A Biocompute Object For FDA-ARGOS Reference Genomes

Heike Sichtig, MS/PhD

Subject Matter Expert
Principal Investigator

Center for Devices
Division of Microbiology Devices
US Food and Drug Administration

2017 HTS Computational Standards for Regulatory Sciences Workshop
Mar 16-17, NIH in Bethesda, MD, USA

Disclaimer

The information in these materials is not a formal dissemination of information by FDA and does not represent agency position or policy.

Opinions are my own

FDA Tools for ID NGS Dx

FDA-ARGOS Database

:microbial reference genomes for regulatory use

✓ New and flexible regulatory pathway
  - ▪ Enable In-silico validation
  - ▪ Reduce testing burden

✓ Reference database

Interagency ID NGS Working Group

: team of NGS agency-wide subject matter experts

✓ ID NGS Dx Advisory Board

✓ Consensus FDA-ARGOS genome vetting

✓ Keep current on state of the art

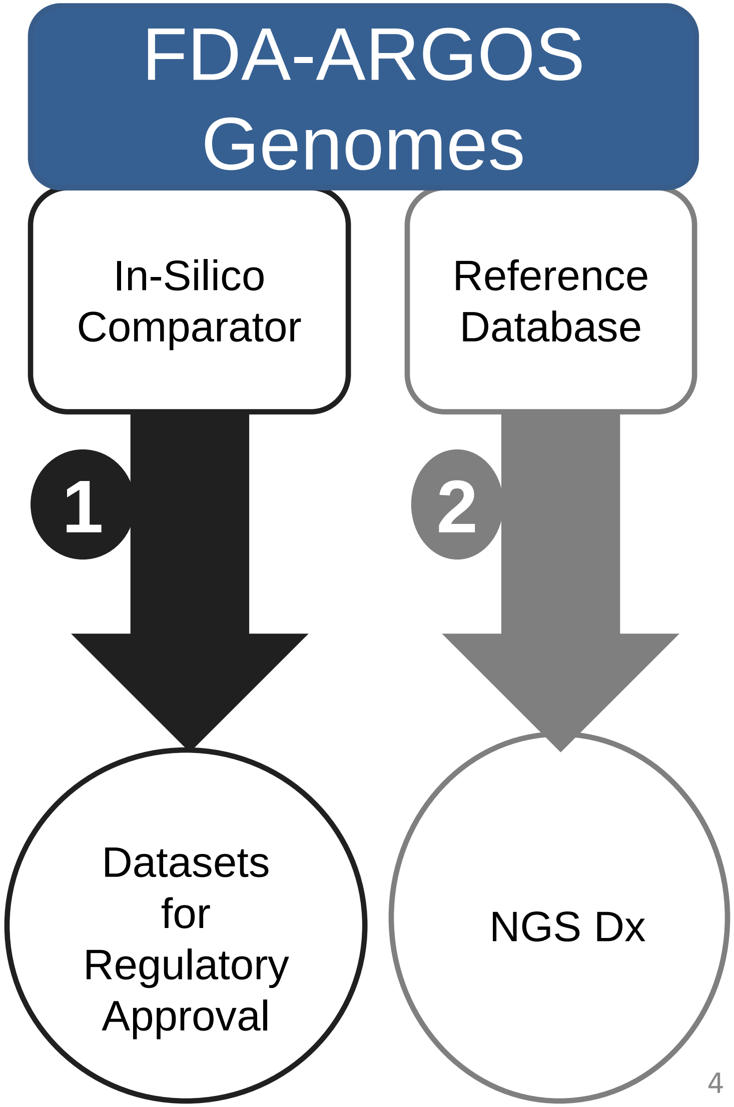
✓ Tackle open questions (i.e. sens/spec)

FDA-ARGOS: Goal and Use

- Public Vetted Resource
- Microbial Reference-Grade Genomes for **Regulatory Use**
- US-Initiated
- Medical Countermeasures
- Common clinical
- Near neighbors

- **Coverage for US Needs**
- **Currently not funded to support *Needs for Developing World* and associated *Global Standards***

NCBI Project **PRJNA231221**



FDA-ARGOS Genomes

In-Silico Comparator

Reference Database

1

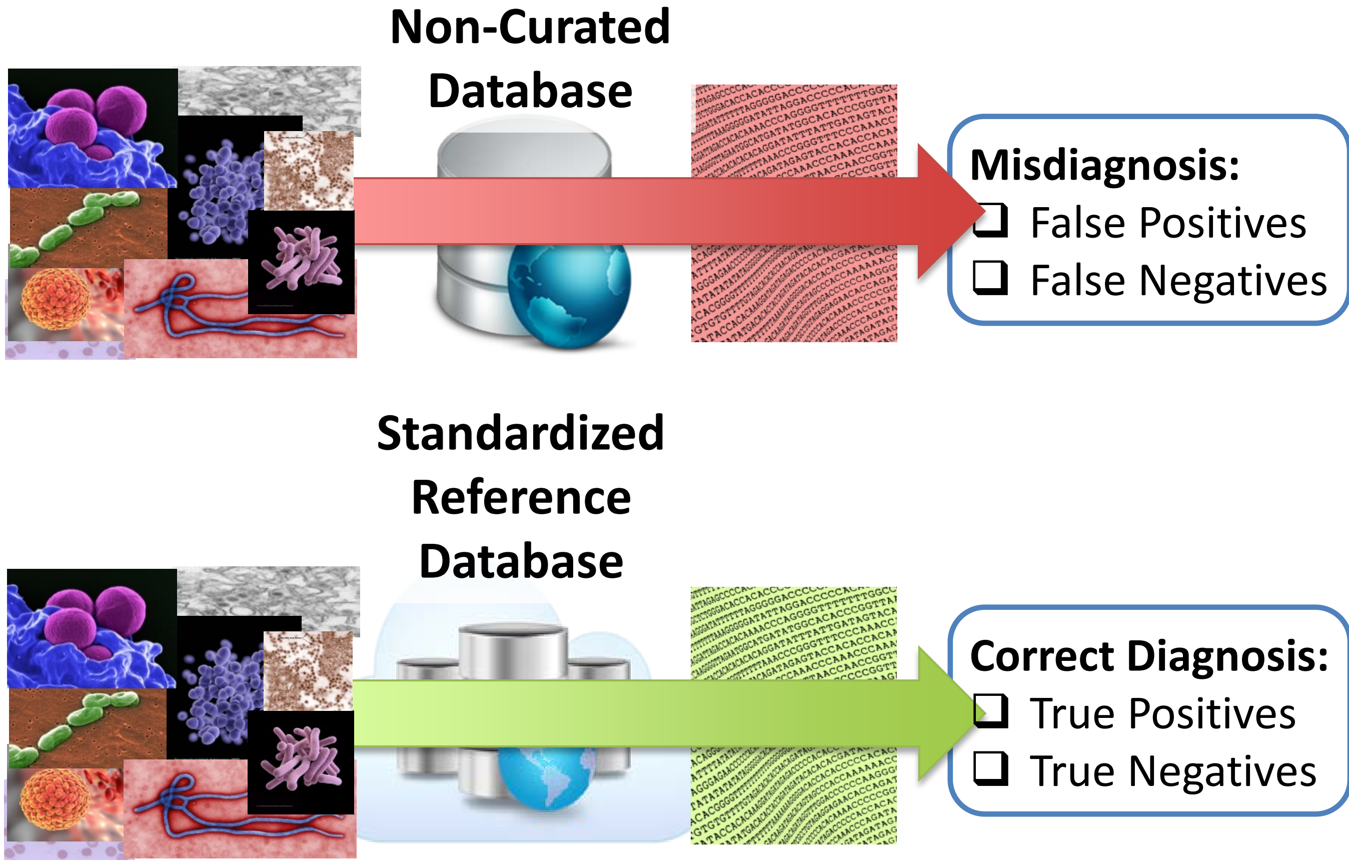
2

Datasets for Regulatory Approval

NGS Dx

Reference Genome Gap: Ebola

Endemic African Diseases
Chikungunya virus
Crimean-Congo Hemorrhagic Fever virus
Dengue virus serotype 1
Dengue virus serotype 2
Dengue virus serotype 3
Dengue virus serotype 4
Ebola virus
Lassa virus
Marburg virus (Angola)
Marburg virus (Ci67)
Plasmodium falciparum
Rift Valley fever virus
West Nile virus
Yellow fever virus
Zika virus



Non-Curated Database

Misdiagnosis:
- ☐ False Positives
- ☐ False Negatives

Standardized Reference Database

Correct Diagnosis:
- ☐ True Positives
- ☐ True Negatives

✓ ***Minimize Misdiagnosis***

✓ ***Evolutionary Change***

✓ ***Rapid* Diagnostics**

In-Silico Comparator Example

DoD Collaboration

- Sequencing-based diagnostic device
- Generate FDA-ARGOS Reference Genomes
- Datasets for Regulatory Approval

> Enable In-Silico Data Analysis

Endemic African Diseases
Chikungunya virus
Crimean-Congo Hemorrhagic Fever virus
Dengue virus serotype 1
Dengue virus serotype 2
Dengue virus serotype 3
Dengue virus serotype 4
Ebola virus
Lassa virus
Marburg virus (Angola)
Marburg virus (Ci67)
Plasmodium falciparum
Rift Valley fever virus
West Nile virus
Yellow fever virus
Zika virus

FDA-ARGOS Genome Pipeline

FDA-ARGOS microbial 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

Phase 3- Vetting and data deposit in NCBI databases

FDA-ARGOS Reference Genome Characteristics:

- High depth of base coverage.
- Placed within a pre-established phylogenetic tree.
- Minimum of 20X over 95 percent of the assembled core genome.
- Sample specific metadata, raw reads, assemblies, annotation and details of the bioinformatics pipeline are available.

Sequencing Approach

Bacteria

- **Hybrid sequencing** approach using Illumina HiSep2000 and the PacBio RSII platform to generate industry standard high quality sequences. Use of multiple assemblers. 3 sets of de novo genome assemblies will be produced 1) Illumina only, 2) PacBio only, and 3) Illumina/PacBio hybrid

Virus

- IGS will use existing and well-established laboratory and bioinformatics pipelines within the Genomic Resource Center. A **three-prong** Illumina sequencing approach followed by customized assembly

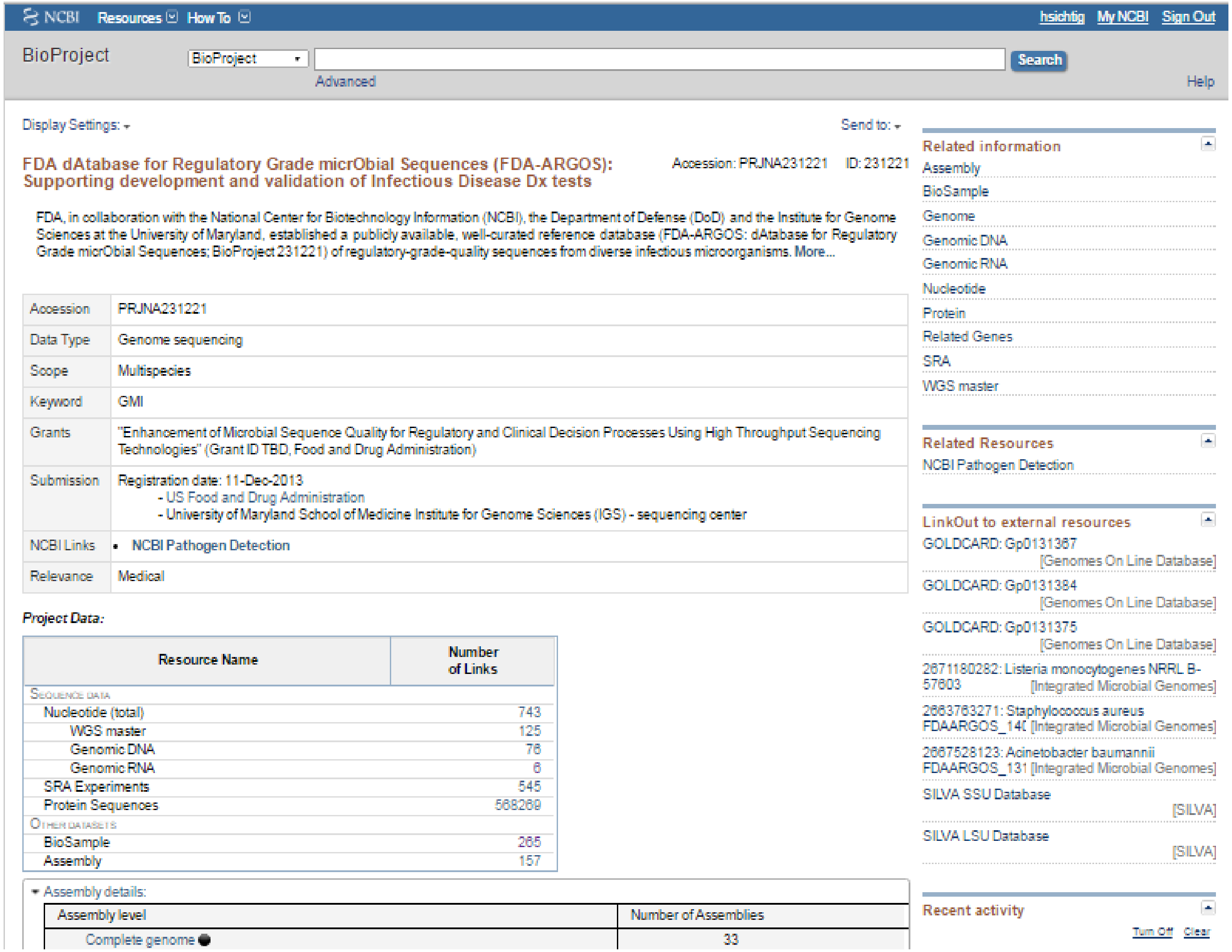
FDA-ARGOS Genome Status

- There are **827 (bacterial, viral)** samples currently at various stages within the FDA-ARGOS sequencing pipeline.
- **322 (bacterial, viral)** genomes from other efforts (i.e. TTC) to be qualified.
- Goal is to collect and sequence **2000** gap organisms

Overall pipeline
Collaborator -> FDA OSEL -> UMD/IGS -> NCBI/FDA
Collaborator -> USAMRIID -> UMD/IGS -> NCBI/FDA

NCBI BioProject 231221

Houses FDA-ARGOS genomes generated with the IGS-UMD Sequencing pipeline



FDA dAtabase for Regulatory Grade micrObial Sequences (FDA-ARGOS): Supporting development and validation of Infectious Disease Dx tests

Accession: PRJNA231221 ID: 231221

FDA, in collaboration with the National Center for Biotechnology Information (NCBI), the Department of Defense (DoD) and the Institute for Genome Sciences at the University of Maryland, established a publicly available, well-curated reference database (FDA-ARGOS: dAtabase for Regulatory Grade micrObial Sequences; BioProject 231221) of regulatory-grade-quality sequences from diverse infectious microorganisms. More...

Accession	PRJNA231221
Data Type	Genome sequencing
Scope	Multispecies
Keyword	GMI
Grants	"Enhancement of Microbial Sequence Quality for Regulatory and Clinical Decision Processes Using High Throughput Sequencing Technologies" (Grant ID TBD, Food and Drug Administration)
Submission	Registration date: 11-Dec-2013 - US Food and Drug Administration - University of Maryland School of Medicine Institute for Genome Sciences (IGS) - sequencing center
NCBI Links	• NCBI Pathogen Detection
Relevance	Medical

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	743
WGS master	125
Genomic DNA	76
Genomic RNA	6
SRA Experiments	545
Protein Sequences	568269
OTHER DATASETS	
BioSample	265
Assembly	157

Assembly details:

Assembly level	Number of Assemblies
Complete genome ●	33

Related information

- Assembly
- BioSample
- Genome
- Genomic DNA
- Genomic RNA
- Nucleotide
- Protein
- Related Genes
- SRA
- WGS master

Related Resources

- NCBI Pathogen Detection

LinkOut to external resources

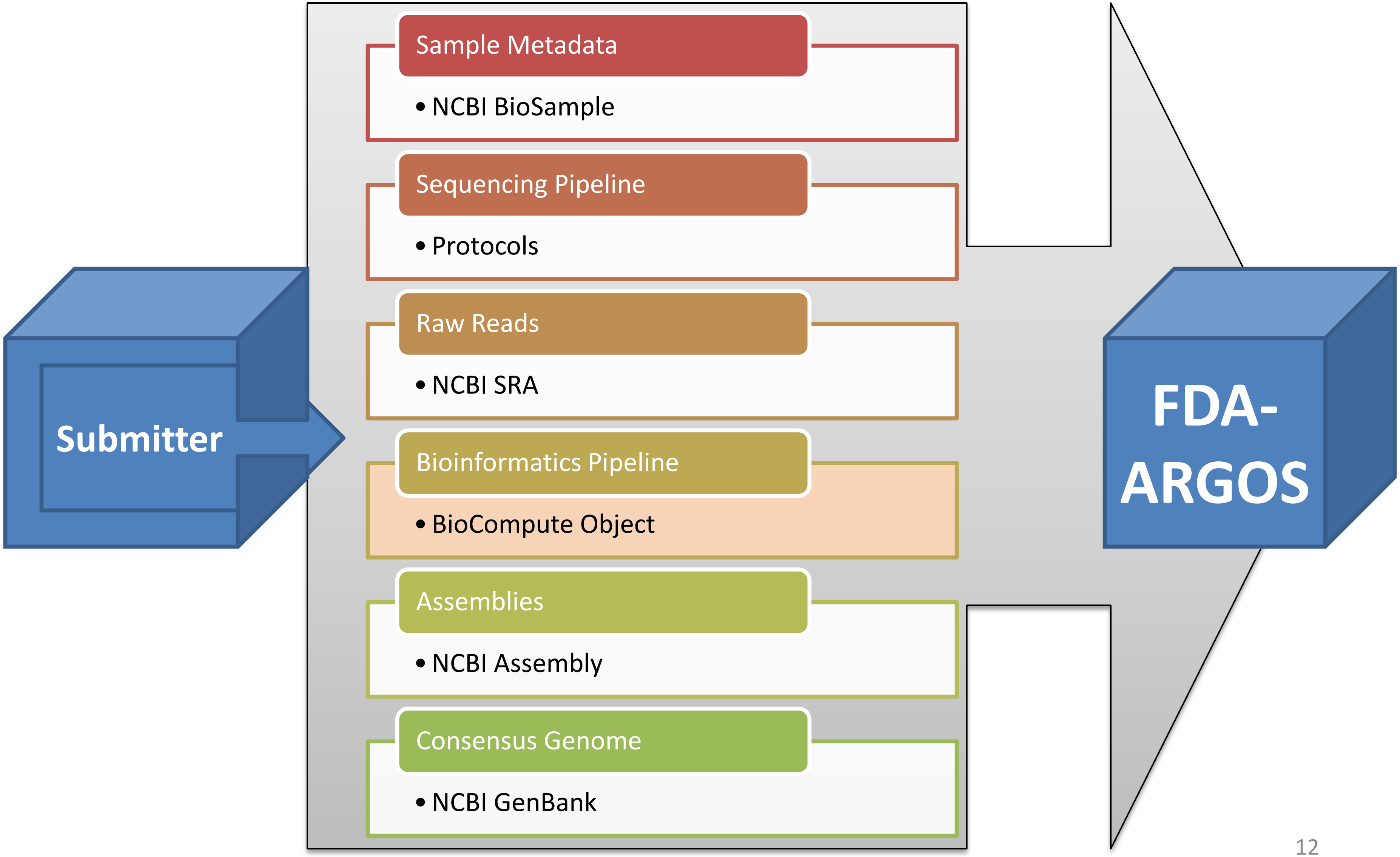
- GOLDCARD: Gp0131367 [Genomes On Line Database]
- GOLDCARD: Gp0131384 [Genomes On Line Database]
- GOLDCARD: Gp0131375 [Genomes On Line Database]
- 2671180282: Listeria monocytogenes NRRL B-57603 [Integrated Microbial Genomes]
- 2663763271: Staphylococcus aureus FDAARGOS_14([Integrated Microbial Genomes]
- 2667528123: Acinetobacter baumannii FDAARGOS_131 [Integrated Microbial Genomes]
- SILVA SSU Database [SILVA]
- SILVA LSU Database [SILVA]

Recent activity

FDA-ARGOS BioCompute Object

- Common language
- Community developed harmonized standard for bioinformatics pipeline
- Considering to use **Biocompute Object** to streamline external genome submission for the FDA ARGOS database

External Genome Submission



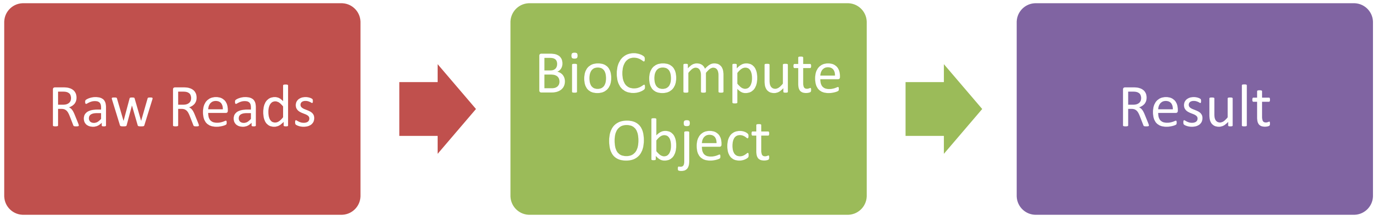
External Genome BioCompute Object

GW Collaborator provided this JSON-format BioCompute Object Example

```
{
   "name": "Bordetella pertussis ",
…
   "authors": [{"name":“Submitter Name"}],
   "description_domain":{
…
"execution_domain": {
      "platform": "unix",
      "pipeline_version": "1.0",
      "env_parameters": ["64-bit processor","2GB RAM" ],
      "driver": "perl5.6",
      "script": "https://github.com/biocomputeobjects//HTSCSRS/tree/master/11_argos/argos.pl",
      "prerequisites": [
         {"name":"Celera","version":"8.2"},
        {"name":"NCBIProkaryoticGenomeAnnotationPipeline","version":"3.1"}
…
   "io_domain": {
      "reference_uri": [ "NA"],
      "input_uri_list": [ "example.fasta” ],
      "output_uri_list": [ "https://www.ncbi.nlm.nih.gov/biosample/SAMN03996260",
                           "https://www.ncbi.nlm.nih.gov/sra?LinkName=biosample_sra&from_uid=3996260",
                           "https://www.ncbi.nlm.nih.gov/nuccore/991852837" ]
   },
}
```

Future Consideration

- NGS data submitted as part of regulatory submission
  - BioCompute Object for bioinformatics pipeline



Raw Reads

BioCompute Object

Result