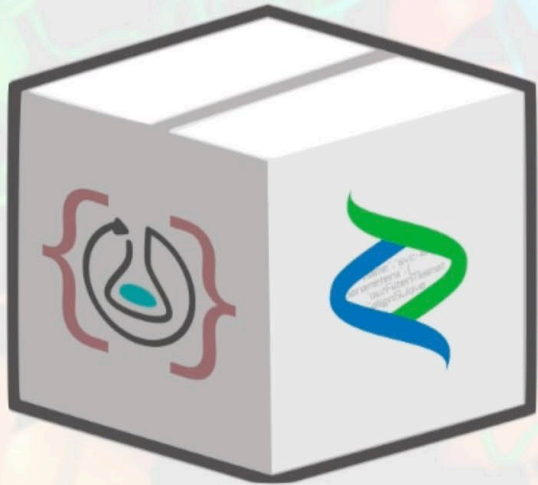


# Describing and packaging workflows using RO-Crate and BioCompute Objects

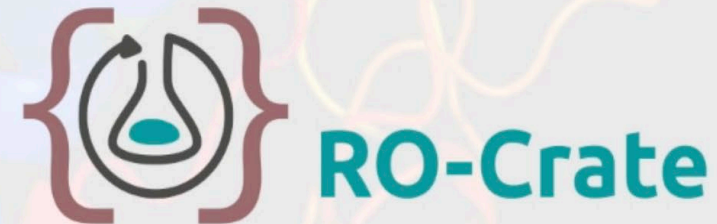


**Stian Soiland-Reyes**

eScience Lab, The University of Manchester

BioExcel Centre of Excellence

 <https://orcid.org/0000-0001-9842-9718>  [@soilandreyes](https://twitter.com/soilandreyes)



# Workflow Preservation and Reproducibility with BCO- Research Objects (RO)



**BioCompute**  
Objects

# Agenda

- Introduction to BioCompute Objects (BCO)
  - Jonathon Keeney
- Introduction to Research Objects (RO)
  - Stian Soiland-Reyes
- Introduction of BCO-RO tutorial
  - Stian Soiland-Reyes
- Q&A

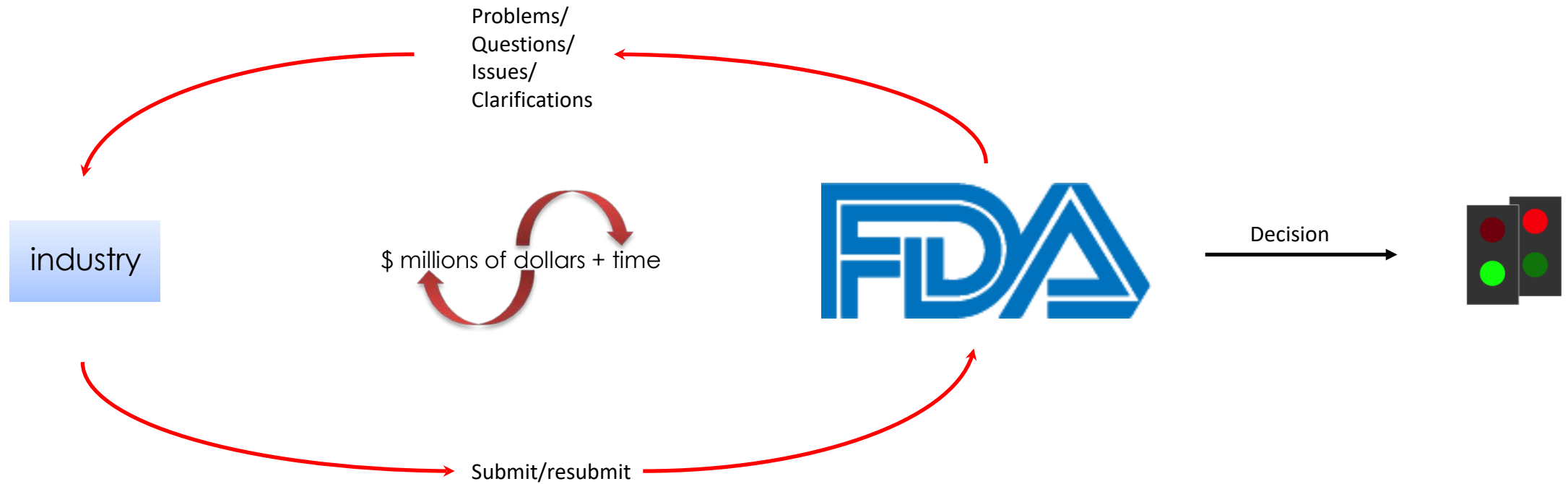
# Agenda

- 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

# Problem: Unclear Communication of Analyses

- What was the purpose of the pipeline?
- What is the context?
- What, exactly, was done?
- What parameters were used?
- What kind of data was being used?

# 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.

- Example of ad hoc solution
- Encompassed recommendations for both data and data processing
- Unstandardized

# 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
- 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

# Solution: BioCompute

Experimental Design

Analysis Steps

Parameters

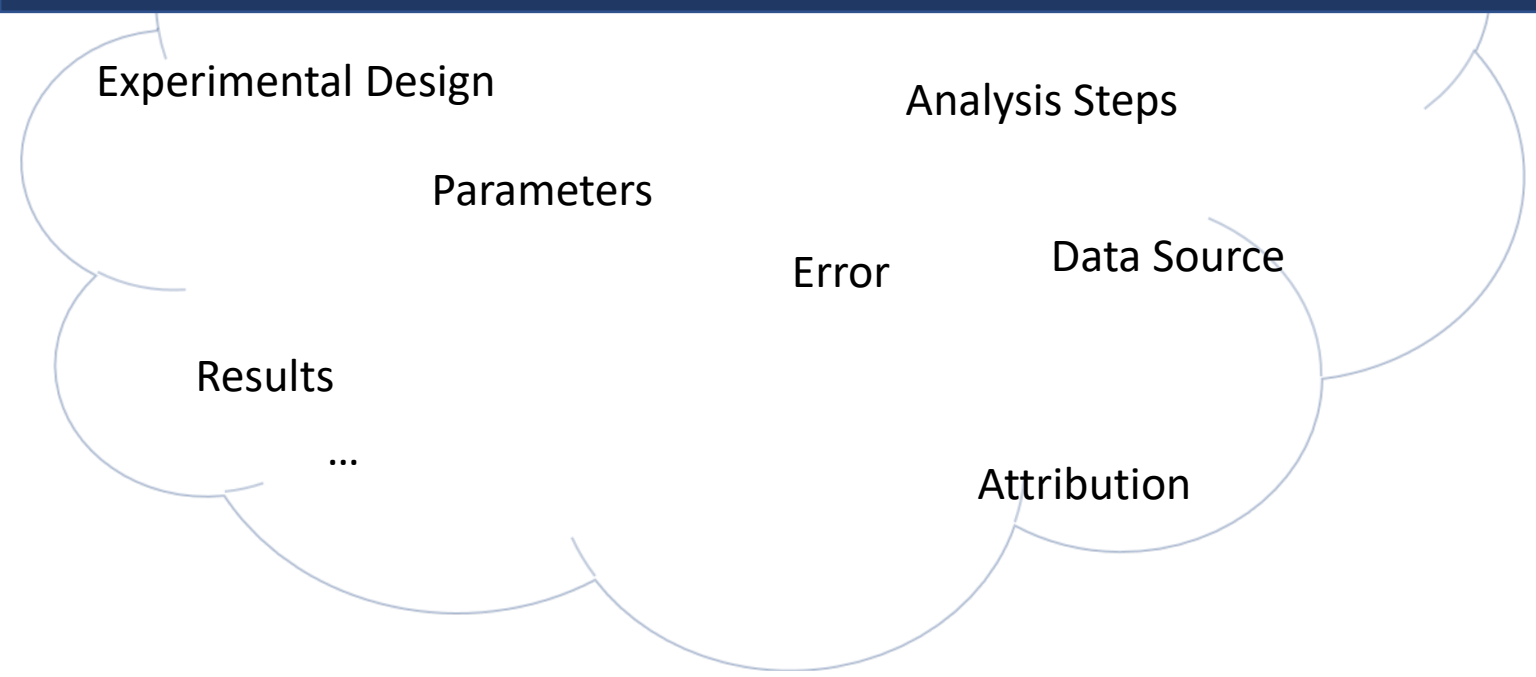
Error

Data Source

Results

...

Attribution



# Solution: BioCompute

Experimental Design

Parameters

Analysis Steps

Error

Data Source

Results

...

Attribution

Standardizes report content

BioCompute streamlines reporting without enforcing any tool, platform, or workflow strategy.

```
spec_version : https://w3id.org/ieee/ieee-2791-schema/
▶ usability_domain [1]
▶ provenance_domain {9}
▼ description_domain {2}
  ▶ keywords [11]
  ▼ pipeline_steps [10]
    ▶ 0 {7}
    ▶ 1 {6}
    ▼ 2 {7}
      name : Spike-In Trim and Filter Reads
      version : 1.0.0
      step_number : 3
      ▶ input_list [1]
      ▶ output_list [1]
```

Machine readability enables customized views

## Metadata

**object\_id** : [https://beta.portal.aws.biochemistry.gwu.edu/bco/BCO\\_00016916](https://beta.portal.aws.biochemistry.gwu.edu/bco/BCO_00016916)  
**spec\_version** : <https://w3id.org/ieeee/ieeee-2791-schema/>  
**etag** : fea7e938e6bdf9a2cfba7fa02f5a5fc3973dcc0b03a64319e1ee29966a5b6b

### provenance\_domain :

embargo :  
created : 2020-08-04T23:50:56.016Z  
modified : 2020-08-04T23:50:56.016Z  
name : Human Healthy Bulk RNA-seq Expression (Bgee)  
version : v-1.0  
obsolete\_after : 2020-04-22T23:57:00.000Z  
contributors :  
contribution :  
createdBy  
name : Amanda Bell  
email : amandab2140@gwu.edu  
affiliation : GW HIVE-Lab  
orcid : <http://orcid.org/0000-0002-9920-565X>  
license : Attribution 4.0 International CC BY 4.0

## Provenance Domain

### description\_domain :

keywords :  
Gene Expression  
Gene Expression Regulation  
Tissue specificity  
xref :  
namespace : ensembl  
name : Ensembl Genome Browser  
ids :  
Ensembl gene ID  
access\_time : 2020-04-22T14:03:00.000Z  
platform :  
OncoMX  
pipeline\_steps :  
step\_number : 1  
name : oncomx server  
prerequisite :  
uri :  
description : Process data  
input\_list :

## Description Domain

### error\_domain :

empirical\_error :  
D168Y: percentage: 0.56, calls: 0.5615, STDEV.P: 0.00075  
algorithmic\_error :  
SCORE\_threshold: 0.5, QUALITY: 25, COVERAGE: 5000

## Error Domain

### parametric\_domain :

param : grep  
value : -r  
step : 1

## Parametric Domain

### execution\_domain :

environment\_variables :  
key : EDITOR  
value : vim  
key : HOSTTYPE  
value : x86\_64-linux  
external\_data\_endpoints :  
url : <https://data.oncomx.org/ONCOMXDS000012>  
name : Human Healthy Bulk RNA-seq Expression (Bgee)  
script :  
uri :  
filename : make-dataset.py  
uri : <http://data.oncomx.org/ln2wwwdata/software/pipeline/integrator/make-dataset.py>  
access\_time : 2020-04-22T14:28:00.000Z  
software\_prerequisites :  
uri :  
filename : shell  
uri : <https://www.python.org/download/releases/2.7.5>  
access\_time : 2020-04-22T14:30:00.000Z  
name : Python  
version : 2.7.5  
script\_driver : Python

## Execution Domain

### io\_domain :

input\_subdomain :  
uri :  
filename : Homo\_sapiens\_UBERON:0000066  
uri :  
[http://data.oncomx.org/ln2wwwdata/downloads/bgee/current/Homo\\_sapiens\\_UBERON:0000066\\_AFFYMETRIX\\_RNA\\_SEQ.tsv](http://data.oncomx.org/ln2wwwdata/downloads/bgee/current/Homo_sapiens_UBERON:0000066_AFFYMETRIX_RNA_SEQ.tsv)  
access\_time : 2020-04-22T20:44:00.000Z  
output\_subdomain :  
uri :  
filename : human\_normal\_expression.csv  
uri : <https://data.oncomx.org/ONCOMXDS000012>  
access\_time : 2020-04-22T20:50:00.000Z  
mediatype : TEXT/CSV

## IO Domain

### extension\_domain :

dataset\_categories :  
category\_value : Homo sapiens  
category\_name : species  
category\_value : normal  
category\_name : disease\_status  
extension\_schema : <https://data.oncomx.org/ONCOMXDS000012>

## Extension Domain

### usability\_domain :

List of human taxid:9606 genes with healthy RNA-Seq and Affymetrix expression data in Bgee; additional documentation available at ([https://github.com/BgeeDB/bgee\\_pipeline/tree/develop/pipeline/collaboration/oncoMX#information-about-the-files-generated-for-oncomx](https://github.com/BgeeDB/bgee_pipeline/tree/develop/pipeline/collaboration/oncoMX#information-about-the-files-generated-for-oncomx)) Only the subset of RNA-Seq data are used to generate the expression profiles for healthy individuals for human used by OncoMX.



## Usability Domain

# BioCompute participants



# Standardization



Institute of Electrical and Electronics  
Engineers Standard

“BioCompute” 2791-2020 approved January  
2020

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



## 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:

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.

### ADDRESSES:

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

# Acknowledgements and Contact



Raja Mazumder, Ph.D., PI  
Professor  
The George Washington University  
[mazumder@gwu.edu](mailto:mazumder@gwu.edu)



Jonathon Keeney, Ph.D., Co-I  
Assistant Research Professor  
The George Washington University  
[keeneyjg@gwu.edu](mailto:keeneyjg@gwu.edu)



Hadley King  
Technical Lead  
The George Washington University



Chris Armstrong  
Development Lead  
The George Washington University



Janisha Patel  
Outreach Lead  
The George Washington University



**BioCompute**  
Objects



<https://orcid.org/0000-0001-9842-9718>

# Stian Soiland-Reyes

## FAIR scholarly communication

Research Object

Research Data Alliance

Open PHACTS

## Linked Data and Web standards

W3C: PROV recommendations

schema.org

ORCID

## Open Source software developer

Apache Software Foundation member

Taverna Workbench

## Reproducibility and scalability

Common Workflow Language (CWL)

Provenance, Containers, BioConda

WorkflowHub.eu



MANCHESTER  
1824

The University of Manchester

eScience Lab, Dept. of Computer Science



<https://elixir-europe.org/>

bioexcel  
Center of Excellence for Computational Biomolecular Research  
<https://bioexcel.eu/>