



BioCompute Objects

*Public Workshop on Opportunities for Accelerating Scientific Discovery:
Realizing the Potential of Advanced and Automated Workflows*

Session on Standards, Governance, and Social Context

Raja Mazumder

The George Washington University

[\(mazumder@gwu.edu\)](mailto:mazumder@gwu.edu)

<https://biocomputeobject.org/>

THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON, DC



Copyrightable parts of slides: Creative Commons
Attribution (CC BY 4.0) License

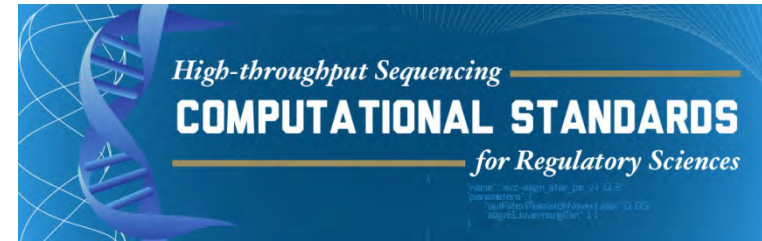
BioCompute Workshops



Purpose: *Community input on creating a standardized framework for computational analyses of HTS (NGS) data for FDA submission.*

This FRAMEWORK is known as a BioCompute object (BCO). We had hundreds workshop participants since first workshop in 2014.

Accurate **communication** is critical for regulatory evaluation of HTS (NGS) based products.

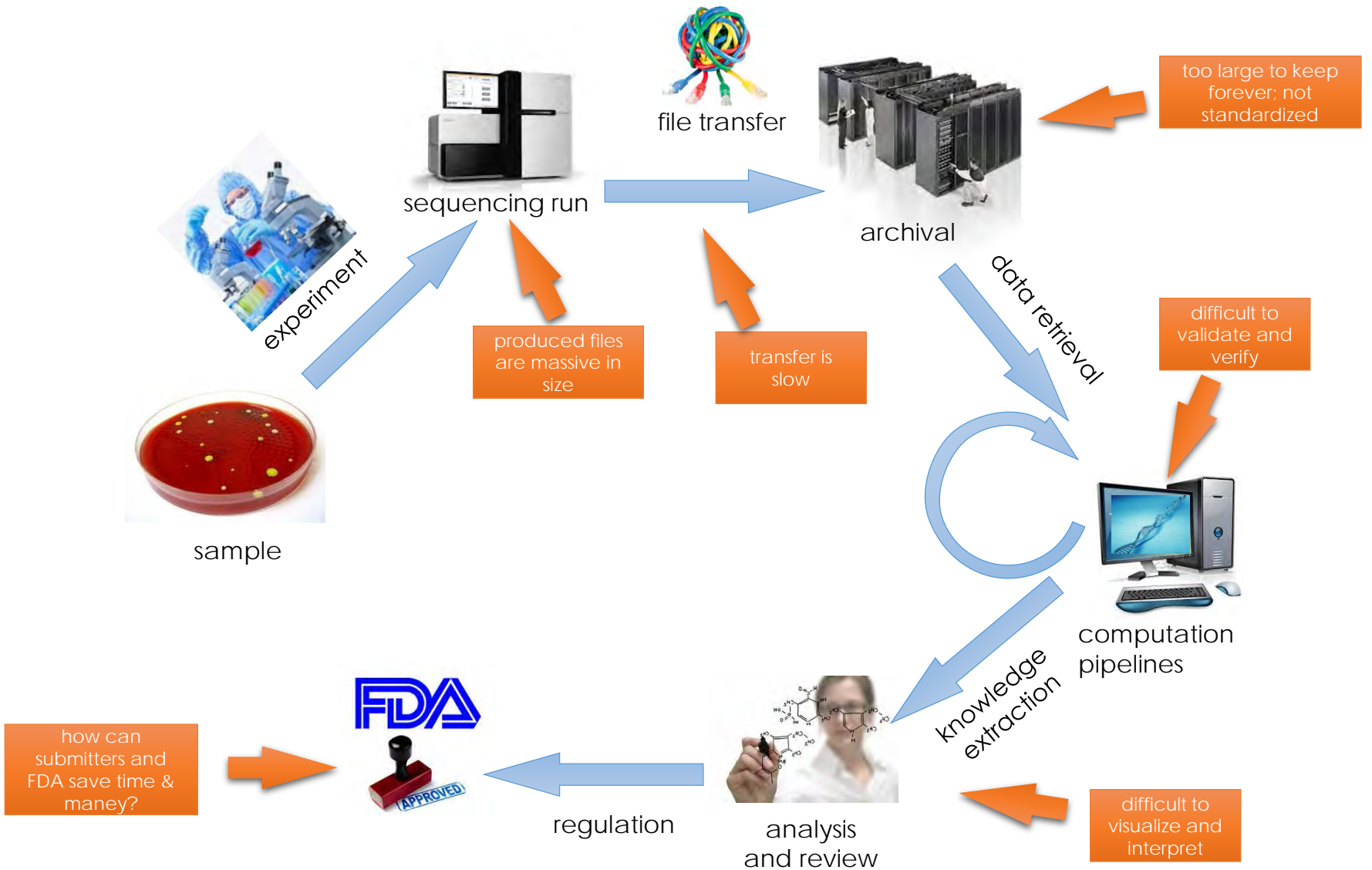


Main outcomes:

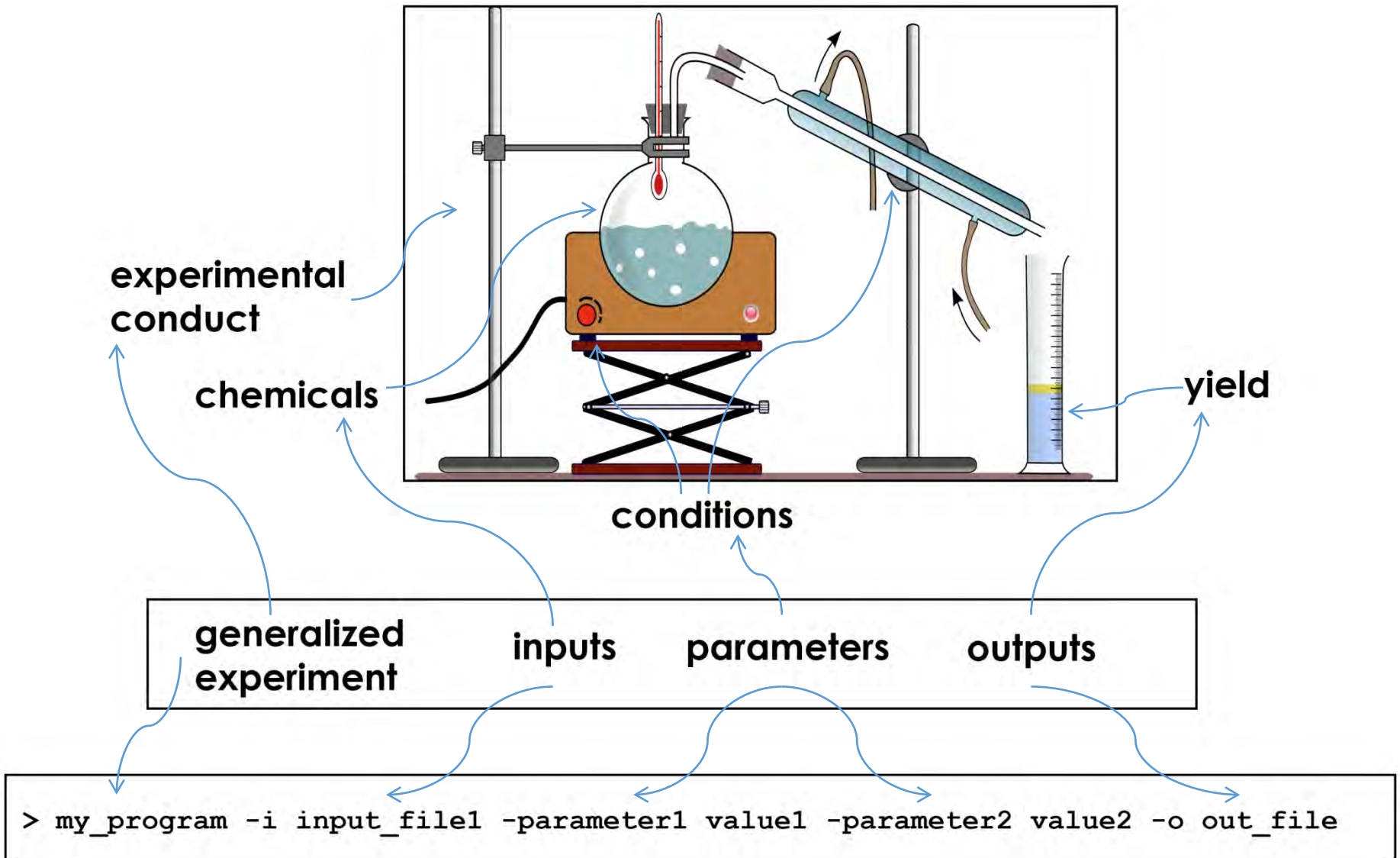
- Creation of the BCO specification document
- Public-private BCO-spec working groups with regular meetings
- BCO demonstration projects (for submitters and software platform developers) with community stakeholders.



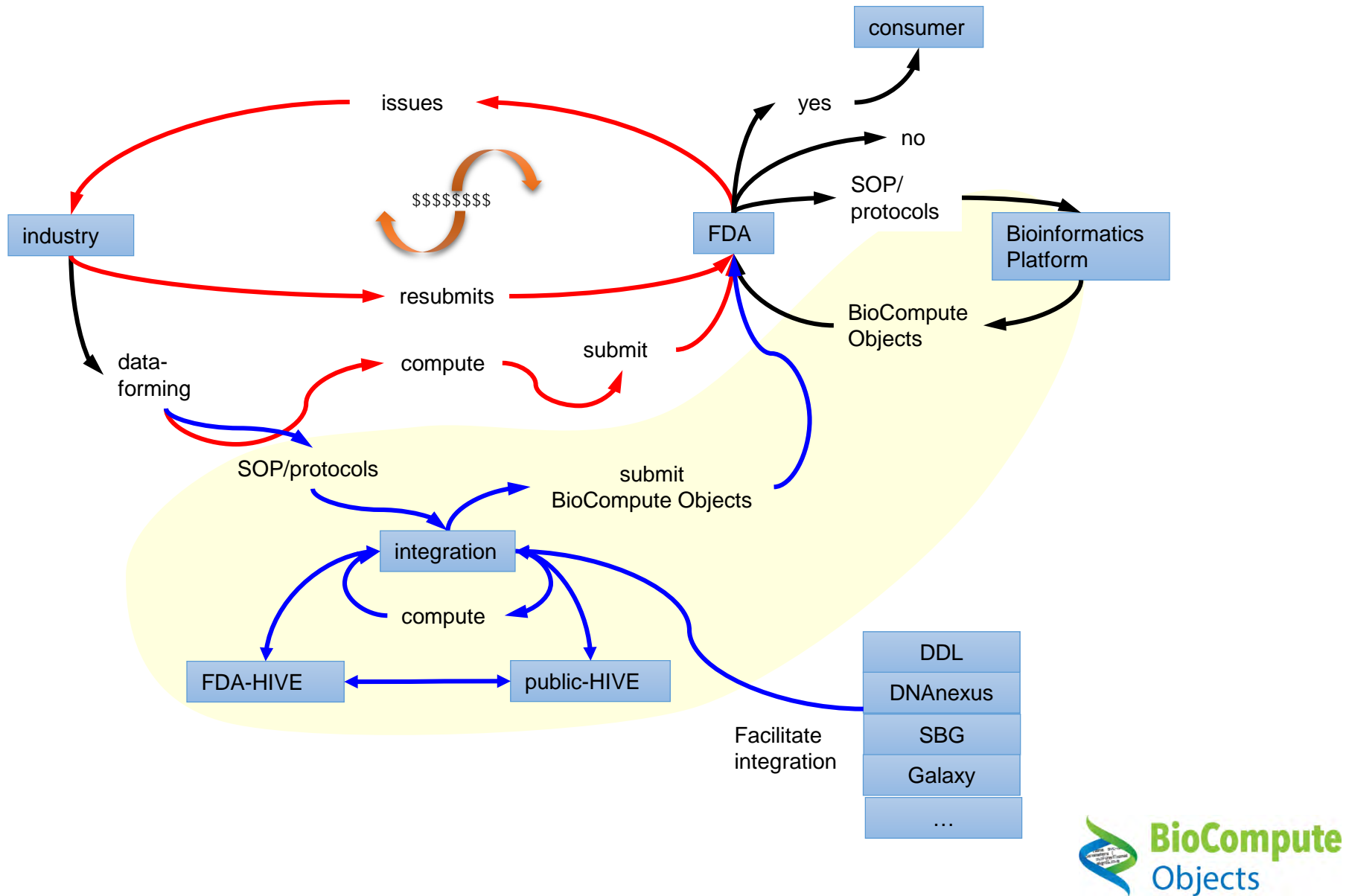
NGS lifecycle: from a biological sample to biomedical research and regulation



BioCompute Object need



BioCompute Object need



A solution should ...

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

802.11 Analogy



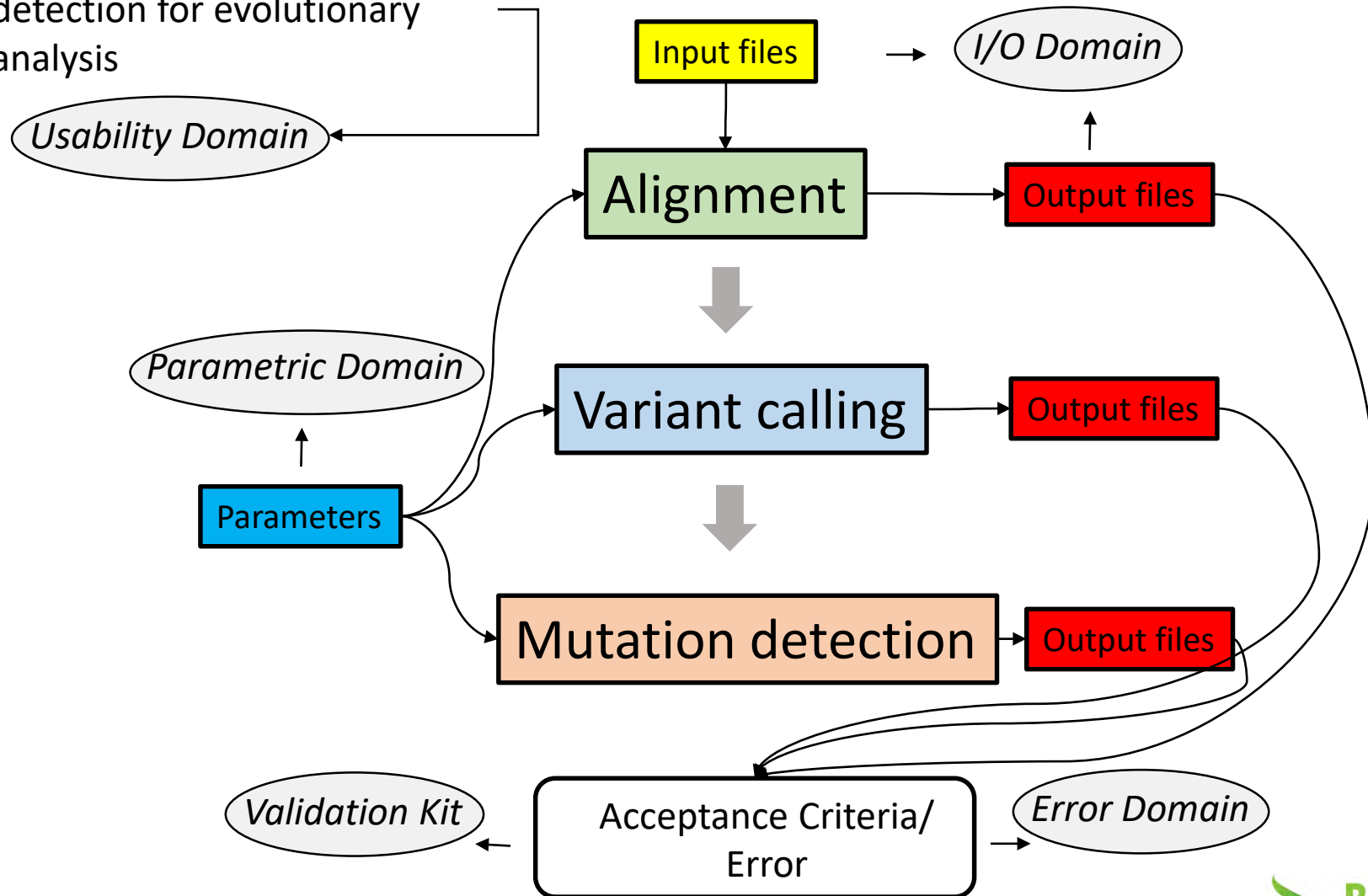
- More commonly called “WiFi.” Does not standardize the platform that information is generated on, the applications that use the information.
- The only thing that it standardizes is how information is collected and communicated between two devices. From there, you can do whatever you want with it.

BioCompute

- Standard for communicating computational analysis workflows
- Acts like an envelope for entire pipeline
 - Can incorporate other ontologies, standards (e.g. CWL, RO, DO, GA4GH, SEQC2...)
- Built with input from FDA, academia and industry
- Human and machine readable
 - Written in JSON
 - Unique IDs for versioning
- Categorized by domains (Usability, Execution, Error Domain etc.); Interoperable; Adaptable; ...
- IEEE approved standard for communicating genomic analysis workflows (<https://standards.ieee.org/standard/2791-2020.html>)

What is a BioCompute Object?

HCoV2 NS2 protein mutation
detection for evolutionary
analysis



Partnerships With Standards Organization

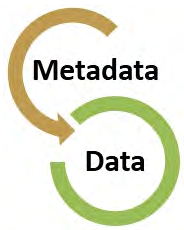
- Institute of Electrical and Electronics Engineers Standard
 - Available January 30th 2020
 - Scheduled for publication March 30th, 2020
- Under review with International Standards Organization (ISO)
 - IEEE/ISO joint agreement for expedited standardization



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, EDNR number Tests are distinguished by manufacturer, FDA submission ID(s), clinical trial ID(s) and PubMed ID(s). 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 nci_biomarker: Searchable terms for gene/Biomarker from NCI Thesaurus (NCIT)	Usability domain Extension domain
Description Domain Keywords: cancer, breast cancer, biomarker, biomarker test, FDA, UniProtKB, EDNR External References: (Name, Namespace, Ids) PubMed; pubmed; UniProt; accession; EDRN; 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

Salient Features of BioCompute

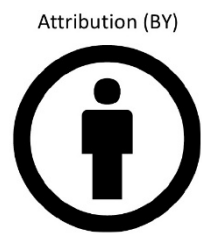
- Provides important **metadata** required for reproducing data



- Provides seamless **communication** and **collaboration** opportunities



- Provides appropriate **attribution**



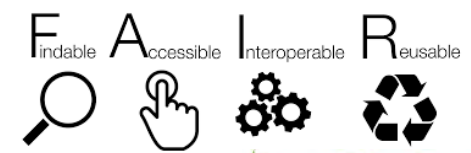
- Available in **machine readable** and **human readable** format



- Provides **license** details that helps others to use the data accordingly



- Data and metadata can follow **FAIR principles**



Acknowledgements



Vahan Simonyan
BioCompute Co-
founder



Jonathon Keeney
BioCompute Lead



Hadley King
BioCompute
Technical Lead

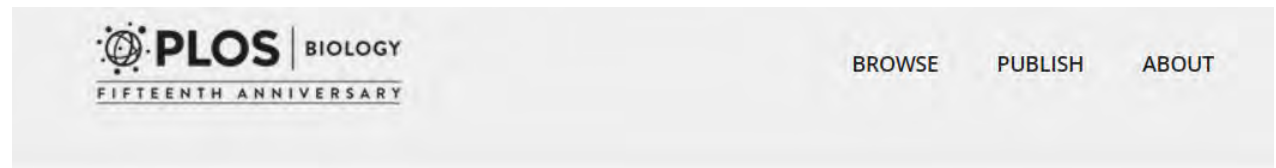


Janisha Patel
BioCompute
Training Lead

> PDA J Pharm Sci Technol, 71 (2), 136-146 Mar-Apr 2017

Biocompute Objects—A Step Towards Evaluation and Validation of Biomedical Scientific Computations

Vahan Simonyan ¹, Jeremy Goecks ², Raja Mazumder ³



OPEN ACCESS

COMMUNITY PAGE

Enabling precision medicine via standard communication of HTS provenance, analysis, and results

Gil Alterovitz, Dennis Dean, Carole Goble, Michael R. Crusoe, Stian Soiland-Reyes, Amanda Bell, Anais Hayes, Anita Suresh, Anjan Purkayastha, Charles H. King, Dan Taylor, Elaine Johanson, Elaine E. Thompson, [...], Raja Mazumder [view all]

Funding: FDA HHSF223201510129C (training); NSF/Internet2 E-CAS (AWS portability Exploring Clouds for Acceleration of Science)