



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

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

Guidelines

1. Please turn off video
2. Please mute
3. Unmute for questions or post in chatbox
4. Please use Internet Explorer for compatibility with Adobe Connect

Thank you!

Agenda

- Introduction to **BioCompute** (20min)
Q/A(5min)
- User Story: Athena DDL Pipeline (10min)
Q/A (5min)
- Mock Evaluation of a Submission (10min)
Q/A (5min)
- Usage Examples
 - Usability Domain (10min)
 - Error Domain (10min)
 - Extension Domain (10min)
- Use Case Gathering (20 min)
- Q&A (15min)

Goals of this Workshop

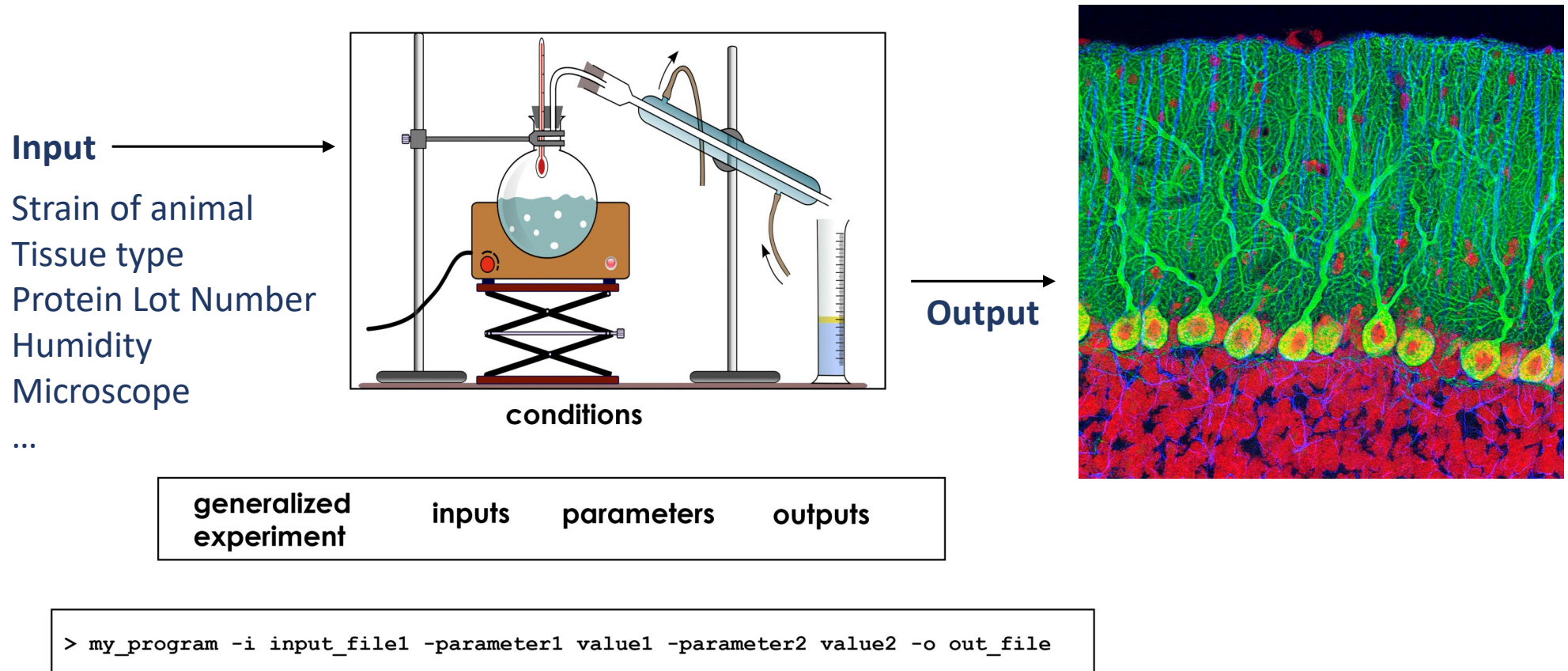
1. Introduce BioCompute Objects (BCO) for computational analysis
2. Explain BioCompute vocabulary
3. Introduce the application and utility of BCOs
4. Demonstrate how BCOs would be used in the context of FDA review of NGS data in regulatory submissions through a mock evaluation of a submission and additional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

Introduction to BioCompute



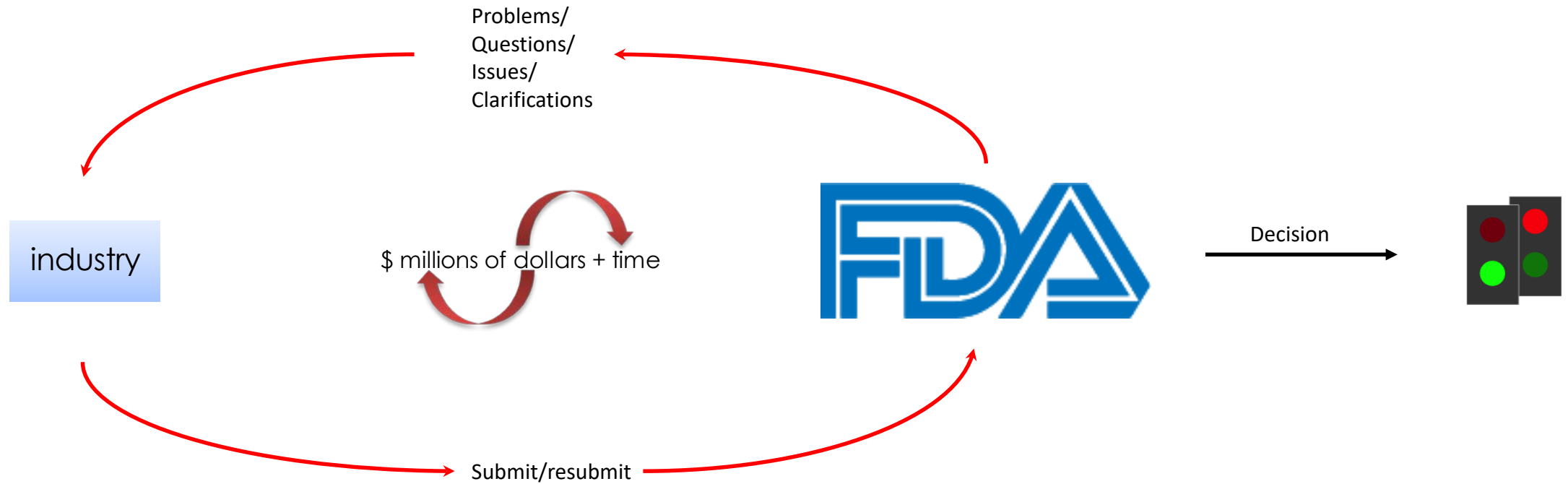
BioCompute
Objects

Challenge: Workflow Communication



Analogy: wet lab experiments

Wasted Time and Money



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

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 Domain**
 - Prevent others from viewing a BCO for any amount of time



BioCompute Object

Top Level

BCO ID: <https://w3id.org/biocompute/1.3.0/examples/FDA-NA-TestsBreastCancer>
Checksum: 06DACE70679F35BA87A3DD6FFED4ED24A4F5B8C2571264C37E5F1B3ADE04A31
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, EDNR 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_manufacturer: FDA-listed patent company for the approved test
test_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 EDNR 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, EDNR
External References: (Name, Namespace, Ids)
PubMed; pubmed;
UniProt; accession;
EDNR; EDNR 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/1xUY7WJNEZHyCgH5YpxEuqAbtgVUUwgr2oc0IWhH28Y/edit#gid=1492026303>

IO
domain

Error Domain

Error domain

BioCompute Schema Files



ieee-2791-schema

Project ID: 116

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

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-2791-schema/	1 month ago
AUTHORS	Update AUTHORS	1 month ago
CONTRIBUTORS	Update CONTRIBUTORS	1 month ago
LICENSE	Update LICENSE	1 month ago

BioCompute Schema Files

iee-2791-schema

Project ID: 116

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

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

master

iee-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-2791-...>

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

ACCESS: Private | NAME: test-workflow | ORG: dnanexus.science | ADDED BY: sam.westreich | ID: workflow-FQ7P7Vj05922F6k6J3b87yQ6

CREATED: 2018-12-10 23:16:23

Buttons: Edit tags, Revision: 1 Latest, Edit, Fork, Export, Run Workflow rev1

Inputs: file Input 1 (REQUIRED) workflow-app-1, file Input 2 (REQUIRED) workflow-app-2

Outputs: file Output 1 (REQUIRED) workflow-app-1, file Output 2 (REQUIRED) workflow-app-2

Identifiers and File name(s): Search | F Queries | Save Query | Copy files to project

Start Query From:

- Case
- File
- Sample
- Portion
- Slide
- Analyte
- Align
- Drug therapy
- Radiation therapy
- Follow up
- New Tumor Event

Workflow Diagram: File (ADD FILTER) -> Data Format (Remove filters) -> Experimental Strategy (Remove filters) -> Disease Type (Remove filters)



Task	Progress
Building histograms	100%
Preparing alignments	100%
Visualizing alignments in track	100%
Feeding alignments	100%
Creating mutation list digest	100%

Results: Taxonomy Details, Convergence, Phylogenetic Tree, Text Tree, Table, Starburst

What's Next? Alignment



Galaxy Administration | Administration | Security | Data | Server | Tool sheds | Form Definitions | Sample Tracking

Repository Actions | Tool Shed Actions

Genome/Exome paired analysis (SNVMix1)

Boxes are red when tools are not available in this repository (this page displays SVG graphics)



BioCompute participants





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. Major features of the BioCompute project include a "Usability Domain" for free text description by the researcher, strong data provenance and user attribution, a "Validation Kit" for quickly verifying the output of an analysis, highly extensible through a user-defined "Extension Domain," and an "Embargo Domain" for sensitive analyses not to be made public yet. See the [About](#) page for more information.

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](#).



powered by aws



powered by aws



powered by aws

BioCompute Portal



Welcome to the BCO Editor, a platform-free, web-based form for creating BioCompute Objects (BCOs). For more information, see the [BioCompute Website](#), the [official IEEE standard](#), and the [open source repository](#) for all schema files.

Sign in

Email address

Password

[SIGN IN NOW](#)

Don't have an account? [Sign up](#)
[Forgot Password?](#)

<https://portal.aws.biochemistry.gwu.edu/sign-in>

BioCompute Object (BCO) App-a-thon

May 14 through October 18 2019



Introduction to BioCompute



Integrating with Other Standards

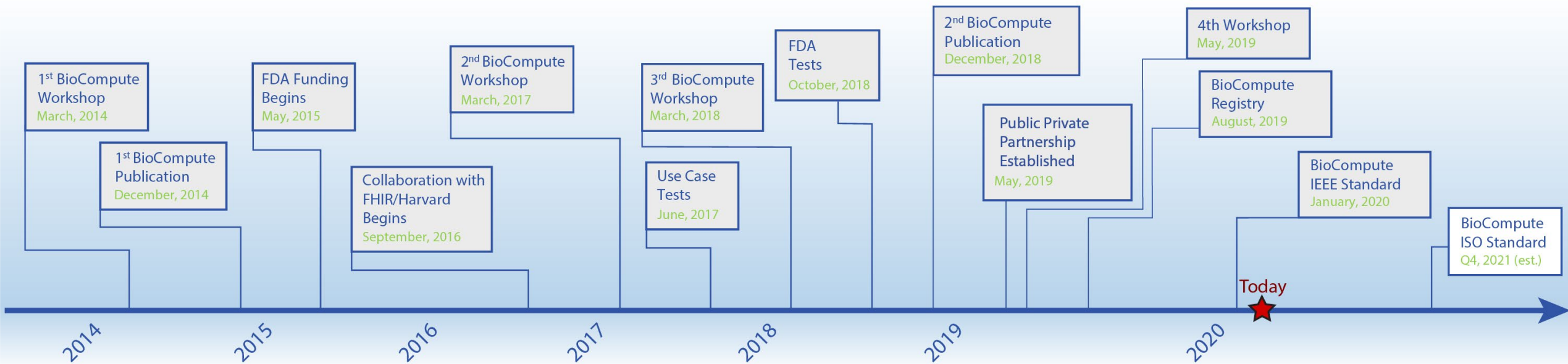


Institute of Electrical and Electronics
Engineers Standard

BioCompute P2791-2020 approved January
2020

<https://standards.ieee.org/content/ieee-standards/en/standard/2791-2020.1>

BCO Timeline



User Story

Athena DDL Pipeline



BioCompute
Objects

DDL Athena NGS pipeline: viral drug resistance mutation analyses



**THE GEORGE
WASHINGTON
UNIVERSITY**

WASHINGTON, DC

Pipeline:

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

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

DDL Athena NGS pipeline: viral drug resistance mutation analyses



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

Mock Evaluation of a Submission



BioCompute
Objects

[Keeney]

Usage Examples



BioCompute
Objects

Usability Domain

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

Step 1 CensuScope – MAPPING

~Manual QC Steps~

Step 2: Hexagon – ALIGNMENT

How should manual QC steps be represented?

```
{
  "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"
    ]
  }
}
```

Error Domain: acceptable range of variability

BioCompute Error Domain is used to evaluate a pipeline's ACCURACY & 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*.

```
"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  
    }  
  }  
}
```

Error Domain: empirical error

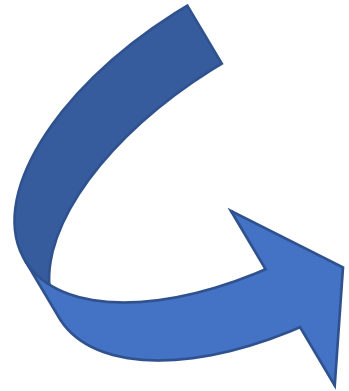
	percentage	# of reads	coverage	Athena%	AthenaQUALITY	STDEV.P
D168A_baseLine	0.0005	80		0		0.00025
D168Y_baseLine	0.011	1768	5126	0.01229	33.56	0.000645
M28T_baseLine	0.01	1608	5111	0.00841	34.09	0.000795
M28S_baseLine	0.08	12861	5111	0.06985	33.97	0.005075
Q30R_baseLine	0.0008	129	5163	0.00136	32	0.00028
M414T_baseLine	0.03	4823	5094	0.02827	33.16	0.000865
S556G_baseLine	0	0		0		0
WT_baseLine	0.8677	139497		0.87982		0.00606

Contains empirically determined values such as:

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

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.

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
```

Verification Kit

The IO and Error Domain compose the **VERIFICATION KIT**

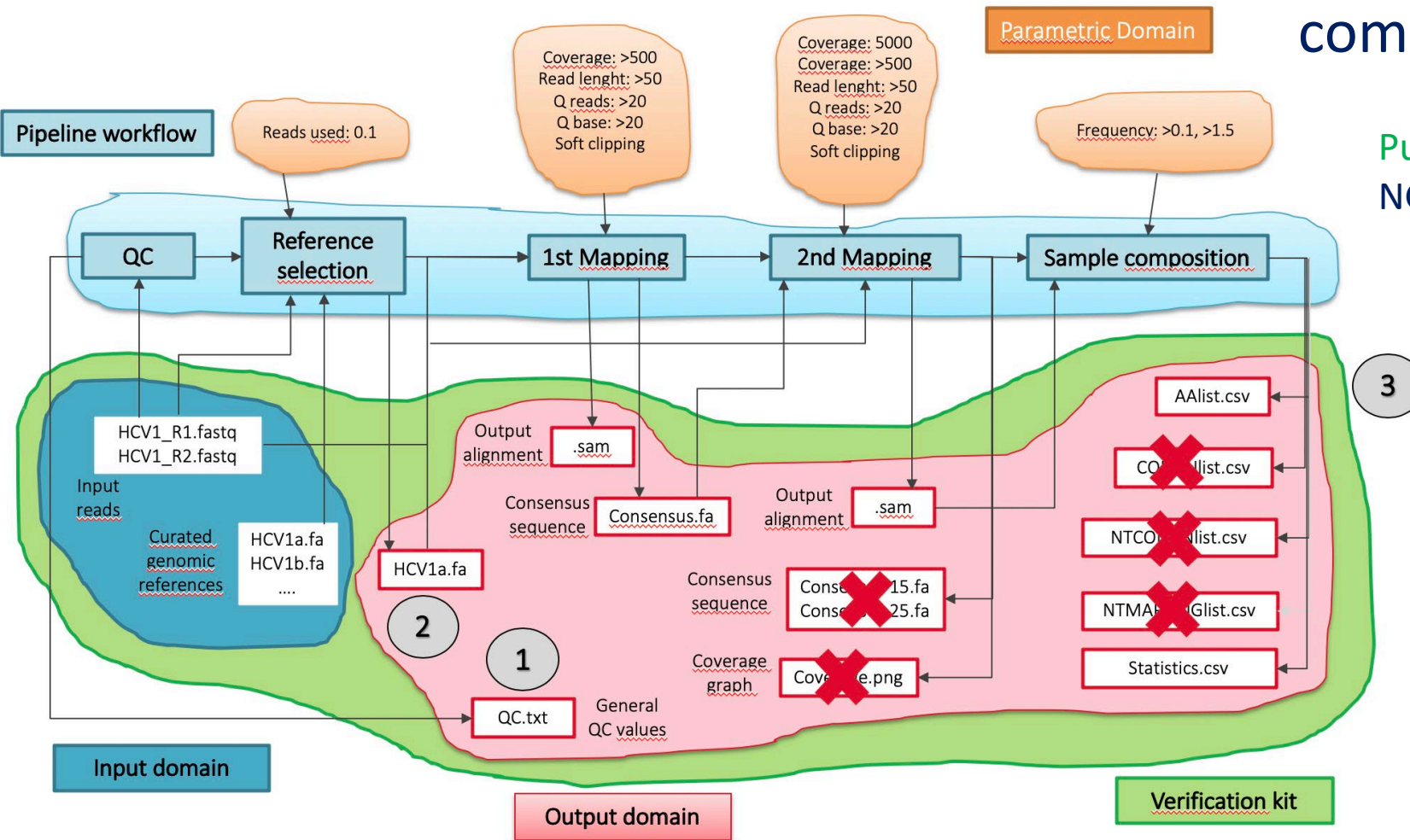
Purpose: to demonstrate the accuracy of NGS data analysis workflow

Includes:

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

Yields:

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



Extension Domain

[Keeney]

Use Case Gathering



BioCompute
Objects

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

```
{
  "bco_id": 416356,
  "bco_spec_version": "1.4",
  "bco_checksum": "cd4dd749525048e117fb9056fe901713daefc68b",
  "provenance_domain": {
    "provenance_name": "Regulatory BCO for hepatitis C virus resistance analysis",
    "provenance_version": "1.0",
    "provenance_review": {
```

Always a unique ID

Should always conform to IEEE specification: 1.4

Top level "Domain"

Nested Domain details are indented

Minor/patch changes may indicate grammatical or other minor fixes

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

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

Discussion: Feedback

The way that information is captured will depend on the environment the analysis is run in. As a Reviewer, what is the best format for representing file structure?

What are the “best practices?”

- E.g. for a spike-in study with multiple versions of the same pipeline, do you prefer multiple BCOs that reference each other? Or a single BCO?

How are manual QC steps represented?

How are files represented in Command Line?

<https://hive.biochemistry.gwu.edu/confluence/display/BUW/BioCompute+Workshop>

Q & A



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

Contact

Jonathon Keeney, Ph.D.
Assistant Research Professor
The George Washington University
keeneyjg@gwu.edu



BioCompute
Objects