BioCompute Database and Transfer Mechanism Development Workshop

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 & Outreach Lead
Technical Writer
janishapatel@gwu.edu

Workshop Guidelines

- If you have questions during the talk, please type them into the chat.

- Q&A
  - Moderator: Charles Hadley King
  - Use “raise hand” feature during Q&A session to ask a question.

- This workshop will be recorded, and the recording will be sent to all attendees shortly after.

Thank you!

Goals of this Workshop

1. Introduce BioCompute Objects (BCO) for computational analysis
2. Provide BioCompute resources for future reference
3. Demonstrate tools available via BioCompute Portal and BCODB

BioCompute Objects

Agenda

- Introduction: Use cases and BioCompute
- BioCompute Portal Walkthrough
- Demo of user account and DB access
- Description of DB and schema
- Transfer from Galaxy, HIVE, & local machine
- Q&A

Use-Case Examples

Test Submission

Tuberculosis Detection

Vaccine Safety

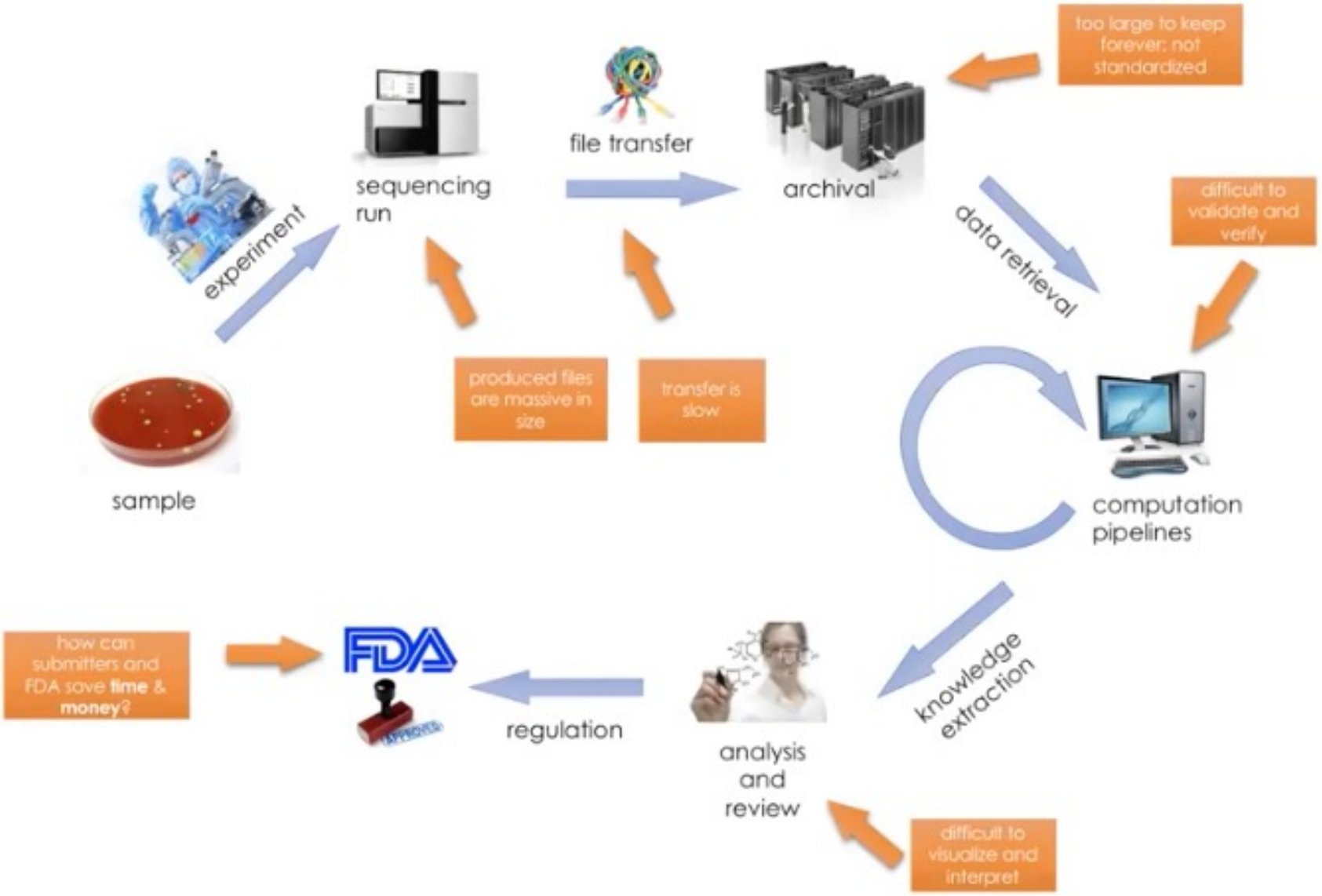
Provenance Domain

Name	R Safety Assessment Algorithm for Aluminum in Infant Vaccines.
Version	1.0
License	https://creativecommons.org/licenses/by/4.0/
Created	2020-04-30T18:03:25.679Z
Modified	2021-12-03T19:23:45.074Z

Contributors

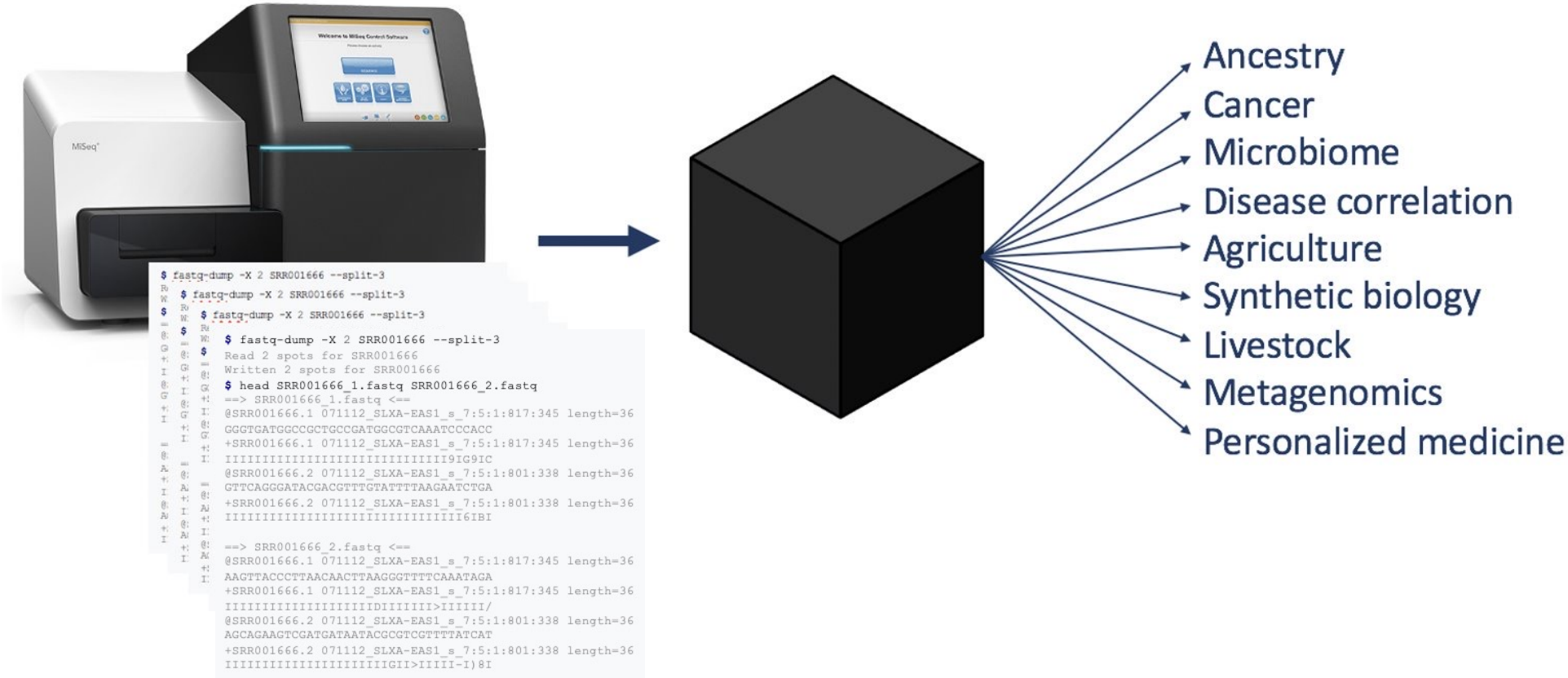
Name	Contribution	Affiliation	eMail
Charles Hadley King	createdBy, curatedBy	George Washington University	hadley_king@gwu.edu
Mark Walderhaug	curatedBy, curatedBy, authoredBy	U.S. Food and Drug Administration	mark.walderhaug@fda.hhs.gov
RJ Mitkus	authoredBy	U.S. Food and Drug Administration	rj.mitkus@fda.hhs.gov

NGS life cycle



Adapted by Vahan Simonyan

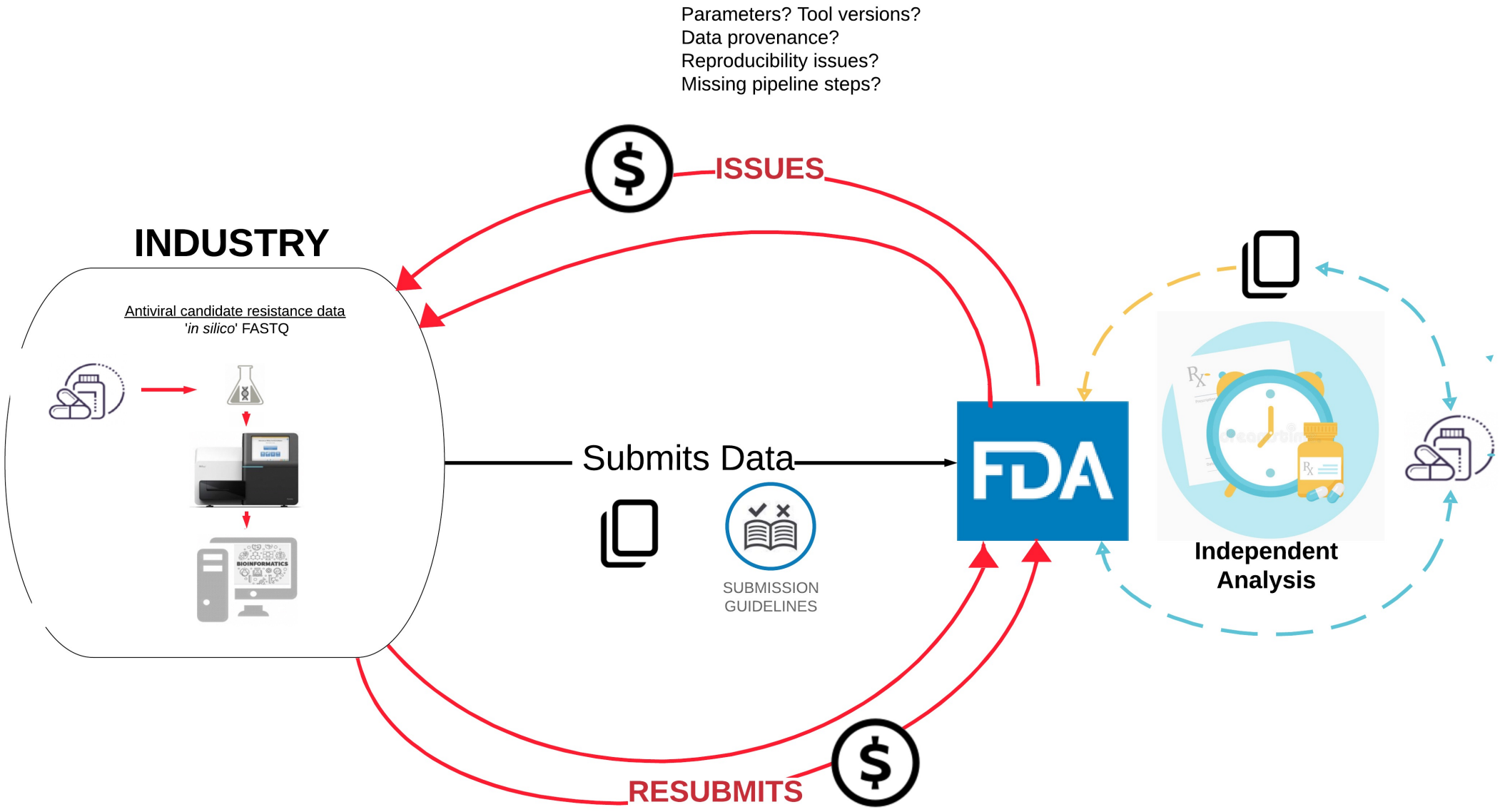
NGS Data Flows



BioCompute Objects

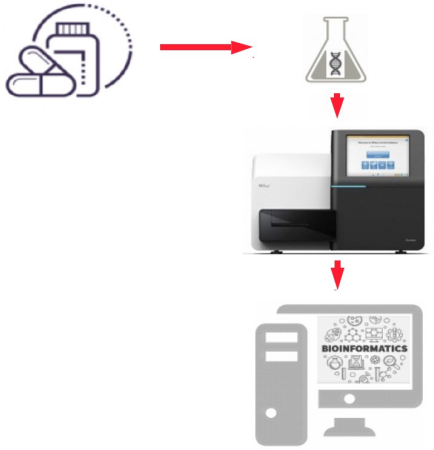
Introduction to BioCompute

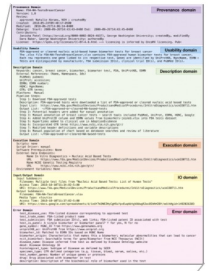
BCO Value in Regulatory Submission



BCO Value in Regulatory Submission

Clinical trial data:
'*in silico*' FASTQ



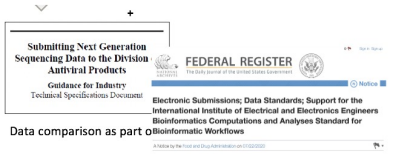


Verification Kit: 'insilico' FASTQ (subset)
AIM: to demonstrate the accuracy of NGS data analysis workflow

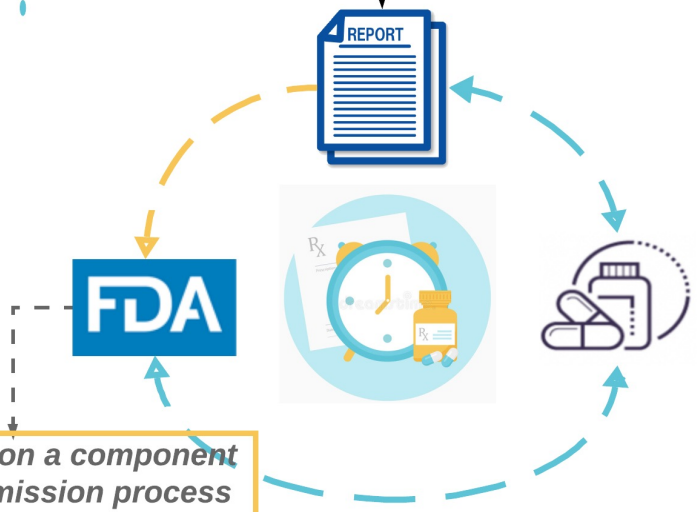
BioCompute Object

+

SUBMISSION GUIDELINES



Data comparison as part o



Data Comparison a component of current submission process

A solution should…

In 2014, the Genomics Working Group convened a special session to discuss the problem, and came up with four key focus areas

1. Be **human readable**: like a GenBank sequence record
2. Be **machine readable**: structured information with predefined fields and associated meanings of values

Written in JSON

3. Contain enough information to understand the computational pipelines, interpret information, maintain records, and reproduce experiments

Categorized by domains

- Adaptable to other schemas

4. Be **immutable**: ensure information has not been altered

Adheres to and encourages F.A.I.R. principles

- Fully open source
- Preserves data provenance (unique IDs for versioning)

BioCompute Objects

```
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=
```

Parametric 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/1xUY7WJNE7HyCgM5sYpxEuqAbtgVURwgR2oc8TWhH28Y/edit#gid=1492026303
```

IO domain

```
Error Domain
```

Error domain

Provenance Domain

Usability Domain

Extension Domain

Description Domain

Execution Domain

Parametric Domain

IO Domain

Error Domain

BioCompute Objects

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

NATIONAL ARCHIVES
FEDERAL REGISTER
The Daily Journal of the United States Government

Notice

Electronic Submissions; Data Standards; Support for the International Institute of Electrical and Electronics Engineers Bioinformatics Computations and Analyses Standard for Bioinformatic Workflows

A Notice by the Food and Drug Administration on 07/22/2020

PUBLISHED DOCUMENT

AGENCY:

Food and Drug Administration, Health and Human Services (HHS).

ACTION:

DOCUMENT

Printed ve
PDF
Publicatio
07/22/20
Agencies:

PUBLISHED DOCUMENT

AGENCY:

Food and Drug Administration, Health and Human Services (HHS).

ACTION:

Notice.

SUMMARY:

The Food and Drug Administration (FDA or Agency) is announcing support for use in regulatory submissions the current version of the International Institute of Electrical and Electronics Engineers (IEEE) bioinformatics computations and analyses standard for bioinformatic workflows (BioCompute) and an update to include this standard in the FDA Data Standards Catalog for the submission of high-throughput sequencing (HTS) data in new drug applications (NDAs), abbreviated new drug applications (ANDAs), biologics license applications (BLAs), and investigational new drug applications (INDs) to the Center for Biologics Evaluation and Research (CBER), Center for Drug Evaluation and Research (CDER), and Center for Food Safety and Applied Nutrition (CFSAN).

DATES:

Submit either electronic or written comments on the notice by August 21, 2020.

DOCUMENT DETAILS

Printed version:
PDF

Publication Date:
07/22/2020

Agencies:
Food and Drug Administration

Dates:
Submit either electronic or written comments on the notice by August 21, 2020.

Comments Close:
08/21/2020

Document Type:
Notice

Document Citation:
85 FR 44304

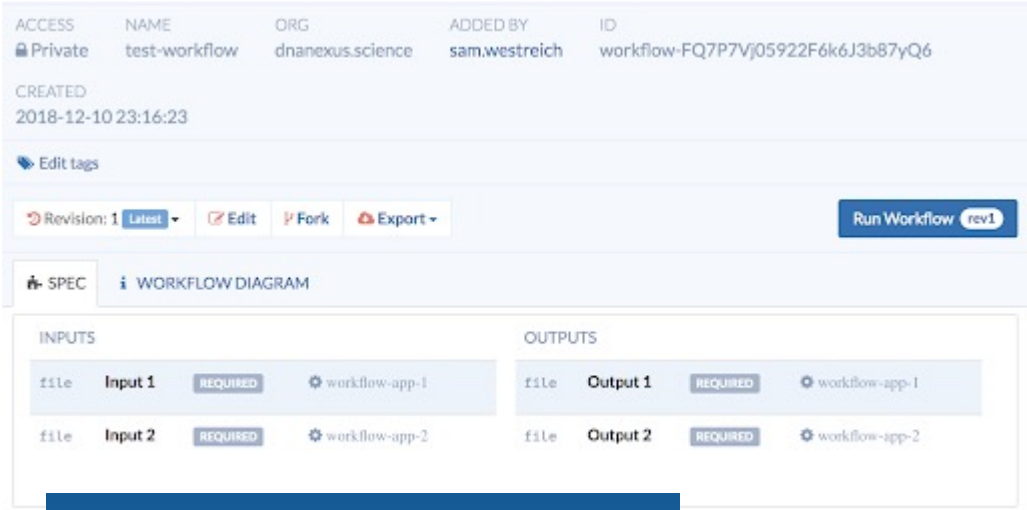
Page:
44304-44305 (2 pages)

Agency/Docket Number:
Docket No. FDA-2020-N-1450

Document Number:
2020-15771

DOCUMENT DETAILS

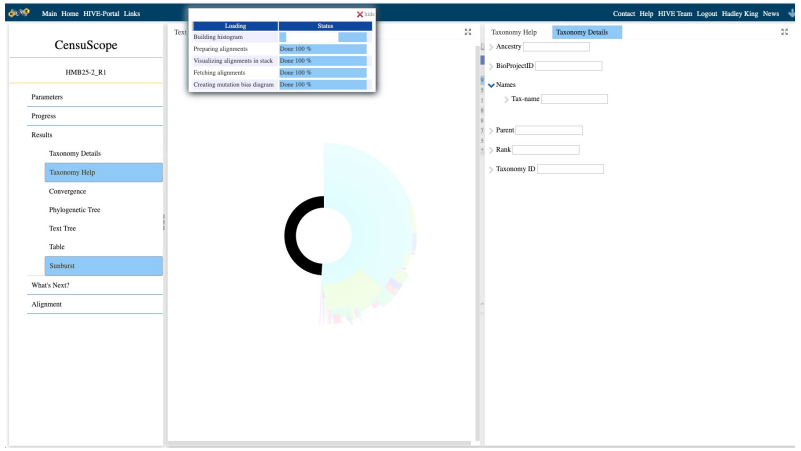
https://www.federalregister.gov/documents/2020/07/22/2020-15771/electronic-submissions-data-standards-support-for-the-international-institute-of-electrical-and

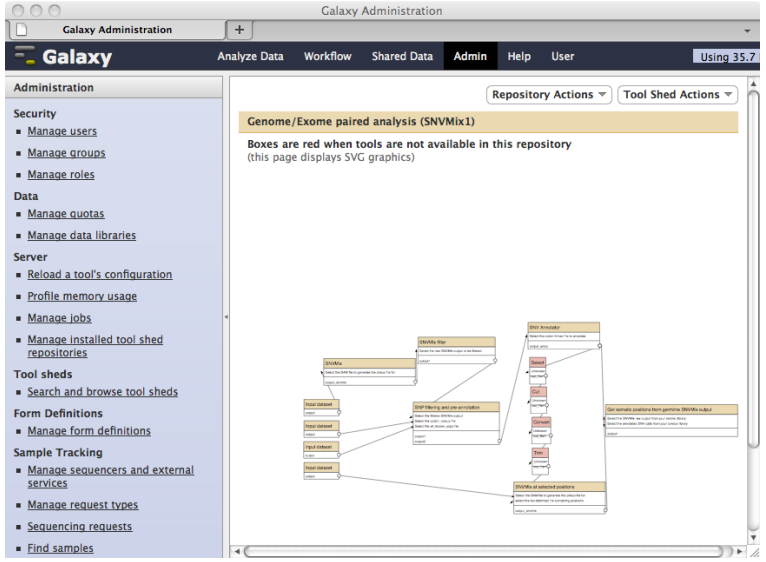


Seven Bridges

precisionFDA

DNAnexus
powering the genomics revolution™





Galaxy PROJECT

HIVE

BioCompute Objects

Introduction to BioCompute

BioCompute participants

Agenda

- Introduction to BioCompute
- BioCompute Portal Walkthrough
- Demo of user account and DB access
- Description of DB and schema
- Transfer from Galaxy, HIVE, & local machine
- Q&A

BioCompute Portal Walkthrough

Demo

How to navigate the Portal?
How to create a BioCompute Object?
How do you edit a BioCompute Object?

biocomputeobject.org

Incognito

Resources | Builder | BioCompute Objects | Log In/Register

BioCompute Documentation

External site

User Guide
Best Practices
SOP
Tutorials

IEEE 2791-2020

IEEE Standard for Bioinformatics Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication

IEEE
Advancing Technology for Humanity

BCO TSC

The Technical Steering Committee of the BioCompute Partnership (TSC) is a body of experienced professionals with BioCompute standard subject matter expertise. See here for the Meeting notes and agenda for all past and the upcomming meetings.

News and Events

FDA Notice on BioCompute

Electronic Submissions; Data Standards; Support for the International Institute of Electrical and Electronics Engineers Bioinformatics Computations and Analyses Standard for Bioinformatic Workflows.

Tweets by @BioComputeObj

BioCompute Retweeted

GW SMHS
@GWSMHS
A consortium led by @GW_HIVE_Lab, @rmazumde, and @Embleema, with @TempleUniv, has received @FDA funding to advance the work against #infectiousdiseases. Read more: bit.ly/3FaDCvY via @TheVDT

The Valdo Daily Tin

Embleema and George Washington University-led Co...
METUCHEN, N.J.--(BUSINESS WIRE)--Oct 28, 2021--
valdostadailytimes.com

Nov 15, 2021

BioCompute Retweeted

Cloud-based tools for BioCompute

See our resources page for additional tools and services.

HIVE
powered by aws

Access AWS HIVE, the High-Performance Integrated Virtual Environment, on AWS. HIVE is a cloud-based environment optimized for the storage and analysis of extra-large data, such as biomedical data, clinical data, next-generation sequencing (NGS) data, mass spectrometry files, confocal microscopy images, post-market

BioCompute Builder

Use the BioCompute Builder or view objects in the database.
The BioCompute Builder is a platform-free, form-based editor. The builder walks a user through building a BCO through text boxes, indicating which entries are required to adhere to the IEEE standard.

Galaxy PROJECT
powered by aws

Use Galaxy on AWS, the open source, web-based platform for data intensive biomedical research. Assemble your [pipeline](/about) in the workspace, designate the outputs in the module boxes, and record the entire pipeline as a BCO.

Portal home

Portal Walkthrough

Demo

User account and DB access

Demo

Register a new account
Build an Object
Create a draft
Edit a draft
Save edits
Validate Object
Publish

Description of DB and Schema

BioCompute Schema Files

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

IEEE.org | IEEE Xplore Digital Library | IEEE Standards | IEEE Spectrum | More Sites

IEEE SA OPEN ≡ Menu

ieee-2791-schema

- Project information
- Repository
- Issues 1
- Merge requests 1
- CI/CD
- Deployments
- Monitor
- Packages & Registries
- Analytics
- Wiki
- Snippets

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 year ago
45683af9

README | Other | Auto DevOps enabled

Name	Last commit	Last update
.gitignore	Creates initial release of BioCompute Obje...	2 years ago
2791object.json	replaces https://w3id.org/2791/ with https:...	1 year ago
AUTHORS	Update AUTHORS	1 year ago
CONTRIBUTORS	Update CONTRIBUTORS	1 year ago
LICENSE	Update LICENSE	1 year ago
README.md	Update README.md	1 year ago
description_domain.json	replaces https://w3id.org/2791/ with https:...	1 year ago
error_domain.json	replaces https://w3id.org/2791/ with https:	1 year ago

Demo

Repositories

Transfer from external sources using Swagger site for API access

biocomputeobject.org/api/docs/

Swagger Supported by SMARTBEAR

https://biocomputeobject.org/api/docs/?format=openapi

Explore

BioCompute Object Data Base API (BCODB API) 1.3.0

[Base URL: biocomputeobject.org/]
https://biocomputeobject.org/api/docs/?format=openapi

A web application that can be used to create, store and edit BioCompute objects based on BioCompute schema described in the BCO specification document.

Terms of service
Contact the developer
MIT License

Schemes
HTTPS

Django Login | Authorize

Filter by tag

Account Management

GET	/api/accounts/activate/{username}/{temp_identifier}	Activate an account	~~api_accounts_activate_read~~
POST	/api/accounts/describe/	Account details	api_accounts_describe_create
POST	/api/accounts/new/	Account creation request	api_accounts_new_create

Group Management

POST	/api/groups/create/	Create group	api_groups_create_create

Swagger site for API access

Demo

Transfer from Galaxy

API tools

Thank you!

Your time and feedback are greatly appreciated.

Acknowledgments

U.S. FOOD & DRUG ADMINISTRATION

Konstantinos Karagiannis
Eric Donaldson
Mark Walderhaug
Carolyn Wilson
Anton Golikov

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Raja Mazumder
Jonny Torcivia

BioCompute Objects

Vahan Simonyan
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
Nan Xiao

UNIVERSITY OF MIAMI

Chris Armstrong

SevenBridges

Dennis A. Dean
Jeffrey Grover
Soner Koc

DNAnexus

Omar Serang
John Didion

Acknowledgements and Contact

Raja Mazumder, Ph.D., PI
Professor
The George Washington University
mazumder@gwu.edu

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

Charles Hadley King
Technical Lead
The George Washington University
hadley_king@gwu.edu

Janisha Patel
Outreach & Training Lead
The George Washington University
janishapatel@gwu.edu

Recording on workshop: <youtube link>

- Introduction: Use cases and BioCompute
- BioCompute Portal Walkthrough
- Demo of user account and DB access
- Description of DB and schema
- Transfer from Galaxy, HIVE, & local machine
- Q&A

Q&A