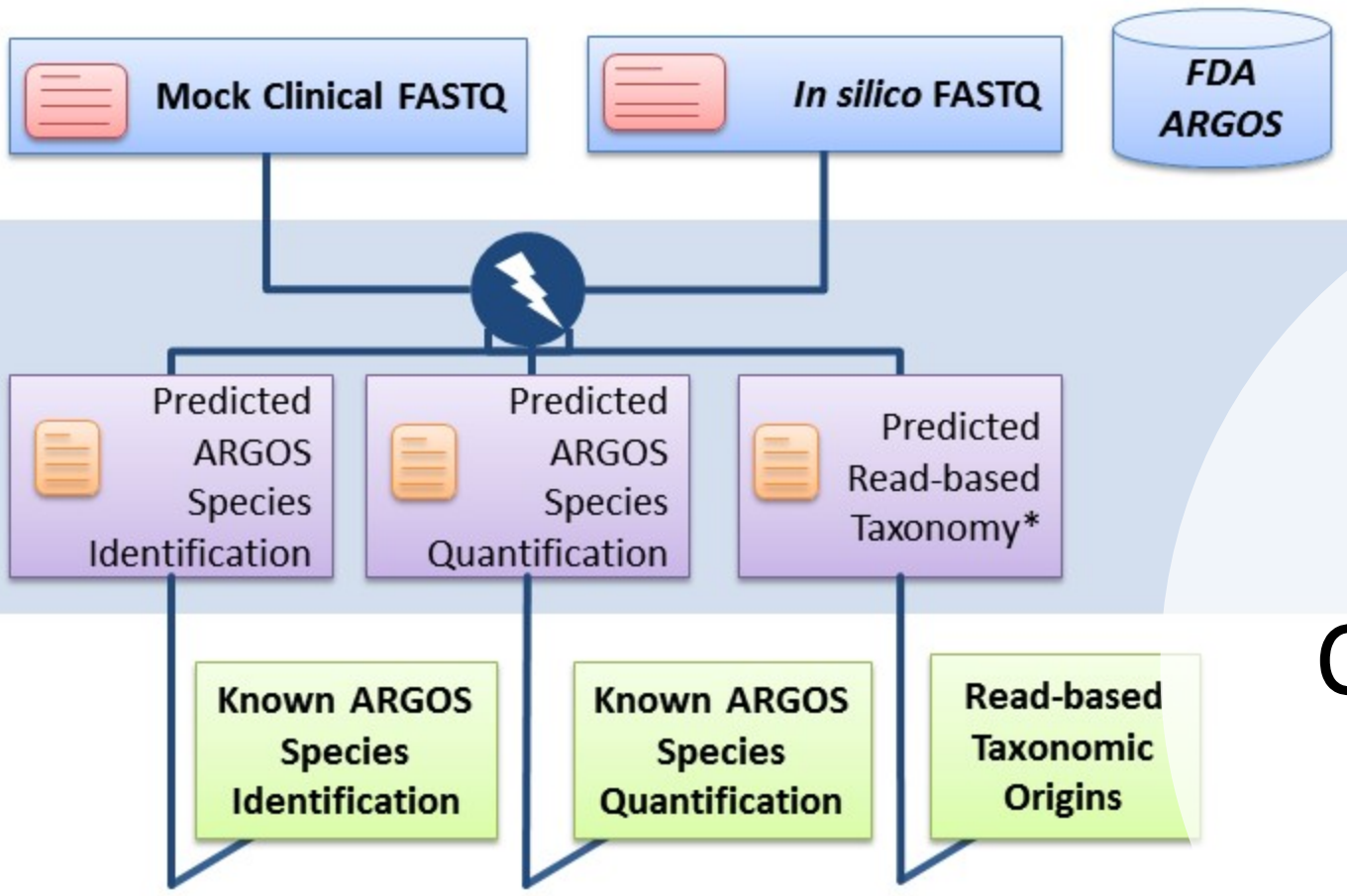
Log in | Request Access →

Biothreat Challenge

Benchmark your detection algorithm on a task to identify and quantify biothreat organisms in clinically relevant metagenomics next generation sequencing (NGS) samples.

*(The reference database is **fixed** in this challenge.)*

precisionFDA
ID-NGS Diagnostics Biothreat Challenge
August 3 through October 18



Challenge Design

Data Sets

- 21 metagenomics samples
  - 9 *in silico* Samples (C1-C9)
  - 12 biological samples (C10-21)
- 517 blinded FDA-ARGOS reference genomes
  - CR_1 – CR_517

Samples	Microbial Species
C01, C02, C03	*Burkholderia thailandensis* *Burkholderia mallei* *Escherichia coli* *Propionibacterium acnes*
C04, C05, C06	Zika virus Chikungunya virus Ross River Valley Virus
C07, C08, C09	Ebola Virus
C10, C11, C12	*Yersinia pestis* *Yersinia pseudotuberculosis* *Escherichia coli*
C13, C14, C15	*Burkholderia thailandensis*
C17, C19, C21	*Staphylococcus aureus*
C16, C18, C20	NA

Expected Results

1. Development of novel computational algorithms for identifying emerging pathogens in clinical matrix, such as the Ebola virus
2. Independent evaluation of ID NGS computational algorithms with a fixed reference database to aid future developers
3. Greater public and scientific engagement in infectious disease detection and surveillance

The **precisionFDA**

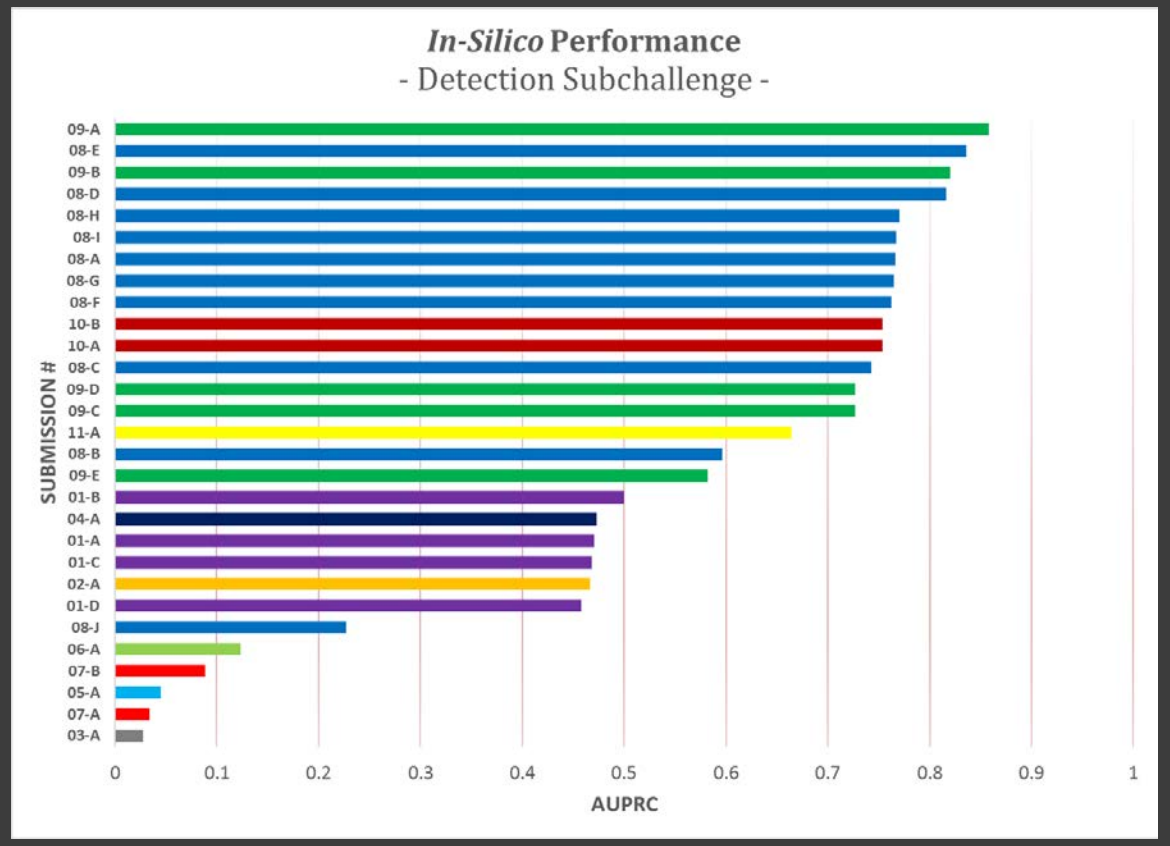
CDRH Biothreat Challenge ran from

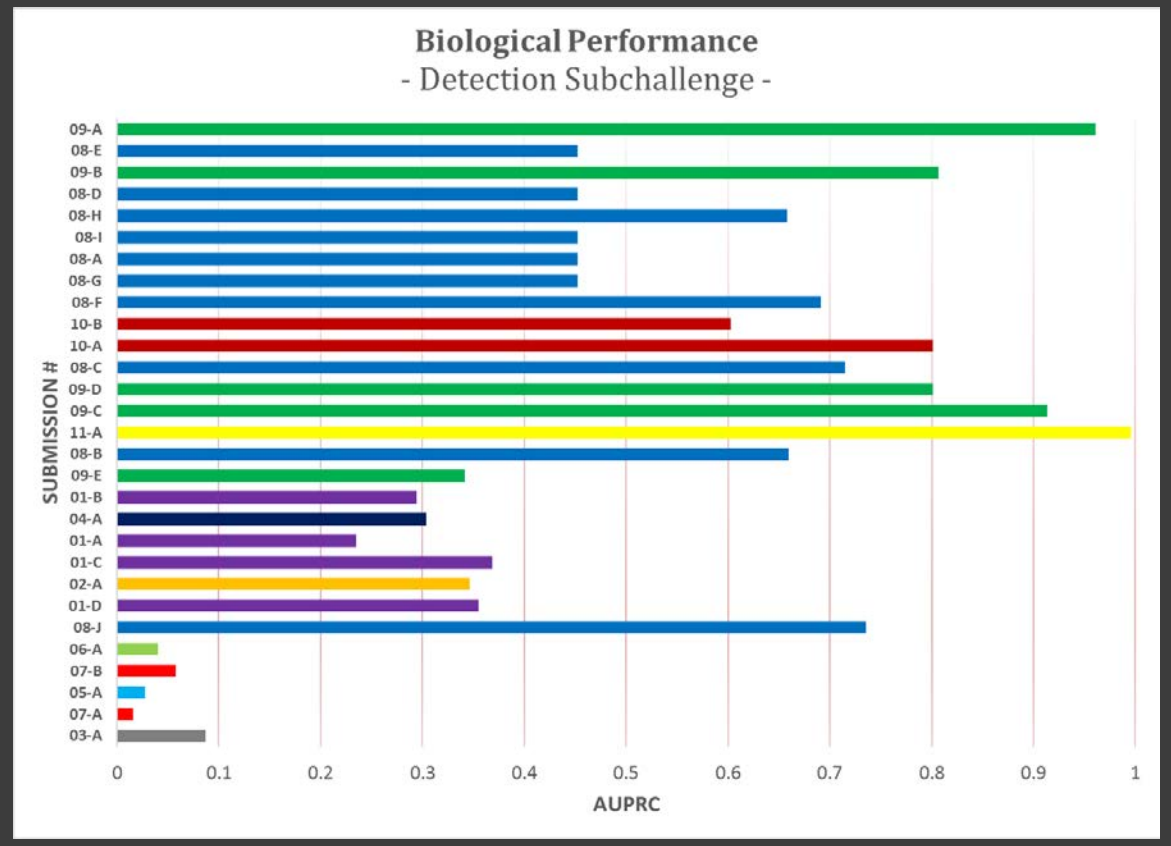
August 3, 2018 to October 18, 2018.

There were 29 valid entries from 11 participants.

Detection Subchallenge Evaluation

The area under the precision-recall curve (AUPRC) was computed by comparing the predicted normalized confidence scores for identified species to the known species.





Behind the Scenes … A First Look

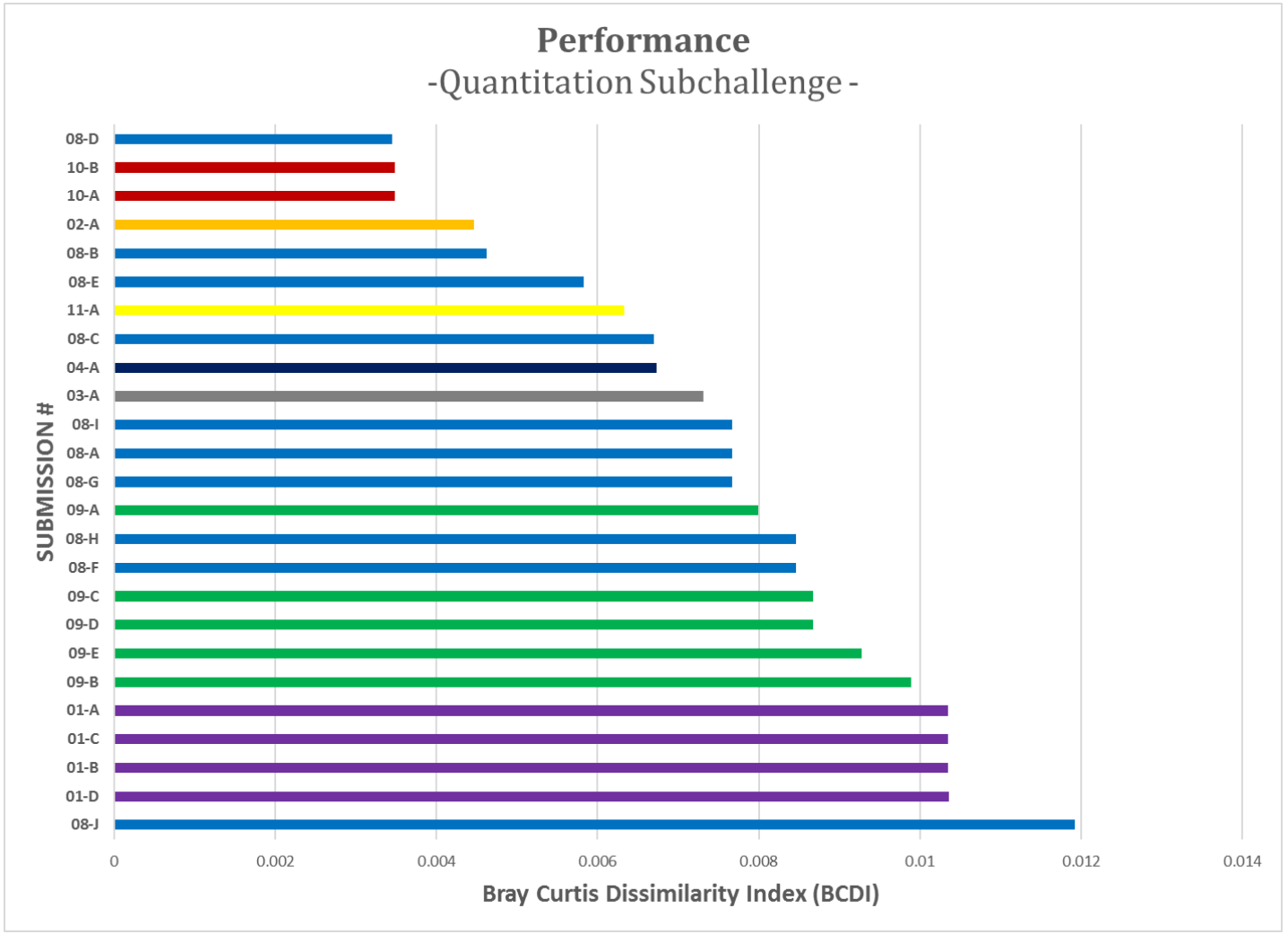
Detection Subchallenge Winners

Rank	Overall	In Silico	Biological
1	Jason Simpson Microbiohm MB_FL_Genome_Identification	Jason Simpson Microbiohm MB_FL_Genome_Identification	Richa Agarwala NCBI submit.identification
2	Jason Simpson Microbiohm MB_S1_UF_Sp_2_Genome_Identification	Chung-Tsai Su Atgenomix m4_Genome_Identification	Jason Simpson Microbiohm MB_FL_Genome_Identification
3	Jason Simpson Microbiohm MB_S1_UF_Genome_Identification	Jason Simpson Microbiohm MB_S1_UF_Genome_Identification	Jason Simpson Microbiohm MB_S1_UF_Sp_2_Genome_Identification
4	Nick Greenfield One Codex OCX_2_ARGOS_Reference_Genome_Identification	Chung-Tsai Su Atgenomix m3_Genome_Identification	Jason Simpson Microbiohm MB_S1_UF_Genome_Identification
5	Jason Simpson Microbiohm MB_S1_UF_Sp_3_Genome_Identification	Chung-Tsai Su Atgenomix m7_Genome_Identification	Nick Greenfield One Codex OCX_2_ARGOS_Reference_Genome_Identification

Quantification Subchallenge Evaluation

The species quantifications were evaluated based on their agreement with the species composition of samples C1 to C9. The Bray Curtis Dissimilarity Index (BCDI) was used to evaluate the agreement between the predicted and known species quantifications.

More Behind the Scenes …



Quantitation Subchallenge Winners

Rank	Overall
1	Chung-Tsai Su Atgenomix m3_Genome_Quantification
2	Nick Greenfield One Codex OCX_ARGOS_Reference_Genome_Quantification
3	Jonathan Jacobs QIAGEN ARGOS_Reference_Genome_Quantification_vSCJJ1
4	Chung-Tsai Su Atgenomix m1_Genome_Quantification
5	Chung-Tsai Su Atgenomix m4_Genome_Quantification

Acknowledgements

These results offer a first glance at our understanding.

We welcome the community to further explore these results and provide insight for the future.

The precisionFDA CDRH Biothreat Challenge team is preparing a scientific manuscript that describes that challenge and challenge results.

Team

- PrecisionFDA: Elaine Johanson, Ruth Bandler
- PrecisionFDA CDRH: Adam Berger (now at NIH), Zivana Tezak
- Booz Allen: Zeke Maier
- DNAnexus: Singer Ma, John Didion
- FDA CDRH: Heike Sichtig, Yi Yan
- USAMRIID: Timothy Minogue, Chris Stefan