BioCompute Workshop for Reviewers:
Tool for Communicating Sequencing Analysis

Raja Mazumder, Ph.D.
Principal Investigator
Professor, GW
Chair, BioCompute Executive Steering Committee
mazumder@gwu.edu

Hadley King, M.S.
Operational Lead
Chair, BioCompute Technical Steering Committee
hadley_king@gwu.edu

Jonathon Keeney, Ph.D.
Co-Investigator
Assistant Research Professor, GW
Managing Director, BioCompute Executive Steering Committee
keeneyjg@gwu.edu

Janisha Patel, M.S.
Training Lead
Technical Writer
janishapatel@gwu.edu

Agenda

- Welcoming Remarks
- Introduction to BioCompute
- User Story: Athena DDL Pipeline
- Mock Evaluation of a Submission
- Usage Examples
  - Usability Domain
  - Extension Domain
  - Error Domain
- BCO Resources
- Use Case Gathering
- Q&A

BioCompute Objects

Goals of this Workshop

1. Introduce BioCompute Objects (BCO) for computational analysis
2. Explain BioCompute vocabulary
3. Introduce the application and utility of BCOs
4. Demonstrate how BCOs would be used in the context of FDA review of NGS data in regulatory submissions through a mock evaluation of a submission and additional use case examples.
5. Provide BioCompute resources for future reference

BioCompute Objects

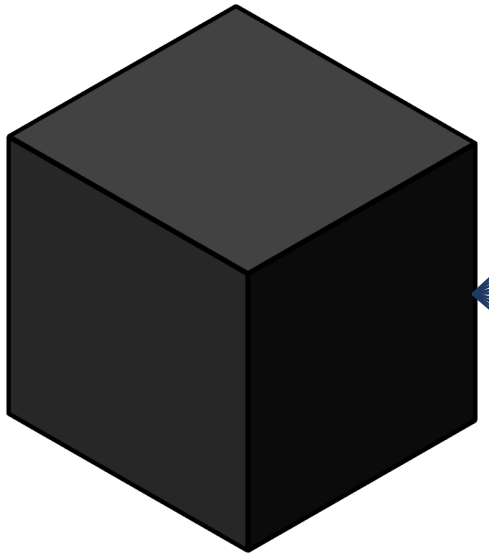
Introduction to BioCompute

BioCompute Objects

NGS Data Flows


```
$ fastq-dump -X 2 SRR001666 --split-3
Read 2 spots for SRR001666
Written 2 spots for SRR001666
$ head SRR001666_1.fastq SRR001666_2.fastq
==> SRR001666_1.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI

==> SRR001666_2.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

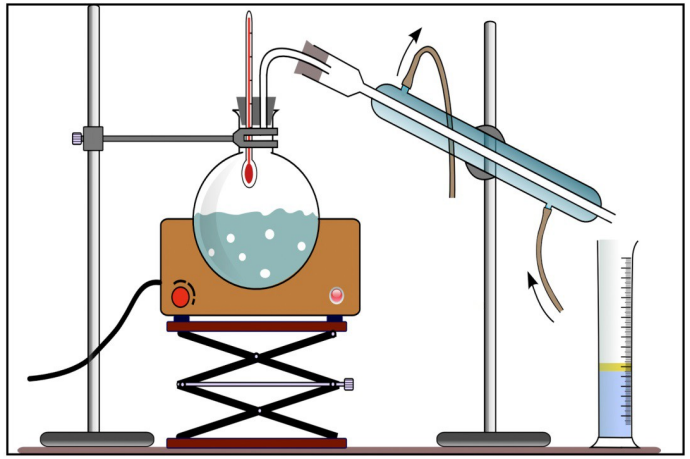


- Ancestry
- Cancer
- Microbiome
- Disease correlation
- Agriculture
- Synthetic biology
- Livestock
- Metagenomics
- Personalized medicine

Challenge: Workflow Communication

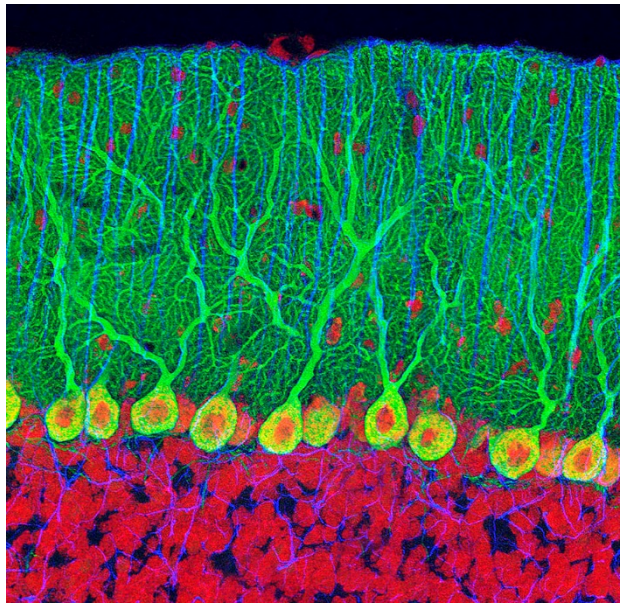
Input

Strain of animal
Tissue type
Protein Lot Number
Humidity
Microscope
…



conditions

Output

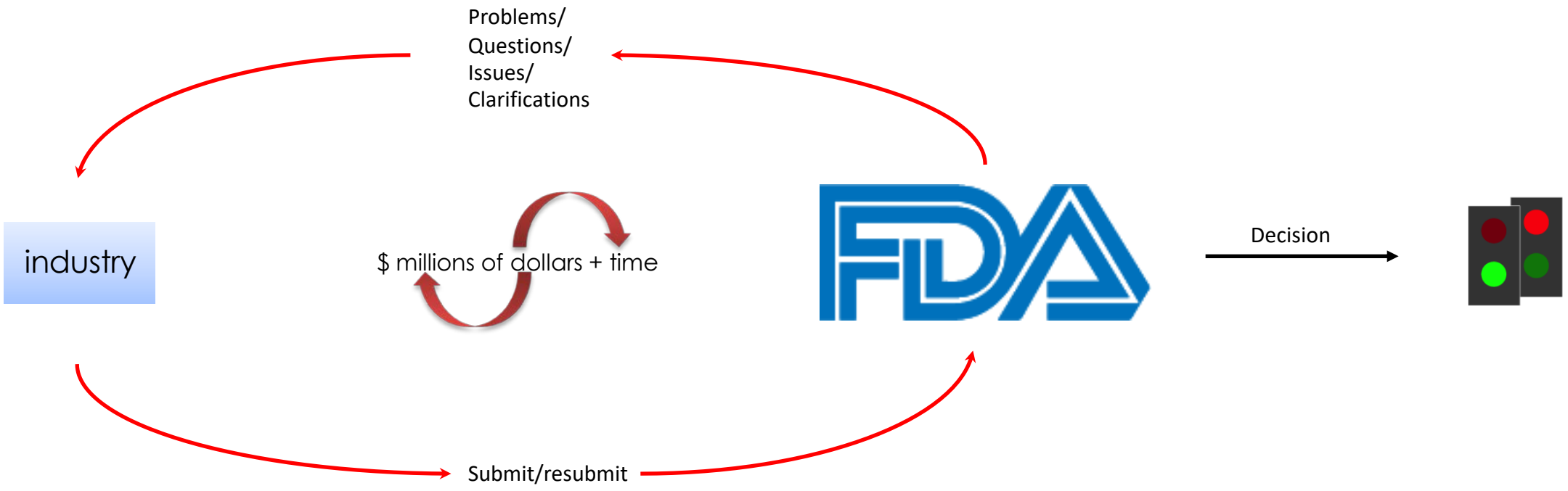


generalized experiment	inputs	parameters	outputs

```
> my_program -i input_file1 -parameter1 value1 -parameter2 value2 -o out_file
```

Analogy: wet lab experiments

Wasted Time and Money



BioCompute Objects

This is not a Guidance Document

DRAFT: Please provide comments and suggestions

Submitting Next Generation Sequencing Data to the Division of Antiviral Products Experimental Design and Data Submission

Acceptable Next Generation Sequencing Platforms

The division will accept Next Generation sequencing data generated from most standard Next Generation Sequencing (NGS) platforms provided the sponsor supplies the appropriate details for the sequencing platform, the protocols to be used for sample preparation, the raw NGS data, and the methods used to analyze the data. We recommend communicating with the division early in the process and providing these details prior to submitting the sequencing data. Please consider the following information when preparing your NGS submissions.

Data Transfer

1. **Portable hard drive**
    a. The raw NGS data in the fastq format should be sent to the division on a secured, portable hard drive following the guidelines outlined in this Guidance: http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM163567.pdf
    b. Please note that only the raw NGS data, the frequency table, and a table of contents should be contained on the hard drive. Additional files, such as those with a .exe extension may result in rejection of the submission. In addition, if the hard drive is password protected (not required or recommended at this time), please consult with the division ahead of time to ensure that the password is provided to the appropriate personnel in the document room.
    c. All additional data should be submitted via the electronic document gateway.

A solution should…

- Be **human readable**: like a GenBank sequence record
- Be **machine readable**: structured information with predefined fields and associated meanings of values
- Contain enough information to understand the computational pipelines, interpret information, maintain records, and reproduce experiments
- Be **immutable**: ensure information has not been altered

Solution: BioCompute

IEEE approved standard for communicating bioinformatic analysis workflows

- Acts like an envelope for entire pipeline
  - Can incorporate other standards (e.g. CWL)
- Built in collaboration with the FDA
- Human and machine readable
  - Written in JSON
- Categorized by domains
- Adheres to and encourages F.A.I.R. principles
  - Fully open source
- Adaptable
  - e.g. to other schemas
- Preserves data provenance
- Unique IDs for versioning

BioCompute Objects

Key Features of a BCO

- Abstract away workflow based on commonalities
  - Platform/tool/protocol independent
- Usability Domain
  - Free text description
- Data provenance
  - Data manifest, track files from beginning to end
  - Track user attribution (authored by, contributed by, reviewed by, etc.)
- Validation Kit
  - Error Domain + IO Domain
  - Sanity check: given the input files and the inherent error, is the output this analysis claims to have gotten valid?
- Extensible
  - Extension Domain
  - Open source repository
- Embargo Field
  - Prevent others from viewing a BCO for any amount of time

BioCompute Objects

BioCompute Object

```
Top Level
  BCO ID:  https://w3id.org/biocompute/1.3.0/examples/FDA-NA-TestsBreastCancer
  Checksum:  06DACE70679F35BA87A3DD6FFFED4ED24A4F5B8C2571264C37E5F1B3ADE04A31
  Specification: https://w3id.org/biocompute/1.3.0/
```

Metadata

```
Provenance Domain
  Name: FDA-NA-TestsBreastCancer
  Version: 1.0
  Review:
    approved: Natalie Abrams, NIH ; createdBy
  Created:  2018-05-24T09:40:17-0500
  Modified:  2018-06-21T14:06:14-0400
  Embargo:  Start: 2000-09-26T14:43:43-0400 End:  2000-09-26T14:43:45-0400
  Contributors:
    Janisha Patel (http://orcid.org/0000-0002-8824-4637), George Washington University; createdBy, modifiedBy
    Dara Baker, George Washington University; authoredBy
  License:  https://spdx.org/licenses/CC-BY-4.0.html --> licensing is inferred by OncoMX licensing. Pub=
```

Extension domain

```
Usability Domain
    FDA-approved or cleared nucleic acid-based human biomarker tests for breast cancer
    The .xlsx file FDA-NA-TestsBreastCancer.xlsx contains FDA-approved human biomarker tests for breast cancer.
    Each row represents one gene linked to its respective test. Genes are identified by      UniProtKB, HgncName, EDRN number
    Tests are distinguished by manufacturer, FDA submission ID(s), clinical trial ID(s) and PubMed ID(s).
```

Usability domain

```
Extension Domain
  Dataset Extension:
    Comment: Unique column headers for the dataset
    Test_disease_use: FDA-listed disease corresponding to approved test
      test_trade_name: FDA-listed product name
      test_manufacturfeer: FDA-listed patent company for the approved test
      sest_submission: FDA submission ID(s), web links; FDA-listed patent ID associated with test
      test_is_panel: A single biomarker or biomarker panel? Y for yes, N for no
      gene_symbol: HGNC_ID from https://www.genenames.org
      uniprotKB_ac: UniProtKB from https://www.uniprot.org
      biomarker_id: Matched to EDRN IDs based on HGNC Name
      biomarker_origin: Characteristic that makes this a biomarker; molecular abnormalities that can lead to cancer
      ncit_biomarker: Searchable terms for gene/Biomarker from NCI Thesaurus (NCIt)
```

Extension domain

```
Description Domain
  Keywords: cancer, breast cancer, biomarker, biomarker test, FDA, UniProtKB, EDRN
  External References: (Name, Namespace, Ids)
    PubMed; pubmed;
    UniProt; accession;
    EDRN; EDRN number;
    HGNC; HgncName;
    GTR; GTR terms;
  Platform:  Manual
  Pipeline Steps:
    Step 1: Download FDA-approved tests
    Description: FDA-approved tests were downloaded a list of FDA-approved or cleared nucleic acid based tests
    Input List:  https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm330711.htm
    Output List:  ~/FDA-approved-or-cleared-NA-based-tests
```

Description domain

```
Execution Domain
  Scripts: none
  Script Driver: manual
  Software Prerequisites: None
  External Data Endpoints:
      Name In Vitro Diagnostics > Nucleic Acid Based Tests
      URL   https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm330711.htm
      Name NCBI Genetic Testing Registry
      URL   https://www.ncbi.nlm.nih.gov/gtr/
  Environment Variables: None
```

Execution domain

```
Parametric Domain
      N/A
```

Parametric domain

```
Input/Output Domain
  Input Subdomain:
    Filename: Multiple test files from "Nucleic Acid Based Tests: List of Human Tests"
    Access Time: 2018-10-10T11:34:02-5:00
    URI: https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm330711.htm
  Output Subdomain:
    Filename: FDA-NA-TestsBreastCancer.xlsx
    Media Type: xlsx/csv
    Access Time: 2018-10-10T11:37:02-5:00
    URI: https://docs.google.com/spreadsheets/d/1xUY7WJNEZHyCgH5sYpxEuqAbtgVUUwgR2oc0IWhH28Y/edit#gid=1492026303
```

IO domain

```
Error Domain
```

Error domain

IEEE Standard

Institute of Electrical and Electronics Engineers Standard

IEEE 2791-2020 approved January 2020

https://standards.ieee.org/content/ieee-standards/en/standard/2791-2020.html

BioCompute Schema Files

https://opensource.ieee.org/2791-object/ieee-2791-schema/

ieee-2791-schema
Project ID: 116

24 Commits **2** Branches **3** Tags **276 KB** Files **276 KB** Storage **1** Release

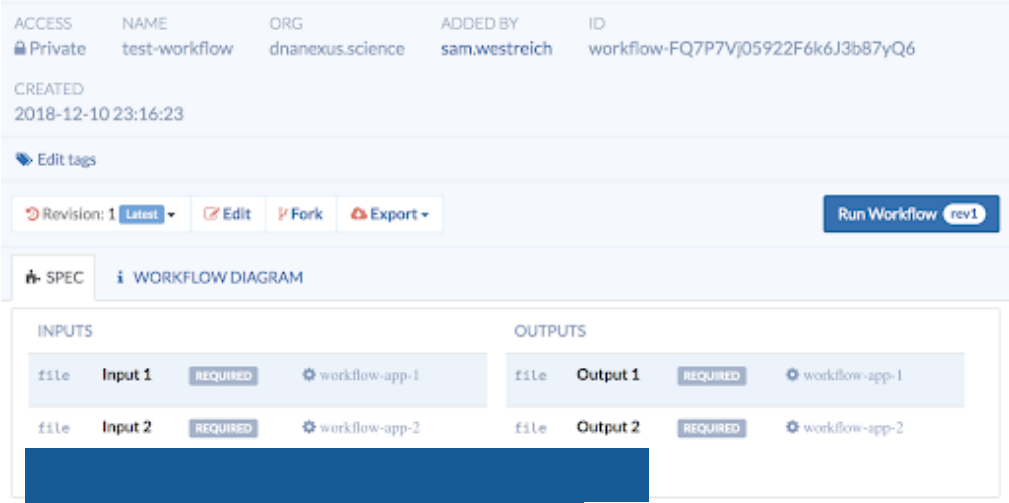
master | ieee-2791-schema | History | Find file | Clone

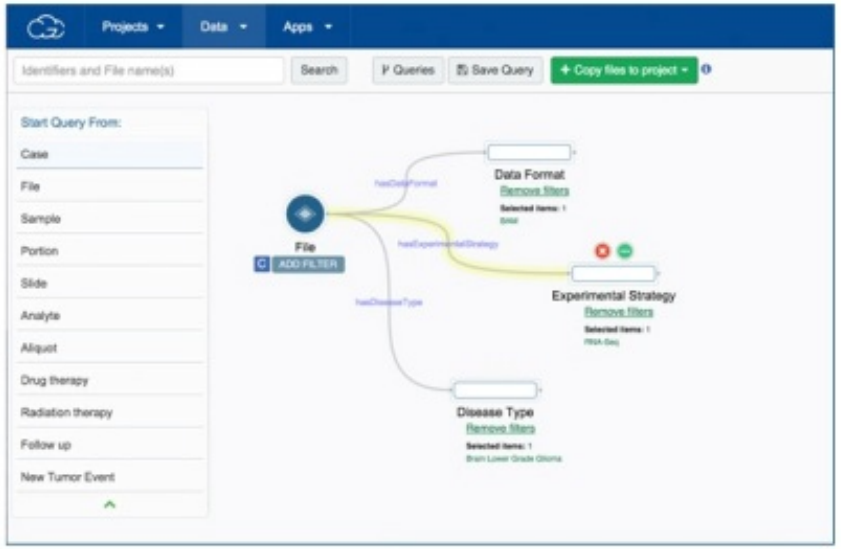
Update README.md
Joshua Gay authored 1 month ago
45683af9

README | BSD 3-clause "New" or "Revised" License

Name	Last commit	Last update
.gitignore	Creates initial release of BioCompute Object Schema in prep for ball...	1 year ago
2791object.json	replaces https://w3id.org/2791/ with https://w3id.org/ieee/ieee-279...	1 month ago
AUTHORS	Update AUTHORS	1 month ago
CONTRIBUTORS	Update CONTRIBUTORS	1 month ago
LICENSE	Update LICENSE	1 month ago

Platforms with BioCompute Integration

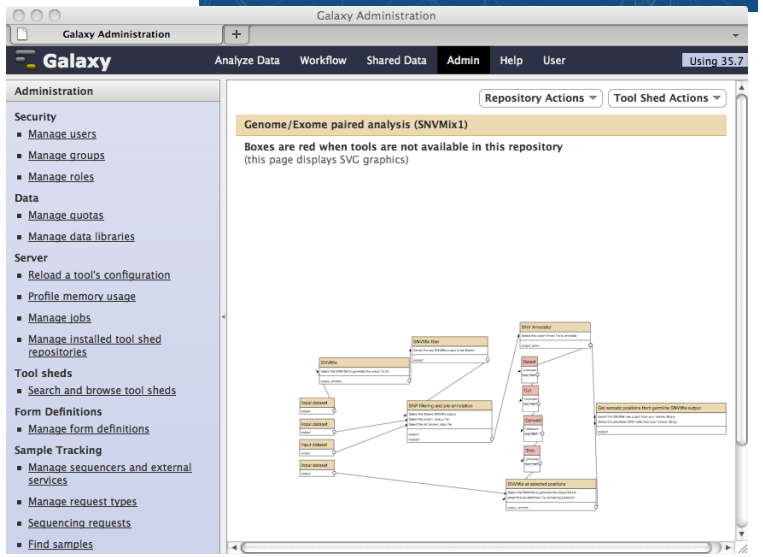


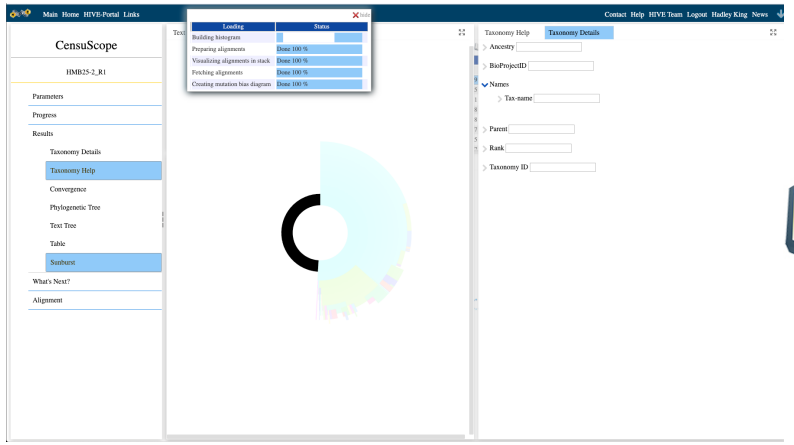


Seven Bridges

precisionFDA

DNAnexus
powering the genomics revolution™



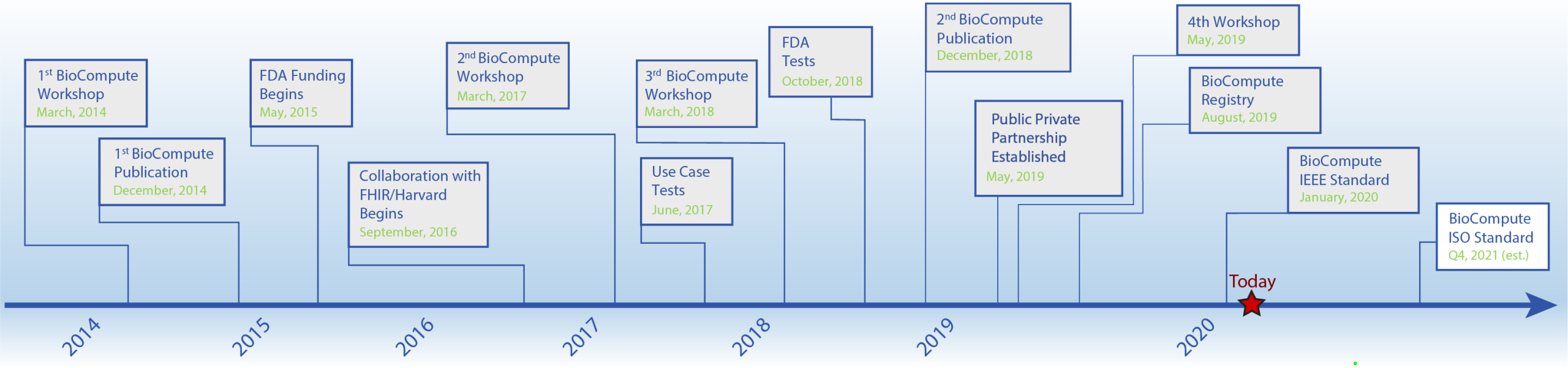


Galaxy PROJECT

HIVE

BioCompute Objects

BCO Timeline



BioCompute Objects

BioCompute participants

Children's National
precisionFDA
ATTAIN
Global Alliance for Genomics & Health
DISEASE ONTOLOGY
EPINOMICS
NIH NATIONAL CANCER INSTITUTE
CUROVERSE
Forum for Collaborative HIV Research
Google
CDC CENTERS FOR DISEASE CONTROL AND PREVENTION
EMBL-EBI
DNAnexus
HARVARD UNIVERSITY
NATIONAL LIBRARY OF MEDICINE
NCBI
FHIR Genomics
paradigm4
Oxford e-Research Centre
UNIVERSITY OF OXFORD
Seven Bridges
NIST
Galaxy PROJECT
NEW YORK GENOME CENTER
GENOME NEXT
CRITICAL PATH INSTITUTE
NIH
research object.org
HL7
CWL
FLATIRON
MANCHESTER 1824 The University of Manchester
elixir UNITED KINGDOM
CUNY

* Complete conformance check details are available on challenge site

BioCompute Object (BCO) App-a-thon

May 14 through October 18 2019

	Beginner Track: Create BCOs	Advanced Track: Create BCO software tools	App-a-thon Launch: May 14, 2019 Challenge details are made available
BCOs and Tools	Complete BCO	Conformance Tool; Writer Tool	**May 14 – October 18, 2019** Participants create BCOs and BCO tools
Conformance Check*	Ability of an FDA reviewer to understand the BCO	Ability to generate accurate BCO; Ability to conduct conformance check on BCOs	**Deadline: October 18, 2019** Evaluation begins
Results			Results released: March 2020

Introduction to BioCompute

HIVE precisionFDA THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC

Mock Clinical Trial

Athena DDL HCV1a Variant Profiling

BioCompute Objects

Proof of Concept

U.S. FOOD & DRUG ADMINISTRATION

DDL DIAGNOSTIC LABORATORY

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Proof of Concept:

Mimic real clinical trial FDA submission to determine if BioCompute could facilitate the submission process by:

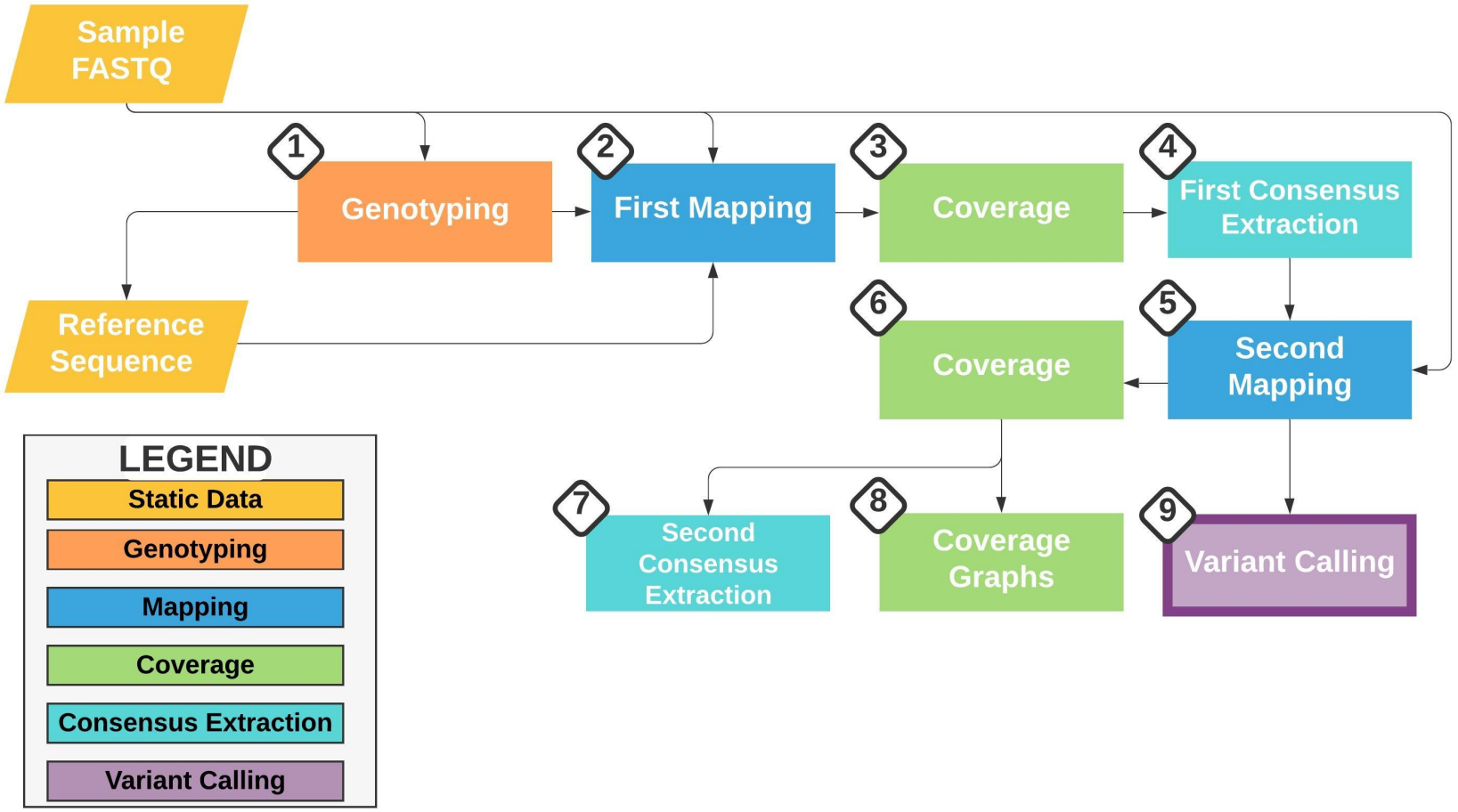
- Clearly communicating with regulatory agencies
- Aid to show the high-quality sequencing results appropriately

User Story

BioCompute Objects

DDL Athena NGS pipeline: viral drug resistance mutation analyses

Clinical trial Next-Gen Sequencing workflow



Pipeline:

MK-3682B in **Hepatitis C** (**GT1** or **GT3**) patients who have failed a DAA (Direct Acting Antiviral Regiment)

User Story

For more information on pipeline: https://clinicaltrials.gov/ct2/show/NCT02613403?term=MK-3682B&cond=HCV&rank=1

DDL Athena NGS pipeline: viral drug resistance mutation analyses

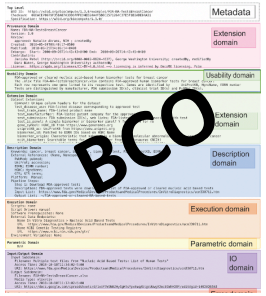
U.S. FOOD & DRUG ADMINISTRATION

DDL DIAGNOSTIC LABORATORY

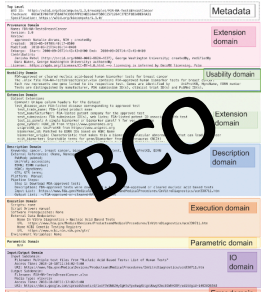
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Methods

- Replicate a real clinical trial using synthetically generated data made to resemble real biological data. Two separate analyses executed to simulate:
  1) pharmaceutical submission to the FDA
  2) simulate the FDA review



1. Athena BCO



2. FDA BCO

- BioCompute was utilized as the tool for communication of analysis and used for comparison of final results

BioCompute Objects

User Story

Mock Evaluation of a Submission

BioCompute Objects

Usage Examples

BioCompute Objects

Usage Example 1

Usability Domain

BioCompute Objects

Usability Domain: searchability and usability

BioCompute Usability Domain is used to **facilitate searchability** and **communicate BCO's usability**

```
    "usability_domain": [
        "Identify baseline single nucleotide polymorphisms (SNPs)[SO:0000694], (insertions)[SO:0000667], and (deletions)[SO:0000045] that correlate with reduced (ledipasvir)[pubchem.compound:67505836] antiviral drug efficacy in (Hepatitis C virus subtype 1)[taxonomy:31646]",
        "Identify treatment emergent amino acid (substitutions)[SO:1000002] that correlate with antiviral drug treatment failure",
        "Determine whether the treatment emergent amino acid (substitutions)[SO:1000002] identified correlate with treatment failure involving other drugs against the same virus",
        "GitHub CWL example: https://github.com/mr-c/hive-cwl-examples/blob/master/workflow/hive-viral-mutation-detection.cwl#L20"
    ],
```

Usability Domain

- Provides a space for the author to define the usability domain of the BCO
- helps determine when and how the BCO can be used.

Usage Example 1

BioCompute Objects

Usability Domain: QC Steps


```
{
    "bco_id": "http://biocomputeobject.org/BCO_000563",
    "e-tag":"853d1471120527093ef2728417d9f9cc1d7275b5f64ab7396e714ebe5d4b6fb8",
    "bco_spec_version": "1.3.0",
    "provenance_domain": {
        "name": "Comparative abundance of microbial strains associated with
diet change in epileptic patients",
        "version": "1.0",
        "license": "https://spdx.org/licenses/CC-BY-4.0.html",
        "created": "2019-12-10T18:30:04.008460",
        "modified": "2019-12-12T20:43:58.007411",
        "review": [
            {
                "status": "reviewed",
                "reviewer_comment": "Approved by GW Staff.",
                "reviewer": {
                    "orcid": "https://orcid.org/0000-0002-8824-4637",
                    "affiliation": "George Washington University",
                    "contribution": [
                        "curatedBy"
                    ],
                    "name": "Janisha Patel",
                    "email": "janishapatel@gwu.edu"
                },
                "date": "2019-03-10"
            }
        ],
```


Comparative abundance of microbial strains associated with diet change in epileptic patients

Step 1: CensuScope – MAPPING

Manual QC Steps → GitHub

Step 2: Hexagon – ALIGNMENT

Usage Example 1

BioCompute Objects

Usability Domain: Metagenomic Analysis


```
    "usability_domain": [
        "This pipeline is part of a larger study to look for a potential causal link between the ketogenic diet (KD) and seizure response, by evaluating metagenomic data. The pipeline described in this BCO identifies the relative abundance of bacterial strains in epileptic patients. The first step of this pipeline uses CensuScope [PMID: 25336203], which identifies taxonomic compositions in each sample, followed by QC steps that can build a reference database. Lastly, Hexagon [PMID: 24918764] is used to align the reads and generate a profile of abundances for each sample. The data generated from this bioinformatics pipeline were subsequently used to construct predictive models that ultimately examine three key questions 1) is there enough signal to separate a person on ketogenic diet (KD) from a person that is not on a KD, 2) are there specific organisms that contribute to the KD signal, and 3) can a patient's response to KD be predicted prior to starting a KD diet?",
```


Manual QC steps can be be represented in both Usability Domain OR Description Steps


```
    "pipeline_steps": [
        {
            "name": "CensuScope",
            "version": "Albinoni 2.2.8",
            "step_number": 1,
            "input_list": [...],
            "output_list": [...],
            "prerequisite": [...],
            "description": "CensuScope is used as a taxonomic identifier of microbial communities."
        },
        {
            "name": "Hexagon",
            "version": "Albinoni 2.2.8",
            "step_number": 2,
            "input_list": [...],
            "output_list": [...],
            "prerequisite": [],
            "description": "Hexagon is used to generate a "hit list" and
```


…provided the data is not changed as a result of these QC steps!

Usage Example 1

Usage Example 2

Extension Domain

BioCompute Objects

Extension Domain: Expose all Parameters

- Reveal all parameters
  - A full parameter list can be used to show all the parameters that were changed, not just the default parameters that are in the base BCO.
  - Can also be used for application-specific needs.

```
“parametric_domain”: [
     {
        “is_default”: true,
        “param”: “Minimum match length”,
        “value”: 50,
        “step”: 1
     },
     {
        “is_default”: false,
        “param”: “Conflict resolution method”,
        “value”: “Markovnikov rule”,
        “step”: 1
     }
 ]
```

Usage Example 3

Extension Domain: Bibliography Domain

- Separate container for references
  - Write for desired style (e.g. APA)

```
“bibliography_domain”: [
        {
            “journal-article-title: {
                "A perspective on judgement and
                choice: Mapping bounded
                rationality.“
            }
        },
        {
            “journal-article-authors: {
                [“Daniel Kahneman”]
            }
        },
        {
            “journal-article-journal_name: {
                “American Psychologist.”
            }
        },
```

Usage Example 3

Extension Domain: Supplementary Domain

- Container for additional content
  - Not required for computational analysis
  - Possibly still relevant for analysis comprehension

```
“supplementary_domain”: [
   {
      “appendix_a”:
https://docs.google.com/spreadsheets/d/1B3BrdD2ypRT0jk1wHyWU9xciVCyExGBruWw-
txazE9s/edit?usp=sharing,
      “description”: “Google drive template for a Gantt Chart”
   }
]
```

Usage Example 3

Usage Example 3

Error Domain

BioCompute Objects

Error Domain: acceptable range of variability


```
"error_domain": {
    "empirical_error": {
        "definitions": {...},
        "M414T_baseLine": {
            "percentage": "0.03",
            "reads_generated": "4823",
            "coverage": "150",
            "mutation_call_prob_Athena": "1",
            "AthenaREADCOUNT": "144",
            "AthenaCOVERAGE": "5094",
            "AthenaPERCENTAGE": "0.02827",
            "AthenaQUALITY": "33.16",
            "AthenaFCOUNT": "66",
            "AthenaRCOUNT": "78",
            "AthenaFRSCORE": "0.1388",
            "STDEV.P": "0.000865"
        },
        "M28T_baseLine": {...},
        "D168Y_baseLine": {...},
        "D168A_baseLine": {...},
        "S556G_baseLine": {...},
        "WT_baseLine": {...},
        "M28S_baseLine": {...},
        "Q30R_baseLine": {...},
        "C316N_baseLine": {...}
    },
    "algorithmic_error": {
        "AthenaFRSCORE_threshold": 0.5,
        "AthenaQUALITY": 25,
        "AthenaCOVERAGE": 5000
    }
}
```


BioCompute Error Domain is used to evaluate a pipeline's **accuracy** and **precision**

- Range of outputs that are within a defined tolerance level
- Can be used to optimize or verify algorithm
- Consists of two subdomains:

  empirical and *algorithmic*

BioCompute Objects

Error Domain: empirical error

	percentage	reads	coverage	READCOUNT	Athena %	QUALITY	STDEV.P
D168A_basel	0.0005	80	2.5		0		0.00025
D168Y_basel	0.011	1768	55	63	0.01229	33.56	0.000645
M28T_baseLi	0.01	1608	50	43	0.00841	34.09	0.000795
M28S_baseLi	0.08	12861	400	357	0.06985	33.97	0.005075
Q30R_baseLi	0.0008	129	4	7	0.00136	32	0.00028
C316N_basel	0	0	0		0		0
M414T_base	0.03	4823	150	144	0.02827	33.16	0.000865
S556G_basel	0	0	0		0		0
WT_baseLine	0.8677	139497	4338.5		0.87982		0.00606

Contains empirically determined values such as:

- limits of detectability
- false positive rates
- false negatives rates
- statistical confidence of outcomes

Usage Example 2

BioCompute Objects

Error Domain: empirical error

	percentage	reads_gene	coverage	AthenaCOVE	AthenaPERCI	AthenaQUAL	STDEV.P
D168A_baseLine	0.0005	80	2.5		0		0.00025
D168Y_baseLine	0.011	1768	55	5126	0.01229	33.56	0.000645
M28T_baseLine	0.01	1608	50	5111	0.00841	34.09	0.000795
M28S_baseLine	0.08	12861	400	5111	0.06985	33.97	0.005075
Q30R_baseLine	0.0008	129	4	5163	0.00136	32	0.00028
M414T_baseLine	0.03	4823	150	5094	0.02827	33.16	0.000865
S556G_baseLine	0	0	0		0		0
WT_baseLine	0.8677	139497	4338.5		0.87982		0.00606


```
"error_domain": {
    "empirical_error": {
        "definitions": {...},
        "M414T_baseLine": {
            "percentage": "0.03",
            "reads_generated": "4823",
            "coverage": "150",
            "mutation_call_prob_Athena": "1",
            "AthenaREADCOUNT": "144",
            "AthenaCOVERAGE": "5094",
            "AthenaPERCENTAGE": "0.02827",
            "AthenaQUALITY": "33.16",
            "AthenaFCOUNT": "66",
            "AthenaRCOUNT": "78",
            "AthenaFRSCORE": "0.1388",
            "STDEV.P": "0.000865"
        },
        "M28T_baseLine": {...},
        "D168Y_baseLine": {...},
        "D168A_baseLine": {...},
        "S556G_baseLine": {...},
        "WT_baseLine": {...},
        "M28S_baseLine": {...},
        "Q30R_baseLine": {...},
        "C316N_baseLine": {...}
    },
```


Can be measured by:

- running the algorithm on multiple data samples of the usability domain
- carefully designed in-silico data.

For example:

In-silico samples run through the pipeline to determine the false positives, negatives, and limits of detection.

Usage Example 2

BioCompute Objects

Error Domain: algorithmic error

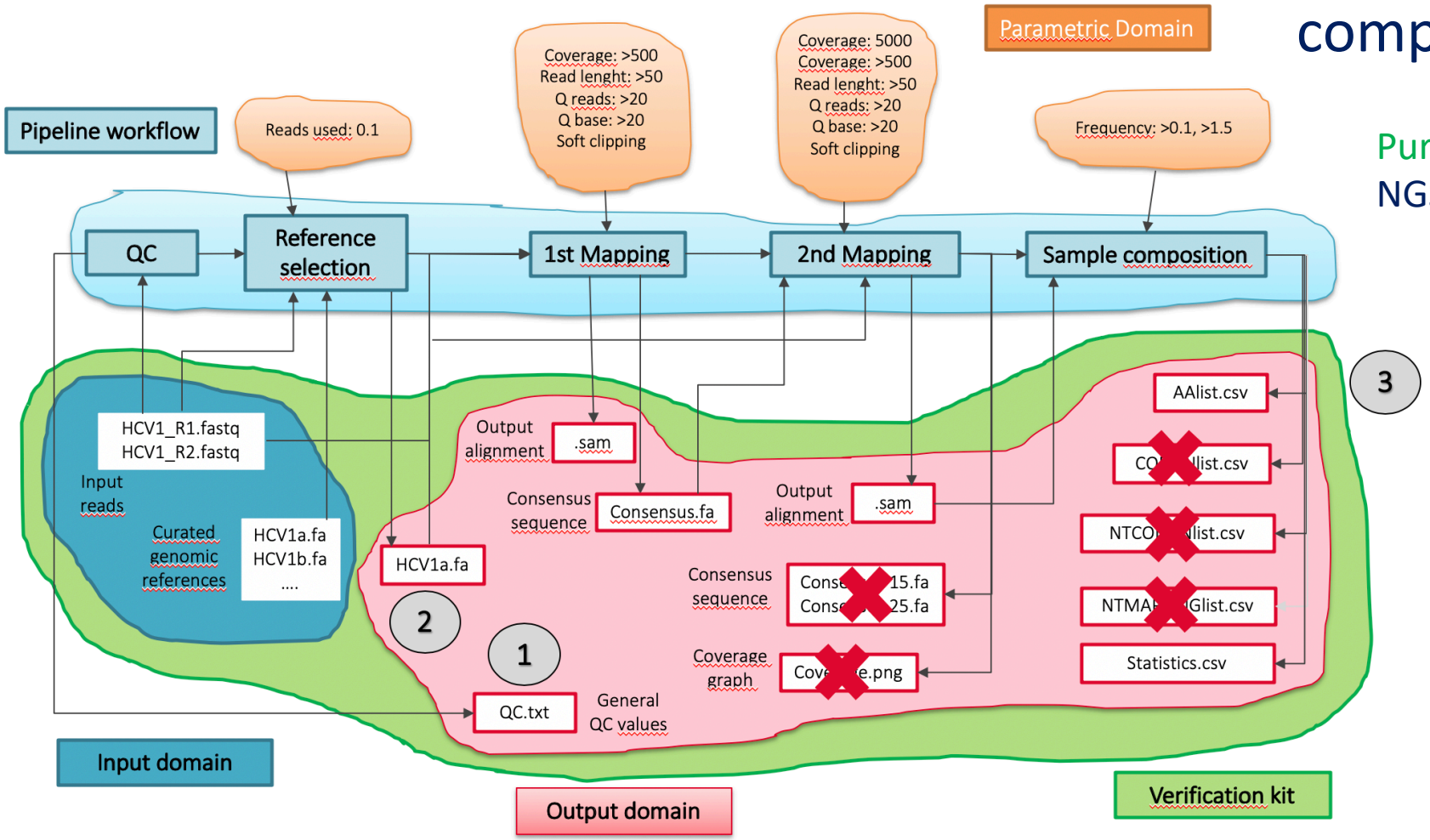
- Descriptive of errors that originate by:
  - fuzziness of the algorithms
  - driven by stochastic processes in dynamically parallelized multi-threaded executions
  - in machine learning methodologies where the state of the machine can affect the outcome.
- This can be measured by:
  - re-running analysis on random subset of the data
  - modeling of accumulated errors to generate confidence values.
- For example, bootstrapping is frequently used with stochastic simulation-based algorithms to estimate statistically significant variability for the results.

```
    "algorithmic_error": {
        "AthenaFRSCORE_threshold": 0.5,
        "AthenaQUALITY": 25,
        "AthenaCOVERAGE": 5000
```

Usage Example 2

BioCompute Objects

Verification Kit



The IO and Error Domain compose the VERIFICATION KIT

Purpose: to demonstrate the accuracy of NGS data analysis workflow

Includes:

- A small set of input and output files
- Complete BCO with Error Domain

Yields:

- An easy way to verify a pipeline for replication and understanding
- Confidence in results reported by pipeline

Usage Example 2

BioCompute Objects

BCO Resources

BioCompute Objects

Cheat Sheet

8 Top Level Domains	
Provenance Domain: Metadata describing the BCO	
Usability Domain: Free text field for researcher to explain the analysis and relevant details.	
Extension Domain: User-defined fields	
Description Domain: Steps of the analysis, external resources needed for the steps, and the relationship of I/O objects	
Execution Domain: Information about the environment in which the analysis was run	
Parametric Domain: Records any parameters that were changed from default values	
Input and Output Domain: A list of global input and output files	
Error Domain: Used for describing errors. Can include the limits of detectability, false positives, false negatives, statistical confidence of outcomes, and description of errors	
Required	**Optional**

Quick Checks


```
{
  "etag": "d516a923967ec1f8ee4bc666a2256bb91b3e035a91a1f5ef64b92e33ad23a104",
  "object_id": "https://beta.portal.aws.biochemistry.gwu.edu/bco/BCO_00016484",
  "spec_version": "https://w3id.org/ieee/ieee-2791-schema/",
  "provenance_domain": {
    "embargo": {},
    "name": "Regulatory BCO for hepatitis C virus resistance analysis",
    "version": "3.0",
```


Always a unique ID

Always this URL

Top level "Domain"

Check for open visibility

Check for modifications

Guidelines

- "bco_id" may have user specific values
- (e.g. "FDA_00001" or "GWU_01A")
- Use Extension Domains to ask for more project specific information
- Use Verification Kit to quickly check validity of results
- Steps that do not transform data (e.g. column sorting) can be described in the Usability Domain instead of as a full step in the Description Domain, at the Reviewer's discretion
- Use IO Domain as a manifest for all data files

Resources:
Website: https://biocomputeobject.org/
Official Standard: https://standards.ieee.org/content/ieee-standards/en/standard/2791-2020.html
Open source repository: https://opensource.ieee.org/2791-object/ieee-2791-schema
Contact: keeneyjg@gwu.edu, hadley_king@gwu.edu, janishapatel@gwu.edu, mazumder@gwu.edu

BCO Cheat Sheet

Front side

For FDA Reviewers and administrators

Examples for User-Defined Extension Domains

The extension domain is a user-defined field that can be used to include additional information not recorded anywhere else in the BCO.

Example 1: Expose All Parameters

A full parameter list can be used to show all the parameters that were changed, not just the default parameters that are in the base BCO. It can also be used for application-specific needs.

```
“parametric_domain”: [
    {
        “is_default”: true,
        “param”: “Minimum match length”,
        “value”: 50,
        “step”: 1
    },
    {
        “is_default”: false,
        “param”: “Conflict resolution method”,
        “value”: “Markovnikov rule”,
        “step”: 1
    }
 ]
```

Example 2: Bibliography Domain

A bibliography domain can be created to keep track of all references, and to give them an identifier to refer to in the BCO.

```
{
  "title": “On the Tendency of Varieties to Depart Indefinitely from the Original Type",
  “book-authors": { “examples”: [“Alfred Russel Wallace, Charles Darwin"] }
},
```

Example 3: Supplementary Domain

A supplementary domain can be created to track any files that are not input/output files, but which may be relevant for understanding the BCO.

```
“supplementary_domain”: [
   {
      “appendix_a”: https://docs.google.com/spreadsheets/d/1B3BrdD2ypRT0jk1wHyWU9xciVCyExGBruWw-txazE9s/edit?usp=sharing,
      “description”: “Google drive template for a Gantt Chart”
   }
]
```

Resources:

Extension Domains are custom designed to specific needs, and will vary heavily by need and user experience with developing and working with schemas. The BioCompute team can help develop custom Extension Domains for specific needs.

Contact: keeneyjg@gwu.edu, hadley_king@gwu.edu, janishapatel@gwu.edu, mazumder@gwu.edu

Back side

BCO Cheat Sheet

For FDA Reviewers and administrators

Ontology for Contributors

Ciccarese *et al*. https://doi.org/10.1186/2041-1480-4-37

"authoredBy"	Agent that originated or gave existence to the work that is expressed by the digital resource.
"contributedBy"	Agent that provided any sort of help in conceiving the work that is expressed by the digital artifact.
"createdAt"	The geo-location of the agents when creating the resource.
"createdBy"	Agent primarily responsible for encoding the digital artifact or resource representation. This creation is distinct from forming the content, which is indicated with *pav:contributedBy* or its subproperties.
"createdWith"	The software/tool used by the creator (pav:createdBy) when making the digital resource.
"curatedBy"	Agent specialist responsible for shaping the expression in an appropriate format. Often the primary agent responsible for ensuring the quality of the representation.
"derivedFrom"	Derived from a different resource.
"importedBy"	An agent responsible for importing data from a source given by *pav:importedFrom*.
"importedFrom"	Original source of imported information.
"providedBy"	Original provider of the encoded information (e.g. PubMed).
"retrievedBy"	Entity responsible for retrieving the data from an external source (usually a software entity).
"retrievedFrom"	The URI where a resource has been retrieved from.
"sourceAccessedBy"	agent who accessed the source.

Versioning

https://semver.org/

2.4.11

Versioning is based on "**Semantic Versioning**"

major.minor.patch

Patch

Use for editorial corrections

Minor

Use for addition of material, like a reviewer block

Major

Not allowed in BCO, value omitted

```
{
  "reviewer": {
    "name": "Josiah Carberry",
    "affiliation": "FDA",
    "email": "jcarberry@fda.hhs.gov",
    "contribution": [
      "curatedBy"
    ],
    "orcid": https://orcid.org/0000-0002-1825-0097
  },
  "status": "reviewed"
}
```

Home About Services Organization BCO Specification Resources Contact

BioCompute Objects

BioCompute is a standardized way to communicate an analysis pipeline. BioCompute substantially improves the clarity and reproducibility of an analysis, and can be packaged with other standards, such as the Common Workflow Language. An analysis that is reported in a way that conforms to the BioCompute specification is called a BioCompute Object (BCO). A BCO abstracts the properties of an analysis away from any specific platform, tool or goal. A BCO is broken down into conceptually meaningful "Domains" for capturing relevant information about the analysis pipeline.

The open source repository for the project can be accessed here. Several tools have been developed to read or write an analysis as a BCO. The most popular ones are below. Other resources can be found here.

HIVE powered by aws

powered by aws

Galaxy PROJECT powered by aws

https://biocomputeobject.org/

Use Case Gathering

BioCompute Objects

Use-Case Examples

Test Submission

- HCV-1a use case using synthesized data
- What data are necessary to make a regulatory decision?
- Are summary data from one analysis pipeline sufficient?
- How will the analysis pipeline be validated?

Tuberculosis Detection

- Tuberculosis (TB) is top infectious killer in the world
- WHO is adopting ReSeqTB pipeline to address the many challenges of detecting TB
- Requires lineage identification, prediction of antibiotic resistance, recurrence of TB in previously treated patients

Embleema

- Embleema is a platform that allows users to take control of their own data
- Marketplace for directly selling personal genome data
- Aggregator for Real World Evidence

Use-case gathering

Q & A

BioCompute Objects

Acknowledgments

U.S. FOOD & DRUG ADMINISTRATION

Eric Donaldson
Mark Walderhaug
Carolyn Wilson
Anton Golikov

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Raja Mazumder
Brian Fochtman

BioCompute Objects

Vahan Simonyan
Dennis Dean
Jeremy Goecks
Gil Alterovitz
Carole Goble
Jonas Almeida
Dan Taylor

Ntino Krampis
Michael Crusoe
Stian Soiland-Reyes
Konstantinos Krampis
Elaine Thompson
Nicola Soranzo
Jason Travis

DDL DIAGNOSTIC LABORATORY

Nuria Guimera
Souvik Das

Contact

Raja Mazumder, Ph.D.
Principal Investigator
Professor, GW
Chair, BioCompute Executive Steering Committee
mazumder@gwu.edu

Jonathon Keeney, Ph.D.
Co-Investigator
Assistant Research Professor, GW
Managing Director, BioCompute Executive Steering Committee
keeneyjg@gwu.edu

Hadley King, M.S.
Operational Lead
Chair, BioCompute Technical Steering Committee
hadley_king@gwu.edu

Janisha Patel, M.S.
Training Lead
Technical Writer
janishapatel@gwu.edu

BioCompute Objects

Thank you!

Your time and feedback are greatly appreciated.
Project specific feedback will be hosted here:
https://hive.biochemistry.gwu.edu/confluence/display/BUW/BioCompute+Workshop

BioCompute Objects